Psychological Review
2009, Vol. 116, No. 4, 661–716

0033-295X/09/$12.00 DOI: 10.1037/a0017201

Theory-Based Causal Induction

Thomas L. Griffiths
University of California, Berkeley

Joshua B. Tenenbaum
Massachusetts Institute of Technology

Inducing causal relationships from observations is a classic problem in scientific inference, statistics, and machine learning. It is also a central part of human learning, and a task that people perform remarkably well given its notorious difficulties. People can learn causal structure in various settings, from diverse forms of data: observations of the co-occurrence frequencies between causes and effects, interactions between physical objects, or patterns of spatial or temporal coincidence. These different modes of learning are typically thought of as distinct psychological processes and are rarely studied together, but at heart they present the same inductive challenge—identifying the unobservable mechanisms that generate observable relations between variables, objects, or events, given only sparse and limited data. We present a computational-level analysis of this inductive problem and a framework for its solution, which allows us to model all these forms of causal learning in a common language. In this framework, causal induction is the product of domain-general statistical inference guided by domain-specific prior knowledge, in the form of an abstract causal theory. We identify 3 key aspects of abstract prior knowledge—the ontology of entities, properties, and relations that organizes a domain; the plausibility of specific causal relationships; and the functional form of those relationships—and show how they provide the constraints that people need to induce useful causal models from sparse data.

Keywords: causal induction, intuitive theories, rational analysis, Bayesian modeling

In 1695, Sir Edmond Halley was computing the orbits of a set of comets for inclusion in Newton's *Principia Mathematica* when he noticed a surprising regularity: The comets of 1531, 1607, and 1682 took remarkably similar paths across the sky, and visited the Earth approximately 76 years apart. Newton had already shown that comets should follow orbits corresponding to conic sections—parabolas, hyperbolas, and ellipses—although no elliptical orbits had yet been observed. Halley inferred that the sightings of these comets were not three independent events, but three consequences of a single common cause: a comet that had visited the Earth three times, travelling in an elliptical orbit. He went on to predict that it would return along the same orbit in 1758. The comet returned as predicted, and has continued to visit the Earth approximately every 76 years since, providing a sensational confirmation of Newton's physics.

Halley's discovery is an example of *causal induction*: inferring causal structure from data. Explaining this discovery requires appealing to two factors: abstract prior knowledge, in the form of a causal theory, and statistical inference. The prior knowledge that guided Halley was the mathematical theory of physics laid out by Newton. This theory identified the entities and properties relevant to understanding a physical system, formalizing notions such as velocity and acceleration, and characterized the relations that can hold among these entities. Using this theory, Halley could generate a set of hypotheses about the causal structure responsible for his astronomical observations: They could have been produced by three different comets, each travelling in a parabolic orbit, or by one comet, travelling in an elliptical orbit. Choosing between these hypotheses required the use of statistical inference. While Halley made no formal computations of the probabilities involved, the similarity in the paths of the comets and the fixed interval between observations convinced him that "it was highly probable, not to say demonstrative, that these were but one and the same Comet" (from the *Journal Book of the Royal Society*, July 1696, reproduced in Hughes, 1990, p. 353).

Causal induction is not just a problem faced by scientists. The capacity to reason about the causes of events is an essential part of cognition from early in life, whether we are inferring the forces involved in physical systems (e.g., Shultz, 1982b), the mental states of others (e.g., Perner, 1991), or the essential properties of natural kinds (e.g., S. A. Gelman & Wellman, 1991). Often, these causal relationships need to be inferred from data. Explaining how people make these inferences is not just a matter of explaining how causation is identified from correlation, but of accounting for

Thomas L. Griffiths, Department of Psychology and Program in Cognitive Science, University of California, Berkeley; Joshua B. Tenenbaum, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology.

While completing this work, we were supported by the James S. McDonnell Causal Learning Research Collaborative, Air Force Office of Scientific Research Grants FA9550-07-1-0351 (to Thomas L. Griffiths) and FA9550-07-1-0075 (to Joshua B. Tenenbaum), and Army Research Office MURI Grant W911NF-08-1-0242 (to Joshua B. Tenenbaum).

We thank Elizabeth Bonawitz, Tevye Krynski, Tania Lombrozo, Sourabh Niyogi, Laura Schulz, and David Sobel for comments on the article and related discussions, Marc Buehner for providing the data from Buehner and Cheng (1997), and Hongjing Lu for providing a version of the datasets considered by Perales and Shanks (2007) that included the valence of the causal relationship. We are grateful to Elizabeth Bonawitz and Andreas Stuhlmuller for their assistance in data collection.

Correspondence concerning this article should be addressed to Thomas L. Griffiths, Department of Psychology, 3210 Tolman Hall, MC 1650, University of California, Berkeley, CA 94720-1650. E-mail: tom_griffiths@berkeley.edu

how complex causal structure is inferred in the absence of (statistically significant) correlation. People can infer causal relationships from samples too small for any statistical test to produce significant results (e.g., Gopnik, Sobel, Schulz, & Glymour, 2001) and solve problems like inferring hidden causal structure (e.g., Kushnir, Gopnik, Schulz, & Danks, 2003) that still pose a major challenge for statisticians and computer scientists. Human causal induction is not always on target: sometimes we miss causal connections that would be most valuable to exploit, or see connections that do not in fact exist. Yet the successes stand out. No conventional statistical recipe or computer learning algorithm can compete with a young child's capacity to discover the causal structure underlying everyday experience—or at least, to come close enough to causal "ground truth" with knowledge that supports such flexible prediction, planning, and action in the world.

In this article, we present a formal framework for explaining how human causal learning works across a wide range of contexts and information sources. Our goal here is not a mechanistic explanation in terms of psychological processing steps or neural machinery. Rather we want to explain how human learners can successfully infer such rich causal models of the world given that the data they observe are so sparse and limited. Our explanations of human causal induction take the form of computational theories, in the sense introduced by Marr (1982) and pursued by Shepard (1987) and Anderson (1990), among others (see Oaksford & Chater, 1998): We identify the abstract computational problem addressed by a cognitive capacity, derive an optimal solution to that problem, and use that solution to explain human behavior.

In our analysis of the computational problem underlying everyday causal induction, the two factors of prior knowledge and statistical inference that we identified in Halley's famous discovery both play central roles. Prior knowledge, in the form of an abstract theory, generates hypotheses about the candidate causal models that can apply in a given situation. Principles of Bayesian inference generate weights for these hypotheses in light of observed data and thus predictions about which causal relations are likely to hold, and which patterns of future events are likely to be observed. To the extent that the prior knowledge is veridical—when people's abstract intuitive theories reflect the way causal systems in their environment tend to work—our rational framework explains how people's inferences about the structure of specific causal systems can be correct, even given very little data. Yet our framework is not strictly normative: In cases where people hold the wrong abstract theories, rational statistical inference may lead them to incorrect beliefs about a novel system even given extensive experience.

The idea that causal induction draws on prior knowledge is not novel—it has been noted in many influential theories (e.g., Cheng, 1997; Lien & Cheng, 2000), discussed in the context of rational statistical inference (Alloy & Tabachnik, 1984), and explored in two separate research programs (Koslowski, 1996; Waldmann, 1996; Waldmann, Hagmayer, & Blaisdell, 2006). However, previous formal models have focused mostly on the effects of specific forms of prior knowledge, such as the plausibility of a causal relationship, painting a relatively simple picture of the role this knowledge plays in learning. Our contribution is a formal framework that provides a way to systematically identify the aspects of prior knowledge that can influence causal induction, to describe this knowledge precisely, and to explain how it is combined with rational mechanisms of statistical inference. We call this framework *theory-based causal induction*. We propose that three aspects of prior knowledge are central in generating hypotheses for causal induction—the ontology of entities, properties, and relations that organizes a domain; the plausibility of specific causal relationships; and the functional form of those relationships—and that these three aspects are the key constituents of people's intuitive causal theories. Mathematical models of causal induction in specific settings can be derived by performing Bayesian inference over the hypothesis spaces generated by appropriate theories of this sort, and they illustrate how relatively complex interactions between prior knowledge and data can emerge.

By viewing causal induction as the result of domain-general statistical inference guided by domain-specific causal theories, our framework provides a unified account of a set of phenomena that have traditionally been viewed as distinct. Different aspects of causal learning have tended to be explained in different ways. Theories of causal learning from contingency data, in which people are provided with information about the frequency with which cause and effect co-occur and are asked to evaluate the underlying relationship, emphasize statistical learning from covariation between cause and effect (e.g., Cheng, 1997; Jenkins & Ward, 1965; Shanks, 1995b). In contrast, analyses of learning about the causal relationships that govern physical systems, such as simple machines, tend to focus on the role of domain-specific knowledge about the nature of possible causal mechanisms (e.g., Bullock, Gelman, & Baillargeon, 1982; Shultz, 1982b). Finally, inferences about causal relationships based on spatial and temporal dimensions of dynamic events—most famously illustrated in Michotte's (1963) classic studies of perceived causality in collisions—are often viewed as the product of a modular, automatic perceptual mechanism, distinct from a general cognitive capacity for statistical inference thought to underlie causal learning from contingency data (e.g., Leslie, 1986; Schlottmann & Shanks, 1992).

From the perspective of theory-based causal induction, these apparently disparate phenomena are not discrete cases requiring separate explanations but rather points on a continuum, where the strength of the constraints provided by prior knowledge gradually increases, and the amount of information required in order to make a causal inference decreases accordingly. Standard experiments on causal induction from covariational data tap into relatively weak prior knowledge and hence require a relatively large number of experienced events (typically, tens of data points) for learners to reach confident causal conclusions. Causal learning in simple physical systems draws on richer knowledge about causal mechanisms, allowing confident causal inferences from only a handful of examples. When even richer (if implicit) knowledge about the spatiotemporal dynamics of physical interactions is involved, as in standard cases of perceptual causality, confident inferences can be made from just a single observed event—a "suspicious coincidence"—with the appropriate spatiotemporal structure.

The plan of the article is as follows. In the next section, we summarize previous work illustrating how different aspects of prior knowledge can influence causal learning. We then discuss the goals of a computational-level analysis of causal induction, focusing on a description of the central inductive problems to be solved by the learner. This description introduces causal graphical models, a formalism for representing and reasoning about causal relationships which has been the basis for previous accounts of human

causal induction, but which crucially does not have the ability to express the forms of abstract prior knowledge that guide human causal learning along such different trajectories in different domains and contexts. We then introduce our framework of theory-based causal induction, building on the foundation of causal graphical models but making explicit the structure and function of the learner's prior knowledge. The bulk of the article consists of the application of this framework to the settings mentioned above: causal induction from contingency data, learning about the properties of physical systems, and inferring causal relationships from coincidences in space and time. In considering these phenomena, we focus on the importance of the two key components of our approach—statistical inference and prior knowledge expressed in the form of a causal theory—in explaining how people can learn about causal relationships from limited data.

How Does Prior Knowledge Influence Causal Induction?

The study of causal induction has a long history, in both philosophy (e.g., Hume, 1739/1978) and psychology (e.g., Inhelder & Piaget, 1958). Detailed reviews of some of this history are provided by Shultz (1982b; Shultz & Kestenbaum, 1985) and White (1990, 1995). This history is marked by a tension between statistical learning and abstract prior knowledge about causality as accounts of human causal induction. Psychological theories about causal induction have tended to emphasize one of these two factors over the other (Cheng, 1997; Newsome, 2003; Shultz, 1982b): In the tradition of Hume (1739/1978), 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 & Novick, 1990, 1992; Shanks & Dickinson, 1987), whereas, in a tradition often traced to Kant (1781/1964; see Shultz, 1982b, for an account of the connection), mechanism-based approaches focus on the role of prior knowledge about the mechanisms by which causal force can be transferred (e.g., Ahn & Kalish, 2000; Shultz, 1982b; White, 1995).

Recently, explanations of human causal learning have begun to explore a middle ground between these positions, looking at how mechanism knowledge might influence learning from covariation between cause and effect (e.g., Lagnado & Sloman, 2004; Lagnado, Waldmann, Hagmayer, & Sloman, 2007; Waldmann, 1996; Waldmann et al., 2006). These accounts are based on a range of results indicating the importance of both of these factors. The analysis that we present in this article can be viewed in part as an attempt to develop a formal framework that can capture the kind of knowledge needed to explain these effects, providing the tools required to define computational models of this "knowledge-based" approach to causal induction (Waldmann, 1996). As a first step toward developing such a framework, we need to identify exactly what aspects of prior knowledge are relevant to causal induction. In this section, we briefly review work that has explored this question, using the example of Halley's discovery for illustration. Following this example, we divide the kind of prior knowledge that is relevant to causal induction into three categories: information about the types of entities, properties, and relations that arise in a domain (the ontology); constraints on the plausible relations among these entities; and constraints on the functional form of such relations.[1]

Ontology

Newton's theory of physics picked out the critical variables for thinking about the motion of objects—their mass, velocity, and acceleration. The question of how entities are differentiated on the basis of their causal properties has been thoroughly explored in developmental psychology, through consideration of the ontological commitments reflected in the behavior of infants and young children. Both infants and young children have strong expectations about the behavior of physical objects, and these expectations are quite different from those for intentional agents (Saxe, Tenenbaum, & Carey, 2005; Shultz, 1982a; Spelke, Phillips, & Woodward, 1995). Similarly, children have different expectations about the properties of biological and nonbiological entities (e.g., Springer & Keil, 1991). Gopnik et al. (2001) have shown that children use the causal properties of entities to determine whether they belong to a novel type—objects that differed in appearance but both activated a "detector" were more likely to both be considered "blickets" than objects with similar appearance that differed in their causal properties.

Research with adults has also explored how the types of entities influence causal inferences. For example, Lien and Cheng (2000) conducted several experiments examining the circumstances under which causal properties are generalized across the members of a category. In a typical experiment, people learned about the tendency of 15 chemicals to produce blooming in a plant. The chemicals could be divided into groups on the basis of their color and shape. Lien and Cheng explored how people used information about color and shape, which provided a basis for identifying different types of chemicals in causal learning. Their conclusion was that people used these types in learning causal relationships: People formed the generalization that chemicals of the type that maximized the strength of the resulting relationship were those that caused the plant to bloom. In related work, Waldmann and Hagmayer (2006) examined how intuitive theories influence whether previously learned categories are transferred to novel causal learning problems. Tenenbaum and Niyogi (2003) also showed that people spontaneously organize objects into types on the basis of their causal properties, forming abstract categories of schematic blocks that cause one another to light up in a computer simulation, and Kemp, Goodman, and Tenenbaum (2007) showed that such categories could carry with them expectations about the strength of causal relationships.

Plausible Relations

Knowledge of the types of entities in a domain can provide quite specific information about the *plausibility* of causal relationships. For example, Newton precisely laid out the kinds of forces by which the properties of one object can influence those of another. Explorations of how plausibility influences causal induction have

[1] Although we refer to expectations about ontologies, plausible relations, and functional form as *prior knowledge*, we mean this to be interpreted as indicating that the knowledge is available prior to a specific instance of causal induction. We are not claiming that this knowledge is innate, and anticipate that in almost all cases it is acquired through experience with the world and in particular through other instances of causal induction, a point that we return to in our discussion of learning causal theories.

examined mainly how children learn about the structure of physical systems (e.g., Shultz, 1982b), although even accounts of causal induction from contingency data that emphasize the importance of covariation between cause and effect recognize a role for *top-down knowledge* (e.g., Cheng, 1993, 1997). In one classic study, Shultz (1982b) demonstrated that young children have strong expectations about the plausibility of different kinds of causal relationships, in part derived from their experience with the properties of these objects in the course of the experiment. For example, he found that children used the knowledge that a lamp is more likely than a fan to produce a spot of light, that a fan is more likely than a tuning fork to blow out a candle, and that a tuning fork is more likely than a lamp to produce resonance in a box.

On the basis of examples like those provided by Shultz (1982b), several authors have equated the plausibility of a causal relationship with the existence of a potential mechanism by which the cause could influence the effect (e.g., Ahn & Kalish, 2000; Schlottmann, 1999). Koslowski and colleagues (Koslowski, 1996; Koslowski & Okagaki, 1986; Koslowski, Okagaki, Lorenz, & Umbach, 1989) have conducted a series of experiments investigating this claim, finding that people consider causal relationships more plausible when supplied with a potential mechanism and less plausible when the most likely mechanisms are ruled out.

Recent work examining causal learning in adults has also noted the importance of prior expectations about the direction of causal relationships, particularly when people are simultaneously learning about multiple relationships. Waldmann and colleagues (Waldmann, 1996, 2000; Waldmann & Holyoak, 1992; Waldmann, Holyoak, & Fratianne, 1995) have conducted a number of studies that suggest that people's expectations about the causal structure among a set of variables can determine how covariational evidence affects their beliefs. For example, Waldman (2000) gave people information that suggested that the relationship among a set of variables was either a "common cause" relationship, with one variable causing several others, or a "common effect" relationship, with several variables all producing a single effect. People's beliefs about the underlying causal structure influence their interpretation of the pattern of covariation among the variables: Only those who believed in the common effect structure took into account competition between causes when evaluating their strength.

Functional Form

In physics, the *functional form* of causal relationships, such as how the velocity of one object depends on its mass and the mass and velocity of another object with which it collides, can be laid out precisely. The knowledge that guides most causal inferences is less precise, but even in the most basic cases of causal induction we draw on expectations as to whether the effects of one variable on another are positive or negative, whether multiple causes interact or are independent, and what type of events (binary, continuous, or rates) are relevant to evaluating causal relationships (Cheng, 1997; Novick & Cheng, 2004).

One setting in which questions about functional form have arisen explicitly is in examining how causes should be assumed to combine. Many theories of animal learning assume that multiple causes of a single effect combine additively, each making a constant contribution to the effect (e.g., Rescorla & Wagner, 1972). A number of researchers, including Shanks, Wasserman, and their colleagues, have advocated these linear models as accounts of human causal learning (e.g., López, Cobos, Caño, & Shanks, 1998; Shanks, 1995a, 1995b; Shanks & Dickinson, 1987; Wasserman, Elek, Chatlosh, & Baker, 1993). However, whether this assumption is appropriate for modeling human judgments seems to be affected by people's beliefs about what aspect of the causes produces the effect. Waldmann (2007) presented a study in which participants were told about a hypothetical experiment that found that drinking a yellow liquid increased the heart rate of animals by 3 points, while drinking a blue liquid increased the heart rate by 7 points. The participants were asked to predict the consequences of drinking a mixture of the two liquids. The results depended upon whether the participants were told that the effect of the drink was a consequence of its taste, or of its strength. More people produced predictions consistent with a weighted average of the effects if they believed the effect was modulated by strength, for which a linear functional form is more appropriate. Recent work has also shown that the magnitude of some effects that assume additivity is affected by the extent to which people believe causes combine additively (Beckers, De Houwer, Pineno, & Miller, 2005; Lovibond, Been, Mitchell, Bouton, & Frohart, 2003).

Perhaps the most comprehensive attempt to characterize the possible ways in which causes could combine is that of Kelley (1973), who suggested that causal induction from small numbers of observations may be guided by *causal schemas*. Kelley distinguished between generative and preventive causes, and he identified three schemas describing the interaction between generative causes: multiple sufficient causes, multiple necessary causes, and compensatory causes. Under the multiple sufficient causes schema, the effect occurs in the presence of any one of the causes (the equivalent of a logical OR function). In the multiple necessary causes schema, the effect occurs only if all of the causes are present (the equivalent of a logical AND function). In the compensatory causes schema, increasing the strength of each cause increases the tendency for the effect to be expressed. All of these schemas constitute different assertions about the functional form of the relationship between cause and effect, and knowing which of these schemas is relevant in a particular situation can facilitate evaluating whether a particular causal relationship exists.

The functional form of causal relationships becomes most important when dealing with causal inferences in physical systems, where the ways in which one object influences another can be quite complex. Developmental psychologists have extensively investigated how well children understand the functional relationships that hold in physical systems. Shultz and Kestenbaum (1985) provided a review of some of this work. One interesting example of this project is provided by Zelazo and Shultz (1989), who investigated whether children understood the different functional relationships between the potency of a cause and the resistance of the effect in two systems: a balance beam, where one object was weighed against another, and a ramp, where one object slid down to displace another. For the balance beam, the magnitude of the effect depends upon the difference in the masses of the two objects, whereas for the ramp, it depends upon the ratio. Zelazo and Shultz (1989) found that although adults were sensitive to this difference, 5-year-olds tended to use a single functional form for both systems.

The functional form of a causal relationship can also determine the temporal coupling between cause and effect. The time between

the occurrence of a potential cause and the occurrence of an effect is a critical variable in many instances of causal induction. Several studies have explored covariation and temporal proximity as cues to causality in children, typically finding that the event that immediately precedes an effect is most likely to be perceived as the cause, even if there is covariational evidence to the contrary (e.g., Shultz, Fisher, Pratt, & Rulf, 1986). Hagmayer and Waldmann (2002) presented an elegant series of studies that showed that different assumptions about the delay between cause and effect could lead to different interpretation of the same set of events, determining which events were assumed to be related. Similar phenomena have recently been investigated in detail by Lagnado and Sloman (2006) and Buehner and colleagues (Buehner & May, 2002, 2003; Greville & Buehner, 2007). Finally, Anderson (1990) provided a computational analysis of data involving the interaction between spatial separation and temporal contiguity in causal induction.

Summary

The three aspects of prior knowledge identified in this section can support strong expectations about possible causal relationships. Having an ontology, knowing the plausibility of relationships among the entities identified within that ontology, and knowing the functional form of those relationships provides information that makes it possible to generalize about the causal relationships among completely new variables. The research we have summarized in this section makes a compelling case for an influence of prior knowledge on causal induction but raises the question of exactly how this knowledge should be combined with the evidence provided by the data observed by learners. Answering this question is the project undertaken in the remainder of the article. Our next step toward obtaining an answer is to understand the computational problem underlying causal induction, which is the focus of the next section.

A Computational-Level Analysis of Causal Induction

The aim of this article is to provide a computational-level analysis of causal induction, in the sense introduced by Marr (1982). This section begins with a discussion of what such an analysis means, clarifying our motivation and methodology. We then turn to the question of how to formulate the computational problem underlying causal induction. Our formulation of this problem makes use of causal graphical models, a formalism for representing, reasoning with, and learning about causal relationships developed in computer science and statistics (Pearl, 2000; Spirtes, Glymour, & Scheines, 1993). We introduce this formalism and use it to clearly state the problem faced by causal learners. We then consider existing rational solutions to this problem, arguing that while they feature one of the two factors that are necessary to explain human causal induction—statistical inference—they do not incorporate the kind of prior knowledge described in the previous section. Reflecting upon the nature of this knowledge leads us to argue that a level of representation that goes beyond causal graphical models will be required.

Analyzing Causal Induction at the Computational Level

Marr (1982) distinguished between three levels at which an information processing system can be analyzed: the levels of computational theory, representation and algorithm, and hardware implementation. Analyses at the first of these levels answer the question "What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?" (Marr, 1982, p. 25). This is a question about the abstract problem that the information processing system is trying to solve and what solutions to that problem might look like. One part of a computational-level analysis is thus considering the form of rational solutions to a problem faced by the learner, a strategy that is also reflected in Shepard's (1987) search for universals laws of cognition—laws that must hold true for any information-processing system due to the structure of the problem being solved—and Anderson's (1990) formulation of rational analysis.

In the context of causal induction, a computational-level analysis seeks to identify the abstract problem being solved when people are learning about causal relationships and to understand the logic that makes it possible to solve this problem. Since our aim is to provide a unifying account of causal induction across a range of settings, we want to define the underlying computational problem in as broad a way as possible, highlighting the fact that a single solution can be applied across these domains. A major challenge of this approach is finding a formal framework that can provide a solution to the underlying computational problem, compounded by the fact that we want more than just any solution: We want a solution that is optimal for the problem being posed.

Developing an account of causal induction that is an optimal solution to the underlying computational problem is attractive not just as a unifying account, but as a way of answering three kinds of questions about human cognition. The first kind of question is a "How possibly?" question: How could people possibly solve the problem of inferring causal relationships from observational data? Many aspects of human learning, including causal induction, seem far better than that of any automated systems. To repeat an example from the introduction, it is commonplace for people to draw correct conclusions about causal relationships from far less data than we might need to do a statistical test. Understanding how we might explain such inferences in rational terms helps us understand how it is that people are so good at them, and what factors—such as domain knowledge—play a role in this success.

The second kind of question we can answer using this kind of analysis is a "How should it be done?" question: Given the computational problems that people face, what should they be doing to solve those problems? Optimal solutions tell us something about the properties that we might expect to see in the behavior of intelligent organisms, and can thus tell us which aspects of that behavior might be purely a consequence of the nature of the problems being solved. This kind of strategy is common in vision science, where "ideal observer" models have helped reveal how much of human perception might be explained as an optimal response to the structure of the environment (Yuille & Kersten, 2006). In the case of causal induction, a critical issue is combining statistical evidence with prior knowledge, and a rational account can indicate how this should be done, and what the consequences should be for the inferences that people make. Although research in judgment and decision-making has illustrated that people often deviate from the predictions of rational models (e.g., Tversky & Kahneman, 1974), the revelations that this work has made about the psychological mechanisms involved were partially made possible by the existence of a well-developed account of how a

rational agent should make decisions. For other complex problems such as causal induction, we are only just beginning to develop these rational accounts, and understanding what people should do when making causal inferences will be a valuable tool in determining how people actually solve this problem.

Finally, a third, related, question we can answer is a "What is necessary?" question: What knowledge or other constraints on hypotheses would an ideal learner need in order to reach the same conclusions as people? Since an ideal learner makes the best use of the available data, the answer to this question places a lower bound on the kind of constraints that human learners might use. Understanding the impact of different kinds of prior knowledge on causal induction by analyzing their effects on an ideal learner gives us a way to predict the role that these kinds of knowledge might play in human causal induction.

Answering these three questions requires not just defining a problem and deriving a solution, but arguing that this problem and solution connect to human causal learning. This connection can be established only by comparing the predictions of models developed within our formal framework to the results of experiments with human participants. The empirical results that will be relevant to this argument are those that are framed at the same level of abstraction as our analysis: results that indicate what conclusions people reach given particular data. As a consequence, we focus mainly on static measurements of beliefs about causal relationships, rather than capturing the dynamics of human learning, although we have explored this topic in the past (Danks, Griffiths, & Tenenbaum, 2003) and view it as an important direction for future research. The goal of our models is to produce the same conclusions from the same data, and our framework will be successful if it allows us to define models that incorporate the kinds of knowledge that make this possible.

In pursuing a computational-level analysis, we are not trying to make claims about the other levels at which causal induction might be analyzed. In particular, we are not asserting that particular representations or algorithms are necessary, or making other commitments as to the mechanisms or the psychological processes involved. Marr (1982) argued that different levels of analysis will provide constraints on one another, with the computational level indicating what kinds of representations and algorithms will be appropriate for solving a problem. In the context of causal induction, we anticipate that many different psychological mechanisms could result in behavior similar to the predictions made by specific models we consider, with associative learning, heuristics, or explicit hypothesis testing being good strategies for individual tasks, and we briefly outline some possible psychological mechanisms in the Discussion. Our general aim, however, is to provide a unifying account at the more abstract level of the underlying problem and its solution, ultimately helping to explain why particular representations and algorithms might be appropriate in a particular task.

Finally, our aim of providing a computational-level account of causal induction also influences the kinds of models that we use for comparison. In this article, our emphasis is on comparison of the predictions of our account to those of other rational models. These models all use the same formal ideas and operate at the same level of analysis, but they differ in their assumptions about the knowledge that informs causal induction or the nature of statistical learning. Comparison with these other rational models thus helps to highlight which components of our framework are relevant to explaining behavior on a given task. We do not doubt that it is possible to define better models of specific tasks, since presumably an accurate model of the actual mechanisms people use to solve these problems will make better predictions than the abstract kind of analyses obtained from our framework. Ultimately, we see the key criterion for the success of our approach to be its usefulness in capturing the effects of prior knowledge on causal induction across a wide range of settings, and it is this criterion that we have in mind when we evaluate the performance of individual models. In this way, we expect that our framework will be evaluated in the same fashion as other general approaches that can be used to define a variety of computational models, such as parallel distributed processing (e.g., McClelland & Rumelhart, 1986) or production systems (e.g., Anderson, 1993).

Causal Graphical Models

Having introduced our motivation and methodology, we now turn to the question of how to formulate the computational problem posed by causal induction. We will do this using causal graphical models, also known as Bayesian networks or Bayes nets. Causal graphical models have recently begun to be used in psychological accounts of causality (e.g., Danks & McKenzie, 2009; Glymour, 1998, 2001; Gopnik et al., 2004; Griffiths & Tenenbaum, 2005; Lagnado & Sloman, 2002; Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2006, 2007, 2008; Rehder, 2003; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003; Tenenbaum & Griffiths, 2001, 2003; Waldmann & Martignon, 1998). In this article, we highlight only the elements of causal graphical models that are relevant to our account. More detailed introductions are provided by Pearl (2000), Heckerman (1998), Glymour (2001), and Sloman (2005).

A causal graphical model has three components: a set of variables, a causal structure defined upon those variables, and a set of assumptions about the functional form of the relationships indicated by this structure. The variables are represented by nodes in a graph. These nodes are connected by arrows, indicating the direction of causal dependencies among the variables. Assumptions about the functional form of causal relationships make it possible to use this graphical structure to reason about the probabilities of different kinds of events. The functional form defines a probability distribution for each variable conditioned on its causes, which is referred to as the *parameterization* of the nodes.

Causal graphical models can be used to compute the probability of observing particular values for the variables and the consequences of interventions. An intervention is an event in which a variable is forced to hold a value, independent of any other variables on which it might depend. Following Pearl (2000), we denote intervention that sets a variable X to value x with $\mathrm{do}(x)$, and in general use uppercase letters to indicate variables and lowercase letters to indicate their values. Probabilistic inference on a modified graph, in which incoming edges to X are removed, can be used to assess the consequences of intervening on X (Pearl, 2000; Spirtes et al., 1993).

The Computational Problem and Existing Solutions

Causal graphical models provide us with the tools to give a precise definition of the computational problem underlying causal

induction. We take the problem of causal induction as that of identifying the causal graphical model—including both structure and parameters—responsible for generating the observed data *D*. This problem has been extensively explored in the literature on causal graphical models in computer science and statistics, and it is typically divided into two parts (e.g., Griffiths & Tenenbaum, 2005; Heckerman, 1998): structure learning and parameter estimation. We discuss these parts in turn, highlighting connections to existing proposals about human causal induction.

Structure learning. Learning the causal structure that relates a large number of variables is a difficult computational problem, as the number of possible structures increases exponentially with the number of variables. Research in computer science and statistics has focused on two strategies for solving this problem. *Constraint-based* algorithms attempt to identify causal structure on the basis of the patterns of dependency exhibited by a set of variables, whereas *Bayesian* methods evaluate the probability that a particular structure generated the observed data.

Constraint-based algorithms for structure learning (e.g., Pearl, 2000; Spirtes et al., 1993) proceed in two steps. First, standard statistical tests such as Pearson's χ^2 test are used to identify which variables are dependent and independent. Since different causal structures should result in different patterns of dependency among variables, the observed dependencies provide constraints on the set of possible causal structures. The second step of the algorithms identifies this set, reasoning deductively from the pattern of dependencies. The result is one or more causal structures that are consistent with the dependencies exhibited by the data. By not making any commitments about the consequences of causal relationships other than statistical dependency, constraint-based algorithms provide a general-purpose tool for causal induction that can be applied easily across many domains. This generality is part of the appeal of these algorithms as psychological theories, as they provide a way to explain the acquisition of causal knowledge without recourse to domain-specific mechanisms (e.g., Gopnik & Glymour, 2002; Gopnik et al., 2004).

The Bayesian approach to structure learning (Cooper & Herskovits, 1992; see Heckerman, 1998) treats causal induction as a special case of the more general statistical problem of identifying the statistical model most likely to have generated observed data. Bayesian inference provides a solution to this problem. The heart of this solution is Bayes' rule, which can be used to evaluate the probability that a hypothetical model *h* was responsible for generating data *D*. The *posterior* distribution, $P(h|D)$, is evaluated by combining *prior* beliefs about the probability that *h* might generate any data, encoded in the distribution $P(h)$, with the probability of *D* under the model *h*, $P(D|h)$, typically referred to as the *likelihood*. Bayes' rule stipulates how these probabilities should be combined, giving

$$P(h|D) = \frac{P(D|h)P(h)}{\sum_{h' \in H} P(D|h')P(h')}, \quad (1)$$

where *H* is the *hypothesis space*, the set of all models that could possibly have produced *D*.

As with any Bayesian inference, this approach requires specifying a prior probability and a likelihood for every hypothesis within a hypothesis space, *H*. In typical applications of this method, *H* consists of all directed graphs defined over the available variables. The data *D* consist of the values that those variables assume as the result of observation and intervention. Standard Bayesian structure-learning algorithms define $P(D|h)$ in a way that makes very weak assumptions about the functional form of the relationship between causes and effects. A separate parameter is used to express the probability of the effect for each configuration of its causes, meaning that the causes can have essentially any kind of influence on the effect—generative or preventive, large or small—and can combine in any imaginable way. $P(D|h)$ is evaluated by defining a distribution over these parameters, and then integrating over the specific values the parameters take on (e.g., Cooper & Herskovits, 1992). This makes it possible to compute the probability of the data given a particular graphical structure without committing to a particular choice of parameter values. The prior over graph structures, $P(h)$, is typically either uniform (giving equal probability to all graphs), or gives lower probability to more complex structures. Algorithms that use these principles differ in whether they then proceed by searching the space of structures to find that with the highest posterior probability (Friedman, 1997), or evaluate particular causal relationships by integrating over the posterior distribution over graphs (Friedman & Koller, 2000). Tenenbaum and Griffiths (2001; Griffiths & Tenenbaum, 2005) developed a model of human causal induction based on the principles of Bayesian structure learning, which we discuss in more detail later in the article.

Parameter estimation. Parameter estimation assumes a fixed causal structure and aims to identify the parameters that specify the probability of a variable given the values of the variables that influence it. The simplest way to parameterize a causal graphical model is to use a separate parameter for the probability of each value of a variable given the values of its causes—something that we refer to as the *generic* parameterization. This generic parameterization is the one typically used in Bayesian structure learning in computer science, as discussed above. An alternative is to make a stronger set of assumptions about the way in which causes combine to produce their effects. One such set of assumptions yields the noisy-OR function, which is widely used in computer science and statistics (Pearl, 1988) and is a key part of a prominent model of human causal induction (Cheng, 1997).

The noisy-OR function results from a natural set of assumptions about the relationship between cause and effect: that causes are generative, increasing the probability of the effect, that the effect occurs in the absence of any causes with a constant probability w_0, that each cause produces the effect with a constant probability w_i, and that the opportunities for the causes to produce the effect are independent (Cheng, 1997). For example, if we had an effect variable *E* and a cause variable *C*, then the conditional probability of *E* given *C* would be

$$P(e^+|c; w_0, w_1) = 1 - (1 - w_0)(1 - w_1)^c, \quad (2)$$

where w_1 is a parameter associated with the strength of *C* and *c* takes on values $c^+ = 1$ in the presence of the cause or $c^- = 0$ in its absence. This expression gives w_0 for the probability of *E* in the absence of *C*, and $w_0 + w_1 - w_0w_1$ for the probability of *E* in the presence of *C*. This parameterization is called a noisy-OR because if w_0 and w_1 are both 1, Equation 2 reduces to the logical OR function: The effect occurs if and only if either some background factor or *C* is present. With w_0 and w_1 in the range [0, 1] it generalizes this function to allow probabilistic causal relationships.

If E had multiple parents $X_1, \ldots, X_n$, we could associate a separate strength w_i with each parent, and the noisy-OR parameterization would give

$$P(e^+|x_1, \ldots, x_n; w_0, w_1, \ldots, w_n) = 1 - (1 - w_0)\prod_i (1 - w_i)^{x_i}, \quad (3)$$

where again $x_i = 1$ if X_i is present, and 0 if X_i is absent.

A simple solution to the problem of estimating the parameters of a causal graphical model is to use maximum-likelihood estimation, choosing the values of the parameters that maximize the probability of the observed data. For the case of the noisy-OR function with a single cause, the maximum-likelihood estimate of w_1 is

$$\text{power} = \frac{P(e^+|c^+) - P(e^+|c^-)}{1 - P(e^+|c^-)}, \quad (4)$$

where $P(e^+|c^+)$ is the empirical conditional probability of the effect given the presence of the cause. We have labeled this quantity *power* as it corresponds to Cheng's (1997) definition of causal power, proposed as a rational model of human causal induction. Glymour (1998) pointed out that the assumptions Cheng (1997) used in deriving this model are equivalent to those underlying the noisy-OR parameterization, and Tenenbaum and Griffiths (2001; Griffiths & Tenenbaum, 2005) showed that causal power is a maximum-likelihood estimator of w_1.

The numerator of Equation 4 has also been proposed in its own right as a model of human causal induction, being known as ΔP. This quantity,

$$\Delta P = P(e^+|c^+) - P(e^+|c^-),$$

reflects the change in the probability of the effect occurring as a consequence of the occurrence of the cause. This measure was first suggested by Jenkins and Ward (1965), was subsequently explored by Allan (1980, 1993; Allan & Jenkins, 1983), and has appeared in various forms in both psychology and philosophy (Cheng & Holyoak, 1995; Cheng & Novick, 1990, 1992; Melz, Cheng, Holyoak, & Waldmann, 1993; Salmon, 1980). ΔP can also be shown to be a rational solution to the problem of estimating the strength of a causal relationship, assuming that causes combine linearly (Griffiths & Tenenbaum, 2005; Tenenbaum & Griffiths, 2001).

Two Challenges for a Formal Framework

This brief survey of existing methods for solving the problem of identifying the causal graphical model that generated observed data—and corresponding rational models of human causal induction—highlights two challenges for the kind of formal framework we aim to develop. First, this framework should naturally capture the effects of knowledge on causal induction. Existing approaches make either weak or generic assumptions about the nature of the knowledge that people use in evaluating causal relationships and are consequently limited in their ability to account for the effects of ontology, plausibility, and functional form outlined in the previous section. Second, the framework should be broad enough to encompass learning of both causal structure and the parameters that describe a given causal relationship. We now discuss these issues in turn, arguing that both can be addressed by adopting a more general Bayesian framework.

Capturing the effects of prior knowledge. The approaches to structure learning and parameter estimation outlined above all make either weak (in the case of structure learning) or general-purpose (in the case of parameter estimation) assumptions about the nature of causal relationships. These assumptions are incompatible with the richness of human knowledge about causal relationships and the corresponding flexibility of human causal induction exhibited in the examples discussed in the previous section. In part, this is a consequence of the context in which these approaches were developed. In statistics and computer science, developing algorithms that make minimal assumptions about the nature of causal relationships maximizes the number of settings in which those algorithms can be used. Psychological models of causal induction have justified making general-purpose assumptions about the nature of causal relationships through the expectation that it will be relatively straightforward to integrate the effects of prior knowledge into the resulting models. For example, Cheng (1997, p. 370) stated:

> The assumption that causal induction and the influence of domain-specific prior causal knowledge are separable processes is justified by numerous experiments in which the influence of such knowledge can be largely ignored . . . . The results of these experiments demonstrate that the induction component can indeed operate independently of prior causal knowledge.

In the few cases where formal accounts of the integration of prior knowledge and data have been explored (e.g., Alloy & Tabachnik, 1984; Lien & Cheng, 2000), these accounts have focused on just one aspect of prior knowledge, such as the plausibility of causal relationships or the level of the ontology at which those relationships should be represented.

Constraint-based structure-learning algorithms are particularly limited in their use of prior knowledge.[2] Again, this is partly by design, being a result of the data-driven, bottom-up approach to causal induction that these algorithms instantiate in a particularly clear way. As these algorithms are defined, they use only a weak form of prior knowledge—the knowledge that particular causal relationships do or do not exist (e.g., Spirtes et al., 1993). They do not use prior knowledge concerning the underlying ontology, the plausibility of relationships, or their functional form. This insensitivity to prior knowledge has previously been pointed out by some critics of constraint-based algorithms in computer science and statistics (Humphreys & Freedman, 1996; Korb & Wallace, 1997). Prior knowledge provides essential guidance to human inferences, making it possible to infer causal relationships from very small samples. Without it, constraint-based algorithms require relatively large amounts of data in order to detect a causal relationship—enough to obtain statistically significant results from a statistical significance test.

[2] It might not be impossible to develop a more global constraint-based framework for causal induction, defined over richer representations of prior knowledge and integrating both bottom-up and top-down information in a more holistic style of inference. However, this would be a major departure from how constraint-based approaches have traditionally been developed (Spirtes et al., 1993; Glymour, 2001).

The need for relatively large amounts of data is compounded by the fact that constraint-based algorithms cannot combine weak sources of evidence or maintain graded degrees of belief. This is a direct consequence of the policy of first conducting statistical tests, then reasoning deductively from the results. Statistical tests impose an arbitrary threshold on the evidence that data provide for a causal relationship. Using such a threshold is a violation of what Marr (1982) termed the *principle of least commitment*, making it hard to combine multiple weak sources of evidence. The binarization of evidence is carried forward by deductively reasoning from the observed patterns of dependency. Such a process means that a particular causal structure can be identified only as consistent or inconsistent with the data, admitting no graded degrees of belief that might be updated through the acquisition of further evidence.

Although Bayesian structure learning can deal with weak evidence and graded degrees of belief, the standard assumptions about priors, likelihoods, and hypothesis spaces mean that this approach is just as limited in its treatment of prior knowledge as constraint-based algorithms. However, it is relatively straightforward to modify this approach to incorporate the effects of prior knowledge. Different assumptions about the functional form of causal relationships can be captured by including models with different parameterizations in the hypothesis space, and the plausibility of causal relationships can be used in defining the prior probability of different graph structures. Recent work in computer science has begun to explore methods that use more complex ontologies, with each type of entity being characterized by a particular pattern of causal relationships with a particular functional form (e.g., Segal, Pe'er, Regev, Koller, & Friedman, 2003). This work is motivated by problems in bioinformatics that, as in many of the settings for human causal induction, require learning complex structures from limited data (e.g., Segal, Shapira, et al., 2003).

A similar strategy can be used to incorporate the effects of prior knowledge in parameter estimation, allowing the expectations of learners to influence their inferences about the strength of causal relationships. Maximum-likelihood estimation finds values for parameters based purely on the information contained in the data. This makes it hard for these models to incorporate the knowledge of learners into the resulting estimates. Bayesian estimation techniques provide a way to combine existing knowledge with data, through a prior distribution on the parameters. For example, when estimating the strength of a cause using the noisy-OR function, we might have a prior expectation that causal relationships will tend to be strong if they exist at all, corresponding to a prior distribution favoring large values of w_1. Lu et al. (2007, 2008) have developed a model of human causal learning based on Bayesian parameter estimation, using a general-purpose prior distribution favoring strong causal relationships.

Learning both causal structures and parameter values. The distinction between structure learning and parameter estimation is valuable when examining the assumptions behind different models of causal induction, but it is clear that both processes are key components of human learning. In previous work (Griffiths & Tenenbaum, 2005) we emphasized the importance of structure learning, in part because it was a component of causal induction that was not reflected in existing models, but we do not deny that people are capable of learning causal strength and that certain tasks are more likely to tap this ability than others. We provide a more detailed discussion of this point when we consider causal induction from contingency data, where the relevant phenomena are perhaps clearest. However, the framework that we develop needs to be sufficiently general that it can capture both of these aspects of human causal induction.

The work of Lu et al. (2007, 2008) illustrates how a Bayesian approach can be applied to the problem of estimating the strength of a causal relationship. This analysis casts the problem in a shared formal language with that of Bayesian structure learning, providing a simple way to develop a unifying framework. In Bayesian parameter estimation, the hypothesis space is the set of values for the parameters of a fixed causal structure. In Bayesian structure learning, the hypothesis space is the set of possible causal structures, evaluated by summing over the parameters. We can define a single framework in which both kinds of inferences can be made by defining our hypothesis space to consist of fully specified causal graphical models, each with both a structure and a full set of parameters. Using this hypothesis space, we can estimate the strength of a relationship by conditioning on a given structure and using the posterior distribution on the strength parameter, as is done by Lu et al. (2007, 2008). We can also answer a question about whether a particular causal relationship exists by summing over all hypotheses—including structure and parameters—and evaluating the probability of those hypotheses consistent with the existence of the relationship, similar to the approach taken by Griffiths and Tenenbaum (2005).

Summary. We have identified two challenges for a computational-level account of causal induction: incorporating prior knowledge and allowing both structure learning and parameter estimation. Both of these challenges seem to be something that can be addressed by adopting a more general Bayesian framework. Within this framework, the hypothesis space consists of a set of fully specified graphical models, each with a structure and a full set of parameters, and the knowledge of the learner is reflected in substantive assumptions about the prior probability of hypotheses, the predictions that hypotheses make about data that are instantiated in the likelihoods, and the selection of the hypotheses that compose the hypothesis space. This leaves us with a new problem: Where do the priors, likelihoods, and hypotheses that are used in making a particular inference come from? Or, more precisely: How can we formalize the knowledge about ontologies, plausible relations, and functional form that allows hypothesis spaces to be constructed? This is the question that we attempt to answer in the remainder of the article. First, however, we argue that this knowledge is something that cannot itself be captured in a causal graphical model.

Beyond Causal Graphical Models

Formulating the problem of causal induction as a Bayesian decision as to which causal graphical model generated observed data provides a precise specification of how prior knowledge could guide this inference. Knowledge about the ontology, plausibility, and functional form of causal relationships should influence the prior, likelihood, and hypothesis space for Bayesian inference. However, expressing this knowledge requires going beyond the representational capacities of causal graphical models. Although this knowledge can be *instantiated* in a causal graphical model, it generalizes over a set of such models, and thus cannot be *expressed* in any one model.

Our inability to express prior knowledge relevant to causal learning in the form of a causal graphical model is partly because of an inherent limitation in the expressive capacity of graphical models. Causal graphical models are formally equivalent to a probabilistic form of propositional logic (e.g., Russell & Norvig, 2002). A causal graphical model can be used to encode any probabilistic logical rule that refers to the properties of specific entities in the domain. However, causal graphical models cannot capture the fact that there are different types of entities, or the way that the types of entities involved in a potential relationship influence our expectations about the plausibility and functional form of that relationship. Such notions require going beyond causal graphical models and considering richer probabilistic logics.

The knowledge that constrains causal learning is at a higher level of abstraction than specific causal structures, just as the principles that form the grammar for a language are at a higher level of abstraction than specific sentences (Tenenbaum, Griffiths, & Niyogi, 2007). The syntactic structure of a single sentence cannot express the grammar of a language, which makes statements about the syntactic structures of the set of sentences that compose that language. More generally, making statements about sets requires defining abstract variables that can be instantiated in a given member of the set and quantifying over the values of those variables. These higher level abstractions and generalizations require adopting a representation that goes beyond that used by any member of the set itself.

The development of probabilistic predicate logic remains an open problem in artificial intelligence research (Friedman, Getoor, Koller, & Pfeffer, 1999; Kersting & De Raedt, 2000; Koller & Pfeffer, 1997; Milch, Marthi, & Russell, 2004; Muggleton, 1997). In the next section, we outline how some of the ideas behind this research can be used to develop a different level of representation for causal knowledge: a set of principles that can be used to guide inferences about the causal structure that was most likely to have generated observed data.

Theory-Based Causal Induction

So far, we have argued that human causal induction is affected by prior knowledge in the form of ontological assumptions, beliefs about the plausibility of causal relationships, and assertions about the functional form of those relationships. Causal graphical models provide us with a language in which we can express the computational problem underlying causal induction and embody a set of domain-general assumptions about the nature of causality (including, for example, the effects of intervening on a variable). However, causal graphical models are not sufficient to represent the domain-specific knowledge that guides human inferences. In this section, we develop a formal framework for analyzing how prior knowledge affects causal induction. First, we argue that the kind of knowledge that influences human causal induction fits the description of an *intuitive theory*, suggesting that the appropriate level of representation for capturing this knowledge is that of a causal theory. We then consider the function and content of such theories, arguing that theories can play the role of hypothesis space generators, and presenting a simple schema for causal theories that makes it easy to specify the information that is needed to generate a hypothesis space of causal graphical models.

Prior Knowledge and Causal Theories

Many cognitive scientists have suggested that human cognition and cognitive development can be understood by viewing knowledge as organized into intuitive theories, with a structure analogous to scientific theories (Carey, 1985a; Gopnik & Meltzoff, 1997; Karmiloff-Smith, 1988; Keil, 1989; Murphy & Medin, 1985). This approach has been used to explain people's intuitions in the biological (Atran, 1995; Inagaki & Hatano, 2002; Medin & Atran, 1999), physical (McCloskey, 1983), and social (Nichols & Stich, 2003; Wellman, 1990) domains and suggests some deep and interesting connections between issues in cognitive development and the philosophy of science (Carey, 1985a; Gopnik, 1996).

Although there are no formal accounts of intuitive theories, there is consensus on what kind of knowledge they incorporate: an ontology, indicating the types of entities that can be encountered in a given domain, and a set of causal laws expressing the relations that hold among these entities. For example, Carey (1985b) stated that:

> A theory consists of three interrelated components: a set of phenomena that are in its domain, the causal laws and other explanatory mechanisms in terms of which the phenomena are accounted for, and the concepts in terms of which the phenomena and explanatory apparatus are expressed. (p. 394)

When discussing causal theories, it is often productive to distinguish among different levels at which a theory might operate. In a philosophical work that has inspired much of the treatment of theories in cognitive development, Laudan (1977) made such a distinction, separating everyday scientific theory from higher level "research traditions." He characterizes a research tradition as consisting of

> an ontology which specifies, in a general way, the types of fundamental entities which exist in the domain or domains within which the research tradition is embedded . . . . Moreover, the research tradition outlines the different modes by which these entities can interact. (p. 79)

This distinction between these different levels of theory has been carried over into research on cognitive development, where Wellman (1990) and Wellman and Gelman (1992) distinguished between "specific" and "framework" theories:

> Specific theories are detailed scientific formulations about a delimited set of phenomena . . . framework theories outline the ontology and the basic causal devices for their specific theories, thereby defining a coherent form of reasoning about a particular set of phenomena. (p. 341)

All of these definitions draw upon the same elements—ontologies and causal laws.

The three aspects of prior knowledge that we have identified as playing a role in causal induction map loosely onto the content of intuitive theories identified in these definitions. The division of the entities in a domain into a set of different types is the role of an ontology, and causal laws identify which relationships are plausible and what form they take. This suggests that we might think of the knowledge that guides causal induction as being expressed in a causal theory. In particular, it is a theory that plays the role of a framework theory, providing a set of constraints that are used in

discovering the causal graphical model that describes a system, the analogue of a specific theory. Causal theories thus constitute a level of representation above that of causal graphical models, answering our question of how knowledge that is instantiated in a set of causal graphical models might be expressed. However, making this connection does not solve all of our problems: In order to have a complete formal framework for modeling human causal induction, we need to give an account of the function and content of these causal theories.

Theories as Hypothesis Space Generators

The Bayesian framework sketched in the previous section leaves us with the problem of specifying a hypothesis space, a prior on that space, and a likelihood for each hypothesis in that space. This problem can be solved by defining a probabilistic procedure for generating causal graphical models. Such a procedure needs to specify probability distributions from which the variables, structure, and parameterization of causal graphical models are drawn. The hypothesis space is the set of causal graphical models that can be generated by sampling from these distributions, the prior is the probability with which a given model is generated by this process, and the likelihood is determined by the parameterization of that model. By limiting which causal structures and parameterizations can be generated, it is possible to impose strong constraints on the hypotheses considered when reasoning about a causal system. We view this as the function of causal theories: They specify a recipe that can be used to generate hypothesis spaces for causal induction.

The commitments and consequences of this claim can be understood by extending the analogy between language comprehension and causal induction introduced in the previous section. Under this analogy, a theory plays the same role in solving the problem of causal induction that a grammar plays in language comprehension: Like a grammar, a theory generates the hypotheses used in induction. A schematic illustration of the correspondence between these two problems is shown in Figure 1. Under this view, the solution to the inductive problem of causal learning has the same character as identifying the syntactic structure of sentences: just as grammars generate a space of possible phrase structures, theories generate a space of possible causal graphical models. Causal learning is thus a problem of "parsing" the states of the variables in a system with respect to a causal theory. If the theory provides strong enough constraints, such parsing can be done swiftly and easily, picking out the causal structure that is most likely to have generated the data. Just as recent work in computational linguistics has emphasized the value of probabilistic approaches in solving such parsing problems (e.g., Chater & Manning, 2006; Manning & Schütze, 1999), the assumption that theories generate hypotheses and hypotheses generate data means that we can view each of these levels of representation as specifying a probability distribution over the level below. The result is a *hierarchical Bayesian model* (Tenenbaum, Griffiths, & Kemp, 2006), supporting probabilistic inference at all of these levels.

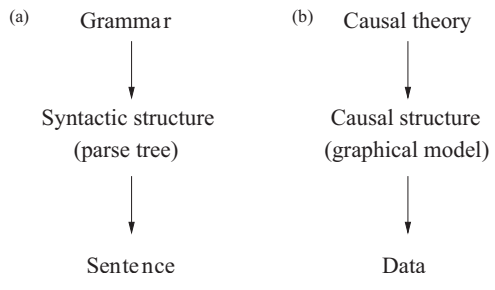

Figure 1. Three levels of representation in (a) language comprehension and (b) causal induction. Each level generates the level below, and language comprehension and causal induction both involve inferring the middle level based upon data below and constraints from above.

Formalizing the Content of Causal Theories

To specify the content of causal theories, we need to identify their basic constituents and explain how these are used to generate causal graphical models. When cognitive scientists appeal to an intuitive theory to explain the inferences that people make in a given domain, they typically mean a structured representation with causal content, similar in spirit to a scientific theory (e.g., Carey, 1985a). As discussed above, accounts in philosophy of science and cognitive development are more precise about the structure and content of such theories, seeing them as constructed from an ontology and causal laws (Carey, 1985b; Gopnik & Meltzoff, 1997; Wellman, 1990; Wellman & Gelman, 1992). Providing a formal treatment of causal theories that captures their richness and complexity, as well as the breadth of inferences that they are supposed to support, is a task that goes beyond the scope of this article. We formalize just the aspects of causal theories relevant to generating hypothesis spaces for causal induction. As a result, the theories we discuss are far less abstract that what is typically described in discussions of framework theories, being just one level above the observable variables. We consider the possibility of more abstract causal theories elsewhere (Tenenbaum et al., 2007).

The causal theories that we present in this article will have three components, corresponding to the three aspects of prior knowledge that influence causal induction identified above, and the three elements of the definition of a causal graphical model. These three components are an *ontology*, a set of principles that identify *plausible relations*, and a statement of the *functional form* of those relations. These three components of a theory each generate one part of a causal graphical model, being the variables, the causal structure, and the parameterization, respectively. We describe these components using a combination of probability statements and first-order logic. Since our aim is to produce a computational-level account of human causal induction, the specific choices we have made in using this formalism are not intended to be interpreted as assertions about the nature of the representations that people actually use when solving these problems, nor are the specific theories we present supposed to capture the full complexity of the information that people have available about these systems. However, we are committed to the level of representation (i.e., using a language that is richer than that of causal graphical models) and the constraints that are embodied in the theories, which are ultimately expressed as distributions over causal graphical models. In this spirit, we have used a variety of different formalisms for causal theories in other presentations of these ideas (Griffiths, Baraff, & Tenenbaum, 2004; Griffiths &

Tenenbaum, 2007b; Tenenbaum & Griffiths, 2003; Tenenbaum & Niyogi, 2003).

To provide some concreteness to our formalism for expressing causal theories, we consider a specific example throughout. Many empirical studies of causal induction (e.g., Buehner & Cheng, 1997; Buehner, Cheng, & Clifford, 2003; Lober & Shanks, 2000) use medical scenarios, such as evaluating the influence of chemicals on gene expression. These studies typically examine learning about a single causal relationship, such as whether injecting a particular chemical into mice causes them to express a particular gene. While simple, even these settings have enough structure that we can identify a nontrivial theory expressing prior knowledge that influences causal learning in this domain. We introduce our general framework in this simplest setting and then move to more complex settings, where richer prior knowledge allows learning from sparser data.

The first component of a theory, the ontology, identifies the types of entities that exist in a domain,[3] the number of entities of each type (or a distribution over this number), and the predicates that can be used to describe these entities. Such an ontology is far more limited than the kinds of ontologies considered in philosophy or computer science but is sufficient to express constraints on causal induction. In our example, where participants discover whether chemicals cause genes to be expressed in mice, there are three types of entities: **Chemical**, **Gene**, and **Mouse**. Any entity in the domain must belong to one of these three types. The number of entities of each type can either be stipulated or treated as a random variable drawn from a specific distribution. For example, we might state that the number of chemicals (N_C), the number of genes (N_G), and the number of mice (N_M) are drawn independently from distributions P_C, P_G, and P_M, respectively, but leave these distributions undefined—in many cases, N_C, N_G, and N_M will be apparent, and we need not be concerned about generating them. The predicates defined on these types state which properties and relations can take arguments of particular types, and what values those predicates can take on. In the example, these would include **Injected(Chemical, Mouse)**, indicating that a particular chemical was injected into a particular mouse, and **Expressed(Gene, Mouse)**, indicating that a particular gene was expressed in a particular mouse. Both of these predicates are Boolean, being either true or false. This ontology is summarized in Figure 2. The ontology required for this example is relatively simple, but the kind of knowledge that people have in other situations may be much more complex. For example, an ontology could be hierarchical, with objects belonging to types at multiple levels and predicates applying based upon the type at each of those levels (Griffiths & Tenenbaum, 2007b).

The second component of an intuitive theory is a set of rules that determine which causal relationships are plausible. These rules can be based upon the types of the entities involved or the predicates that apply to them. In the cases we consider, the rules will be based purely on types.[4] In our example, the structure of the problem is such that injecting chemicals does not cause injections of other chemicals, and neither does gene expression. The only relationships with which we concern ourselves are those between chemicals and genes. Figure 2 states a rule by which the plausibility of such relationships might be expressed, assigning a probability p to the existence of a causal relationship between a particular chemical and a particular gene, regardless of the mouse involved. All other causal relationships have probability 0.

The final component of an intuitive theory is a statement of the functional form that causal relationships are expected to possess. This requires specifying a parameterization (or distribution over parameterizations) for each predicate identified in the ontology. For the example, we need to define the probability that a particular mouse receives an injection of a particular chemical. This probability will not influence any of our subsequent analyses and thus is not specified: The theory indicates that this is a Bernoulli event, being true with some probability, but does not give the probability. In contrast, **Expressed(G,M)** is identified as a Bernoulli event with parameter ν, where ν is computed using a noisy-OR function, allowing each cause—in this case **Injected(C,M)** for some **C**—to have an independent opportunity to influence the effect with probability w_i. The parameters w_i are all assumed to be drawn from a uniform distribution, reflecting a lack of expectations about the strengths of the causes, and making our hypothesis space contain a continuum of causal graphical models in which the strength of the causal relationships varies between 0 and 1.

Ontology:

Types	Number
Chemical	$N_C \sim P_C$
Gene	$N_G \sim P_G$
Mouse	$N_M \sim P_M$

Predicates	Values	
Injected(**Chemical**, **Mouse**)	**Boolean:**	{**T**, **F**}
Expressed(**Gene**, **Mouse**)	**Boolean:**	{**T**, **F**}

Plausible relations:
Injected(**C**, **M**) → **Expressed**(**G**, **M**)
True for all **M** with probability p for each **C**, **G** pair

Functional form:

Injected(**C**, **M**) ∼ Bernoulli(·)
Expressed(**G**, **M**) ∼ Bernoulli(ν) for ν from a noisy-OR:

Cause	Strength
(Background)	$w_0 \sim \text{Uniform}(0, 1)$
Injected(**C**, **M**)	$w_i \sim \text{Uniform}(0, 1)$

Figure 2. Theory for causal induction from contingency data in a medical setting.

Generating a Hypothesis Space

The process by which a causal graphical model is generated from a theory is as follows:

1. *Generate variables.* Sample the number of entities of each type from the distribution specified in the **Ontology**.

[3] The term *type* is used here in the technical sense associated with a typed or many-sorted logic (e.g., Enderton, 1972). Types restrict quantifiers and the application of predicates, with each predicate being applicable only to entities of particular types.

[4] Defining the rules based purely on type results in simpler theories. More generally, we could allow predicates to play a role in determining whether causal relationships are plausible. In fact, this is done implicitly even when only type is used, since a typed logic can be reduced to standard propositional logic by introducing predicates that indicate type (e.g., Enderton, 1972). Pursuing this strategy requires distinguishing between predicates that participate in causal relationships and predicates that are used just to determine the plausibility of those relationships. The former are used to generate the variables of the causal graphical models, whereas the latter define the prior probability of each model (see Griffiths & Tenenbaum, 2007b, for an example of this).

Generate the complete set of grounded predicates for these entities. This is the set of variables that form the nodes of the graph.

2. *Generate structure.* Sample links between nodes using the probabilistic procedure stated in the **Plausible relations** component of the theory.

3. *Generate parameterization.* For each node, sample a parameterization as specified in the **Functional form** component of the theory, including the values of the relevant parameters.

This generative process defines a hypothesis space of fully specified causal graphical models, together with a prior probability and, via the parameterization, a likelihood for each model in that space.

We illustrate how this generative process works by using the theory given in Figure 2. We assume that the number of chemicals, genes, and mice involved in a particular experiment is known, and implicitly condition on this information. For example, we might have a single chemical **c**, a single gene **g**, and N_M mice $\mathbf{m}_1, \ldots, \mathbf{m}_{N_M}$. The set of grounded predicates is constructed by substituting all possible entities for the arguments of each predicate in the ontology. In our case, this set consists of N_M statements indicating whether **Injected(c,m$_i$)** holds of mouse $\mathbf{m}_i$, and N_M statements indicating whether **Expressed(g,m$_i$)** holds of mouse $\mathbf{m}_i$. We then have to consider possible causal structures on these $2N_M$ variables. Since the constraints on plausible relations are such that if **Injected(c,m$_i$)** causes **Expressed(g,m$_i$)** for some mouse $\mathbf{m}_i$, then it does so for all mice, we will simply draw an arrow between two variables C and E to indicate that the relationship between **Injected(c,M)** and **Expressed(g,M)** holds for all mice **M**. The constraints on plausible relations imply that the only possible causal relationship in this graphical model is that from **Injected(c,m$_i$)** to **Expressed(g,m$_i$)**, and that this relationship holds with probability p. The hypothesis space H thus consists of two causal structures: one in which **Injected(c,M)** causes **Expressed(g,M)**, which has prior probability p, and one in which **Injected(c,M)** does not cause **Expressed(g,M)**, which has prior probability $1 - p$. These are Graph 1 and Graph 0, respectively, shown at the top of Figure 3. We obtain our full hypothesis space of causal graphical models by augmenting these structures with parameters w_0 and w_1 drawn from the appropriate prior distribution.

The same procedure can be used to generate a hypothesis space of causal graphical models for any number of entities, as illustrated in Figure 3. For example, with two chemicals and two genes, the hypothesis space contains 16 causal graphical models, with the prior probabilities determined by the number of causal relationships expressed in the graph. The same causal theory can be used to define a hypothesis space for five chemicals and 10 genes, or 50 chemicals and a thousand genes, simply by applying the same abstract principles.

Summary

In this section we have outlined some of the key components of causal theories and explained how they can be used to generate hypothesis spaces of causal graphical models. In the following sections, we present a series of case studies illustrating how these components vary in different settings and how the constraints that they provide are essential to understanding how people learn about causal relationships. These case studies help to illustrate how our theory-based approach can provide a unifying account of a broad range of phenomena related to causal induction: The basic framework of statistical inference informed by causal theories remains constant, being a domain-general strategy, but the knowledge contained within the theories varies, allowing us to explain domain-specific inferences. The influence of the relevant prior knowledge increases as we go through the examples, starting with causal induction from contingency data where constraints from prior knowledge are typically weak, but can nonetheless have interesting consequences.

Causal Induction From Contingency Data

The aspect of human causal induction that has been studied most extensively by psychologists is that of inferring a single causal relationship from contingency data. Given information about the frequencies with which C and E co-occur, as summarized in Table 1, people are asked to assess the extent to which C causes E. A number of mathematical models have been proposed to explain how people use contingency data to evaluate causal relationships (e.g., Allan, 1980; Anderson, 1990; Anderson & Sheu, 1995; Cheng, 1997; Cheng & Novick, 1990, 1992; Jenkins & Ward, 1965; López et al., 1998; Shanks, 1995b). These models tend to downplay the role of prior knowledge, assuming that such knowledge serves to provide a set of candidate causes, but contingency data are used to evaluate those causes.

We provide an account of human causal induction from contingency data within our theory-based framework. Most experiments using contingency data select candidate causes and effects for which causal relationships are plausible. This uniformity of plausibility underlies claims about the separability of causal induction and prior knowledge and means that the aspect of causal theories that determines the plausibility of relationships will not be as relevant here as in other settings discussed later in the article. However, as our framework emphasizes, prior knowledge is not restricted to plausibility: It also determines assumptions about functional form. Our framework thus makes two claims about causal learning from contingency data: that variation in the assumed functional form should produce variation in human judgments and that causal inferences can be understood as Bayesian inferences about the causal models most likely to have generated the observed data.

The plan of this section is as follows. First, we analyze the problem of causal induction from contingency data using our theory-based framework. We then compare the results of this analysis with four rational models of human judgments, using this comparison to highlight the assumptions behind our analysis that are critical for predicting human judgments. We go on to test a prediction produced by this analysis, examining whether different tasks result in different treatment of the statistical problem underlying causal induction. Finally, we discuss how approaches formulating the problem of causal induction in terms of structure learning and parameter estimation can both be captured within our

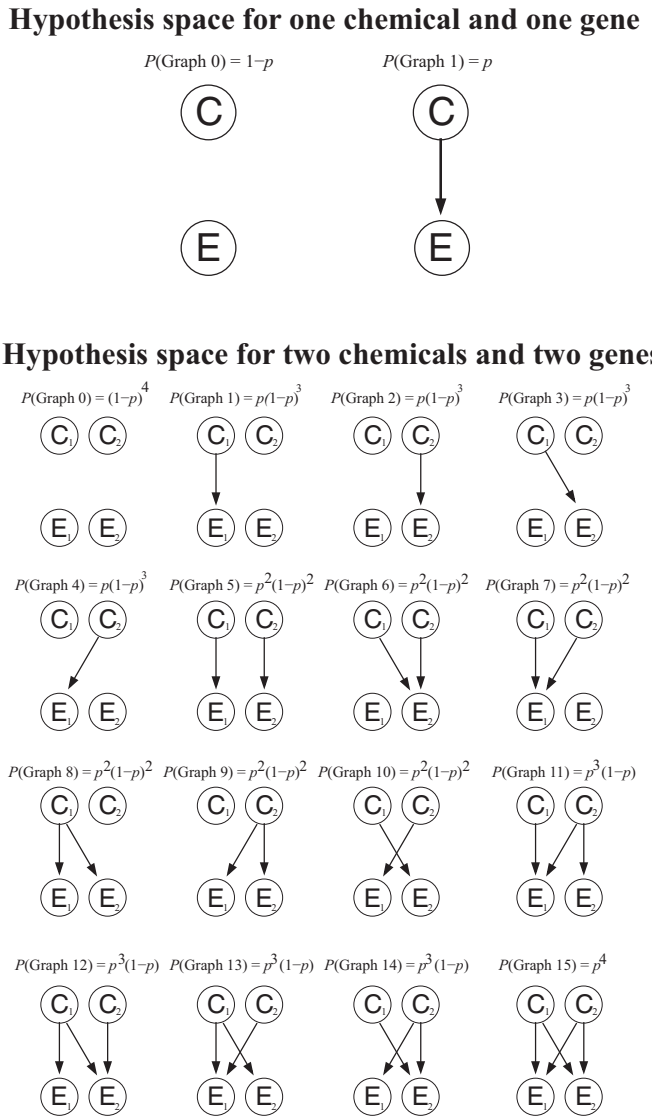

Figure 3. Hypothesis spaces generated by the theory shown in Figure 2. The top of the figure shows the hypothesis space for one chemical and one gene, which includes only two causal structures. With two chemicals and two genes, the hypothesis space includes 16 causal structures, as shown in the lower portion of the figure. In the graphs, C corresponds to **Injected(C,M)** for Chemical *C* and E corresponds to **Expressed(G,M)** for Gene *G*, both for some mouse *M*. C_1, C_2, E_1, and E_2 should be interpreted similarly. p indicates the probability of a causal relationship existing between a given chemical and gene, as outlined in the theory shown in Figure 2.

framework, consider different strategies that may manifest in the data, and compare our model to a simple heuristic account.

Theory-Based Causal Induction

Buehner and Cheng (1997; see also Buehner, Cheng, & Clifford, 2003) conducted an experiment in causal induction from contingency data using a medical scenario similar to that outlined in the previous section: People were asked to rate how strongly they thought the particular rays cause mutation on a scale from 0 (*the ray does not cause mutation at all*) to 100 (*the rays cause mutation every time*). Rays and viruses play roles that are directly analogous to chemicals and genes in our example, and expectations about their causal interactions can be captured using the same causal theory. The experiment used a design in which 15 sets of contingencies expressed all possible combinations of $P(e^+|c^-)$ and ΔP in increments of .25. Experiments were conducted with both generative causes, for which *C* potentially increases the frequency of *E*, and preventive causes, for which *C* potentially decreases the frequency of *E*. The results of Buehner and Cheng (1997, Experiment 1B), which used generative causes, are shown in Figure 4.

The theory given in Figure 2 can be used to generate a hypothesis space of causal models expressing the different kinds of

Table 1
Contingency Table Representation Used in Causal Induction

	Effect	
Cause	Present (e^+)	Absent (e^-)
Present (c^+)	$N(e^+c^+)$	$N(e^-c^+)$
Absent (c^-)	$N(e^+c^-)$	$N(e^-c^-)$

Note. $N(\cdot)$ denotes the frequency of a particular event.

structure that might explain the observed data, substituting rays and viruses for chemicals and genes appropriately. For each set of contingencies we have a single ray and a single virus, so the hypothesis space H contains the two models Graph 0 and Graph 1 shown at the top of Figure 3. The first contribution that the theory makes to this inference is thus to narrow down the hypothesis space of possible causal structures to just two, ruling out other structures in which viruses cause the application of rays or hidden common causes exist. According to the theory, these two models should use the noisy-OR parameterization, with each cause independently having the chance to influence the effect. In fact, each of these graphical structures corresponds to a continuum of causal models, each with different values of w_0 and w_1, with the prior being uniform over these values.

We can explore a variety of ways of answering the question that participants were asked in this experiment. While the question is nominally about the strength of the causal relationship, it is relatively ambiguous and previous work has suggested that people can still produce responses consistent with making an inference about causal structure when faced with such a question (Griffiths & Tenenbaum, 2005). For consistency with the other models we consider in this article, we focus for now on modeling these results as the consequence of a structural inference, although we return to the issue of the roles of structure and strength in causal induction later in the section.

Under this hypothesis space, the probability that a causal relationship exists is the probability that Graph 1 was responsible for generating the data, D, being the frequencies with which cause and effect co-occurred. We thus want to compute

$$P(\text{Graph } 1|D) = \frac{P(D|\text{Graph } 1)P(\text{Graph } 1)}{P(D|\text{Graph } 1)P(\text{Graph } 1) + P(D|\text{Graph } 0)P(\text{Graph } 0)}, \quad (5)$$

which is simply Bayes' rule, applied to the two hypotheses Graph 0 and Graph 1. The theory defines $P(\text{Graph } 1)$ and $P(\text{Graph } 0)$ via the parameter p, which we set to .5. Computing $P(D|\text{Graph } 1)$ and $P(D|\text{Graph } 0)$ requires integrating over the values of the parameters of these models. The details of computing such integrals are discussed in Griffiths and Tenenbaum (2005), where essentially the same model is used to account for a range of phenomena in causal induction from contingency data.[5]

Figure 4 displays $P(\text{Graph } 1|D)$ for different sets of contingencies D, under this noisy-OR model. There is a remarkably close correspondence between the predictions of the model and people's judgments. The rank-order correlation between the probabilities and human responses is $\rho = .971$. The model also seems to capture all of the important trends in the data, including the curious trend shown by the leftmost five stimuli in the figure: People's judgments increase as $P(e^+|c^-)$ increases, despite the fact that there is no difference in the probability of the effect in the presence and absence of the cause. This phenomenon is known as the *frequency illusion* or the *outcome density effect* (Allan & Jenkins, 1983; Shanks, López, Darby, & Dickinson, 1996), and, as we shall see shortly, presents a challenge for other models of causal induction.

Alternative Accounts

We can gain insight into the critical features of this account of causal induction from contingency data by comparing its predictions to those of other models. Our theory-based account assumes that people approach causal induction as a decision between different causal structures and that their intuitive theory involves a functional form equivalent to the noisy-OR. We consider two classes of models that can be used to evaluate these assumptions. One class treats causal induction as a problem of estimating the strength of a causal relationship, rather than inferring the underlying causal structure. The two models in this class are the rational models of causal induction from contingency data introduced above, causal power (Cheng, 1997), and ΔP (Jenkins & Ward, 1965). The other class of models focuses on learning causal structure but does not make strong assumptions about the functional form of the underlying causal relationship. The two models in this class correspond to the two approaches to structure learning that have been pursued by computer scientists and statisticians: χ^2, the frequentist test of statistical dependence used in constraint-based algorithms, and Bayesian inference comparing causal models with a generic parameterization. Since these models have not been fully introduced, the next two paragraphs provide the relevant mathematical details.

The first step of constraint-based algorithms for learning causal structure is to evaluate the dependencies among a set of variables. This is typically done using the standard frequentist analysis of contingency tables, Pearson's (1904/1948) χ^2 test for independence. The use of the χ^2 test as a model of human causal judgment was suggested in the psychological literature (e.g., Allan, 1980) but was rejected on the grounds that it neglects the kind of asymmetry that is inherent in causal relationships, providing information solely about the existence of statistical dependency between the two variables (Allan, 1980; López et al., 1998; Shanks, 1995b). χ^2 also makes no commitment about the functional form of the relationship between cause and effect: It simply detects any kind of statistical dependency between C and E.

[5] The basic model we explore in this section expresses the assumptions behind the causal support model introduced by Griffiths and Tenenbaum (2005) in our formal framework. Causal support was compared to human ratings using a power function of the log-likelihood ratio, $\log\frac{P(D|\text{Graph } 1)}{P(D|\text{Graph } 0)}$, rather than the posterior probability $P(\text{Graph } 1|D)$. We choose to use the posterior probability here for simplicity, and for consistency with the other examples discussed in the article. The posterior probability can also be obtained via a logistic transformation of the log-likelihood ratio, so the two models are related by a monotonic transformation.

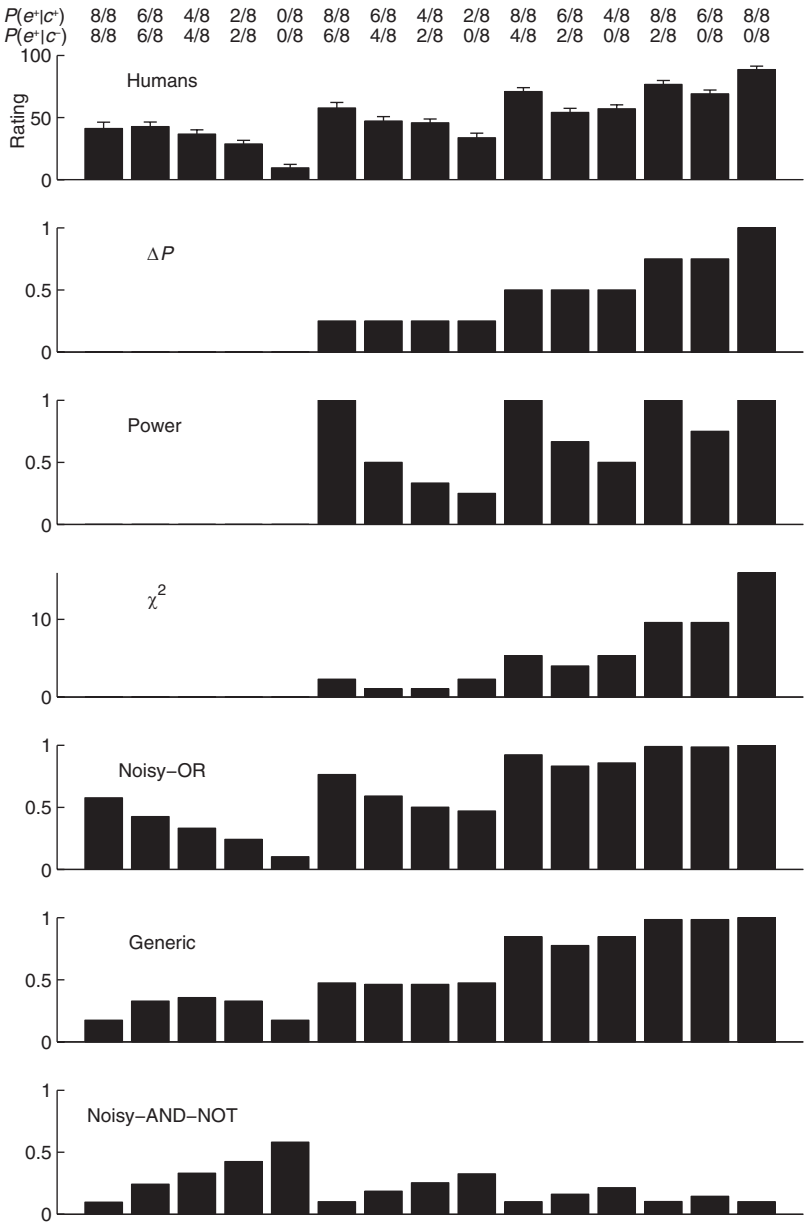

Figure 4. The top graph shows the results of Buehner and Cheng (1997, Experiment 1B). The other graphs show predictions for these contingencies for different models—ΔP, causal power (Cheng, 1997), and the χ^2 test for independence—together with our Bayesian model under three choices of parameterization (noisy-OR, generic, and noisy-AND-NOT). $P(e^+|c^+)$ and $P(e^+|c^-)$ indicate the probability of the effect in the presence and absence of the cause, respectively.

Our theory-based account assumes that the relationship between C and E uses the noisy-OR parameterization. This assumption can be relaxed by using the "generic" parameterization introduced above, simply defining the probability of the effect in the presence and absence of the cause by separate parameters, with $P(e^+|c^-) = w_0$ and $P(e^+|c^+) = w_1$. This parameterization makes no assumptions about the nature of the causal relationship between C and E, postulating simply that the two variables are dependent. By defining priors on these parameters, which we assume are uniform, and then integrating out the parameters, it is possible to compute $P(\text{Graph } 1|D)$ as in Equation 5. As in our theory-based account, we assume that $P(\text{Graph } 1) = .5$. This generic parameterization of a set of causal graphical models is widely used in Bayesian methods for learning causal structure (see Cooper & Herskovits, 1992; Heckerman, 1998) and is closely related to the rational model of causal induction described by Anderson (1990).

The importance of the assumptions of our theory-based account can be revealed by comparing its performance to that of these

different models. Causal power uses the functional form identified in Figure 2, but it models judgments as strength estimation. Comparing our model with causal power thus reveals the role of characterizing human inferences as an inference about causal structure rather than an inference about causal strength. χ^2 and the generic Bayesian approach are both concerned with causal structure but make weak assumptions about functional form. Comparison with these models reveals the importance of the assumptions about functional form embodied in the causal theory that underlies our model.

The predictions of all of these models are shown in Figure 4, and the rank-order correlations between models and human judgments are summarized in Table 2. None of the models had any free parameters, and the predictions were not transformed to fit the data because the rank-order correlation is unaffected by monotonic transformations. Our theory-based model matched or exceeded the performance of the other models. Both ΔP and causal power accounted for some of the trends in the data—for example, ΔP captures the increase in judgments as $P(e^+|c^+) - P(e^+|c^-)$ increases, and causal power generally captures the increase in judgments as $P(e^+|c^-)$ increases—but these trends are essentially orthogonal, and consequently neither model provides a complete account of the data. The theory-based noisy-OR model predicts both of these trends and is the only model that accurately predicts human judgments for the five stimuli on the left of the figure—those for which $\Delta P = 0$. People's assessment of the extent to which it seems that C causes E decreases as $P(e^+|c^-)$ decreases. The fact that none of the other models predicts these judgments reveals that these predictions are the direct consequence of combining a particular functional form (the noisy-OR) with Bayesian structure learning. We now examine why the theory-based model makes this prediction.

The Importance of Functional Form

The fact that human judgments decrease as $P(e^+|c^-)$ decreases when $\Delta P = 0$ seems counterintuitive: There is no difference in the probability of the effect in the presence or absence of the cause for any of these stimuli, so why should people change their belief that a causal relationship exists? In our theory-based account, the explanation for these predictions is not that there is decreasing evidence for a causal relationship as $P(e^+|c^-)$ decreases, but rather that there is no evidence for or against a causal relationship when $P(e^+|c^-) = 1$, and increasing evidence against a causal relationship as $P(e^+|c^-)$ decreases. This account depends on the assumption that the causal relationship—if it exists—is generative (increasing the probability of the effect, rather than preventing it). This assumption is a key part of the theory shown in Figure 2 and manifests in the use of the noisy-OR parameterization. At one extreme, when $\{P(e^+|c^+), P(e^+|c^-)\} = \{8/8, 8/8\}$, all viruses mutated irrespective of ray exposure, and it is clear that there is no evidence for a causal relationship. But there can also be no evidence against a (generative) causal relationship, because of a complete "ceiling" effect: It is impossible for the cause to increase the probability of E occurring above its baseline value when $P(e^+|c^-) = 1$. This uncertainty in causal judgment when $P(e^+|c^-) = 1$ and $\Delta P = 0$ is predicted by both our model, in which $P(\text{Graph } 1|D)$ is close to .5, and also (as Cheng, 1997, pointed out) by causal power, which is undefined for these contingencies.

Only the noisy-OR model, however, predicts the gradient of judgments as $P(e^+|c^-)$ decreases. $P(\text{Graph } 1|D)$ decreases as the ceiling effect weakens and the observation that $\Delta P = 0$ provides increasing evidence against a generative causal relationship. At the other extreme, when $P(e^+|c^-) = 0/8$, no unexposed viruses mutated, and there are eight opportunities for a causal relationship to manifest itself in the exposed viruses if such a relationship in fact exists. The fact that the effect does not appear in any exposed viruses, $P(e^+|c^+) = 0/8$, suggests that the ray does not cause virus mutation. The intermediate cases provide intermediate evidence against a causal relationship. The contingencies {2/8, 2/8} offer six chances for the cause to have an effect, and the fact that it never does so is slightly weaker evidence against a relationship than in the {0/8, 0/8} case, but more compelling than for {6/8, 6/8}, where

Table 2
Rank-Order Correlations for Different Models of Causal Induction

	Buehner and Cheng (1997)		Functional form experiment		
Model	Generative	Preventive	Generative	Difference	Preventive
Bayesian models					
Noisy-OR	**.961**	−.814	**.971**	.850	−.721
Generic	.876	.760	.957	**.975**	.930
Noisy-AND-NOT	−.868	**.893**	−.732	−.336	**.971**
ΔP					
Positive	.883	−.769	.968		−.946
Absolute value	.883	.769	.968	.968	.946
Negative	−.883	.769	−.968		.946
Power					
Generative	.942	−.531	.949		−.800
Preventive	−.698	.884	−.858		.971
χ^2	.880	.761	.966	.958	.940

Note. Boldface indicates the highest correlation in each column. Many of the correlations listed in the table represent combinations of models and tasks that have not been advocated in the literature, such as using the preventive version of causal power to model inferences about generative causes. We include these correlations for completeness, and because they provide a clear verification of the need for models assuming different functional forms for different tasks.

the cause has only two chances to manifest itself and the observation that $\Delta P = 0$ could easily be a coincidence. This gradient of uncertainty shapes the Bayesian structural inference that underlies our model, but it does not affect the maximum-likelihood parameter estimates underlying causal power or ΔP.

This explanation depends upon two aspects of our theory-based account: the general assumption that causal induction is a statistical inference comparing causal structures and the specific assumption that the appropriate functional form is the noisy-OR. The importance of these assumptions is supported by the failure of causal power, an estimate of the strength of a causal relationship under the noisy-OR, and the generic Bayesian structure-learning model. One prediction of this account is that people should alter their judgments in circumstances where a different functional form is appropriate. An opportunity to test this prediction comes from preventive causes. Since the noisy-OR only allows causes to increase the probability of their effects, a different parameterization is required to capture the properties of causes that decrease the probability of their effects. Buehner and Cheng (1997, Experiment 1A) used a design similar to that already described for generative causes to assess people's judgments about preventive causes. The results of this experiment are shown in Figure 5. From the figure, it can be seen that the trend at $\Delta P = 0$ is reversed for preventive causes: A decrease in $P(e^+|c^-)$ results in an increase in people's judgments.

Modeling these data requires making an assumption about the functional form of preventive causes. A simple theory of this kind would have the same content as the theory for generative causes shown in Figure 2, except for the assumptions about functional form, replacing the noisy-OR with the noisy-AND-NOT parameterization (Cheng, 1997; Novick & Cheng, 2004; Pearl, 1988). This parameterization follows from a set of assumptions similar to those made in the noisy-OR. In the case of Graph 1, these assumptions are that E occurs in the absence of C with probability w_0, and C independently prevents E from occurring with probability w_1. The resulting parameterization generalizes the logical statement that E will occur if the background factors are present and not C, allowing the influence of these factors to be probabilistic. The conditional probability can be written as

$$P(e^+|c; w_0, w_1) = w_0(1 - w_1)^c, \tag{6}$$

which gives w_0 for the probability of E in the absence of C and $w_0(1 - w_1)$ when C is present. As with the noisy-OR, both w_0 and w_1 are constrained to lie in the range [0, 1], and the function can be generalized to accommodate the influence of multiple parents.

Under a theory using the noisy-AND-NOT, the posterior probabilities of different causal structures can still be evaluated via Equation 5, but the different parameterization results in different probabilities for $P(D|\text{Graph 1})$ and $P(D|\text{Graph 0})$. First, we note that comparing structures with the noisy-AND-NOT parameterization provides a poor account of human judgments for generative causes, as shown in Figure 4 and Table 2. This model is actually strongly anticorrelated with human judgments for generative causes. However, the noisy-AND-NOT model gives a good account of human judgments for preventive causes, with $\rho = .893$. The predictions of this model are shown in Figure 5.

The alternative models discussed above can also be applied to this preventive setting. Rather than using ΔP to predict judgments about preventive causes, we use $-\Delta P$. Cheng (1997) defined causal power for preventive causes, following from a set of assumptions similar to those made for generative causes. The causal power for a preventive cause is

$$\text{power} = \frac{-\Delta P}{P(e^+|c^-)}. \tag{7}$$

The effect of $P(e^+|c^+)$ on causal power is the reverse of that for generative causes, with ΔP having a greater influence when $P(e^+|c^+)$ is small. Tenenbaum and Griffiths (2001; Griffiths & Tenenbaum, 2005) showed that this form of causal power is a maximum-likelihood estimate of the w_1 parameter for the noisy-AND-NOT. χ^2 and Bayesian structure learning with the generic parameterization are unaffected by the valence of a causal relationship and can be applied just as in the generative case.

The predictions of these models are shown in Figure 5, and their correlations with human judgments are given in Table 2. Bayesian structure learning with the noisy-AND-NOT outperformed all other models and gave a better account of human judgments than Bayesian structure learning with either the noisy-OR or the generic parameterization. This included accurately predicting the trend shown by the stimuli for which $\Delta P = 0$, which goes in the opposite direction from that for generative causes, but has the same statistical explanation—since causes decrease the probability of effects in the preventive setting, {8/8, 8/8} gives the most opportunities to discover that the cause does not influence the effect.

Inferences about generative causes are best captured using a noisy-OR parameterization, and inferences about preventive causes are best captured by the noisy-AND-NOT. This suggests that people make different assumptions about the functional form of generative and preventive causes and is consistent with the analysis of generative and preventive causal relationships given by Cheng (1997) and Novick and Cheng (2004). A further question raised by these results is whether there are any circumstances under which people's judgments will be best captured by the generic parameterization. The generic parameterization makes no assumptions about the nature of the relationship between cause and effect. Bayesian structure learning with this parameterization amounts to assessing whether there is any difference in the probability with which the effect occurs in the presence and absence of the cause. Consequently, we would expect that explicitly asking people to assess whether there is a difference in the probability with which the effect occurs under two different conditions will produce responses consistent with the generic parameterization.

To test this hypothesis, we ran an experiment in which people were asked to evaluate a set of contingencies as either evidence for a *generative* causal relationship between injecting a chemical into a mouse and a gene being expressed, evidence for a *preventive* causal relationship between injecting a chemical and a mouse catching a virus, or evidence for a *difference* in the probability of gene expression between two species of mice. The generative condition used the contingencies from Buehner and Cheng (1997, Experiment 1B), as given in Figure 4; the preventive condition used the contingencies from Buehner and Cheng (1997, Experiment 1A). These contingencies differ only in whether $P(e^+|c^+)$ is greater than $P(e^+|c^-)$ or vice versa. The difference condition used the same contingencies, randomizing whether the probability of the effect was higher for the first species or the second. In each case, these stimuli were presented on a one-page survey, which

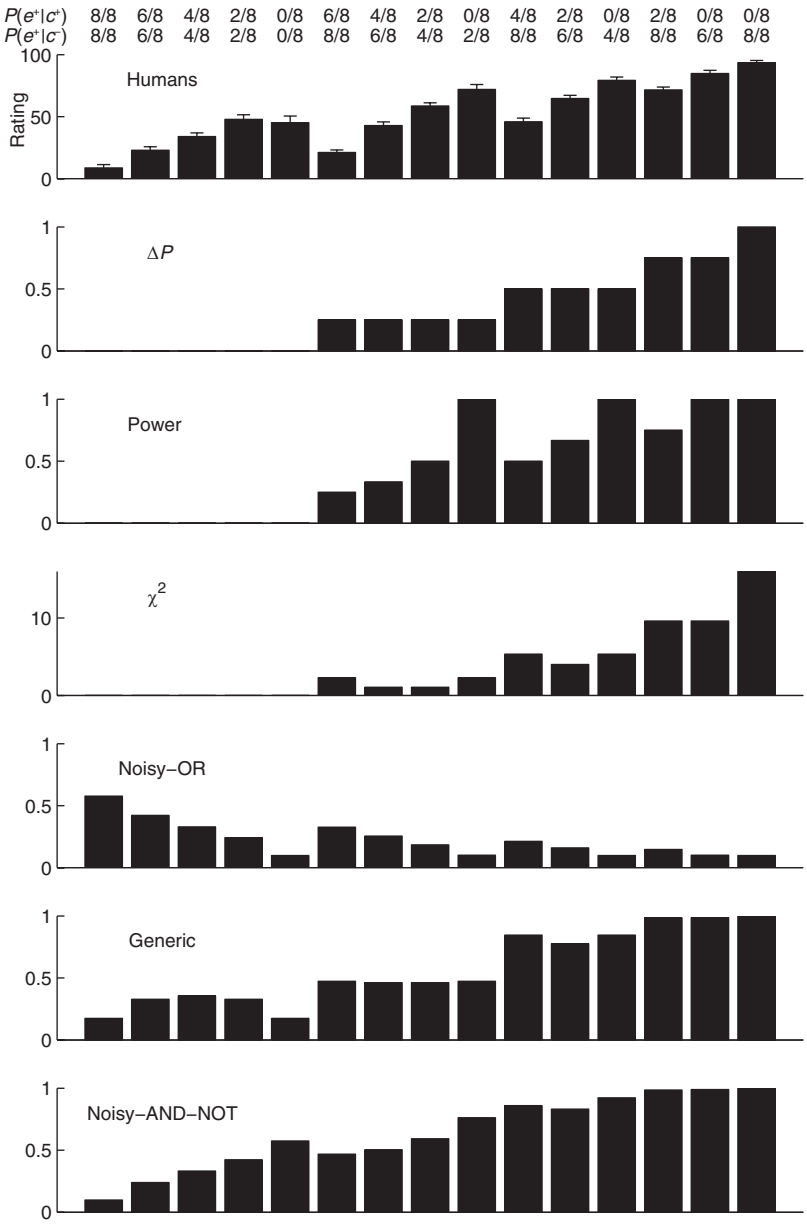

Figure 5. The top graph shows the results of Buehner and Cheng (1997, Experiment 1A). The other graphs show predictions for these contingencies for different models—ΔP, causal power (Cheng, 1997), and the χ^2 test for independence—together with our Bayesian model under three choices of parameterization (noisy-OR, generic, and noisy-AND-NOT). $P(e^+|c^+)$ and $P(e^+|c^-)$ indicate the probability of the effect in the presence and absence of the cause, respectively.

outlined the medical cover story and then asked people to evaluate the evidence provided by the contingencies, stating that each set of contingencies indicated the results of a laboratory study (the full scenarios appear in Appendix A). In the generative condition, participants were given the following instructions:

> For each study, write down a number between 0 and 100 representing the degree to which the chemical causes the gene to be expressed. A rating of 0 indicates that the chemical DOES NOT CAUSE the gene to be expressed at all, and a rating of 100 indicates that the chemical DOES CAUSE the gene to be expressed every time. Use intermediate ratings to indicate degrees of causation between these extremes.

In the preventive condition, the instructions read as follows:

> For each study, write down a number between 0 and 100 representing the degree to which the chemical prevents a virus being caught. A rating of 0 indicates that the chemical DOES NOT PREVENT the virus at all, and a rating of 100 indicates that the chemical DOES PREVENT the virus every time. Use intermediate ratings to indicate degrees of prevention between these extremes.

Finally, the difference condition gave these instructions:

> For each study, write down a number between 0 and 100 representing how likely you think it is that the two species differ in their probability of expressing that gene. A rating of 0 indicates that the two species DEFINITELY have THE SAME probability of expressing the gene, while a rating of 100 indicates that the two species DEFINITELY have DIFFERENT probabilities of expressing the gene. Use intermediate ratings to indicate degrees of likelihood between these extremes.

Participants completed the survey as part of a booklet of unrelated surveys. There were 73 participants in the generative condition, 47 in the difference condition, and 67 in the preventive condition.

People's judgments in the generative and preventive conditions replicated the results of Buehner and Cheng (1997), having a linear correlation of $r = .993$ and $r = .989$ with Experiments 1B and 1A, respectively (we used linear rather than rank-order correlations here since there was no need to take into account possible nonlinear transformations). The correlations between Bayesian structure-learning models and these judgments are shown in Table 2. Also included in the table are the predictions of ΔP, causal power, and χ^2 (the absolute value of ΔP was used for the difference condition, but there was no appropriate analogue of causal power). As predicted, the noisy-OR model gave the best account of responses in the generative condition, the generic model did best in the difference condition, and the noisy-AND-NOT did best in the preventive condition. Human judgments for the five stimuli for which $\Delta P = 0$ are shown in Figure 6. ΔP, causal power, and χ^2 all predict that there should be no variation in responses across these stimuli. Contrary to these predictions, the effect of $P(e^+|c^-)$ on judgments was statistically significant in both the generative, $F(4, 288) = 5.32$, $MSE = 216.92$, $p < .001$, and preventive conditions, $F(4, 264) = 2.63$, $MSE = 313.25$, $p < .05$, and approached significance in the difference condition, $F(4, 184) = 2.40$, $MSE = 345.74$, $p = .051$. There was also a statistically significant interaction between $P(e^+|c^-)$ and condition, $F(8, 736) = 3.68$, $MSE = 283.68$, $p < .001$. As can be seen from the figure, this variation was exactly as should be expected if people are performing Bayesian structure learning with the appropriate parameterization.

Structure, Strength, and Strategies

The models we have presented in this section are based on the assumption that people are making a judgment about whether or not a causal relationship exists, rather than estimating its strength. As we noted when we first defined our framework, this is not a necessary assumption: By taking our hypotheses to include both a causal structure and the values of the appropriate parameters, it is also possible to define models that correspond to Bayesian strength estimation of the kind endorsed by Lu et al. (2007, 2008). Whether structure or strength is relevant will be determined by the demands of the task. For example, Buehner et al. (2003) have shown that people tend to produce judgments that are more consistent with causal power when they are asked a counterfactual question, such as "What is the probability that a mouse not expressing the gene before being injected will express it after being injected with the chemical?" As pointed out by Griffiths and Tenenbaum (2005, pp. 374–375) this finding is consistent with the analysis we have presented in this section, since the rational answer to this question should be an estimate of w_1. The Bayesian treatment of strength estimation presented by Lu et al. (2007, 2008) provides a way to incorporate the effects of prior knowledge into these strength estimates.

Even in cases where people's judgments seem to be better characterized in terms of structure learning, there is room for variation in the assumptions we make about the knowledge that informs these judgments. The model we have presented in this section makes the simplest possible assumption about the prior on the strength parameters w_0 and w_1, considering each value of these parameters to be equally likely. However, this assumption has an effect on the model's predictions both about the probability that a causal relationship exists and about the strength of that relationship. Lu et al. (2006, 2007, 2008) have shown that adopting a more

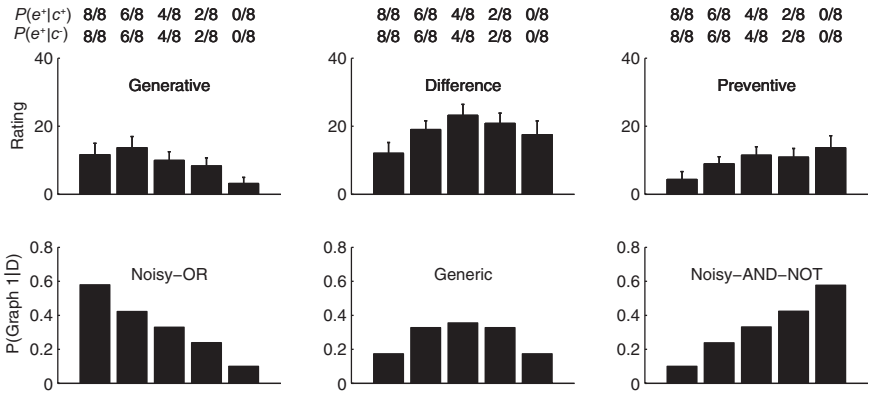

Figure 6. Effect of assumptions about functional form on causal induction. The top row shows people's judgments for a set of stimuli for which $\Delta P = 0$, under three different kinds of instructions, as described in the text. The bottom row shows the predictions of our theory-based account under three different assumptions about the functional form of a causal relationship. There appears to be a direct correspondence between task instructions and assumed functional form. $P(e^+|c^+)$ and $P(e^+|c^-)$ indicate the probability of the effect in the presence and absence of the cause, respectively.

informative prior on w_0 and w_1, in which causes are assumed to be strong when they exist and the values of w_0 and w_1 are inversely correlated, can improve the performance of this model for both structure learning and strength estimation. This modification can be incorporated into our framework through a small change in the part of the theory that specifies the functional form, replacing the uniform prior with this more informative version.

Whether people approach a task as structure learning or strength estimation will depend on their interpretation of the task and the strategies that they adopt. More generally, the richer characterization of the knowledge that people can bring to bear on a causal induction task provided by our framework opens up some interesting opportunities in understanding individual differences. For example, Buehner et al. (2003) noted that there appeared to be significant individual differences in the judgments that people produced in the experiment that we used to introduce our model. They performed a cluster analysis and found two clusters, one of which correlated well with causal power ($\rho = .902$) and the other with ΔP ($\rho = .968$). The mean ratings for the two clusters are shown in Figure 7. On the basis of these findings, other researchers have considered the possibility of modeling human judgments as a linear combination of ΔP and causal power (Perales & Shanks, 2007). However, our framework suggests another possibility: that these individual differences were not the consequence of different strategies, but different assumptions about the nature of causal relationships. Figure 7 shows that these two clusters of subjects actually seem to correspond more closely with the predictions of the Bayesian model using two different parameterizations: a noisy-OR for the first cluster and a generic parameterization for the second cluster. Rank-order correlation coefficients of $\rho = .946$ and $\rho = .980$, respectively, strongly support these accounts. It appears that everyone in this experiment is judging the evidential support for a causal relationship, but one group appears to be thinking of that causal relationship in terms of a generative mechanism whereas the other makes a weaker assumption that the cause is just a "difference maker" in the probability of the effect. Exploring the extent to which individual differences can be explained as a consequence of variation in the construal of the assumptions that people make about causal induction tasks is an interesting direction for future work.

Comparing to Other Models

Causal induction from contingency data has been the target of an unusually large number of computational models, and our discussion so far has focused on a comparison with just four models. These models were selected to highlight the contributions of the key components of our framework as clearly and briefly as possible, since the focus of this article is on capturing the effects of prior knowledge rather than developing the best model of causal induction from contingency data. However, two recent meta-analyses that compared the predictions of several models to human judgments across several experiments provide a way to get a sense for how well the model we have used as the basis for our analysis performs in comparison to other models of causal induction.

Hattori and Oaksford (2007) compared 41 models against human judgments from 143 sets of contingencies derived from five sets of experiments including their own, including only sets of contingencies consistent with generative causes. Models were assessed by computing the linear correlation within each experiment (or set of experiments in one case), and then aggregating these correlations to compute an overall average. Our Bayesian approach gave the fifth best performance of the 34 models without free parameters, despite the use of an arbitrary monotonic transformation of the model predictions in order to compute the correlation. Two of the models that gave higher correlations were heuristic accounts based on combining the frequencies that appear in the cells of the contingency table, using weights established through previous experiments. The other two models were variations on the scheme that Hattori and Oaksford (2007) were advocating in their article. Consistent with the results presented in this section, our Bayesian model performed better than both causal power and ΔP.

Perales and Shanks (2007) performed a similar meta-analysis, including 114 sets of contingencies derived from eight sets of experiments, with some overlap with those considered by Hattori and Oaksford (2007). Unlike Hattori and Oaksford, the experiments testing these sets of contingencies used instructions that asked people to evaluate generative causes, preventive causes, or left the valence of the cause unspecified. The models considered included causal power, ΔP, a weighted mixture of causal power and ΔP, our Bayesian account, a heuristic based on a weighted combination of cell frequencies, and two associative learning models. Models were assessed by computing a single linear correlation with human judgments across the full set of contingencies. In their analysis, our Bayesian approach gave the worst performance of all models being considered, with an overall correlation of $r = .81$, worse than ΔP ($r = .89$), causal power ($r = .92$), and the heuristic model ($r = .94$).

The criteria by which Perales and Shanks (2007) chose datasets for inclusion in their analysis were not particularly favorable to rational models, since they emphasized online learning, mixed preventive and generative causes, and allowed response scales ranging from -100 (indicating a strong preventive relationship) to 100 (indicating a strong generative relationship). As they note in their article, their application of our Bayesian model did not resolve the problems created by mixing types of causes (which result in different predictions from the model) or provide a well-motivated solution to the problem of mapping to a scale from -100 to 100. One way to solve both of these problems is to define a hypothesis space that includes variation in functional form. For example, with just one potential cause C and effect E we can define a hypothesis space containing three structural hypotheses: Graph 0, in which no causal relationship exists; Graph $+$, in which a generative (noisy-OR) relationship exists; and Graph $-$, in which a preventive (noisy-AND-NOT) relationship exists. All three structural hypotheses are also associated with a continuum of values for w_0 and w_1, defining a complete hypothesis space of fully specified graphical models. We can then use Bayes' rule to compute the posterior distribution over these hypotheses and take the predicted response to be $-100P(\text{Graph} -|D) + 100P(\text{Graph} +|D)$, where we obtain the probabilities for each graph by integrating over the values of w_0 and w_1 as before. This response rule reduces to the posterior probability that a causal relationship exists (expressed as a percentile) when the valence of the causal relationship is known, and it is thus consistent with the other Bayesian structure-learning models presented in this section.

We applied this revised model to the sets of contingencies considered by Perales and Shanks (2007). When the valence of the

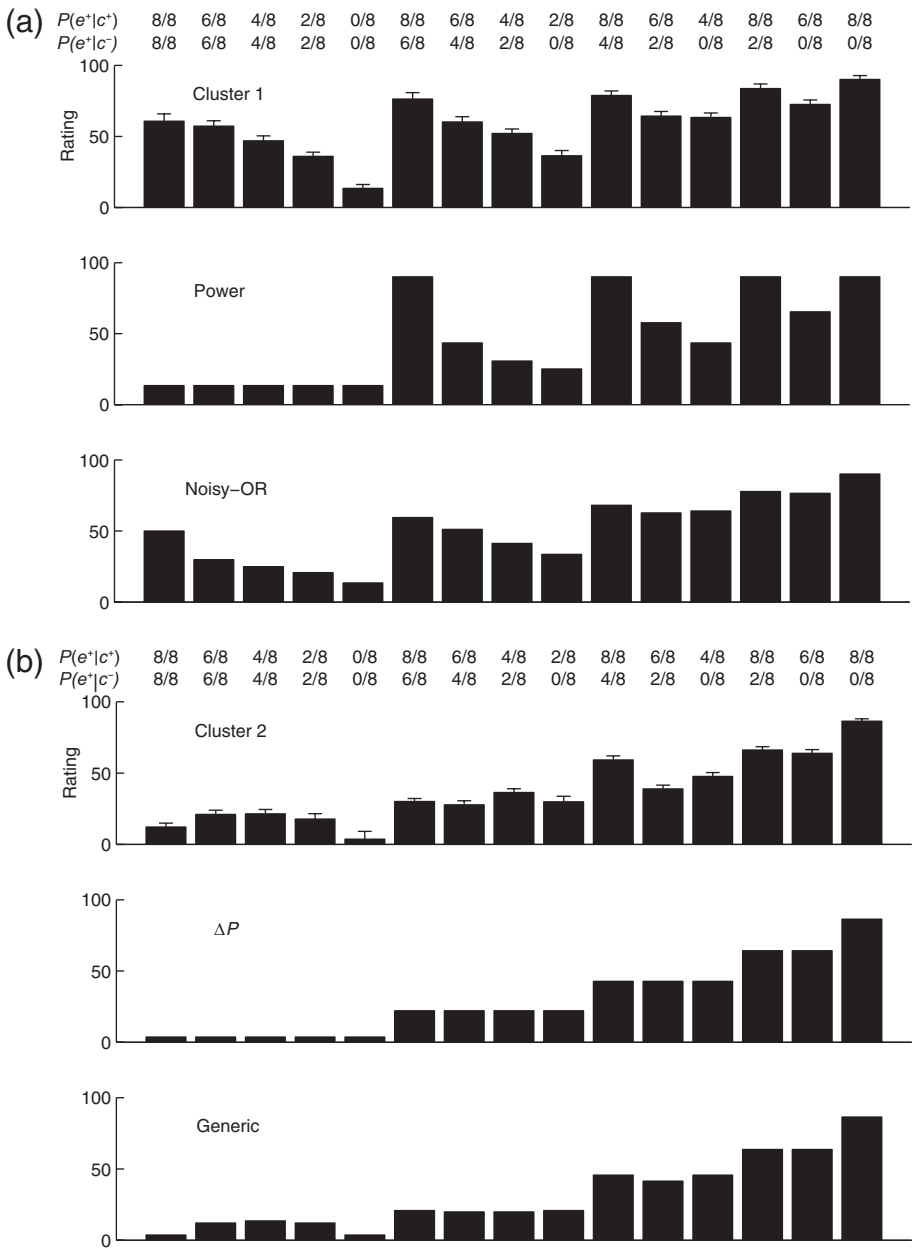

Figure 7. Mean ratings for two clusters of participants in Buehner and Cheng (1997, Experiment 1B), based on the analysis performed by Buehner et al. (2003). (a) The first cluster was interpreted as reflecting causal power, but also seems consistent with Bayesian structure learning using a noisy-OR parameterization. (b) The second cluster was interpreted as corresponding to ΔP, but also seems consistent with Bayesian structure learning using a generic parameterization. $P(e^+|c^+)$ and $P(e^+|c^-)$ indicate the probability of the effect in the presence and absence of the cause, respectively.

cause was unambiguous, established through experimental instructions and the use of a rating scale that allowed only a single sign, the Bayesian structure-learning model with the appropriate functional form was used. When the valence was ambiguous, the model allowing both functional forms was used. The prior was taken to be uniform over all hypotheses in each case. Resolving the issues of determining the appropriate functional form and response mapping resulted in a large improvement in fit, with the model producing $r = .93$. Thus, when the consequences of mixing generative and preventive causes are taken into account, the model produces a correlation that is higher than that of both causal power and ΔP, and very close to that of the best-performing heuristic model. Lu et al. (2008) have shown that a similarly high correlation can be obtained by using a Bayesian strength estimation model using a

uniform prior, and that an even higher correlation results from use of an informative prior. Given that the fit of the heuristic model also depends on setting several unmotivated free numerical parameters, we take these results on balance to favor our Bayesian approach to causal learning, contrary to the conclusions of Perales and Shanks.

Summary

Learning a single causal relationship from contingency data is arguably the most basic form of causal learning and is certainly that which has inspired the most psychological research. The results discussed above suggest that causal theories play a subtle but important role in guiding causal induction from contingency data. Models developed using our theory-based approach outperform leading rational models—ΔP and causal power—as well as models based upon standard algorithms for causal learning developed in computer science and statistics. The success of these models can be traced to two factors: the use of a causal theory that narrows down the hypothesis space and postulates the correct functional form, and the formulation of causal induction as a Bayesian inference about the causal model most likely to have generated the observed data. Varying the context of causal learning alters the functional form that people assume and the way that they construe the statistical problem, resulting in variations in behavior that can be explained within this framework.

Although we have focused on the effects of just three different parameterizations in the context of learning about a single causal relationship, the particular parameterization that learners should use will depend on the context in which learning takes place. Different settings will result in different expectations about the nature of causal relationships. Consequently, there are many directions in which this analysis could be extended. A simple next step would be to explore the contexts in which functional forms corresponding to the kinds of causal schemas that Kelley (1973) identified are used, and how people might learn that such schemas are appropriate. Understanding how to generalize the approach we have taken here to predict other behavioral data, such as acquisition curves (e.g., Shanks, 1987), is also important, and will require working out how to translate this kind of rational analysis into predictions about the process of learning (see Danks et al., 2003, for some initial steps in this direction). More ambitiously, it will be important to take this approach beyond the case of a single causal relationship in order to make predictions about causal induction in contexts where causes interact (e.g., Dickinson, Shanks, & Evenden, 1984; Novick & Cheng, 2004; Price & Yates, 1993). Now we turn to a different setting, in which more complex causal structures are considered but theories provide even stronger constraints on causal models, making it possible to infer causal relationships from far fewer observations.

Inferences From Small Samples

Although causal learning from contingency data typically involves presenting a large number of observations of cause and effect, much of human causal learning proceeds on far smaller samples. For example, flipping a switch once and observing a light come on is normally sufficient to convince us of the existence of a causal relationship. The human ability to infer causal relationships from small samples is at odds with the idea that people infer causal relationships purely from contingencies, and for standard algorithms for causal learning. For Hume (1748/1977, p. 52), causal induction required "many uniform instances." Similarly, the statistical tests that scientists use to evaluate causal claims, and which are at the heart of many generic algorithms for learning causal graphical models (e.g., Pearl, 2000; Spirtes et al., 1993), require large samples to achieve significance. One of the key challenges for a computational account of human causal inferences is thus explaining how it is possible for people to learn so much from so little.

Explaining rapid causal learning might not seem problematic for advocates of the idea that human causal induction can be explained in terms of associative learning (e.g., Shanks, 1995b). Models of associative learning, such as the Rescorla-Wagner model (Rescorla & Wagner, 1972), typically incorporate a parameter that plays the role of a *learning rate*, determining how much people modify their estimates of the strength of a causal relationship on the basis of their experiences in a single trial of learning. A high learning rate results in rapid learning, with only a few trials being sufficient to produce a large change in the estimate of the strength of a causal relationship. Human inferences from small samples might thus be accounted for by saying that people have a high learning rate. However, this kind of account seems unsatisfying. In explaining how it is possible for people to learn so much from so little, we would also hope to explain why this seems to take place in some situations but not others. Flipping a switch once might seem to provide definitive evidence about the existence of a causal relationship, but running an experiment in which a single participant produces a single response that is consistent with our predictions is unlikely to convince us that we have discovered a new causal relationship. Saying that we have a high learning rate for light switches and a low learning rate for psychology experiments seems like a redescription of the data. A more satisfying explanation would be a systematic account that predicts under what circumstances we should expect causal induction to be rapid and under what circumstances we might be willing to believe in a causal relationship only after collecting many observations.

An intuitive answer to the question of how people learn so much from so little is that their inferences are informed by prior knowledge, and that this knowledge is itself acquired through experience. This is essentially the answer that we will provide, arguing that causal induction from small samples can be explained as the result of a combination of strong constraints from relatively sophisticated causal theories with statistical inference. The challenge here is in showing that this prior knowledge interacts with the observed data in a nontrivial way. We show that the kind of knowledge that is relevant to producing correct causal inferences from small samples goes beyond forming generalizations based on more abstract categories of causes (Lien & Cheng, 2000) or the a priori plausibility of causal relationships (Alloy & Tabachnik, 1984), although these are obviously both important ways in which prior knowledge can affect people's inferences. In particular, we argue that assumptions about functional form are one of the key factors in enabling causal induction from small samples. Thus, we can draw strong conclusions from flipping a switch but not from the response of a single experimental participant because we expect the former causal relationship to be

deterministic but the latter to be probabilistic.[6] This argument boils down to learners having expectations about the variability associated with causal relationships, which will influence the amount of data that they need to see in a way that generalizes to other inductive problems as well (for a similar argument, see Holland, Holyoak, Nisbett, & Thagard, 1986).

We support our analysis by examining the inferences that children and adults make in a context where they have strong prior knowledge: reasoning about the behavior of a machine for detecting the hidden properties of objects. Gopnik and Sobel (2000) introduced a novel paradigm for investigating causal inference in children, in which participants are shown a number of blocks, along with a machine—the "blicket detector." The blicket detector "activates"—lights up and makes noise—whenever a "blicket" is placed on it. Some of the blocks are blickets, others are not, but their outward appearance is no guide. Participants observe a series of trials, on each of which one or more blocks are placed on the detector and the detector activates or not. They are then asked which blocks have the power to activate the machine.

Gopnik and Sobel have demonstrated various conditions under which children successfully infer the causal status of blocks from just one or a few observations (Gopnik et al., 2001; Sobel et al., 2004). Two experiments of this kind are summarized in Table 3. In these experiments, children saw two blocks, **a** and **b**, placed on the detector either together or separately across a series of trials. On each trial the blicket detector either became active or remained silent. We encode the placement of **a** and **b** on the detector with variables *A* and *B*, respectively, and the response of the detector with the variable *E*. After seeing a series of trials, children were asked whether each object was a blicket. Table 3 gives the proportion of 4-year-olds who identified **a** and **b** as blickets after several different sequences of trials.

Theory-Based Causal Induction

The inferences that both adults and children draw about blickets and blicket detectors will be explained with reference to a simple causal theory. Such a theory should reflect people's intuitive expectations about how machines (and detectors) work and be informed by the instructions provided in the experiment. In the experiments we discuss (Gopnik et al., 2001, Experiment 1; Sobel et al., 2004, Experiment 2), children were introduced to the blicket detector by being told that it was a "blicket machine" and that "blickets make the machine go," and they saw blocks that activated the machine being identified as blickets and blocks that did not activate the machine being identified as nonblickets. A theory expressing this information is sketched in Figure 8.

Table 3
Probability of Identifying Blocks as Blickets for 4-Year-Old Children

Condition	Stimuli	**a**	**b**
One cause	$e^+\|a^+b^-$ $e^-\|a^-b^+$ $2e^+\|a^+b^+$	.91	.16
Two causes	$3e^+\|a^+b^-$ $2e^+\|a^-b^+$ $e^-\|a^-b^+$	.97	.78
Indirect screening-off	$2e^+\|a^+b^+$ $e^-\|a^+b^-$	.00	1.00
Backwards blocking	$2e^+\|a^+b^+$ $e^+\|a^+b^-$	1.00	.34
Association	$e^+\|a^+b^-$ $2e^+\|a^-b^+$	.94	1.00

Note. The one cause and two causes conditions are from Gopnik et al. (2001, Experiment 1); indirect screening-off, backwards blocking, and association conditions are from Sobel et al. (2004, Experiment 2). In describing the stimuli, e^+ and e^- indicate the presence and absence of the effect, respectively, with similar notation used to indicate the presence and absence of the objects **a** and **b** on the detector.

Following the schema introduced above, this theory has three parts: an ontology, a set of plausible relations, and the functional form of those relations. The ontology identifies the kinds of entities in our domain, which are divided into two types: **Block** and **Detector**. The number of entities of each type, N_B and N_D, are specified by distributions P_B and P_D, respectively. This ontology is hierarchical, with **Block** being divided into **Blicket** and **NonBlicket**. The probability that a **Block** is a **Blicket** is set by a parameter p. The ontology also identifies a set of predicates that apply to these entities: **Contact(b,d,t)** indicates that **Block b** is on **Detector d** in **Trial t**, and **Active(d,t)** indicates that **Detector d** is active on **Trial t**.

The plausible relations state that only blickets can cause detectors to activate, and every blicket can activate every detector. The functional form gives the probabilities of different kinds of events, stating how causal relationships influence these probabilities. The theory indicates that contact between a block and a detector is a relatively rare event. While the specific probabilities given here will not have any impact on our analysis, they could be used to make predictions about other kinds of experiments. The critical piece of information supplied by the functional form concerns how activation of a detector is affected by its causes. The theory indicates that the probability of activation follows a noisy-OR distribution (Equation 2). w_i is the "causal power" of blicket i (cf. Cheng, 1997)—the probability that blicket i will cause the detector to activate. w_0 represents the probability that the detector will activate without any blickets being placed upon it. The parameter ε indicates the reliability of the detector, determining the probability that it makes an error. Taking $\varepsilon = 0$, we can define a theory that expresses two important assumptions. First, the detector cannot activate unless a blicket is in contact with it ($w_i = 0$). Second, the probability with which a blicket will activate the detector is $w_i = 1 - \varepsilon$. If we take $\varepsilon = 0$, we obtain $w_i = 1$. These two assumptions make this the *deterministic detector* theory, embodying a simple *activation law* (Sobel et al., 2004): Only blickets can activate the blicket detector, and they always do so.[7]

[6] Although we distinguish between "probabilistic" and "deterministic" causal relationships in defining these theories, both are consistent with an underlying causal determinism: The probabilistic component of all of our theories can be interpreted as reflecting the influence of unknown factors, rather than a metaphysical commitment to a stochastic universe.

[7] The deterministic detector theory is an attractive simplification of the more general causal theory, and one that recent results on beliefs about causal determinism in children (Schulz & Sommerville, 2006) would seem to support as a reasonable default assumption. The model of Lu et al. (2006, 2007, 2008) makes similar claims about people's expectations concerning functional form in learning from contingency data, expressing the idea that people expect causes to have a high probability of bringing out their effects.

Ontology:

Types	Number	Predicates	Values
Block	$N_B \sim P_B$	**Contact(Block,Detector,Trial)**	**Boolean:** {**T**, **F**}
Blicket	p	**Active(Detector,Trial)**	**Boolean:** {**T**, **F**}
NonBlicket	$1-p$		
Detector	$N_D \sim P_D$		
Trial	$N_T \sim P_T$		

Plausible relations:
Contact(**B**, **D**, **T**) → **Active**(**D**, **T**)
Relation holds over all **T** for any **D** if **B** is a **Blicket**

Functional form:

Contact(**B**, **D**, **T**) ∼ Bernoulli(·)
Active(**D**, **T**) ∼ Bernoulli(ν) for ν from a noisy-OR:

Cause	Strength
(Background)	$w_0 = \epsilon$
Contact(B,D,T)	$w_i = 1 - \epsilon$

Figure 8. Theory for causal induction with blicket detectors. In the deterministic detector theory, ε = 0. The probabilistic detector theory allows nonzero values of ε.

In tasks involving the blicket detector, participants usually know the number of blocks (N_B), the number of detectors (N_D), and the number of trials (N_T). However, they usually do not know which blocks are blickets. The question of whether a block is a blicket comes down to whether that block causes the blicket detector to activate, since only blickets can cause activation of the detector. This question can be addressed via a Bayesian inference over causal networks: The posterior probability that a block is a blicket is simply the posterior probability that there is a causal relationship between placing that block on a detector and the activation of the detector.

The deterministic detector theory generates a hypothesis space H of causal networks for any events involving blocks and detectors. Assuming that we know that we have two blocks, **a** and **b**, and a single detector, **d**, H would consist of four graph structures, as shown in Figure 9. We use the variables A and B to indicate **Contact(a,d,T)** and **Contact(b,d,T)**, respectively, and E to indicate **Active(d,T)**, all for the same trial, indicated by the logical variable **T**. The prior probabilities of these models, P(Graph i), are set by the causal theory. The likelihood of a set of trials under these models can be evaluated using the probabilities given by the noisy-OR. The posterior probability distribution over this set of causal models can be evaluated for each set of trials shown in Table 3, denoting the set of trials D and applying Bayes' rule. The probability that a particular block is a blicket can be evaluated by summing the posterior probability of the models in which such a causal relationship exists. For example, to evaluate the probability that A causes E, we would add P(Graph 2|D) and P(Graph 3|D).

The predictions of this account are given in Table 4. These predictions provide a strong qualitative correspondence with the judgments of the children in the experiments. The most interesting case is that of backwards blocking, where the Bayesian model predicts that the probability that a causal relationship exists between B and E after the series of trials is p, the prior probability that such a relationship exists. The analysis of this experiment is as follows. After the $e^+|a^+b^+$ trials (which we denote D_1), at least one block must be a blicket: The consistent hypotheses are Graphs 1, 2, and 3. After the $e^+|a^+b^-$ trial (with the accumulated data being denoted D_2), only Graphs 2 and 3 remain consistent. The nonzero posterior probabilities are then given as follows (all others are zero):

$$P(\text{Graph } 2|D_1) = P(\text{Graph } 1|D_1) = \frac{p(1-p)}{p^2 + 2p(1-p)},$$

$$P(\text{Graph } 3|D_1) = \frac{p^2}{p^2 + 2p(1-p)},$$

$$P(\text{Graph } 2|D_2) = \frac{p(1-p)}{p^2 + p(1-p)} = 1 - p,$$

$$P(\text{Graph } 3|D_2) = \frac{p^2}{p^2 + p(1-p)} = p.$$

Consequently, the probability that A causes E is 1, and that B causes E is p.

However, this deterministic detector theory cannot explain all of the inferences that children make about blickets. In particular, it cannot explain the two causes condition in Experiment 1 of Gopnik et al. (2004). This condition was used as a control for the one cause condition, demonstrating that children drew quite different inferences about the causal relationships among a set of objects when the same associative relations (the frequency with which cause and effect co-occurred) were maintained but the structure of the trials manifesting those relations was modified. This control experiment involved showing children a block (**b**) which activated the detector on two out of three occasions. Such an event cannot be explained by our deterministic theory, under which a block either causes a detector to activate all the time or never. A set of trials in which a block activates a detector on two out of three occasions has zero probability under all causal models.

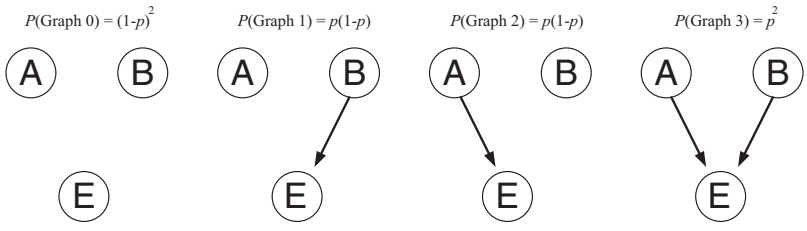

Figure 9. Causal structures generated by the theory for the blicket detector with two blocks, **a** and **b**, and one detector, **d**. A and B indicate the truth value of **Contact(a,d,T)** and **Contact(b,d,T)** for **Blocks a** and **b** and **Detector d**, while E indicates the truth value of **Active(d,T)**. Causal relationships hold over all trials, **T**. p indicates the probability of a given block being able to activate the detector, as outlined in the theory shown in Figure 8.

Table 4
Predictions of Probabilistic Theory and Alternative Models

	Deterministic		Probabilistic		Generic		Noisy-OR	
Condition	**a**	**b**	**a**	**b**	**a**	**b**	**a**	**b**
One cause	1.00	.00	.99	.07	.54	.27	.65	.25
Two causes	?	?	1.00	.81	.32	.32	.52	.28
Indirect screening-off	.00	1.00	.13	.90	.33	.50	.29	.56
Backwards blocking	1.00	*p*	.93	.41	.33	.25	.49	.40
Association	1.00	1.00	.82	.98	.27	.27	.38	.43

Note. The question marks indicate that the posterior distributions are undefined for these cases. p = probability that a causal relationship exists after the series of trials.

Consequently, the posterior distribution is undefined for this case, indicated by the question marks in Table 4.

The problem raised by the two causes condition can be addressed by relaxing one of the assumptions of the deterministic detector theory. If we allow blickets to activate detectors only some of the time, then inconsistent patterns of activation like that exhibited by block **b** are possible. We can make this change by allowing the parameter ε in the theory shown in Figure 8 to take values other than 0, where ε determines the error rate of the detector. This *probabilistic detector* theory gives the same predictions as the deterministic detector theory in the limit as $\varepsilon \rightarrow 0$ but also predicts that both **a** and **b** are blickets with probability 1.00 in the two causes condition. Different values of ε give different predictions. The predictions of this theory with $\varepsilon = .1$ and $p = 1/3$ are shown in Table 4. This model captures some of the finer details of children's judgments that are not captured by the deterministic detector, such as the fact that **b** is judged less likely to be a blicket than **a** in the two causes condition.

Alternative Accounts

As before, considering alternative accounts of these data provides insight into the assumptions that allow the theory-based approach to succeed. We compare this account with three alternatives: constraint-based algorithms, Bayesian structure learning with the generic parameterization, and a Bayesian model in which the noisy-OR parameters w_0 and w_1 can take any value between 0 and 1, as was used in the account of causal induction from contingency data presented above. All of these alternatives approach human inferences as a decision between causal structures but differ from the theory-based account in their assumptions about the functional form of the relationship between cause and effect. Examining these alternative accounts reveals that the strong expectations about functional form embodied in our deterministic and probabilistic theories are necessary to explain how children can infer causal relationships with high certainty from small samples.

Constraint-based algorithms. Gopnik et al. (2004) argued that children's inferences about blicket detectors can be explained by standard constraint-based algorithms for learning causal graphical models. They point out that, given appropriate information about the dependencies between the variables *A*, *B*, and *E*, these algorithms will infer the appropriate causal structure—for example, that **a** is a blicket in the one cause condition. However, there are two significant problems with this account: inferring the dependencies and using probabilistic prior knowledge.

The first step in applying a constraint-based algorithm is to identify the statistical dependencies that hold among a set of variables. Typically, this is done using frequentist tests such as the χ^2 test for independence. These tests impose no constraints on the functional form of the relationships between variables, and deciding that two variables are dependent requires imposing some criterion of statistical significance on the results of the tests. This raises a problem: The inferences that children make in these experiments are the result of only a handful of observations—far fewer than would be required to produce statistically significant results. Gopnik et al. (2004) addressed this issue by suggesting that "the sample size is given a large fictitious multiplier" (p. 21). As with explaining rapid inferences in terms of high learning rates, the fundamental problem with appealing to fictitious multipliers is that it is clear that such multipliers should not be used indiscriminately. For example, applying a large fictitious multiplier to the stimuli seen in experiments on causal induction from contingency data would result in a strong conviction that a relationship exists for almost all stimuli, which is at odds with people's performance on the task.

Allowing the sample size to be multiplied by some fictitious amount (or postulating a high learning rate for an associative model) leaves us with the problem of why a multiplier is appropriate in some contexts but not others. Under our account, the reason why small samples are so compelling in the case of the blicket detector is that children have strong expectations about the functional form of the relationship between placing blickets on the detector and the detector activating—namely that "blickets make the machine go," and that the machine does not go in the absence of blickets. The χ^2 test makes far weaker assumptions about functional form, and thus requires more information to identify a relationship. Fictitious multipliers thus act as a proxy for the prior knowledge that children are exploiting when making their inferences.

A second problem with explaining these results using constraint-based algorithms is that these algorithms do not exploit probabilistic prior knowledge. While particular structures can be ruled out on the basis of prior knowledge, it is difficult to see how the knowledge that the probability that a block is a blicket is p can be used by these algorithms. In contrast, the theory-based account predicts that such knowledge should be useful in the backwards blocking condition, where the probability that **b** is a blicket (under the deterministic theory) is p. Furthermore, by reasoning deductively from a pattern of dependencies, constraint-based algorithms cannot maintain degrees of uncertainty: A causal structure is either consistent or inconsistent with the data. Both the backwards block-

ing and the two causes conditions illustrate that people exhibit graded degrees of belief in the existence of a causal relationship.[8]

Bayesian structure learning with the generic parameterization. Standard Bayesian structure-learning algorithms assume that the relationship between A, B, and E can be expressed using the generic parameterization, using a separate parameter to define the probability of E for each combination of values of A and B. This approach faces one of the problems identified above for constraint-based algorithms: It makes weak assumptions about functional form and consequently requires large samples to identify the existence of a causal relationship. This is illustrated in Table 4, which shows the predictions obtained by applying Bayesian structure learning using the generic parameterization (with a uniform distribution over parameters) to the stimuli given in Table 3. The assumptions used to generate the predictions are exactly those of the theory in Figure 8 (with $p = 1/3$), except for the functional form. The predictions often deviate from human judgments—for example, in the two causes condition, **a** and **b** have a probability of being blickets that is scarcely different from the prior, because there is no strong evidence that contact between **a** and **b** and the blicket detector affects the probability with which the detector activates. In the cases where predictions move in the same direction as children's inferences, the probability that any block will be identified as a blicket remains close to the prior probability in all cases. Both of these issues are consequences of using the generic parameterization: If identifying a causal relationship requires determining that two variables are dependent, small samples can produce only small changes in beliefs about causal relationships.

Relaxing the noisy-OR parameters. As a final comparison, we can try to explain these inferences by applying the theory developed for chemicals and genes in the previous section (see Figure 2) to blicket detectors. This theory assumes that the functional form is a noisy-OR but that the noisy-OR parameters (including the baseline probability w_0) are drawn from a uniform distribution on [0, 1]. Thus, like the deterministic and probabilistic detector theories outlined above, blickets can only increase the probability that the detector activates, but unlike those theories, blickets can vary dramatically in their strengths, and the detector can activate in the absence of any blickets. The predictions under this noisy-OR theory are shown in Table 4. The assumption of generativity is not sufficient to explain children's inferences: Just assuming a noisy-OR does not place sufficiently strong constraints on the functional form of the relationship between blickets and the activation of blicket detectors. The model gives predictions that are slightly more consistent with human judgments than the generic parameterization, but small samples still do not produce dramatic changes in the extent to which a block is believed to be a blicket.

Priors and Ambiguous Evidence

The theory-based account explains how children can infer causal relationships from small amounts of data, positing strong constraints on the relationships considered plausible and the functional form of those relationships. It also makes two further predictions about human performance on these tasks that discriminate it from alternative accounts such as constraint-based algorithms. The first prediction is that prior beliefs, in the form of expectations about the probability that a block is a blicket (the parameter p in the theory), should influence people's judgments. Specifically, in the backwards blocking condition, the posterior probability that **b** is a blicket is just p. The second prediction is that people should be able to maintain graded degrees of belief in the face of ambiguous evidence. Again, the backwards blocking experiment provides one example of this, but the theory-based account predicts that people should be able to infer that a block is a blicket despite never obtaining definitive evidence, such as seeing it light up the detector all on its own. A series of studies have been conducted to test these predictions with both adults and children (Griffiths, Sobel, Tenenbaum, & Gopnik, 2009; Sobel et al., 2004). We summarize the results of these experiments and show how they can be explained by the theory-based account. We focus on the adult experiments, which provide data about the beliefs of the participants after each trial, but the same qualitative effects hold with 4-year-old children.

Griffiths et al. (2009, Experiment 1) explored the extent to which people's judgments were affected by prior probabilities by conducting an analogue of the backwards blocking condition of Sobel et al. (2004, Experiment 1) but varying the probability of a causal relationship. The experiment was done with adults, using a "super-pencil" detector which functioned exactly like a blicket detector, but identified whether golf pencils contained a special kind of lead. Participants were randomly assigned to five conditions, which determined how they were introduced to the detector. In all five conditions, participants saw 12 pencils placed on the detector, one after the other. The conditions varied in how many of those pencils activated the detector, with 2, 4, 6, 8, or 10 pencils being identified as containing super-lead. This training phase was intended to establish prior beliefs about p, the probability that a pencil contains super-lead, with the different conditions corresponding to $p = 1/6$, $1/3$, $1/2$, $2/3$, and $5/6$, respectively.

The test phase of the experiment had three stages. First, participants were shown two new pencils, **a** and **b**, and were asked to rate the probability that they were super-pencils. They then saw **a** and **b** placed on the detector together, and the detector activating, and were again asked to rate the probability that they were super-pencils. Finally, just**a** was placed on the detector, and the detector activated. Once again, participants rated the probability that **a** and **b** were super-pencils. The mean ratings in the different conditions are shown in Figure 10. Manipulating the frequency with which pencils were identified as super-pencils had the expected effect on people's baseline judgments, indicating a difference in prior beliefs. It also affected the judgments that people made after each trial. As shown in the figure, the pattern of judgments is perfectly predicted by Bayesian inference guided by the deterministic detector theory (or the probabilistic detector theory with $\omega = 1 - \varepsilon$ as $\varepsilon \rightarrow 0$): The probability of **a** and **b** being super-pencils should increase after the first trial, and then the second trial should provide unequivocal evidence that **a** is a super-pencil while the probability that **b** is a super-pencil should return to the prior p. The model predictions shown in the figure were obtained by setting p to the baseline probability given by the participants and yield a

[8] The probabilities shown in Table 3 were computed from responses that children made on multiple trials, and in cases where these probabilities are in the middle range children sometimes varied in their responses across trials. This variation is consistent with uncertainty about the existence of a causal relationship.

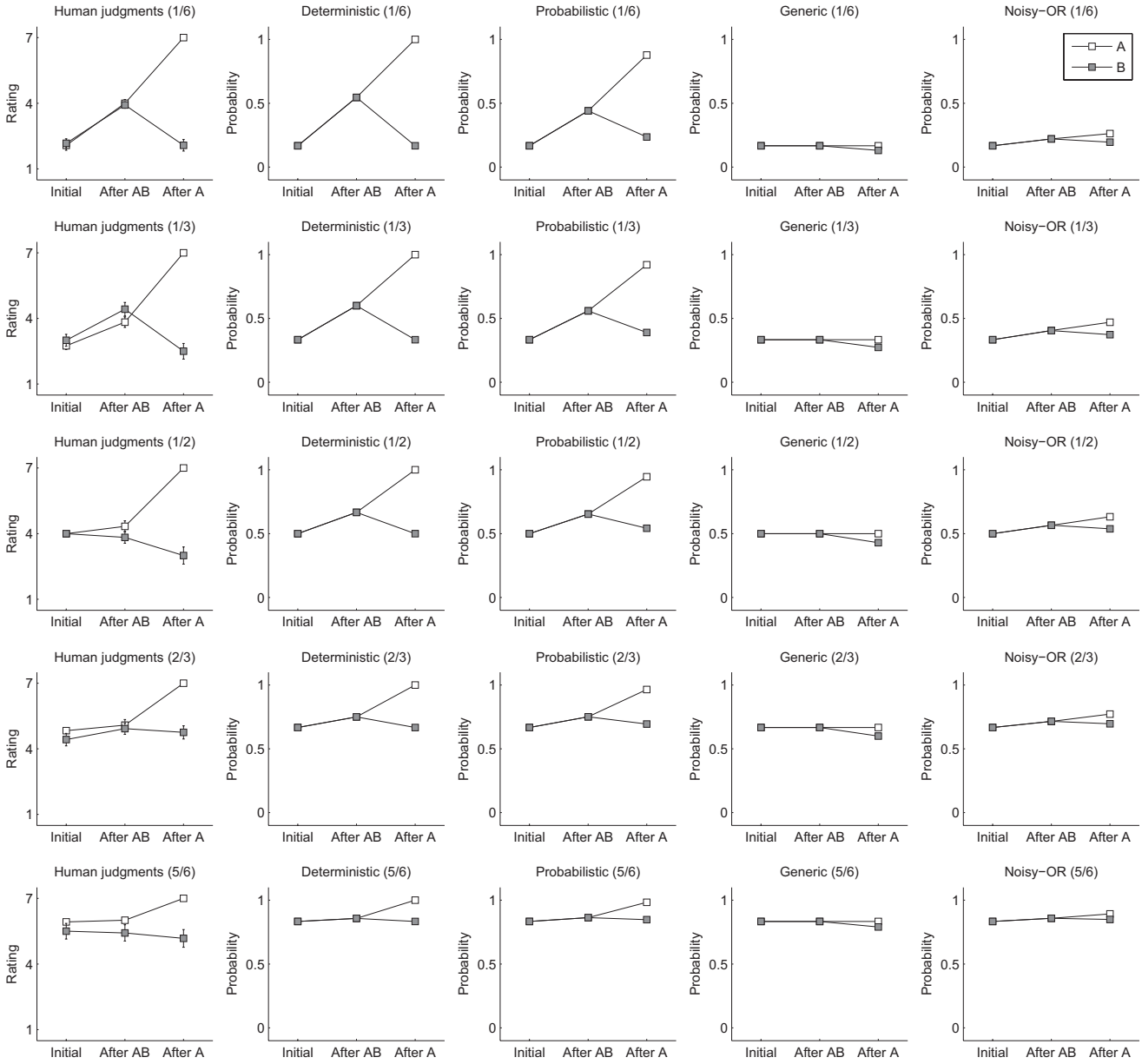

Figure 10. Adult judgments and model predictions for causal induction with "super-pencils," an analogue of the blicket detector task from Griffiths et al. (2009, Experiment 1). The model predictions show the consequences of modifying assumptions about the functional form of the causal relationship, including a deterministic OR, a probabilistic noisy-OR in which pencils have super-lead, a generic parameterization that makes no assumption about the nature or direction of a causal relationship, and a noisy-OR with a uniform distribution over the strength of the causal relationship. In this case, people gave initial ratings, saw a trial on which two pencils, A and B, caused the detector to activate, and then saw a further trial on which A alone caused the detector to activate. In the experiment and the data, the prior probability that a pencil would activate the detector, p, varied from 1/6 to 5/6. Error bars represent 1 *SE*.

correlation of $r = .973$ with the human data. Sobel et al. (2004, Experiment 3) obtained similar results with 4-year-old children.

To demonstrate the importance of the assumptions made in the blicket detector theory, and in particular the assumption of a functional form consistent with a noisy-OR in which blickets are very likely to activate the detector, we also evaluated the predictions of the three Bayesian models making different assumptions about functional form introduced above: a model consistent with the probabilistic detector theory, taking $\varepsilon = .1$; a model assuming a generic functional form, as described above; and a model using a noisy-OR parameterization, with a uniform distribution over w_i, as used in our analysis of inferences from contingency data. The

predictions of these three models are also shown in Figure 10. The three models gave correlations of $r = .950$, $r = .570$, and $r = .648$ with the human data, with the generic and noisy-OR models being far less sensitive to the information provided by the observed data than the human participants in the experiment. These results show how the assumption of a near-deterministic causal relationship increases the amount of information provided by a small number of observations.

This experiment illustrates that people's causal inferences are affected by their prior beliefs in exactly the way the theory-based account predicts. Griffiths et al. (2009, Experiment 2) also showed that people could make inferences from ambiguous evidence in a fashion consistent with theory-based Bayesian inference. This experiment was also conducted with super-pencils, and people saw the detector activated by two out of 12 pencils before beginning the critical trials. They were shown three new pencils, **a**, **b**, and **c**, and were asked to rate the probability that these pencils were super-pencils. They then saw **a** and **b** placed on the detector together, causing the detector to activate, and gave new ratings. Finally, they saw **a** and **c** placed on the detector together, causing the detector to activate, and were asked to rate the probability that each of the pencils was a super-pencil. The mean ratings are shown in Figure 11.

In this experiment, people received no unambiguous clues that a particular pencil was a super-pencil: There were no trials on which a single pencil caused the detector to activate. Nonetheless, people were able to infer that **a** was quite likely to be a super-pencil, whereas **b** and **c** were less likely to be super-pencils, but more likely than they had been at the start of the experiment. Similar results were obtained with 4-year-old children (Griffiths et al., 2009, Experiment 3). The hypothesis space generated by the deterministic detector theory with three blocks and one detector is shown in Figure 12. The hypothesis space H generated by the theory consists of eight causal graphical models, indicating all possible sets of causal relationships between the three blocks and the detector. The predictions produced by applying Bayesian inference with this hypothesis space are shown in Figure 11, setting p to the baseline probability given by the participants. The model predicts four quantitatively different levels of belief for different pencils at different points in the experiment: the baseline probability, the probability that **a** and **b** are super-pencils after the first trial, the probability that **a** is a super-pencil after the second trial, and the probability that **b** and **c** are super-pencils after the second trial. People also show these four levels of graded belief in the existence of a causal relationship, and there is a close correspondence between model and data, with $r = .979$. The three alternative models discussed above—the probabilistic detector, the generic functional form, and the noisy-OR with a uniform prior on strength—gave correlations of $r = .886$, $r = .053$, and $r = .983$. The predictions of these models are also shown in Figure 11 and reflect a similar trend to that seen in the previous experiment: Small amounts of data have a weaker effect on the predictions of the models that have fewer constraints on the functional form of the relationship. Although the correlation between the noisy-OR model and the human judgments was high, the absolute value of the probabilities predicted by this model were very different from those expressed in the human judgments.

Learning the Right Theory

We have outlined two different theories for the blicket detector—the deterministic detector theory and the probabilistic detector theory. In some cases, such as the one cause and two causes conditions, it seems that the probabilistic detector theory provides a better characterization of children's inferences. However, the instructions the children received suggested that the deterministic theory might be more appropriate. This raises an important question: How might a child learn the appropriate causal theory? This question returns to one of the most interesting aspects of the tale of Halley's comet: that the return of the comet provided an indication of the validity of Newton's theory, the theory which had made it possible for Halley to recognize the causal structure responsible for his observations.

Learning causal theories can be modeled naturally within the theory-based causal induction framework. Indeed, this is one of the great strengths of the framework. In the case of the blicket detector, the problem is quite constrained, being a matter of choosing between the deterministic and the probabilistic theory. This decision can be made by using Bayes' rule, treating theories T as hypotheses

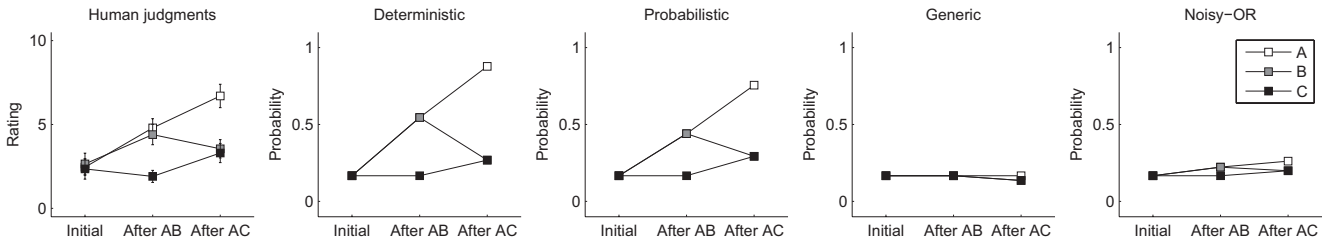

Figure 11. Adult judgments and model predictions for causal induction with "super-pencils," an analogue of the blicket detector task, from Griffiths et al. (2009, Experiment 2). The model predictions show the consequences of modifying assumptions about the functional form of the causal relationship, including a deterministic OR, a probabilistic noisy-OR in which pencils have super-lead, a generic parameterization that makes no assumption about the nature or direction of a causal relationship, and a noisy-OR with a uniform distribution over the strength of the causal relationship. In this experiment, people received purely ambiguous evidence about causal relationships, seeing one trial on which two objects, A and B, activated the detector together, and then a second trial on which A and a novel object C activated the detector together. Error bars represent 1 *SE*.

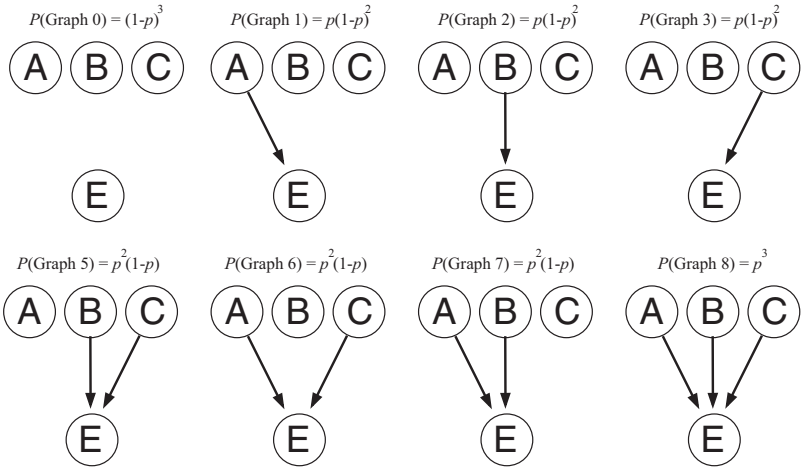

Figure 12. Causal structures generated by the theory for the blicket detector with three blocks, **a**, **b**, and **c**, and one detector, **d**. *A*, *B*, and *C* indicate whether contact between the appropriate block and the detector occurred on a particular trial, while *E* indicates whether the detector activated. These causal relationships hold for all trials **T**. *p* indicates the probability of a given block being able to activate the detector, as outlined in the theory shown in Figure 8.

$$P(T|D) = \frac{P(D|T)P(T)}{P(D)}, \tag{8}$$

where $P(D) = \sum_T P(D|T)P(T)$. The critical probabilities in this expression are of the form $P(D|T)$, being the probability of the data D under a theory T. These probabilities can be computed by summing over all causal graphical models generated by T, being the members of the hypothesis space H_T. Thus we have

$$P(D|T) = \sum_{i=1}^{|H_T|} P(D|\text{Graph } i)P(\text{Graph } i|T),$$

which can be computed using just probabilities defined above: the probability of the data under a particular causal graphical model and the prior probability of such a model under the theory. Other probabilities, such as the probability of a particular causal structure, or that an object is a blicket, can be evaluated by summing over theories T.

Figure 13 illustrates how this learning process can operate concurrently with inferring the causal properties of the entities in a domain. The figure shows the posterior distribution over the two theories—deterministic and probabilistic—and the probability that blocks **a** and **b** are blickets as data D accumulates. In this case, the data are the trials used in the two causes condition. The prior gives a probability of .99 to the deterministic theory and .01 to the probabilistic theory, p is set to .3, and ε is set to .1. The first three data points are all $e^+|a^+b^-$, being events in which **a** is placed on the detector and the detector activates. This is sufficient to iden-

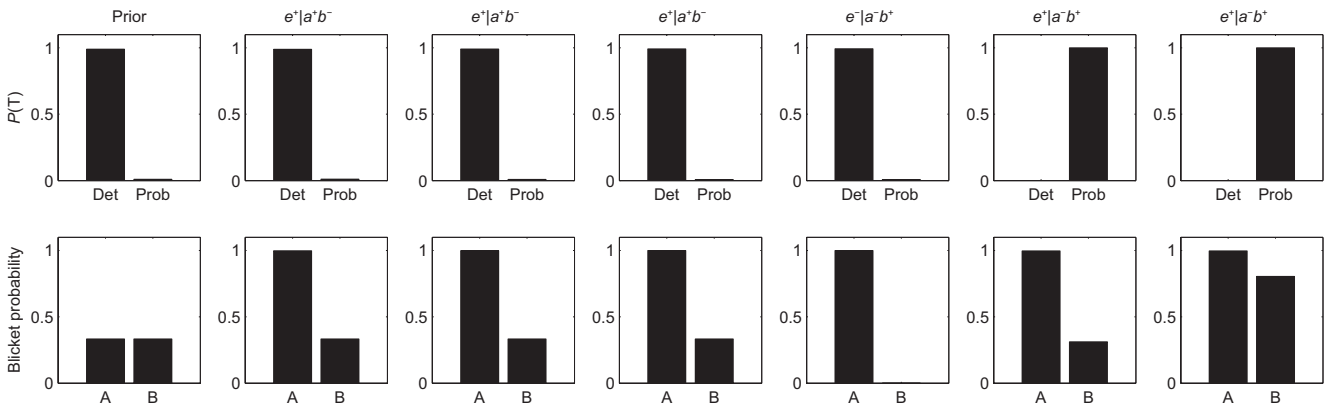

Figure 13. Learning a causal theory. The bar graphs along the top of the figure show the probabilities of the two theories, with "Det" indicating the deterministic detector theory, and "Prob" indicating the probabilistic detector theory. The bar graphs along the bottom show the probabilities that the blocks **a** and **b** are blickets. The probabilities after successive trials are shown from left to right. In describing the stimuli, e^+ and e^- indicate the presence and absence of the effect, respectively, with similar notation used to indicate the presence and absence of the objects a and b on the detector.

tify **a** as a blicket under either theory, and weakly favors the deterministic theory. The fourth data point is $e^-|a^-b^+$, with **b** placed on the detector and the detector not activating. Under the deterministic theory, **b** is definitely not a blicket. Under the probabilistic theory, there remains a small chance that **b** is a blicket, and since the probabilistic theory is still viable, the probability that **b** is a blicket is nonzero. The fifth data point is $e^+|a^-b^+$, the activation of the detector when **b** is placed upon it. The fourth and fifth data points are mutually contradictory under the deterministic theory and have a probability of zero. This event can be explained only by the probabilistic theory, in which **b** is definitely a blicket, and consequently the posterior probability of the probabilistic theory and of **b** being a blicket both become 1.00. By the end of the trials in the two causes condition, one should be firmly convinced that blicket detectors are probabilistic.

Learning a causal theory from evidence provides a possible explanation for why children in the one cause and two causes conditions produced responses consistent with the probabilistic detector theory, whereas children in the other conditions acted in a fashion more consistent with the deterministic detector theory. Under the deterministic theory, children in the one cause condition should never say that **b** is a blicket, but children did so on 16% of trials. Because the one cause and two causes conditions were presented within-subjects, one possibility is that some of the children saw the two causes trials, inferred that the probabilistic detector theory was appropriate, and then used this theory when they subsequently experienced the one cause condition. Examination of the data of Gopnik et al. (2001) provides tentative support for this conclusion: All of the children who identified **b** as a blicket in the one cause condition had seen the two causes condition beforehand.[9]

Griffiths et al. (2009, Experiment 4) followed up on this observation with a further experiment, explicitly manipulating the information provided to their participants about the nature of the detector. In the probabilistic detector condition, participants saw a series of trials in which objects activated the detector on only a subset of the trials where they were placed on the detector. In the deterministic detector condition, objects either always activated the detector or never activated the detector. These trials provide enough information to identify the appropriate causal theory, and participants in both conditions then observed a series of trials matching the one cause condition described above. The results are shown in Figure 14, together with the predictions of the two-theory model described above. As predicted by the model, participants who saw evidence that the detector was probabilistic were still willing to believe that **b** could cause the detector to activate, even after seeing it fail to activate the detector once. The model also predicted a small reduction in certainty that **a** was a blicket. People also showed a trend in this direction, slightly larger than predicted by the model. The correlation between the model predictions and human judgments was $r = .99$ with p set to 5/6, corresponding to the number of objects identified as blickets during the familiarization trials, and ε set to .1.[10] Griffiths et al. (Experiment 5) also found similar results with 4-year-old children.

Summary

When learning about the properties of simple physical systems, both adults and children are able to infer a causal relationship from only a few observations. Our theory-based approach explains this as the result of powerful statistical inference guided by a causal theory that provides strong constraints on the functional form of a causal relationship. This principle accounts for how children can identify blickets on the basis of just a few observations. Taken together with our account of causal learning from contingency data, this case study lays the groundwork that allows us to consider more challenging problems of causal induction involving richer causal structures, such as those with hidden causes, and more complex theories, such as those characterizing dynamical physical systems.

Observations, Interventions, and Hidden Causes

The formalism that underlies causal graphical models distinguishes between *observations*—events in which the values of a set of variables are observed—and *interventions*—events in which some of those variables are set to particular values as a result of actions outside the system (Pearl, 2000). This distinction has important consequences for how these different kinds of data should influence causal induction. In particular, interventions can be far more diagnostic of causal structure. This prediction has been evaluated by several different researchers (e.g., Gopnik et al., 2004; Hagmayer, Sloman, Lagnado, & Waldmann, 2007; Lagnado & Sloman, 2004; Steyvers et al., 2003). Here, we focus on how observations and interventions can be useful in revealing the presence of hidden causes.

Everyday reasoning draws on notions that go far beyond the observable world, just as modern science draws upon theoretical constructs beyond the limits of measurement. The richness of our intuitive theories is a direct result of our ability to identify hidden causal structure. The central role of hidden causes in intuitive theories makes the question of how people infer hidden causal structure fundamental to understanding human reasoning. Psychological research has shown that people can take into account the effects of hidden causes (Cheng, 1997), infer the existence of hidden causes from otherwise unexplained events (Luhmann & Ahn, 2003), learn about the strength of those causes (Luhmann & Ahn, 2007), reason about the impact of hidden causes when evaluating other causal relationships (Hagmayer & Waldmann, 2007), discover that functions are mediated by hidden variables (Busemeyer, McDaniel, & Byun, 1997), and determine hidden causal structure from very little data (Kushnir et al., 2003).

In this section, we provide a case study in the learning of hidden causes, examining how people infer the causal structure that underlies the behavior of a simple physical system—the stick-ball

[9] We thank David Sobel for making these data available.

[10] High correlations are produced for any reasonable (i.e., small) value of ε. To demonstrate this, we sampled 10,000 values of ε uniformly at random from 0 to .2 and computed the correlation for each. The mean correlation was $r = .980$. While we report only four data points here, this was just one condition of three in the original experiment, and the model described in this section gave a high correlation with the results of all three conditions.

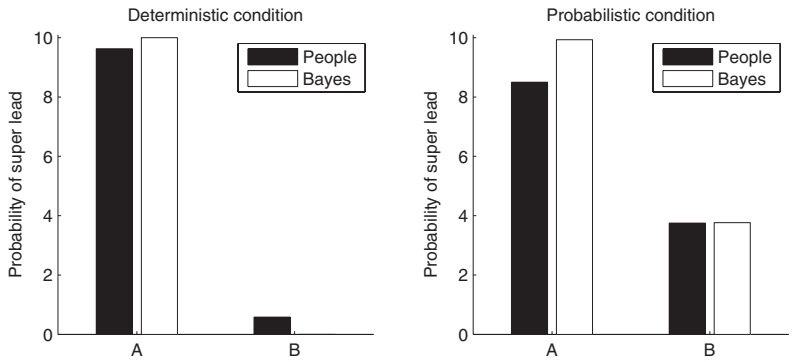

Figure 14. Results from Griffiths et al. (2009, Experiment 4). When adults are exposed to a series of trials in which objects inconsistently activate the detector (the probabilistic condition) rather than always activating the detector (the deterministic condition) they are still willing to consider the possibility that object **b** possesses super-lead, even after it fails to activate the detector once.

machine (Gopnik et al., 2004; Kushnir et al., 2003). First, we introduce this apparatus and summarize the results of experiments that used the stick-ball machine to investigate whether children and adults could combine evidence from observations and interventions, and if they could infer hidden causes (Gopnik et al., 2004; Kushnir et al., 2003). We then present a causal theory that can be used to explain these inferences. As in the case of the blicket detector, using this theory makes it possible to identify causal structure from only a small amount of data (observations or interventions). The theory extends our previous analyses by allowing for the possibility of hidden causal structure.

The Stick-Ball Machine

The stick-ball machine, also known as the puppet machine, is a physical system consisting of a number of colored balls mounted on sticks which can move up and down on a box (see Figure 15a). The mechanical apparatus moving the balls is concealed by the box, keeping the actual causal relationship unknown. The balls can either move on their own or be moved by the experimenter. Different patterns of observations and interventions lead people to believe in different underlying causal structures. Studying which structures are inferred for different stimuli provides the opportunity to understand how people make such inferences.

Gopnik et al. (2004) described a series of experiments using the stick-ball machine to assess causal induction in children. Table 5 summarizes the results of these experiments. In all cases, children were familiarized with the machine and told that "special" balls caused other balls to move. We discuss conditions in which children saw two balls, **a** and **b**, move in various patterns, using the variables A and B to indicate the motion of **a** and **b** on a given trial. In the common effects condition of Experiments 1–3, children saw **a** and **b** move together several times, $5a^+b^+$, then saw the experimenter intervene to move **b** without **a** moving, $a^-|\text{do}(b^+)$. Most children inferred that **a** was special and causing **b** to move. In the association condition of Experiment 2, it was established that this inference made use of the difference between observations and interventions, with children seeing $5a^+|\text{do}(b^+)$ followed by $a^-|\text{do}(b^+)$. These stimuli differed from the common effects stimuli only in the use of intervention on the $a^+|\text{do}(b^+)$ trials, but they produced quite different responses, with the majority of children favoring the hypothesis that **b** was special and causing **a** to move. In the common cause condition of Experiment 3, children saw $5a^+b^+$, followed by two interventions: $a^-|\text{do}(b^+)$ and $b^-|\text{do}(a^+)$. They were asked why the balls were moving together, and the majority of the children referred to an unobserved variable as the cause of these events.[11]

The experiments reported by Gopnik et al. (2004) suggest that children discriminate between observations and interventions when assessing causal relationships and that they are capable of recognizing the presence of hidden causes. Kushnir et al. (2003) conducted two experiments that extend these results to adults. In both experiments, participants were familiarized with the machine, told that if one ball caused the other to move it did so "almost always," and saw the two balls move together four times. There were three test conditions in Experiment 1, seen by all participants. In the common unobserved cause condition, participants saw $4a^+b^+$, then four trials in which the experimenter intervened, twice moving **a** with no effect on **b**, $2b^-|\text{do}(a^+)$, and twice moving **b** with no effect on **a**, $2a^-|\text{do}(b^+)$. In the independent unobserved cause condition, participants saw $2a^+b^-$, $2a^-b^+$, a^+b^+, $2a^-|\text{do}(b^+)$, and $2b^-|\text{do}(a^+)$. In the one observed cause condition, participants saw $4b^+|\text{do}(a^+)$ and $2b^-|\text{do}(a^+)$. Experiment 2 replicated the common unobserved cause condition and compared this with a pointing control condition in which interventions were replaced with observations where the experimenter pointed at the moving ball ($4a^+b^+$, $2a^-b^+$, $2a^+b^-$). On each trial, participants identified the causal structure they thought responsible by indicating images similar to those shown in Figure 15b. The results of both experi-

[11] One condition (Experiment 1, common cause) is not included in the table. This condition examined inferences involving a three-ball machine, being a version of the common cause condition of Experiment 3 in which the common cause was observable. For three balls **a**, **b**, and **c**, the stimuli consisted of several $a^+b^+c^+$ trials, followed by $b^-c^-|\text{do}(a^+)$ and $a^-b^-|\text{do}(c^+)$. Children inferred that the motion of both **a** and **c** was caused by the motion of **b**. This inference can be explained by the theory-based account, under the assumption that $\alpha < \omega$, but we do not discuss it in detail here.

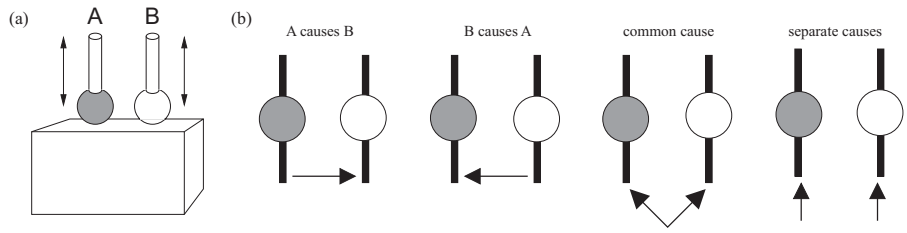

Figure 15. The stick-ball machine (Kushnir et al., 2003). (a) A two-ball machine. (b) Schematic diagrams indicating possible causal structures for the stick-ball machine (after Kushnir et al., 2003).

ments are shown in Table 6. In each condition, the majority of people indicated a single structure—common cause in the common unobserved cause condition, separate causes in the independent unobserved causes condition, **a** causes **b** in the one observed cause condition, and separate causes in the pointing control condition.

Theory-Based Causal Induction

Explaining the inferences of children and adults about the stick-ball machine requires addressing three challenges: accounting for the difference between observations and interventions, explaining how it is possible to identify that hidden causes are at work, and justifying the fact that so little data are required to identify relatively complex causal structures. These three challenges can be addressed by a theory-based account, using the causal theory shown in Figure 16.

The theory shown in Figure 16 differs from the theories considered in previous sections in incorporating a type of entity that is unobserved—the **HiddenCause**. The number of entities of this type is unbounded, representing the fact that there could be arbitrarily many such hidden causes. This is possible because hidden causes not connected to balls have no influence on the probability with which events involving those balls occur. The way that hidden causes are connected to balls, and to each other, is also unaffected by the fact that the number of such causes is unbounded.

In the case of a physical system such as the stick-ball machine, specifying the plausible relations among a set of variables involves identifying the possible physical structures that could be responsible for the motion of the elements of the system. The ontology divides the components of these physical structures into two types: components of the type **HiddenCause** are the *prime movers* in the system, the source of the force that is ultimately responsible for any observed motion, while components of the type **Ball** are passive elements, which can transfer force but not generate it. A graph structure defined on the predicates **Moves** and **Active** applied to these components indicates how force flows through the system, with a **HiddenCause** becoming active causing a **Ball** to **Move**, and the potential for a **Ball** to cause another **Ball** to **Move**. The parameters p and q determine how likely it is that force is able to flow from one ball to another, and from a hidden cause to a ball.

The process by which the hidden cause connected to a ball is selected deserves some further explanation. If it is decided that a ball is connected to a hidden cause, then the particular hidden cause is selected by sampling from a distribution in which each hidden cause **h** that is connected to at least one ball is chosen with probability proportional to the number of other balls to which **h** is connected, and a new hidden cause is chosen with probability proportional to a constant s. This procedure allows balls to share the same hidden causes, or to have independent hidden causes, and does not impose an upper bound on the number of hidden causes that appear in a physical system. The sampling scheme is that of the *Chinese restaurant process* (Aldous, 1985; Pitman, 2002), which is commonly used in nonparametric Bayesian models (e.g., Blei, Griffiths, Jordan, & Tenenbaum, 2004; Navarro, Griffiths, Steyvers, & Lee, 2006) and is formally equivalent to the system involving a *coupling probability* used in Anderson's (1990) rational model of categorization (Neal, 1998; Sanborn, Griffiths, & Navarro, 2006). The distribution that results from this process is *exchangeable*, meaning that the order in which the hidden causes are chosen does not affect the probability of a particular configuration of connections. When s is small, the scheme favors structures in which many balls have the same hidden cause. When s is large, it is more likely that balls will have independent hidden causes.[12]

The set of all structures defined on two balls that is generated by the theory is shown in Figure 17. This set includes all simple causal structures one might identify as possible descriptions of a physical system like the stick-ball machine. Graph 0 is a system in which balls are disconnected from hidden causes and from one another, and thus will never move. Graph 1 is a system in which moving **a** causes **b** to move, but neither **a** nor **b** will move on its own. Graphs 3, 7, 11, 15, and 19 all indicate a bidirectional causal relationship between A and B. Causal graphical models do not usually allow such relationships, being restricted to acyclic directed graphs. We describe how these relationships are dealt with in Appendix B.

The functional form identified by the causal theory summarizes a set of expectations about the interactions between physical objects. It states that no object moves without a cause, and objects are likely to move when caused to do so. Such a functional form results in strong constraints upon the kind of data that one might expect to see under different causal structures. For example, any data in which either ball moves without being intervened upon provides evidence against the causal structure shown in Graph 0. The probabilities of all events involving two balls under all causal

[12] The theory that we use here allows a maximum of one hidden cause per ball. This simplifies the mathematical description of the theory, and results in a smaller hypothesis space. Similar theories can be defined that allow multiple causes per ball, using different nonparametric Bayesian priors (e.g., Griffiths & Ghahramani, 2005).

Table 5
Modal Inferences by Children and Bayes for Two-Ball Machines

Experiment	Condition	Stimuli	Children	Bayes
1, 2, 3	Common effects	$5a^+b^+$, $a^-\|\text{do}(b^+)$	**a** is special	**a** is special
2	Association	$a^+\|\text{do}(b^+)$, $a^-\|\text{do}(b^+)$	**b** is special	**b** is special
3	Common cause	$5a^+b^+$, $a^-\|\text{do}(b^+)$, $b^-\|\text{do}(a^+)$	Hidden cause	Hidden cause

Note. Experiment numbers and conditions refer to Gopnik et al. (2004). In describing the stimuli, a^+ and a^- indicate that ball **a** moves and does not move, respectively. do(a^+) indicates that an external intervention causes ball **a** to move.

structures for a two-ball machine are summarized in Griffiths (2005).

Using this theory, we can compute a posterior distribution over Graphs 0–19 for any data *D*. This posterior distribution can be connected to the results discussed by Gopnik et al. (2004) and Kushnir et al. (2003) by defining an appropriate mapping between the causal structures generated by the theory and the responses possible in the experiment. For the experiments described by Gopnik et al. (2004), possible responses were that **a** was special, **b** was special, and (in Experiment 3) that a hidden cause was involved. We computed the probability that **a** was special by summing over all graphs in which there is a link from *A* to *B*, did likewise for **b** being special, and equated the probability of a hidden common cause with the probability of Graph 12. For the experiments described by Kushnir et al. (2003), there were four responses, corresponding to the structures shown in Figure 15b. The probability of "**a** causes **b**" was evaluated by summing over Graphs 1, 5, 9, and 17, and likewise with the complementary structures for "**b** causes **a**." The common cause and separate causes structures were equated with Graphs 12 and 16.

The theory shown in Figure 16 has five parameters: p, q, s, α, and ω. ω was set empirically, via a small experiment (Griffiths et al., 2004). Ten participants were shown a computer simulation of the stick-ball machine and reproduced the familiarization trials used by Kushnir et al. (2003): Participants were told that when **a** causes **b** to move, it makes it move "almost always," and were shown that **a** moved **b** on four of six trials. They were then asked how often they expected **a** would move **b**. The mean and median response was that **a** would move **b** on 75% of trials, so $\omega = .75$ was used. p was also fixed at .01, since any small value should be sufficient, and the remaining parameters were optimized to provide the best fit to the results of Kushnir et al. Table 6 shows the predictions of the model with $q = .035$, $s = 4$, and $\alpha = .36$. The model captures the major trends in the data, predicting the majority response in each condition, and gives a correlation of $r = .96$. The values of the parameters indicate that it is more likely that a ball will have the power to move on its own than that it will be connected to another ball ($q > p$), that balls are relatively unlikely to move (α is low), and that balls are quite likely to have independent causes ($s > 1$), and the model produces a high correlation with the data for a range of parameter values provided these qualitative constraints are satisfied.[13] With these parameter settings, the model also predicted the pattern of responses shown in Table 5 for the stimuli used by Gopnik et al. (2004).

Alternative Accounts

Potential alternative accounts of how people learn hidden causal structure can be found in the literature on causal induction in both psychology and computer science. We review these different kinds of alternative accounts in turn.

Hidden causes in psychological models. Most models of human causal induction (primarily those coming from a tradition of associative learning) do not make a distinction between observations and interventions or consider learning about hidden causes. However, two recent models do give an explicit treatment of hidden causes and bear further discussion. First, in her development of the idea of causal power, Cheng (1997) explicitly considered the role of hidden causes, in the guise of alternative factors that could bring about the observed effect, and provided a treatment of the difference between observations and interventions in discussing experimental design. Second, Luhmann and Ahn (2007) built on the assumptions behind causal power to develop a model that explains how people might estimate the strength of hidden causes. This model, called BUCKLE, consists of a learning algorithm that computes the probability that a hidden cause is present on a given trial and then updates the strength of the relationship between that cause and the effect.

The models introduced by Cheng (1997) and Luhmann and Ahn (2007) lay the groundwork for understanding how people infer hidden causes, and the approach they take is entirely consistent with the work we have presented in this section. However, the account we present here goes significantly beyond this previous work in making it clear how people might learn not just about a single hidden cause but about the causal structure that relates multiple hidden causes to observable variables. Causal power and BUCKLE provide ways to estimate the strength of a single hidden cause in the presence of a single observable effect and would thus need to be generalized considerably to be able to capture the phenomena we have considered in this section. Our treatment of learning hidden causal structure can be viewed as one way these models might be generalized to incorporate multiple hidden causes. The advantage of framing this generalization within the language of causal structure learning is that it can also be extended to allow inferences about hidden causes to be made from coincidences in space and time, as we discuss in the next section.

[13] To illustrate this, we evaluated the performance of the model with 10,000 random values for p, q, s, and α generated according to the following constraints: p varied uniformly from 0 to .1; q varied uniformly from p to .1; s was taken to be $1/r$ where r was distributed uniformly from 0 to 1; and α varied uniformly from 0 to ω, which was taken to be .75 based on the experiment summarized in the text. Performance was evaluated by calculating the correlation with the results of Kushnir et al. (2003). Despite choosing the parameter values at random, the mean correlation was .80, no parameter values produced a negative correlation, and 26.6% of the correlations were greater than .90.

Table 6
Probability of Choosing Different Causal Structures in Kushnir et al. (2003)

Condition	**a** causes **b**		**b** causes **a**		Common cause		Separate causes	
Common unobserved cause	.00	(.12)	.01	(.12)	**.65**	**(.71)**	.34	(.05)
Independent unobserved causes	.00	(.00)	.00	(.00)	.04	(.01)	**.96**	**(.99)**
One observed cause	**.65**	**(.67)**	.06	(.00)	.08	(.00)	.21	(.33)
Pointing control	.00	(.04)	.04	(.04)	.17	(.16)	**.79**	**(.76)**

Note. Numbers in parentheses are predictions of the Bayesian model. Boldface values indicate majority.

Approaches from computer science. Constraint-based algorithms provide a natural alternative account of these results, particularly because Gopnik et al. (2004) and Kushnir et al. (2003) suggested that their results can be explained by algorithms such as that proposed by Spirtes et al. (1993). Constraint-based algorithms are capable of using information derived from both observations and interventions, as well as identifying hidden causes. However, they cannot explain the data discussed in this section.

Explaining the inferences that people make about stick-balls in terms of these algorithms faces the same objections as arose with blicket detectors: small samples and graded degrees of belief. The experiments described above illustrate that children and adults can identify the causal structure that holds among a set of variables on the basis of only a handful of observations, far fewer than might be required to obtain statistically significant results from standard statistical tests of dependency. The data that people use to make causal inferences are not sufficient to infer that two variables are dependent. A constraint-based algorithm can thus identify the appropriate causal structure in the independent unobserved causes condition, simply because there is not enough information to conclude that a causal relationship exists. However, it will not identify the relationships that do exist in the other conditions. For example, in the association condition of Gopnik et al. (2004) and the one observed cause condition of Kushnir et al. (2003), all of the stimuli involve intervention on B and suggest that A will occur with high probability under such circumstances. These data are insufficient to justify the inference that A and B are dependent: It might just be that A occurs with high probability in general. The inference that B causes A requires an expectation that A is unlikely to occur on its own, and that if B causes A, then A is likely to occur when B does.

In the preceding analysis, we focused on the information that participants received in the trials of the experiment, consistent with the implicit information provided to participants that the only opportunities for balls to move were during these trials. However, one way to provide more data for statistical tests would be to assume that people observe n additional trials on which neither A nor B move, simply through observing that the balls are stationary most of the time. A and B are still not statistically significantly dependent in the independent unobserved causes condition provided $n < 174$, so the algorithm will still reach the correct conclusion in that case.[14] The small amounts of data available in the other two conditions result in no unique solution. In the common unobserved cause condition, a sufficiently large n makes A and B significantly dependent ($p = \frac{4!n!}{(n+4)!}$), but with only four trials on which B was observed and two trials of intervention on B, there is insufficient data to conclude $P(A|B) \neq P(A|\text{do}(B))$ or $P(B|A) \neq P(B|\text{do}(A))$, $p = .067$. And, as noted above, in the one observed cause condition, we have no observations of A, only interventions, and thus cannot even conduct a significance test to check $P(B|A)$ against $P(B|\text{do}(A))$.

The data of Kushnir et al. (2003) also illustrate that people can maintain graded degrees of belief about causal structures. For example, the stimuli in the independent unobserved causes and pointing control conditions both seem to suggest that separate causes are responsible, but this impression is much stronger for the former than the latter, something that is reflected in people's judgments. Because constraint-based algorithms simply identify causal structures that are consistent with patterns of dependency, they cannot capture the subtle variation in degrees of belief that are exhibited by human subjects.

Summary

Experiments with the stick-ball machine reinforce the fact that people can identify causal structure from small samples, and they indicate that this ability extends to complex causal structures, such as those involving hidden causes. Explaining the results of these experiments requires appealing to a simple theory that expresses some of the content of intuitive mechanics. Some of the key assumptions in this theory concern the functional form of causal relationships, placing constraints on the probability of events that have a strong qualitative correspondence to Newton's laws of motion. By exploring causal inferences about other physical systems, we have the opportunity to gain a deeper understanding of the intuitive principles of physics that guide causal induction.

Causal Induction From Patterns of Spatial and Temporal Coincidence

We have presented analyses of how people infer causal structures of increasing complexity, from a single causal relationship to hidden common causes. However, all of these analyses have been based upon settings in which events can be described as discrete trials on which cause and effect co-occur. There is another kind of information that provides very strong evidence for hidden causal structure: patterns of coincidence in data varying along continuous dimensions, such as the time at which events occur or their location in space. With just a few events coincidentally aligned in space or time, and a theory about the factors that could give rise to those patterns, we can induce hidden causal structure.

[14] We use Fisher's exact test to compute these numbers, since the small cell entries lead Pearson's χ^2 to deviate from the χ^2 distribution.

Ontology:

Types	Number	Predicates	Values	
Ball	$N_B \sim P_B$	**Moves(Ball,Trial)**	**Boolean:**	$\{\mathbf{T}, \mathbf{F}\}$
HiddenCause	$N_H = \infty$	**Active(HiddenCause,Trial)**	**Boolean:**	$\{\mathbf{T}, \mathbf{F}\}$
Trial	$N_T \sim P_T$			

Plausible relations:

$\mathbf{Moves}(\mathbf{B}_1, \mathbf{T}) \rightarrow \mathbf{Moves}(\mathbf{B}_2, \mathbf{T})$
True for all $\mathbf{T}$ with probability p for each $\mathbf{B}_1 \neq \mathbf{B}_2$ pair

$\mathbf{Active}(\mathbf{H}, \mathbf{T}) \rightarrow \mathbf{Moves}(\mathbf{B}, \mathbf{T})$
Each $\mathbf{B}$ has an edge from some $\mathbf{H}$ with probability q. If such an edge exists, then the particular $\mathbf{H}$ is chosen based upon the number of existing edges:

$$P(\mathbf{Active}(\mathbf{H}, \mathbf{T}) \rightarrow \mathbf{Moves}(\mathbf{B}, \mathbf{T})) \propto \begin{cases} M_{\mathbf{H},i} & M_{\mathbf{H},i} > 0 \\ s & \mathbf{H} \text{ is new} \end{cases}$$

where $M_{\mathbf{H},i}$ is the number of edges from $\mathbf{H}$ when the edges are chosen for the ith ball.

Functional form:

$\mathbf{Active}(\mathbf{H}, \mathbf{T}) \sim$ Bernoulli(α)
$\mathbf{Moves}(\mathbf{B}_1, \mathbf{T}) \sim$ Bernoulli(ν) for ν from a noisy-OR:

Cause	Strength
(Background)	$w_0 = 0$
$\mathbf{Moves}(\mathbf{B}_2, \mathbf{T})$	$w_i = \omega$
$\mathbf{Active}(\mathbf{H}, \mathbf{T})$	$w_i = \omega$

Figure 16. Theory for causal induction with the stick-ball machine.

The role of spatial and temporal contiguity as cues to causal relationships has been of central interest since Hume (1739/1978) drew attention to it. For example, when a stoplight changes color at the exact moment we turn on the car radio, or the lights in a room are extinguished at the exact moment we lean against the wall, we experience a brief illusion of a causal relationship. The strength of this illusion depends on contiguity in space and time: If even a few seconds pass between one event and the other, or if the stoplight that changes is a long way away, the illusion disappears. Research in developmental psychology has explored the effects of spatial and temporal contiguity on causal induction. Shultz (1982b; Mendelson & Shultz, 1976) pitted spatial and temporal contiguity against contingency information, constructing tasks in which children learned about novel causal relationships from data that were structured so that contiguity pointed toward one relationship and contingency pointed to another. Children were strongly influenced by contiguity, with contingency guiding inferences only when contiguity was weak or controlled. Bullock et al. (1982) obtained similar results with a different kind of causal system, finding that contiguity could overwhelm contingency, particularly when violating spatial contiguity made a simple physical mechanism seem implausible.

While the work discussed in the preceding paragraph shows that there are many interesting dimensions to be explored in analyzing causal induction from spatial and temporal information, in this section we focus on the basic phenomenon of coincidences in space and time leading to a sense of causation. Sensitivity to spatial and temporal coincidences can be explained under a Bayesian account of causal induction, since we would be unlikely to see such coincidences if there were no underlying causal relationship. In the remainder of this section, we discuss how inferences from coincidences in space and time can be formalized within our framework.

Coincidences in Space

Coincidences in space can provide good clues that an underlying causal relationship exists. For example, it was the coincidence in the locations of the orbits of the three comets that suggested to Halley that they might be related. Another famous historical example is the inference that the epidemiologist John Snow made concerning the origins of an outbreak of cholera in the streets of London, described in Snow (1855). By examining the spatial distribution of a set of cases of cholera in Soho, it became clear that they were clustered about a particular region in Broad Street, and subsequent investigation revealed an infected pump at that location. Here, a coincidence in spatial contiguity suggested an underlying causal relationship. A similar sensitivity to spatial coincidences has often led contemporary doctors and patients to become suspicious that a hidden cause may be at work when multiple cases of a rare disease occur in a small geographic area, most famously in "cancer clusters" (Dawes & Hastie, 2001). Reports of cancer clusters frequently turn out to be false alarms, even when people feel very confident about the inferences they are making. What is the basis for this confidence, and does it have some grounding in a rational inductive inference?

Griffiths and Tenenbaum (2007a) examined people's sensitivity to coincidences in spatial data in a series of experiments exploring how this sense of coincidence is related to Bayesian inference about the presence of hidden causes. For example, in one study (Experiment 4), participants were shown distributions of points inside a rectangle and were told they were seeing data on the locations of a set of lemur colonies, collected during a field trip to Madagascar. They were asked to rate how likely they thought it was (on a scale from 1 to 10) that an external factor was influencing where the lemurs chose to live. The stimuli that participants saw were generated by sampling from a mixture of a uniform and a Gaussian distribution, varying the total number of points, the ratio of points drawn from the Gaussian, the location of the

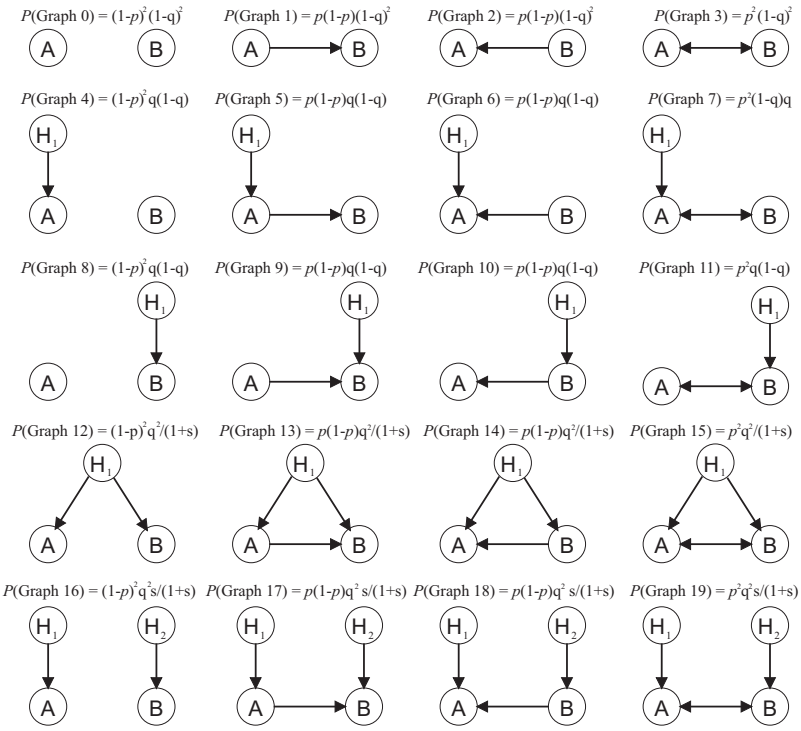

Figure 17. Hypothesis space for a two-ball stick-ball machine. *A* and *B* indicate **Moves(a,T)** and **Moves(b,T)** for **Ball a** and **b**, respectively, while H_i indicates **Active($\mathbf{h}_i$,T)** for the **HiddenCause** $\mathbf{h}_i$. These causal relationships hold for all trials **T**. *p* indicates the probability of one ball causing another to move, *q* indicates the probability of a hidden cause being able to move a ball, and *s* indicates the tendency of balls to have unique hidden causes, as outlined in the theory shown in Figure 16.

Gaussian, and its variance. The results are shown in Figure 18. People's judgments were strongly influenced by manipulations of the number of colonies and the ratio that seemed to fall within a cluster and showed a clear difference between cases where a regularity existed and those in which the locations of colonies were all drawn from a uniform distribution. Griffiths and Tenenbaum (2007a) showed that these judgments were consistent with a Bayesian analysis of the probability that a hidden Gaussian cause is responsible for a subset of the observed data, as measured against the null hypothesis alternative that all data were generated by sampling uniformly at random over the whole observed area.

The experiment in Griffiths and Tenenbaum (2007b) presented a scenario about which people had little prior knowledge (lemur colonies) and stimuli with between 20 and 200 spatial events each. The results show that people's judgments about the existence of a hidden common cause are in line with Bayesian inference, but they do not test these judgments in the setting of greatest interest for theory-based causal induction: when people see only a very small number of events but can draw on their abstract knowledge about how causes work to pick out even a slightly suspicious coincidence as evidence for a hidden cause. To test judgments in this setting, we ran a similar experiment explicitly based on the real-world scenario of cancer clusters and with the goal of identifying hidden causes responsible for some or all cases of a rare disease. Each stimulus showed a much smaller number of spatial events, between three and 12, with most stimuli showing either six or eight events. Participants made judgments over the Internet using a web-based form which asked them to judge the probability of an underlying environmental cause given 12 different patterns of spatial events. There were a total of 255 participants, divided roughly evenly into three groups who saw slightly different versions of the stimuli. The instructions presented to the participants and details of the stimuli appear in Appendix C. As in Griffiths and Tenenbaum (2007a), we manipulated the number of observations, the ratio of those that appear in the cluster to those that do not, and the spread of the cluster. We also showed several stimuli with fully random distributions. The results appear in Figure 19, organized to reflect several interesting comparisons. In particular, the last two lines of plots in the figure show how people can pick up on clusters as small as two observations in a set of three, or eight in a set of 12.

People can infer hidden causes from spatial coincidences even in very small data sets for two reasons: locations are continuous observations, potentially carrying a great deal of information in their relative positions, and people have strong expectations about the effects that hidden causes can have, in this case producing unimodal clusters. We now consider how this kind of inference can be captured within our framework, formalizing the intuitive theory that was implicit in the Bayesian model used by Griffiths and Tenenbaum (2007a).

Theory-Based Causal Induction

Figure 20 shows a theory that can be used to generate a hypothesis space for causal induction from spatial data. The same theory

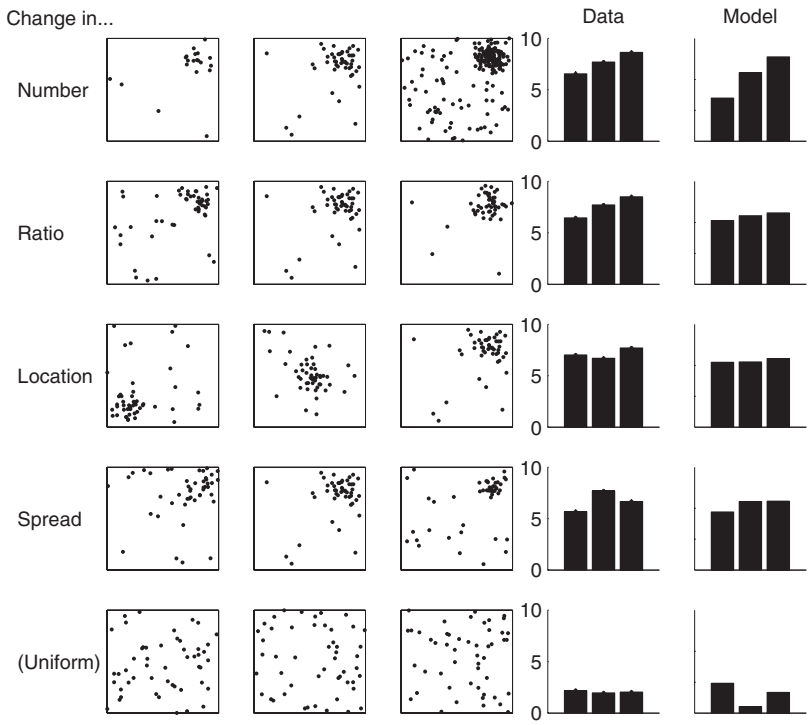

Figure 18. Inferences about hidden causes from spatial coincidences in the lemur colonies scenario. Each line shows the three stimuli used to test the effects of manipulating one of the statistical properties of the stimulus, together with the mean judgments of strength of coincidences from human participants and the predictions of the Bayesian model. Predictions are a transformed version of the log-likelihood ratio in favor of a hidden cause (see Griffiths & Tenenbaum, 2007a, for details).

applies to both lemur colonies and cancer cases, although we use lemur colonies as our running example for simplicity. The theory allows two type of entities—lemur colonies and hidden causes—each associated with a spatial predicate—a location or nexus—constrained to lie within some region R. At most one hidden cause exists, and that hidden cause has a probability p of influencing the location of a given colony. If it does so, the location of the colony follows a Gaussian distribution around the nexus of the hidden cause. If not, the location of the colony is uniformly distributed within R.

This theory generates a rich hypothesis of causal graphical models differing in whether a hidden cause exists (via the value of N_C) and which colonies are influenced by that cause. We can thus ask whether a hidden cause exists by summing the posterior probability of those hypotheses for which $N_C = 1$. We also integrate over the values of p and Σ, using a uniform prior on the former and an Inverse-Wishart prior on the latter (for details, see Griffiths & Tenenbaum, 2007a). The result can be used to predict human judgments and has no free parameters. The likelihood ratio in favor of a hidden cause gives a rank-order correlation of $\rho = .93$ with the human data shown in Figure 18, and a rank-order correlation of $\rho = .80$ with the human data shown in Figure 19, producing all of the qualitative effects shown in the figures.

Coincidences in Time

A classic example of the effects of temporal contiguity on the perception of a causal relationship is Michotte's (1963) extensive investigation of the perception of collisions. In these studies, a mechanical device was used to generate the impression of two objects interacting. Typically, these displays showed one object at rest while another approached it. If the resting object began to move at the moment when the approaching object came into contact with it, people reported the impression that the motion of one object had caused the motion of the other. This impression proved very sensitive to the timing of the motion of the objects, with separation between the end of one object's motion and the start of the other reducing the sense of causality.

More recently, several researchers have begun to examine how the temporal structure of events interacts with contingency information in guiding human causal induction. Shanks, Pearson, and Dickinson (1989) established that delays between cause and effect greater than 2 s significantly impair the ability to infer causal relationships. Buehner and colleagues (Buehner & May, 2002, 2003; Buehner & McGregor, 2006) showed that this effect can be attenuated by providing background information that leads people to expect a delay. Greville and Buehner (2007) have built on these results, providing a quantitative analysis of the interaction between contingency and contiguity in causal induction. Finally, Lagnado and Sloman (2006) examined how people use temporal information in inferring causal relationships, finding that people relied strongly on the temporal order in which a series of events occurred when trying to identify the underlying causal structure.

The experiment we use to illustrate the ability to explain coincidences in time uses a scenario that features a fictitious highly

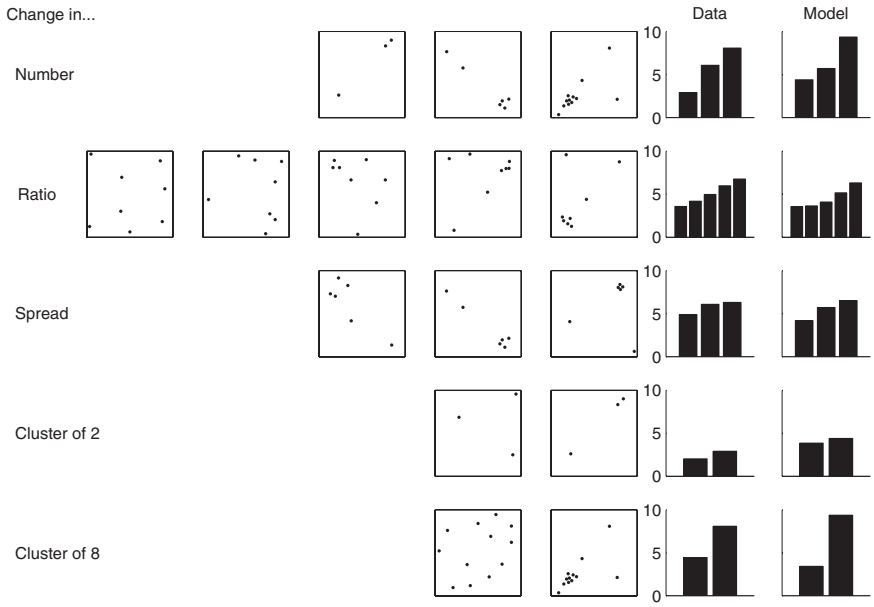

Figure 19. Inferences about hidden causes from spatial coincidences in the cancer clusters scenario. Each line shows a set of stimuli used to test the effects of manipulating one of the statistical properties of the stimulus, together with the mean judgments of strength of coincidences from human participants and the predictions of the Bayesian model. Predictions are the log-likelihood ratio in favor of a hidden cause, transformed linearly to match the mean and variance of the human judgments.

explosive compound called Nitro X, which is stored in cans. Nitro X is so unstable that cans will sometimes explode spontaneously, all on their own (presumably as the consequence of some kind of internal chemical process). Furthermore, an exploding can often causes neighboring cans to explode. The explosions exhibited by a set of cans are thus the result of a complex causal process, and identifying the elements of this process is a significant inductive challenge. Since explosions propagate by invisible force waves, the timing of explosions provides the only information about the underlying causal structure and is not accompanied by any other perceptual cues.

Griffiths et al. (2004) conducted an experiment using (computer simulated) cans of Nitro X to examine whether people could identify causal structure from the timing of explosions. In the experiment, people were introduced to Nitro X and learned about its dynamics: that cans of Nitro X could explode spontaneously and could detonate one another after a time delay that was a linear function of spatial separation, as would be expected from the slow propagation of pressure waves. They then saw a number of different patterns of explosions and were asked about the underlying causal structure.

The patterns of explosions were presented using a display involving several cans placed side by side, as shown in Figure 21. After each trial, participants were asked to indicate whether they had seen a chain reaction, with the first can exploding spontaneously and each subsequent explosion resulting from the explosion of a neighboring can; a spontaneous explosion, with each can exploding all on its own; or something else. If they decided that the

Ontology:

Types	Number	Predicates	Values
$\mathbf{Colony/Cancer}$	$N_C \sim P_C$	$\mathbf{Location(Colony/Cancer)}$	$\mathbf{Space}: \mathcal{R} \subset \mathbb{R}^2$
$\mathbf{HiddenCause}$	$N_H \in \{0, 1\}$	$\mathbf{Nexus(HiddenCause)}$	$\mathbf{Space}: \mathcal{R} \subset \mathbb{R}^2$

Plausible relations:

$\mathbf{Nexus(H)} \rightarrow \mathbf{Location(C)}$
True with probability p for each $\mathbf{H}$, $\mathbf{C}$ pair.

Functional form:

$\mathbf{Nexus(H)} \sim \text{Uniform}(\mathcal{R})$
$\mathbf{Location(C)} \sim$ Gaussian with parameters:

Condition	Mean	Variance
$\mathbf{Nexus(H)} \rightarrow \mathbf{Location(C)}$	$\mathbf{Nexus(H)}$	Σ
otherwise	$(0, 0)$	∞

Figure 20. Theory for inducing hidden causes from spatial coincidences in the lemur colonies or cancer cluster scenarios.

explosion was the result of something else, they were asked to write in a description of what they thought had happened. The first two trials showed chain reactions. On the third trial, participants saw the simultaneous explosion of several cans, as shown in Figure 21b, without the delays characteristic of pressure waves propagating from one can to the next. The number of cans used in the final trial was varied, to see whether this had an effect on people's inferences.

Over 95% of participants correctly identified the causal chain in the first two trials. For the third trial, the responses of people who chose the third option, that something other than a causal chain or spontaneous explosion had occurred, were coded by two raters. The proportion of participants identifying a hidden cause behind the simultaneous explosion is shown in Figure 22. There was a statistically significant effect of N_C: The number of cans influenced whether people inferred hidden causal structure, with most people seeing two simultaneously exploding cans as independent but six such cans as causally related. This provides a compelling illustration of the fact that people are able to infer hidden causes from small samples—in this case, only a single observation—as well as a challenge for computational models: How might such an inference be explained?

Theory-Based Causal Induction

We can explain people's inferences about Nitro X within the theory-based causal induction framework. There are two aspects of these inferences that require explanation: how people were able to identify the causal structure as a chain reaction in the first two test trials, and how they were able to infer a hidden cause in the third test trial. Both of these aspects of human behavior can be accounted for as Bayesian inferences informed by the causal theory sketched in Figure 23. The ontology used in this theory has much the same structure as the theory for the stick-ball machine (see Figure 16), with the key difference being that rather than having a set of discrete trials, the time at which events occur is a continuous quantity. There is also a connection between the plausible relations and functional form assumed by the two theories, although there are important differences that reflect the physical structure of the two systems.

The plausible relations identified by the theory are defined using the same schema as in the stick-ball theory, but they differ in their probabilities. For Nitro X, we know that the explosion of a can is able to cause the explosion of any other can, and that each can is able to explode spontaneously. As a consequence, links exist between the **ExplosionTime** of any two cans and between the

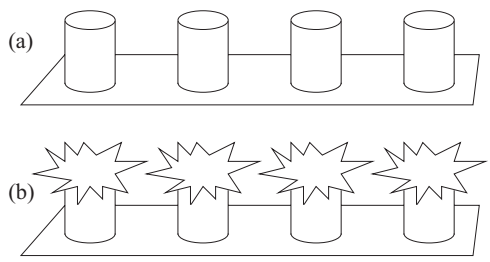

Figure 21. Nitro X. (a) Four dormant cans, as displayed in the experiment described in the text. (b) A simultaneous explosion.

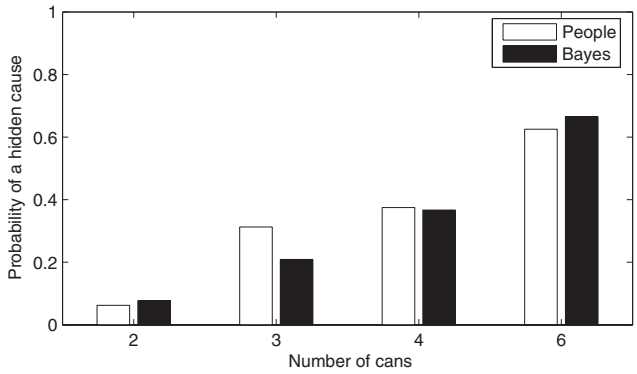

Figure 22. Causal induction from temporal coincidences. The proportion of people who identify the presence of a hidden cause increases with the number of cans that explode simultaneously, in accord with the predictions of the theory-based Bayesian model (data from Griffiths et al., 2004).

ExplosionTime of each can and the **ActivationTime** of some hidden cause. As with the stick-ball theory, the hidden cause responsible for each can is selected in proportion to the number of other cans that are currently influenced by that cause.

The functional form assumed by the theory expresses the same qualitative commitments as the theory of the stick-ball machine but translates these commitments into a domain where time is continuous. The **ActivationTime** of a hidden cause follows an exponential distribution with parameter α, which is the continuous limit of the case where a hidden cause activates on each trial with probability α, as in the stick-ball theory.[15] The distribution of the **ExplosionTime** of cans is similar to the movement of balls, except that the causal relationship between cans is subject to a further constraint: that there should be a delay between explosions commensurate with the distance between cans. As with the stick-ball theory, the assumptions of the Nitro X theory are consistent with those of Newtonian physics: Cans do not explode without a cause and are likely to explode when such a cause manifests itself.

We use a continuous version of the noisy-OR to define the distribution of the **ExplosionTime** of each can. The time at which the can explodes is assumed to be the time of the first arrival in a nonhomogeneous Poisson process. This process defines some rate function $\lambda(t)$ indicating the mean rate of arrivals at each point in time t, with the probability of the first arrival occurring at time t^* being

$$P(\text{first arrival at } t^*) = \lambda(t^*)\exp\{-\int_0^{t^*} \lambda(t)dt\}, \qquad (9)$$

where the integral takes into account the probability of the event not occurring before time t^*. We define the rate function of the Poisson process, $\lambda(t)$, to be a mixture of delta functions

[15] Here, we assume that each hidden cause can become active only once for any set of cans. This is not a necessary assumption—Griffiths et al. (2004) explained these results using a model in which hidden causes could have multiple activations. This assumption produces the same qualitative predictions as our account, but requires more complex mathematical analysis. As with the stick-ball theory, it is also possible to allow multiple hidden causes to act on a single can.

Ontology:

Types	Number	Predicates	Values
Can	$N_C \sim P_C$	**ExplosionTime**(**Can**)	**Time:** $\mathbb{R}^+$
HiddenCause	$N_H = \infty$	**ActivationTime**(**HiddenCause**)	**Time:** $\mathbb{R}^+$
		Location(**Can**)	**Space:** $\mathbb{R}$

Plausible relations:

ExplosionTime($\mathbf{C}_1$) → **ExplosionTime**($\mathbf{C}_2$)
True with probability 1 for each $\mathbf{C}_1 \neq \mathbf{C}_2$ pair

ActivationTime(**H**) → **ExplosionTime**(**C**)
Each **C** has an edge from some **H** with probability 1. The particular **H** is chosen based upon the number of existing edges:

$$P(\textbf{ActivationTime}(\mathbf{H}) \rightarrow \textbf{ExplosionTime}(\mathbf{C})) \propto \begin{cases} M_{\mathbf{H},i} & M_{\mathbf{H},i} > 0 \\ s & \mathbf{H} \text{ is new} \end{cases}$$

where $M_{\mathbf{H},i}$ is the number of edges from **H** when the edges are chosen for the ith can.

Functional form:

ActivationTime(**H**) ∼ Exponential(α)

ExplosionTime($\mathbf{C}_1$) ∼ Exponential($\lambda(t)$) for $\lambda(t)$ from a continuous noisy-OR:

Cause	Strength	Times
(Background)	$\lambda_0 = 0$	
ActivationTime(**H**)	$\lambda_i = \omega$	**ActivationTime**(**H**)
ExplosionTime($\mathbf{C}_2$)	$\lambda_i = \omega$	**ExplosionTime**($\mathbf{C}_2$)+ $\lvert$**Location**($\mathbf{C}_2$) − **Location**($\mathbf{C}_1$)$\rvert/\mu$

Figure 23. Theory for coincidences in explosion times. **ExplosionTime(C)** is the time at which can **C** explodes, and **ActivationTime(H)** the time at which a hidden cause becomes active. Cans can be caused to explode by the hidden causes or by the explosion of other cans. μ is the rate at which force propagates from one explosion to the next.

$$\lambda(t) = \sum_i w_i \delta(t, t_i), \tag{10}$$

where w_i and t_i are the weight and time associated with the ith cause, respectively. This formulation generalizes the noisy-OR, as each cause has an independent opportunity to produce the effect (see Griffiths & Tenenbaum, 2005, for details).

The hypothesis space generated by this theory with $N_C = 4$ is shown in Figure 24. The **ActivationTime** of a **HiddenCause** $\mathbf{h}_i$ is indicated with a variable H_i, and the **ExplosionTime** of a **Can** $\mathbf{c}_i$ is indicated with a variable C_i. All cans can influence the explosions of other cans, in a generalization of the bidirectional relationship between A and B shown in Graph 3 of the hypothesis space for stick-balls (see Figure 17). Likewise, all cans are influenced by a hidden cause. The only difference between the graphs shown in Figure 24 is the configuration of the hidden causes, determining which cans will be correlated in their explosion times. The 15 graphs shown in the figure correspond to all partitions of four objects into different sets, where the objects are cans and all cans within a set share a hidden cause. The distribution over these partitions is provided by the Chinese restaurant process.

This theory can be used to explain how people are able to identify the causal sequence responsible for a particular set of explosions. If we assume that s is much larger than 1, then P(Graph 0) is greater than the probability of any other graph, and the data provide only evidence in support of this conclusion. Consequently, to simplify our analysis we will outline how it proceeds just in Graph 0. Under this causal structure, the first explosion is always the consequence of a hidden cause becoming active and is hence spontaneous. There are two possible explanations for each subsequent explosion occurring at a time appropriate to its distance from previous explosions: that it occurred spontaneously, or that it was caused by those previous explosions. Under this model a set of N_C appropriately spaced explosions is much more probable under a causal chain than a spontaneous explosion (see Griffiths, 2005, Appendix D, for a detailed proof).

Identifying a set of explosions as the consequence of a single hidden common cause requires evaluating the posterior distribution over causal graphical models. In particular, it requires concluding that some structure other than Graph 0 is appropriate. Using the variable C_j to indicate the explosion time of the jth can, the theory in Figure 23 gives the probability of a simultaneous explosion ($C_j = t$ for all j, denoted $C = t$) under Graph i as

$$P(C = t | \text{Graph } i) = \omega^{N_C} (\alpha \exp\{-\alpha t\})^k,$$

where k is the number of hidden causes influencing cans in Graph i. The ω^{N_C} results from the requirement that the N_C cans all explode at that moment, while the remainder of the expression is the probability that the k hidden causes become active at that moment (and not before). Combining this probability with the prior defined by the theory, as shown in Figure 24, gives

$$P(\text{Graph } 0 | C = t) = \frac{\zeta^{N_C}}{\prod_{j=0}^{N_C-1} (j + \zeta)},$$

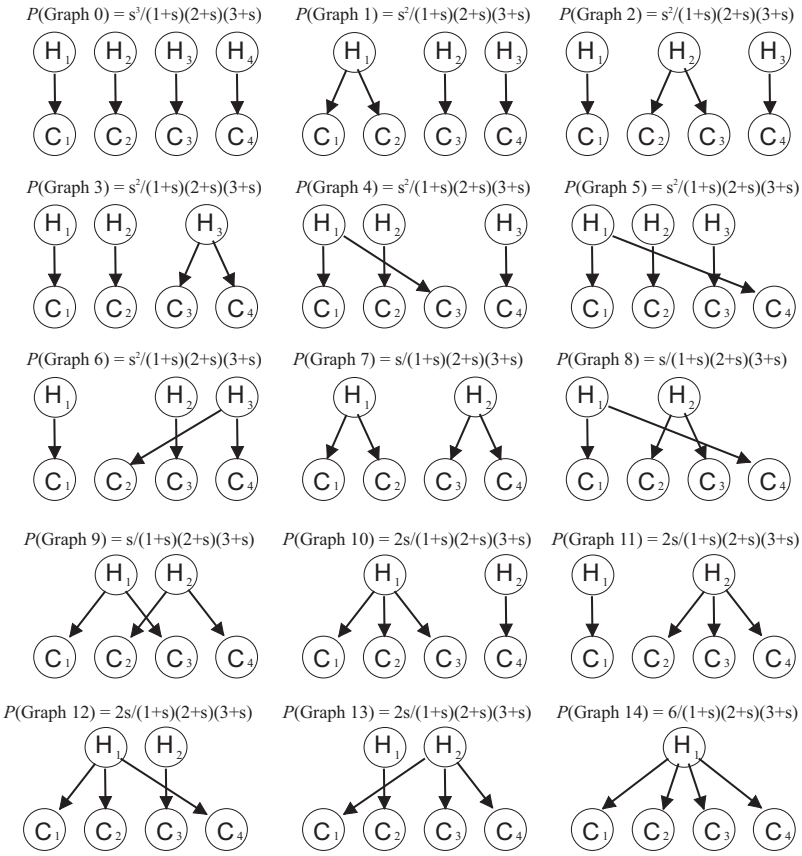

Figure 24. Hypothesis space for four cans of Nitro X. C_i indicates **ExplosionTime(c$_i$)** for **Can c$_i$**, while H_i indicates **ActivationTime(h$_i$)** for **HiddenCause h$_i$**. The dependence of **ExplosionTime(c$_i$)** on **ExplosionTime(c$_j$)** and **Position(c$_i$)** is suppressed. s indicates the tendency for cans to have unique hidden causes, as outlined in the theory shown in Figure 23.

where $\zeta = s\alpha \exp\{-\alpha t\}$. This equation results from the fact that Graph 0 requires N_C hidden causes, and thus has posterior probability proportional to $(s\alpha \exp\{-\alpha t\})^{N_C}$, with the term in the denominator being a normalizing constant for the resulting posterior distribution. The probability of the existence of some hidden cause, being $1 - P(\text{Graph } 0 | C = t)$, increases as N_C increases for any choice of s and α, and using $\zeta = 12$ produces the predictions shown in Figure 22. These predictions correlate with the experimental results at $r = .97$, with equivalently high correlations being produced for similar values of ζ.

Alternative Accounts

Standard algorithms for learning causal graphical models cannot explain these results. If we imagine that time is broken into discrete intervals, and a can either explodes or does not explode in each interval, then we can construct a contingency table for each pair of cans. Statistical significance tests will identify pairwise dependencies among all cans that explode simultaneously, provided appropriate numbers of nonexplosion trials are included. The existence of a hidden common cause is consistent with such a pattern of dependency. However, as a result of reasoning deductively from this pattern, the evidence for such a structure does not increase with N_C: A hidden common cause is merely consistent with the pattern for all $N_C > 2$.

This experiment also illustrates that people are willing to infer hidden causal structure from very small samples—just one data point—and from observations alone. Standard constraint-based algorithms cannot solve this problem: While a hidden common cause is consistent with the observed pattern of dependency, causal structures in which the cans influence one another cannot be ruled out without intervention information. People do not consider this possibility because they have learned that the mechanism by which cans influence one another has a time delay.

Summary

The analyses presented in this section show how spatial and temporal data can be incorporated into our formal framework, illustrated through consideration of how causal relationships can be inferred from coincidences in space and time. Our theory-based approach makes it clear that these cases differ from causal induction from contingency data only in the assumptions that they make about the functional form of causal relationships, allowing these relationships to be expressed in spatial and temporal properties of the observed data as well as in the co-occurrence of cause and

effect. By separating the elements of prior knowledge that influence causal induction, we can emphasize those aspects that remain the same—notions of causality, instantiated in causal graphical models, and expectations about plausible causal structures—and identify those aspects that differ.

The examples presented in this section address only the most basic questions about causal induction from spatial and temporal data, and we hope to have the opportunity to test the predictions of these accounts further. The theories presented in each case make interesting new predictions. For example, the Nitro X theory predicts that the evidence for a hidden common cause should be greater for a simultaneous explosion that occurs earlier. These predictions deserve further investigation, as do more complex instances of causal induction in dynamic physical systems, such as those studied by Michotte (1963), which we have recently begun to address (Sanborn, Mansinghka, & Griffiths, 2006). In addition, we anticipate that the framework for modeling causal induction from temporal data can be used to provide a deeper understanding of the recent results concerning the relationship between temporal contiguity and contingency data that were summarized earlier in this section (Buehner & May, 2002, 2003; Lagnado & Sloman, 2006).

Causation Across Domains

We have presented a number of examples in which people's intuitions about causal relationships can be explained as the result of rational statistical inference applied to causal structures generated by an appropriate causal theory. Since these theories express knowledge relevant to a particular setting, we should expect that they will be *domain sensitive*, having content that is affected by the domain in which causal inferences are made. In this section, we examine how the influence of domain knowledge on causal induction can be captured within our framework.

The definition of a domain is a problematic issue (e.g., Hirschfeld & Gelman, 1994), but the term is commonly used to refer to a set of phenomena that can be explained by appealing to a coherent set of causal principles. Physics, biology, and psychology all involve quite different causal principles, such as force, growth, and desire. Even young children are sensitive to this variation, having different expectations about the causal relationships participated in by biological and nonbiological (Springer & Keil, 1991) and social and nonsocial (R. Gelman & Spelke, 1981; Shultz, 1982a) entities. The early manifestation of domain-specific causal inferences, such as knowledge of the causal properties of objects (e.g., Spelke, Breinlinger, Macomber, & Jacobson, 1992), has led to claims that these inferences are the result of distinct and specialized cognitive modules (e.g., Leslie, 1994).

In attempting to formalize the content of intuitive theories relevant to causal induction, we have the opportunity to consider how the components of these theories might be affected by the domain in which causal inferences are made. In particular, we might expect that domain would have a strong influence on expectations about functional form and the plausibility of causal relationships. We consider how domain affects each of these aspects of causal relationships in turn, using the analysis of experiments conducted by Schulz and Gopnik (2004) to support our claims.

The Effect of Domain on Functional Form

Causal relationships in different domains involve very different causal mechanisms. For example, you would probably use different methods to move a heavy box a foot to the left than to move a friend a foot to the left. However, this difference in the mechanism by which effects are brought about need not be reflected in a difference in the assumed functional form of the underlying relationship. If attempting to drag a heavy box and asking your friend to move are both successful about 90% of the time, these two relationships can be described by a similar functional form. The mapping from domain-specific mechanism to functional form is many-to-one, with a variety of different mechanisms reducing to the same set of qualitative assumptions about functional form. Consequently, causal induction involving systems in quite different domains can have much the same character: Even if the *content* of theories differs, the *constraints* they imply for causal relationships can be the same.

An experiment conducted by Schulz and Gopnik (2004, Experiment 3) illustrates this point. In this experiment, children learned about causal relationships in two different domains: biology and psychology. In the biology domain, children were asked to infer which flowers caused a toy monkey to sneeze, and in the psychology domain, they learned which animals scared a toy rabbit. There were two conditions in each domain. Using A, B, and C to indicate the presence of each of three flowers (or animals) and E to indicate a sneezing monkey or a scared rabbit, the test condition consisted of four events: $e^-|a^+b^-c^-$, $e^-|a^-b^+c^-$, $e^+|a^-b^-c^+$, and $e^+|a^+b^+c^+$. The control condition featured four different events: $e^+|a^+b^-c^-$, $e^+|a^-b^+c^-$, $e^+|a^-b^-c^+$, and $e^+|a^+b^+c^+$. Children made quite similar inferences across the two domains, as shown in Table 7, identifying C as the cause in the test condition, and all of A, B, and C as causes in the control condition.

Schulz and Gopnik (2004) used the results of this experiment to argue that children's ability to infer causal relationships is domain independent. A different interpretation of these results is that they indicate that the same functional form can be assumed in different domains (and for different underlying causal mechanisms). If the theories characterizing two systems are isomorphic, then causal inferences using those theories will be identical. A pair of isomorphic theories for sneezing and scaring is shown in Figures 25 and 26. There is a direct correspondence between the types of entities identified by these theories and the predicates applied to those entities, with **Flower** and **Beast**, **Monkey** and **Rabbit**, and **Sneezes** and **Scared** all playing the same roles. The theories are

Table 7
Effect of Domain on Functional Form

Condition	*C*		All		Other	
Test (biology)	**.78**	**(.90)**	.11	(.00)	.11	(.10)
Control (biology)	.05	(.06)	**.89**	**(.86)**	.05	(.08)
Test (psychology)	**.67**	**(.90)**	.28	(.00)	.05	(.10)
Control (psychology)	.05	(.06)	**.83**	**(.86)**	.11	(.08)

Note. Numbers indicate the proportion of children identifying C as the cause, indentifying A, B, and C as causes, or producing some other response, from Schulz and Gopnik (2004, Experiment 3). Predictions of the Bayesian model are given in parentheses. Boldface values indicate majority.

Ontology:

Types	Number	Predicates	Values	
Flower	$N_F \sim P_B$	**Present(Flower,Trial)**	**Boolean:**	$\{\mathbf{T}, \mathbf{F}\}$
Monkey	$N_M \sim P_M$	**Sneezes(Monkey,Trial)**	**Boolean:**	$\{\mathbf{T}, \mathbf{F}\}$
Trial	$N_T \sim P_T$			

Plausible relations:

$\mathbf{Present}(\mathbf{F}, \mathbf{T}) \rightarrow \mathbf{Sneezes}(\mathbf{M}, \mathbf{T})$
True for all $\mathbf{T}$ with probability p for each $\mathbf{F}, \mathbf{M}$ pair

Functional form:

$\mathbf{Present}(\mathbf{F}, \mathbf{T}) \sim$ Bernoulli($\cdot$)
$\mathbf{Sneezes}(\mathbf{M}, \mathbf{T}) \sim$ Bernoulli(ν) for ν from a noisy-OR:

Cause	Strength
(Background)	$w_0 = \epsilon$
$\mathbf{Present}(\mathbf{F}, \mathbf{T})$	$w_i = 1 - \epsilon$

Figure 25. Theory for causal induction with "biology" (sneezing monkeys).

identical in their assumptions about the plausibility of causal relationships and the functional form of those relationships. For the stimuli shown to the children, both theories generate the same hypothesis space of causal graphical models.

Under the theories shown in Figures 25 and 26, the influences of multiple causes on both sneezing and scaring are described by the noisy-OR parameterization. The use of the same functional form across the two theories is a consequence of the applicability of the same set of assumptions about the nature of causal relationships in the domain: that causes influence their effects probabilistically, and that each of these influences has an independent opportunity to do so. This shared functional form results in the same predictions for the two conditions, as shown in Table 7. These predictions used $p = .5$ and $\varepsilon = .05$ for both biology and psychology.

The Effect of Domain on Plausibility

In many of the examples discussed in the previous sections, assumptions about the plausibility of a causal relationship played a less important role than assumptions about the underlying ontology and the functional form of causal relationships. In part, this is because the stimuli used in experiments on causal learning tend to involve variables among which causal relationships are quite plausible. Studying problems of causal learning involving variables from different domains provides an opportunity to explore the effect of domain on the plausibility of causal relationships. In particular, one might expect that plausible relationships would be restricted to causes that use forces appropriate to that domain. For example, asking a box to move is far less likely to be successful than dragging it. Using examples similar to this, Schulz and Gopnik (2004) have recently investigated how children assess the plausibility of causal relationships across domains and how this assessment interacts with statistical evidence.

Schulz and Gopnik (2004, Experiment 4) introduced children to causal systems in two domains. The physical domain involved a machine that made noise, with the candidate causes of the activation of the machine being two magnetic buttons (the in-domain objects **a** and **b**) and speech (the out-domain object **c**). The psychological domain involved reasoning about what might make a person giggle. The in-domain objects **a** and **b** were silly faces, and the out-domain object **c** was a switch. In each domain, children

Ontology:

Types	Number	Predicates	Values	
Beast	$N_B \sim P_B$	$\mathbf{Present}(\mathbf{Beast}, \mathbf{Trial})$	**Boolean:**	$\{\mathbf{T}, \mathbf{F}\}$
Rabbit	$N_R \sim P_R$	$\mathbf{Scared}(\mathbf{Rabbit}, \mathbf{Trial})$	**Boolean:**	$\{\mathbf{T}, \mathbf{F}\}$
Trial	$N_T \sim P_T$			

Plausible relations:

$\mathbf{Present}(\mathbf{B}, \mathbf{T}) \rightarrow \mathbf{Scared}(\mathbf{R}, \mathbf{T})$
True for all $\mathbf{T}$ with probability p for each $\mathbf{B}, \mathbf{R}$ pair

Functional form:

$\mathbf{Present}(\mathbf{B}, \mathbf{T}) \sim$ Bernoulli($\cdot$)
$\mathbf{Scared}(\mathbf{R}, \mathbf{T}) \sim$ Bernoulli(ν) for ν from a noisy-OR:

Cause	Strength
(Background)	$w_0 = \epsilon$
$\mathbf{Present}(\mathbf{B}, \mathbf{T})$	$w_i = 1 - \epsilon$

Figure 26. Theory for causal induction with "psychology" (scared rabbits).

were first asked which objects were likely to produce the effect, and they unanimously identified the in-domain causes. They then saw a series of trials exactly the same as those used in the test condition of Schulz and Gopnik's (2004) Experiment 3, discussed above. As shown in Table 8, the majority of the children now identified the out-domain object **c** as the cause, despite its low initial plausibility.

A simple theory that characterizes both the physical and the psychological stimuli used by Schulz and Gopnik (2004) is shown in Figure 27. Under this theory, both in-domain and out-domain objects can influence the effect, but the plausibility of such relationships differs. The probability of an in-domain relationship is set by p, while the probability of an out-domain relationship is set by q. The hypothesis space generated by this model for objects **a**, **b**, and **c** is functionally equivalent to that shown in Figure 12, using A to indicate the presence of **a**, B to indicate the presence of **b**, C to indicate the presence of **c**, and E to indicate the activation of the effect.

The initial responses of the children in Schulz and Gopnik's (2004) experiment indicates that q is much less than p. The predictions of the model with $p = .4$, $q = .1$, and $\varepsilon = .1$ are shown in Table 8. The model identifies **c** as a cause, despite its low plausibility, because of the strong assumptions about functional form. This effect can be best understood by considering the limit as $\varepsilon \rightarrow 0$. If E never occurs in the absence of a cause, then seeing E occur in the presence of **c** provides unambiguous evidence that C causes E. Thus, provided q takes on some value greater than zero, the probability that C causes E will be 1.00. Allowing ε to take on values greater than zero increases the influence of q on the outcome. In particular, if ε is somewhat greater than q, it becomes more likely that A and B are causes of E, and the causal relationship simply failed to manifest on the trials when **a** and **b** were present.

Schulz and Gopnik (2004) interpreted this experiment as indicating that children are aware of domain-specific constraints on causal relationships but that these constraints can be overridden by domain-general principles of causal learning. In the analysis above, the probability of an out-domain variable being involved in a causal relationship, q, has little effect on the predictions of the model: The assumptions about the functional form of the causal relationship mean that C will be identified as the cause even if q is very small. The model predicts that the value of q would have a greater effect given ambiguous evidence. For example, seeing $e^+|a^+b^-c^+$ and $e^+|a^-b^+c^+$ would suggest that A and B cause E if q is small, and that C causes E if q is large. Examining inferences from ambiguous evidence thus provides an opportunity to explore whether children really assign lower prior probability to cross-domain causal relationships.

Bonawitz, Griffiths, and Schulz (2006) tested the prediction that ambiguous evidence should trade off with priors based on domain, presenting preschoolers with ambiguous evidence involving either in-domain or cross-domain causes. The experiment was conducted with 4- and 5-year-old children, who were assigned to either a baseline or an evidence condition. Children in both conditions saw both within-domain and cross-domain storybooks. In the baseline condition, the storybooks featured the events of a single day, in which a character interacted with two potential causes (**a** and **b**) of an effect. For example, in the within-domain storybook, a deer ran in two places and got itchy spots. In the cross-domain story book, a bunny rabbit was scared (cause **a**) and ate something (cause **b**), then got a tummy ache. Children were then asked to select which of the causes they thought was responsible for the effect. In the evidence condition, children saw 7 days in the life of the character, and one of the two causes (**a**) recurred on each day, together with a new potential cause each day. This provides ambiguous evidence that **a** is the cause, as it is also possible that each of the seven other potential causes produces the effect. At the end of the story, children were asked which of the causes was responsible for the effect. The results are shown in Figure 28 together with the predictions of the theory-based Bayesian model outlined above, with $p = .4$ and $q = .1$ and $\varepsilon = .001$. Children's responses match the qualitative predictions of this model, with the cross-domain cause initially being given less credence than the within-domain cause, but becoming increasingly acceptable as a result of the provision of the ambiguous evidence. The model provides a reasonable quantitative match to the data, with $r = .84$, although the model predicts a stronger effect of the ambiguous evidence than was observed with children. A replication of the cross-domain condition (Experiment 2 of Schulz, Bonawitz, & Griffiths, 2007) found that 60% of children chose A, expressing a preference more consistent with that of the model. However, in general, children were more conservative than the model predicts, being more affected by the prior and less affected than the data.

Table 8
Effect of Domain on Plausibility

Condition	**c**		**a**, **b**		Other	
Physical	**.75**	**(.72)**	.00	(.09)	.25	(.19)
Psychological	**.81**	**(.72)**	.00	(.09)	.19	(.19)

Note. Numbers indicate the proportion of children identifying **c** as a cause, identifying **a** and/or **b** as a cause, or producing some other response, from Schulz and Gopnik (2004, Experiment 4). Predictions of the Bayesian model are given in parentheses. Boldface values indicate majority.

Summary

Different domains operate by different causal principles, a fact that can be captured in the theory-based framework by using different causal theories and by allowing relationships to differ in their plausibility. However, such differences need not always result in different behavior: As shown by the results of Schulz and Gopnik (2004, Experiment 3), if the theories that describe causal systems in two domains imply the same constraints on causal graphical models, then we should expect causal inferences to have much the same character across those systems. Likewise, while knowledge of the causal principles by which different domains operate can influence the plausibility of causal relationships, strong assumptions about functional form can overwhelm the effects of plausibility, as in Schulz and Gopnik (2004, Experiment 4). These experiments thus leave open the possibility that domain-specific knowledge guides causal induction. Some support for this idea is provided by subsequent experiments, which suggest that causes that cross domains are assigned lower prior probability (Bonawitz et al., 2006; Schulz et al., 2007).

Ontology:

Types	Number
Cause	$N_C \sim P_C$
InDomain	·
OutDomain	·
Effect	$N_E \sim P_O$
Trial	$N_T \sim P_T$

Predicates	Values	
Present(**Cause**, **Trial**)	**Boolean:**	$\{\mathbf{T}, \mathbf{F}\}$
Active(**Effect**, **Trial**)	**Boolean:**	$\{\mathbf{T}, \mathbf{F}\}$

Plausible relations:

$\mathbf{Present}(\mathbf{C}, \mathbf{T}) \rightarrow \mathbf{Active}(\mathbf{E}, \mathbf{T})$
True for all $\mathbf{T}$ with probability p for each $\mathbf{C}, \mathbf{E}$ pair where $\mathbf{C}$ is an **InDomain** cause, and with probability q for each $\mathbf{C}, \mathbf{E}$ pair where $\mathbf{C}$ is an **OutDomain** cause

Functional form:

$\mathbf{Present}(\mathbf{C}, \mathbf{T}) \sim$ Bernoulli(·)
$\mathbf{Active}(\mathbf{E}, \mathbf{T}) \sim$ Bernoulli(ν) for ν from a noisy-OR:

Cause	Strength
(Background)	$w_0 = \epsilon$
$\mathbf{Active}(\mathbf{C}, \mathbf{T})$	$w_i = 1 - \epsilon$

Figure 27. Theory for causal induction across domains.

Discussion

We have presented a computational framework for the explanation of human causal induction—theory-based causal induction. The case studies presented above illustrate how this framework can be used to provide a unifying framework in which explanations of causal learning in a broad range of settings can be presented. The combination of powerful statistical inference with constraints provided by prior knowledge makes it possible to explain some of the most remarkable aspects of human causal learning: that we can identify causal structure from only a few observations, that we can identify complex hidden causal structures with ease, and that we can make inferences from spatial and temporal coincidences. None of the alternative accounts we have considered are capable of explaining all of these phenomena.

Expressing a wide range of causal induction problems within this framework makes their common structure apparent and helps to clarify their differences. Rather than being separate phenomena requiring different kinds of explanations, causal induction from contingency data, learning about physical causal systems, and perceptual causality can be seen as lying on a continuum expressing the strength of constraints that prior knowledge imposes on causal inference. In contingency learning, relatively weak prior knowledge results in a need for relatively large samples, but expectations about the form of causal relationships still influence people's judgments. With simple physical systems such as the blicket detector or stick-ball machine, strong prior knowledge (such as the assumption that physical devices are near deterministic) make it possible for both adults and children to reach strong conclusions from small amounts of evidence. Finally, when making inferences from spatial or temporal data, strong expectations about the spatiotemporal characteristics of causal relationships mean that a single event can be sufficient to reveal the underlying causal structure.

In the remainder of the article, we deal with a set of questions examining the adequacy of this account. The first question concerns the characterization of the role of mechanism knowledge in causal induction. We attempt to articulate exactly which aspects of mechanism knowledge influence causal induction and identify the relationship between the notion of causal mechanism and the causal theories that form the basis of our account. We then turn to two questions concerning our treatment of theories—how the

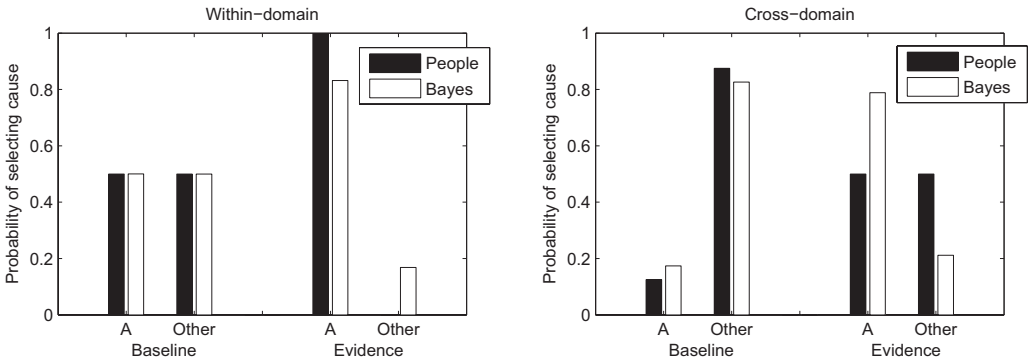

Figure 28. Results from Bonawitz et al. (2006). When presented with ambiguous evidence for a cross-domain causal relationship (cause **a**), 4- and 5-year-old children initially prefer a within-domain cause (here, labeled "other") and then gradually shift to favor the cross-domain cause as evidence mounts.

relatively low-level theories that we have discussed can be extended and how such theories might be learned—before briefly considering the kinds of psychological processes that might support rational theory-based causal induction.

Causal Mechanisms and Causal Theories

As mentioned at the start of the article, psychological theories about causal induction have traditionally 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 & Novick, 1990, 1992; Shanks & Dickinson, 1987), whereas mechanism-based approaches focus on the role of prior knowledge about the mechanisms by which causal force can be transferred (e.g., Ahn & Kalish, 2000; Shultz, 1982b; White, 1995). Recently, these two approaches have begun to be brought together through theories that consider how prior knowledge can be combined with covariational evidence (Koslowski, 1996; Waldmann, 1996; Waldmann et al., 2006). Our theory-based approach formalizes the interaction between prior knowledge and statistical learning in causal induction. In this section we attempt to clarify how the theory-based prior knowledge appealed to by our approach connects to the notion of causal mechanism.

When researchers refer to *causal mechanism*, they typically mean the chain of events mediating between cause and effect, as illustrated in Figure 29a (e.g., Bullock et al., 1982; Glymour, & Cheng, 1998; Shultz, Pardo, & Altmann, 1982; see Shultz & Kestenbaum, 1985, for a discussion of different kinds of mechanism). However, a detailed understanding of the mechanisms mediating between cause and effect is clearly not necessary for causal induction—if one possessed such knowledge there would be nothing to learn. Furthermore, recent studies investigating the limits of people's understanding of causal systems suggests that in fact, our mechanism knowledge may look more like Figure 29b. For example, Rozenblit and Keil (2002) found that when asked to explain how mechanical systems like crossbows and helicopters work, people radically overestimated the extent of their mechanism knowledge. It seems that, in general, our causal knowledge identifies the fact that a mechanism exists but does not necessarily articulate all of the steps that connect cause and effect (Keil, 2003).

Results like those of Rozenblit and Keil (2002) raise an interesting question: If our knowledge of causal mechanisms is as shallow as it appears to be, how is it possible for this knowledge to inform causal induction? Our theory-based account provides an answer to this question. Under our account, prior knowledge plays two important roles in causal induction: identifying which relationships are plausible and characterizing the functional form of those relationships. The shallow mechanism knowledge described by Keil (2003) is sufficient to fulfill these roles. Whether a causal relationship seems plausible is affected by mechanism knowledge, but the key determinant in this decision is not the particular details of the causal mechanism, but whether such a mechanism could exist. Similarly, evaluating the functional form of a causal relationship does not require knowing every step between cause and effect, but knowing what kind of relationship those steps might produce.

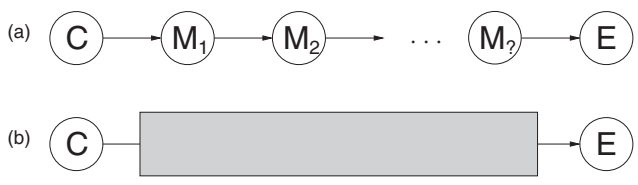

Figure 29. Two conceptions of causal mechanism knowledge. (a) The causal mechanism specifies the chain of events mediating between cause *C* and its effect *E*. (b) Often, people know that some mechanism (M) exists, but not the details.

Neither assessing plausibility nor specifying functional form requires a detailed account of a chain of events from cause to effect. As mentioned previously, a number of studies have suggested that simply believing a mechanism could exist makes a causal relationship seem more plausible (Koslowski, 1996; Koslowski & Okagaki, 1986; Koslowski et al., 1989). Lack of complete mechanism knowledge is likewise no impediment to reasoning about the functional form of a causal relationship. This point is illustrated through our use of Halley's inference as an example of a case in which prior knowledge guided causal induction. Newton's theory was notoriously amechanistic, departing from its forebears by introducing forces unmediated by particles (e.g., Westfall, 1977). Physicists are still engaged in the project of providing a mechanistic account of Newton's ideas, and in particular the force of gravity. Although Halley did not know the means by which the masses of stars and planets influenced the orbits of comets, he was still able to use information about the form of this influence to reason about the cause of the events that he observed. Indeed, in introducing his own account of causality, Pearl (1996) reduced mechanisms to "nothing but ordinary physical laws, cast in the form of deterministic equations" (p. 432), being no more than the specification of the functional form of the relationship between two variables.

The causal theories used in theory-based causal induction can thus be viewed as expressing the consequences of the shallow knowledge people possess of the mechanisms that operate in different domains. However, these theories do not express that knowledge directly: They are just as amechanistic as Newton's theory of physics, characterizing the possible relationships among entities and their form. The theories we have described are the constraints on the functional relationships among variables that can be the consequence of mechanism knowledge rather than the knowledge itself. Many mechanisms can imply the same set of constraints, as illustrated in the discussion of domain specificity in the previous section.

Distinguishing between theories and mechanisms provides an important insight into how causal induction is possible. A major challenge for mechanism-based accounts of causal induction is explaining how new causal mechanisms might be learned: If all causal induction requires mechanism knowledge, one can never discover a relationship that suggests a new mechanism (e.g., Cheng, 1993). If theory and mechanism are distinct, it becomes possible to learn a set of causal relationships without knowing their underlying mechanism. The existence of these relationships can then encourage the search for a mechanism that accounts for them, and the discovery of such a mechanism justifies further inferences about possible causal relationships. Such a pattern is extremely common in science—most of the suspicious coincidences that suggest new causal relationships are followed by a search for an explanatory mechanism (cf. Griffiths & Tenenbaum, 2007a).

Higher Level Causal Theories

The causal theories that we have used throughout this article have the same constituents as the intuitive theories discussed in accounts of cognition and cognitive development—ontologies and causal laws—but are far more limited in scope. Our analogy to Wellman's (1990; Wellman & Gelman, 1992) notion of a framework theory is useful in indicating the way that multiple levels of causal knowledge interact, but misleading in suggesting that the causal theories we describe are at the same level of generality as framework theories. Framework theories are supposed to provide the fundamental principles used in organizing an entire domain, such as intuitive physics, biology, or psychology. The theories we have presented characterize the principles that underlie very simple causal systems, with well-delineated boundaries. Although our theory of Nitro X makes predictions for any array of arbitrarily many cans of explosive, it hardly provides a complete theory of all physical systems.

The theories described in the article constitute the lowest level of causal theory, identifying the causal principles involved in specific kinds of systems. Doing justice to the notion of a framework theory requires postulating the existence of higher level theories, which express principles common to many systems. Following our definition of theories as hypothesis space generators, these higher level theories should generate hypothesis spaces that consist of different theories of particular causal systems (Tenenbaum et al., 2007). As discussed in the next section, such hypothesis spaces can be used to determine which theory best characterizes the forces at work in a system.

Identifying the kind of principles that should be included in these higher level theories requires investigating the assumptions that guide causal induction across a wide range of systems. By examining inferences involving different systems in physics, biology, and psychology, it should become possible to pick out the domain-specific principles that generate the theories of these systems that people use. Our investigation of different physical systems, such as the stick-ball machine and Nitro X, suggests what some of these principles might be for the domain of physics. The theories that we used to explain people's inferences about these systems had much the same character, using hidden causes (*prime movers*) to inject mechanical energy into the system, and having rules stating that no object changes state without a cause, and that causes produce changes in state with high probability. These principles may be a part of the higher level theory that organizes knowledge about all physical systems, constituting a part of our intuitive physics.

Learning Causal Theories

We have argued that human causal induction—the inference to causal structure from data—can be viewed as the result of a statistical inference comparing hypotheses generated by a causal theory. This approach explains how people are able to infer causal relationships from small samples and identify complex causal structures. However, it raises a new problem: explaining how people learn causal theories.

Our account of how people learn causal theories will generally have the same character as our account of how people learn causal structure. As stated above, the three levels of representation used in our account—theories, hypotheses, and data—and the assumption that each level generates the one below define a hierarchical Bayesian model (Tenenbaum et al., 2006). This model can be used to make inferences about any level in the hierarchy. As mentioned in our discussion of the blicket detector, the data D can be used to make inferences about the theory T—this just requires applying Bayes' rule at the level of theories (Equation 8). However, in the case of the blicket detector, there were only two theories to compare. In general, the problem of learning a theory is much more complex.

Applying Bayes' rule at the level of theories, T, requires having a prior distribution over such theories, $P(T)$. Just as people have strong expectations about the causal relationships that might hold in a given system, they have strong expectations about the kind of causal relationships that could operate in a domain. For example, theories of the stick-ball machine are constrained by beliefs about how any physical system could work. These constraints are expressed in higher level domain theories. Assuming that such higher level theories generate theories of particular systems, such higher level theories provide a hypothesis space of theories, T, and a prior on those theories.

Developing and testing this account of how theories are learned provides an important direction for future research. Our discussion of the blicket detector example illustrates how people might learn the functional form of a causal relationship, showing that assumptions about deterministic relationships can be easily overruled. Empirical investigation of the predictions of this account would reveal whether people's judgments can change in the same way. The question of how types of entities might be inferred, and how this interacts with plausibility, has recently been explored by Tenenbaum and Niyogi (2003), who showed that people could learn about the existence of different types of entities purely on the basis of the causal relationships in which they participated. Kemp, Griffiths, and Tenenbaum (2004) have developed a computational model that explains how such learning can take place, using Bayesian inference to simultaneously identify the number of types, the types themselves, and the plausibility of causal relationships among entities of those types, and Kemp et al. (2007) showed how this approach could be applied to human learning of simple theories.

Providing an account of how people learn causal theories of specific systems that appeals to higher level causal theories raises a new problem: explaining how these higher level theories are learned. Changes in the causal theories of domains constitute some of the most interesting phenomena in cognitive development, and in the history of science. However, at this point, concerns about an infinite recursion, providing no ultimate solution to the question of how people learn causal relationships, seem justified. There are three reasons not to be concerned by such a recursion. First, the mechanism by which the inference is performed at each step is the same—regardless of the level of representation, inferences about the level above can be made using Bayesian inference. There is thus no mysterious new force of learning that enters at any point. Second, this inference can become simpler at each level, with potentially fewer hypotheses about the more abstract principles that organize a domain, and a broader range of observed data that provide information about those principles. Third, the recursion is not infinite. At some point, it grounds out in a set of basic assumptions about the nature of causality, which provide constraints on the most general domain theories. Some of the implications of this idea of a recursive hierarchy of causal theories are explored in Tenenbaum et al. (2007).

Psychological Mechanisms Supporting Rational Causal Induction

Our focus in this article has been on providing a computational-level analysis of causal induction, rather than an account of the representations and algorithms that people use to solve this problem, or the way in which those representations and algorithms are implemented. However, Marr's (1982) levels of analysis were not intended to be completely independent, with constraints from the different levels influencing one another. In this case, we can ask what constraints our computational-level analysis might impose on the kinds of psychological processes we might expect to support causal induction.

The most fundamental question that our analysis raises is that of how human minds might even be capable of performing the kind of computations that are involved in the models of causal induction that we have considered. The number of causal graphical models that can be expressed using a given set of variables grows exponentially in the number of variables, and the computational cost of performing Bayesian inference increases linearly with the size of the hypothesis space. Consequently, it would seem that we would rapidly run into serious computational limitations in the kind of causal learning we might expect people to be able to perform.

Our answer to this problem comes in two parts. The first part is built into our framework, through the idea that theories provide constraints on causal learning. These constraints effectively reduce the size of the hypothesis space involved in causal learning—something that makes it easier to learn from small amounts of data, but that also reduces the amount of computation required in learning. As we have argued throughout the article, the strong constraints provided by prior knowledge are the key to why people are so good at causal induction. Indeed, recent work in machine learning has begun to explore ways of constraining hypotheses about causal structure in order to be able to learn from smaller amounts of data and to do so with less computation (e.g., Segal, Pe'er, et al., 2003; Segal, Shapira, et al., 2003).

In considering the psychological plausibility of this kind of constrained combinatorial computation, it is worth returning to our linguistic analogy. Under a standard (generative grammar inspired) account of sentence processing, when people hear a sentence they make an inference about the syntactic structure underlying the words in the sentence. This is exactly the same kind of computation as we are proposing the human mind performs in causal learning: The space of possible syntactic structures increases exponentially in the length of the sentence, and the grammar provides constraints that combine with the words to yield the most likely interpretation. If people can interpret sentences, even doing so online, in a way that is robust to interference, then they have all the computational resources required to perform theory-based causal induction.

The second part of our answer is that we expect the mind not to be performing all of the computations required for Bayesian inference but to be using an efficient approximation. One source of heuristic approximations for the kind of computations involved in causal inference is Monte Carlo simulation: rather than considering all possible hypotheses, considering a randomly generated subset. Monte Carlo techniques are commonly used in the machine learning literature on learning the structure of graphical models (e.g., Friedman & Koller, 2000) and can also be found in solutions to the problem of statistical parsing we have been using in our analogy (e.g., Johnson, Griffiths, & Goldwater, 2007). In particular, a technique known as a *particle filter* provides a way to maintain only a finite number of hypotheses over time, updating those hypotheses as new data become available in order to approximate the new posterior distribution (Doucet, Freitas, & Gordon, 2001). Particle filters have been proposed as psychological models in other settings, including categorization (Sanborn, Griffiths, & Navarro, 2006), change detection (Brown & Steyvers, 2009), and sentence processing (Levy, Reali, & Griffiths, 2009).

We remain agnostic about the specific psychological processes underlying causal induction and open to proposals of ways in which simple heuristics and associative learning might be used to approximate the statistical computations at the heart of our account. However, our computational-level analysis provides clearly specified goals for what such algorithms and representations should aim to approximate, and we suspect that considering strategies for approximating these computations will provide a productive means of generating hypotheses about psychological mechanisms that could support rational theory-based causal induction.

Conclusion

The human ability to infer causal structure from observed data has traditionally been studied in several very different settings, with correspondingly different kinds of cognitive capacities invoked to explain this ability. Causal learning from contingency data has been viewed as the result of associative learning or intuitive statistical reasoning processes. Learning about causality in simple physical systems has been explained in terms of mechanistic reasoning. Detecting hidden causal structure in patterns of spatial and temporal coincidence has been attributed to innate, modular perceptual processes. However, all of these different species of causal induction have at their heart the same computational problem, which suggests common principles by which they might operate.

The common challenge is the classic problem of induction: to infer the true causal relations that hold among a set of events, given a small finite set of observations that could have been produced by (and hence are logically consistent with) many possible causal structures. This problem can be approached as a statistical inference, but learning from small amounts of data can be successful only if constrained by strong and appropriate prior knowledge. We have argued that in order to capture the background knowledge that guides human causal learning, we need a representational framework that is richer and more abstract than the language of causal graphical models or Bayesian networks that has recently become influential in psychology. We have shown how to formalize causal background knowledge in terms of probabilistic logical theories and shown how such theories can be integrated with a Bayesian framework for causal learning. The theory generates a hypothesis space of candidate causal structures that could explain a set of observed data, along with a prior distribution over those hypotheses which serves to favor some candidate explanations over others.

We have applied our theory-based approach to a number of case studies in causal induction, covering a broad range of different kinds of causal systems, and a variety of settings in which people learn about causal relationships. Learning in each of these settings can be explained as a Bayesian inference constrained by a different causal theory. The variation in human judgments across settings provides insight into the roles that background knowledge plays in guiding human causal learning: As background knowledge be-

comes stronger, both our models and human observers require less data to make confident causal inferences. By exploring the constraints that are necessary to explain causal inferences across a range of settings—from the orbits of comets to the workings of gears, patterns of explosions and patterns of disease—we can begin to understand the common principles governing how abstract knowledge is organized and used in solving some of the most difficult inductive problems at the heart of human cognition.

References

Ahn, W., & Kalish, C. (2000). The role of covariation vs. mechanism information in causal attribution. In R. Wilson & F. Keil (Eds.), *Cognition and explanation* (pp. 227–253). Cambridge, MA: MIT Press.

Aldous, D. (1985). Exchangeability and related topics. In P.-L. Hennequin (Ed.), *Ecole d'été de probabilités de Saint-Flour XIII–1983* (pp. 1–198). Berlin, Germany: Springer.

Allan, L. G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society, 15,* 147–149.

Allan, L. G. (1993). Human contingency judgments: Rule based or associative? *Psychological Bulletin, 114,* 435–448.

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

Alloy, L. B., & Tabachnik, N. (1984). Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information. *Psychological Review, 91,* 112–149.

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

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

Anderson, J. R., & Sheu, C.-F. (1995). Causal inferences as perceptual judgments. *Memory & Cognition, 23,* 510–524.

Atran, S. (1995). Classifying nature across cultures. In E. E. Smith & D. N. Osherson (Eds.), *Thinking: An invitation to cognitive science* (Vol. 3, pp. 131–174). Cambridge, MA: MIT Press.

Beckers, T., De Houwer, J., Pineno, O., & Miller, R. R. (2005). Outcome additivity and outcome maximality influence cue competition in human causal learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 238–249.

Blei, D. M., Griffiths, T. L., Jordan, M. I., & Tenenbaum, J. B. (2004). Hierarchical topic models and the nested Chinese restaurant process. In S. Thrun, L. K. Saul, & B. Schölkopf (Eds.), *Advances in neural information processing systems 16* (pp. 17–24). Cambridge, MA: MIT Press.

Bonawitz, E. B., Griffiths, T. L., & Schulz, L. (2006). Modeling cross-domain causal learning in preschoolers as Bayesian inference. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 89–94). Mahwah, NJ: Erlbaum.

Brown, S. D., & Steyvers, M. (2009). Detecting and predicting changes. *Cognitive Psychology, 58,* 49–67.

Buehner, M., & Cheng, P. W. (1997). Causal induction: The Power PC theory versus the Rescorla-Wagner theory. In M. Shafto & P. Langley (Eds.), *Proceedings of the 19th Annual Conference of the Cognitive Science Society* (pp. 55–61). Hillsdale, NJ: Erlbaum.

Buehner, M. J., Cheng, P. W., & Clifford, D. (2003). From covariation to causation: A test of the assumption of causal power. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 1119–1140.

Buehner, M. J., & May, J. (2002). Knowledge mediates the timeframe of covariation assessment in human causal induction. *Thinking & Reasoning, 8,* 269–295.

Buehner, M. J., & May, J. (2003). Rethinking temporal contiguity and the judgment of causality: Effects of prior knowledge, experience, and reinforcement procedure. *Quarterly Journal of Experimental Psychology, 56*(A), 865–890.

Buehner, M. J., & McGregor, S. (2006). Temporal delays can facilitate causal attribution: Towards a general timeframe bias in causal induction. *Thinking & Reasoning, 12,* 353–378.

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, NY: Academic Press.

Busemeyer, J., McDaniel, M. A., & Byun, E. (1997). The abstraction of intervening concepts from experience with multiple input–multiple output causal environments. *Cognitive Psychology, 32,* 1–48.

Carey, S. (1985a). *Conceptual change in childhood.* Cambridge, MA: MIT Press.

Carey, S. (1985b). Constraints on semantic development. In J. Mehler (Ed.), *Neonate cognition* (pp. 381–398). Hillsdale, NJ: Erlbaum.

Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences, 10,* 335–344.

Cheng, P. W. (1993). Separating causal laws from causal facts: Pressing the limits of statistical relevance. In D. L. Medin (Ed.), *The psychology of learning and motivation* (Vol. 30, pp. 215–264). San Diego, CA: Academic Press.

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

Cheng, P. W., & Holyoak, K. J. (1995). Complex adaptive systems as intuitive statisticians: Causality, contingency, and prediction. In H. L. Roitblat & J.-A. Meyer (Eds.), *Comparative approaches to cognitive science* (pp. 271–302). Cambridge, MA: MIT Press.

Cheng, P. W., & Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology, 58,* 545–567.

Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review, 99,* 365–382.

Cooper, G., & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning, 9,* 308–347.

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* (pp. 67–74). Cambridge, MA: MIT Press.

Danks, D., & McKenzie, C. R. M. (2009). *Learning complex causal structures.* Manuscript in preparation.

Dawes, R. M., & Hastie, R. (2001). *Rational choice in an uncertain world: The psychology of judgment and decision making.* Thousand Oaks, CA: Sage.

Dickinson, A., Shanks, D., & Evenden, J. (1984). Judgment of act-outcome contingency: The role of selective attribution. *Quarterly Journal of Experimental Psychology, 36*(A), 29–50.

Doucet, A., Freitas, N. de, & Gordon, N. (2001). *Sequential Monte Carlo methods in practice.* New York, NY: Springer.

Enderton, H. B. (1972). *A mathematical introduction to logic.* New York, NY: Academic Press.

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

Friedman, N., Getoor, L., Koller, D., & Pfeffer, A. (1999). Learning probabilistic relational models. In T. Dean (Ed.), *Proceedings of the 16th International Joint Conference on Artificial Intelligence* (pp. 1300–1309). San Francisco, CA: Morgan Kaufmann.

Friedman, N., & Koller, D. (2000). Being Bayesian about network structure. In *Proceedings of the 16th Annual Conference on Uncertainty in Artificial Intelligence* (pp. 201–202). San Francisco, CA: Morgan Kaufman.

Gelman, R., & Spelke, E. S. (1981). The development of thoughts about animate and inanimate objects: Implications for research in social cognition. In J. H. Flavell & L. Ross (Eds.), *The development of social cognition in children* (pp. 43–66). Cambridge, England: Cambridge University Press.

Gelman, S. A., & Wellman, J. M. (1991). Insides and essence: Early understandings of the nonobvious. *Cognition, 38,* 213–244.

Glymour, C. (1998). Learning causes: Psychological explanations of causal explanation. *Minds and Machines, 8,* 39–60.

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

Glymour, C., & Cheng, P. W. (1998). Causal mechanism and probability: A normative approach. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 295–313). Oxford, England: Oxford University Press.

Gopnik, A. (1996). The scientist as child. *Philosophy of Science, 63,* 485–514.

Gopnik, A., & Glymour, C. (2002). Causal maps and Bayes nets: A cognitive and computational account of theory-formation. In P. Carruthers, S. Stich, & M. Segal (Eds.), *The cognitive basis of science*. Cambridge, England: Cambridge University Press.

Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review, 111,* 1–31.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories.* Cambridge, MA: MIT Press.

Gopnik, A., & Sobel, D. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child Development, 71,* 1205–1222.

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

Greville, W. J., & Buehner, M. J. (2007). The influence of temporal distributions on causal induction from tabular data. *Memory & Cognition, 35,* 444–453.

Griffiths, T. L. (2005). *Causes, coincidences, and theories.* Unpublished doctoral dissertation, Stanford University.

Griffiths, T. L., Baraff, E. R., & Tenenbaum, J. B. (2004). Using physical theories to infer hidden causal structure. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (pp. 446–451). Mahwah, NJ: Erlbaum.

Griffiths, T. L., & Ghahramani, Z. (2005). *Infinite latent feature models and the Indian buffet process* (Tech. Rep. No. 2005–001). Gatsby Computational Neuroscience Unit, University College London, England.

Griffiths, T. L., Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2009). *Bayes and blickets: Effects of knowledge on causal induction in children and adults.* Manuscript submitted for publication.

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

Griffiths, T. L., & Tenenbaum, J. B. (2007a). From mere coincidences to meaningful discoveries. *Cognition, 103,* 180–226.

Griffiths, T. L., & Tenenbaum, J. B. (2007b). Two proposals for causal grammars. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 323–346). Oxford, England: Oxford University Press.

Hagmayer, Y., Sloman, S. A., Lagnado, D. A., & Waldmann, M. R. (2007). Causal reasoning through intervention. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 86–100). Oxford, England: Oxford University Press.

Hagmayer, Y., & Waldmann, M. R. (2002). How temporal assumptions influence causal judgments. *Memory & Cognition, 30,* 1128–1137.

Hagmayer, Y., & Waldmann, M. R. (2007). Inferences about unobserved causes in human contingency learning. *Quarterly Journal of Experimental Psychology, 60,* 330–355.

Hattori, M., & Oaksford, M. (2007). Adaptive noninterventional heuristics for covariation detection in causal induction: Model comparison and rational analysis. *Cognitive Science, 31,* 765–814.

Heckerman, D. (1998). A tutorial on learning with Bayesian networks. In M. I. Jordan (Ed.), *Learning in graphical models* (pp. 301–354). Cambridge, MA: MIT Press.

Hirschfeld, L. A., & Gelman, S. A. (Eds.). (1994). *Mapping the mind: Domain specificity in cognition and culture.* Cambridge, England: Cambridge University Press.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference, learning and discovery.* Cambridge, MA: MIT Press.

Hughes, D. W. (1990). Edmond Halley: His interest in comets. In N. J. M. Thrower (Ed.), *Standing on the shoulders of giants: A longer view of Newton and Halley* (pp. 324–372). Los Angeles, CA: University of California Press.

Hume, D. (1977). *An enquiry concerning human understanding*. Indianapolis, IN: Hackett. (Original work published 1748)

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

Humphreys, P., & Freedman, D. A. (1996). The grand leap. *British Journal of the Philosophy of Science, 47,* 113–123.

Inagaki, K., & Hatano, G. (2002). *Young children's thinking about biological world.* New York, NY: Psychology Press.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence.* London, England: Routledge and Kegan Paul.

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs, 79.*

Johnson, M., Griffiths, T. L., & Goldwater, S. (2007). Bayesian inference for PCFGs via Markov chain Monte Carlo. In *Proceedings of the North American Conference on Computational Linguistics* (pp. 139–146). Rochester, NY: Association for Computational Linguistics.

Kant, I. (1964). *Critique of pure reason.* London, England: Macmillan. (Original work published 1781)

Karmiloff-Smith, A. (1988). The child is a theoretician, not an inductivist. *Mind and Language, 3,* 183–195.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development.* Cambridge, MA: MIT Press.

Keil, F. C. (2003). Folkscience: Coarse interpretations of a complex reality. *Trends in Cognitive Sciences, 7,* 368–373.

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

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2007). Learning causal schemata. In D. McNamara & G. Trafton (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (pp. 389–394). New York, NY: Erlbaum.

Kemp, C., Griffiths, T. L., & Tenenbaum, J. B. (2004). *Discovering latent classes in relational data* (AI Memo 2004–019). Cambridge, MA: Massachusetts Institute of Technology.

Kersting, K., & De Raedt, L. (2000). Bayesian logic programs. In J. Cussens & A. Frisch (Eds.), *Proceedings of the work-in-progress track at the 10th International Conference on Inductive Logic Programming* (pp. 138–155). Berlin, Germany: Springer.

Koller, D., & Pfeffer, A. (1997). Object-oriented Bayesian networks. In *Proceedings of the 13th Annual Conference on Uncertainty in Artificial Intelligence* (pp. 302–313). San Francisco, CA: Morgan Kaufman.

Korb, K. B., & Wallace, C. S. (1997). In search of the Philosopher's Stone: Remarks on Humphreys and Freedman's critique of causal discovery. *British Journal of the Philosophy of Science, 48,* 543–553.

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

Koslowski, B., & Okagaki, L. (1986). Non-Humean indices of causation in problem-solving situations: Causal mechanism, analogous effects, and the status of rival alternative accounts. *Child Development, 57,* 1100–1108.

Koslowski, B., Okagaki, L., Lorenz, C., & Umbach, D. (1989). When covariation is not enough: The role of causal mechanism, sampling method, and sample size in causal reasoning. *Child Development, 60,* 1316–1327.

Kushnir, T., Gopnik, A., Schulz, L., & Danks, D. (2003). Inferring hidden causes. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the 25th*

Annual Conference of the Cognitive Science Society (pp. 699–703). Boston, MA: Cognitive Science Society.

Lagnado, D. A., & Sloman, S. (2002). Learning causal structure. In W. D. Gray & C. Schunn (Eds.), *Proceedings of the 24th Annual Conference of the Cognitive Science Society.* Mahwah, NJ: Erlbaum.

Lagnado, D., & Sloman, S. A. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 856–876.

Lagnado, D. A., & Sloman, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 451–460.

Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., & Sloman, S. A. (2007). Beyond covariation: Cues to causal structure. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 86–100). Oxford, England: Oxford University Press.

Laudan, L. (1977). *Progress and its problems.* Berkeley, CA: University of California Press.

Leslie, A. M. (1986). Getting development off the ground: Modularity and the infant's perception of causality. In P. van Geert (Ed.), *Theory building in developmental psychology* (pp. 406–437). Amsterdam, the Netherlands: North Holland.

Leslie, A. M. (1994). ToMM, ToBY, and agency: Core architecture and domain specificity. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 119–148). Cambridge, England: Cambridge University Press.

Levy, R., Reali, F., & Griffiths, T. L. (2009). Modeling the effects of memory on human online sentence processing with particle filters. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems 21* (pp. 937–944). Cambridge, MA: MIT Press.

Lien, Y., & Cheng, P. W. (2000). Distinguishing genuine from spurious causes: A coherence hypothesis. *Cognitive Psychology, 40,* 87–137.

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

López, F. J., Cobos, P. L., Caño, A., & Shanks, D. R. (1998). The rational analysis of human causal and probability judgment. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 314–352). Oxford, England: Oxford University Press.

Lovibond, P. F., Been, S., Mitchell, C. J., Bouton, M. E., & Frohart, 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 28th 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. McNamara & G. Trafton (Eds.), *Proceedings of the 29th 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.

Luhmann, C. C., & Ahn, W.-K. (2003). Evaluating the causal role of unobserved variables. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp. 734–739). Boston, MA: Cognitive Science Society.

Luhmann, C. C., & Ahn, W. (2007). BUCKLE: A model of unobserved cause learning. *Psychological Review, 114,* 657–677.

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing.* Cambridge, MA: MIT Press.

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

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

McCloskey, M. (1983). Intuitive physics. *Scientific American, 284,* 114–123.

Medin, D. L., & Atran, S. (Eds.). (1999). *Folkbiology.* Cambridge, MA: Bradford Books.

Melz, E. R., Cheng, P. W., Holyoak, K. J., & Waldmann, M. R. (1993). Cue competition in human categorization: Contingency or the Rescorla-Wagner rule? Comments on Shanks (1991). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 1398–1410.

Mendelson, R., & Shultz, T. R. (1976). Covariation and temporal contiguity as principles of causal inference in young children. *Journal of Experimental Child Psychology, 22,* 408–412.

Michotte, A. (1963). *The perception of causality.* New York, NY: Basic Books.

Milch, B., Marthi, B., & Russell, S. (2004). BLOG: Relational modeling with unknown objects. In T. Dietterich, L. Getoor, & K. Murphy (Eds.), *ICML 2004 Workshop on Statistical Relational Learning and Its Connections to Other Fields* (pp. 67–73). Berlin, Germany: Springer.

Muggleton, S. (1997). Learning from positive data. In *Selected papers from the 6th International Workshop on Inductive Logic Programming* (pp. 358–376). Berlin, Germany: Springer-Verlag.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92,* 289–316.

Navarro, D. J., Griffiths, T. L., Steyvers, M., & Lee, M. D. (2006). Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology, 50,* 101–122.

Neal, R. M. (1998). *Markov chain sampling methods for Dirichlet process mixture models* (Tech. Rep. No. 9815). Department of Statistics, University of Toronto, Toronto, Ontario, Canada.

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

Nichols, S., & Stich, S. (2003). *Mindreading: An integrated account of pretense, self-awareness and understanding other minds.* Oxford, England: Oxford University Press.

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

Oaksford, M., & Chater, N. (Eds.). (1998). *Rational models of cognition.* Oxford, England: Oxford University Press.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems.* San Francisco, CA: Morgan Kaufmann.

Pearl, J. (1996). *Structural and probabilistic causality*. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *The psychology of learning and motivation: Causal learning* (Vol. 34, pp. 393–403). San Diego, CA: Academic Press.

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

Pearson, K. (1948). On the theory of contingency and its relation to association and normal correlation. In *Karl Pearson's early statistical papers* (pp. 443–475). Cambridge, England: Cambridge University Press. (Original work published 1904)

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

Perner, J. (1991). *Understanding the representational mind.* Cambridge, MA: MIT Press.

Pitman, J. (2002). *Combinatorial stochastic processes* [Notes for Saint Flour Summer School]. http://dx.doi.org/10.1007/b11601500

Price, P. C., & Yates, J. F. (1993). Judgmental overshadowing: Further evidence of cue interaction in contingency judgment. *Memory & Cognition, 21,* 561–572.

Rehder, B. (2003). A causal-model theory of conceptual representation and

categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 1141–1159.

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

Rozenblit, L. R., & Keil, F. C. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science, 26,* 521–562.

Russell, S. J., & Norvig, P. (2002). *Artificial intelligence: A modern approach* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Salmon, W. C. (1980). *Scientific explanation and the causal structure of the world.* Princeton, NJ: Princeton University Press.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2006). A more rational model of categorization. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 726–731). Mahwah, NJ: Erlbaum.

Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2006). *A Bayesian framework for modeling intuitive dynamics*. Poster presented at the 47th Annual Meeting of the Psychonomic Society, Houston, TX.

Saxe, R., Tenenbaum, J. B., & Carey, S. (2005). Secret agents: Inferences about hidden causes by 10- and 12-month-old infants. *Psychological Science, 16,* 995–1001.

Schlottmann, A. (1999). Seeing it happen and knowing how it works: How children understand the relation between perceptual causality and underlying mechanism. *Developmental Psychology, 35,* 303–317.

Schlottmann, A., & Shanks, D. R. (1992). Evidence for a distinction between judged and perceived causality. *Quarterly Journal of Experimental Psychology, 44*(A), 321–334.

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.

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

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

Segal, E., Pe'er, D., Regev, A., Koller, D., & Friedman, N. (2003). Learning module networks. In *Proceedings of the 19th Annual Conference on Uncertainty in Artificial Intelligence* (pp. 525–534). San Francisco, CA: Morgan Kaufmann.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., & Friedman, N. (2003). Module networks: Identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics, 34,* 166–176.

Shanks, D. R. (1987). Acquisition functions in causality judgment. *Learning and Motivation, 18,* 147–166.

Shanks, D. R. (1995a). Is human learning rational? *Quarterly Journal of Experimental Psychology, 48*(A), 257–279.

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

Shanks, D. R., & Dickinson, A. (1987). Associative accounts of causality judgment. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 21, pp. 229–261). San Diego, CA: Academic Press.

Shanks, D. R., López, F. J., Darby, R. J., & Dickinson, A. (1996). Distinguishing associative and probabilistic contrast theories of human contingency judgment. In D. L. Medin, D. R. Shanks, & K. J. Holyoak (Eds.), *The psychology of learning and motivation* (Vol. 34, pp. 265–312). San Diego, CA: Academic Press.

Shanks, D. R., Pearson, S. M., & Dickinson, A. (1989). Temporal contiguity and the judgment of causality by human subjects. *Quarterly Journal of Experimental Psychology, 41*(B), 139–159.

Shepard, R. N. (1987, September 11). Towards a universal law of generalization for psychological science. *Science, 237,* 1317–1323.

Shultz, T. R. (1982a). Causal reasoning in the social and nonsocial realms. *Canadian Journal of Behavioural Science, 14,* 307–322.

Shultz, T. R. (1982b). Rules of causal attribution. *Monographs of the Society for Research in Child Development, 47*(Serial No. 194).

Shultz, T. R., Fisher, G. W., Pratt, C. C., & Rulf, S. (1986). Selection of causal rules. *Child Development, 57,* 143–152.

Shultz, T. R., & Kestenbaum, N. R. (1985). Causal reasoning in children. *Annals of Child Development, 2,* 195–249.

Shultz, T. R., Pardo, S., & Altmann, E. (1982). Young children's use of transitive inference in causal chains. *British Journal of Psychology, 73,* 235–241.

Sloman, S. (2005). *Causal models: How people think about the world and its alternatives.* Oxford, England: Oxford University Press.

Snow, J. (1855). *On the mode of communication of cholera.* London, England: Churchill.

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.

Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological Review, 99,* 605–632.

Spelke, E. S., Phillips, A. T., & Woodward, A. L. (1995). Infants' knowledge of object motion and human action. In A. J. Premack, D. Premack, & D. Sperber (Eds.), *Causal cognition: A multidisciplinary debate* (pp. 44–77). Oxford, England: Clarendon Press.

Spirtes, P., Glymour, C., & Schienes, R. (1993). *Causation prediction and search.* New York, NY: Springer-Verlag.

Springer, K., & Keil, F. C. (1991). Early differentiation of causal mechanisms appropriate to biological and nonbiological kinds. *Child Development, 62,* 767–781.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E. J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science, 27,* 453–489.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Structure learning in human causal induction. In T. Leen, T. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems 13* (pp. 59–65). Cambridge, MA: MIT Press.

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* (pp. 35–42). Cambridge, MA: MIT Press.

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

Tenenbaum, J. B., Griffiths, T. L., & Niyogi, S. (2007). Intuitive theories as grammars for causal inference. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 301–322). Oxford, England: Oxford University Press.

Tenenbaum, J. B., & Niyogi, S. (2003). Learning causal laws. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp. 1153–1157). Hillsdale, NJ: Erlbaum.

Tversky, A., & Kahneman, D. (1974, September 27). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131.

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

Waldmann, M. R. (2000). Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 53–76.

Waldmann, M. R. (2007). Combining versus analyzing multiple causes: How domain assumptions and task context affect integration rules. *Cognitive Science, 31,* 233–256.

Waldmann, M. R., & Hagmayer, Y. (2006). Categories and causality: The neglected direction. *Cognitive Psychology, 53,* 27–58.

Waldmann, M. R., Hagmayer, Y., & Blaisdell, A. P. (2006). Beyond the

information given: Causal models in learning and reasoning. *Current Directions in Psychological Science, 15,* 307–311.

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General, 121,* 222–236.

Waldmann, M. R., Holyoak, K. J., & Fratianne, A. (1995). Causal models and the acquisition of category structure. *Journal of Experimental Psychology: General, 124,* 181–206.

Waldmann, M. R., & Martignon, L. (1998). A Bayesian network model of causal learning. In M. A. Gernsbacher & S. J. Derry (Eds.), *Proceedings of the 20th Annual Conference of the Cognitive Science Society* (pp. 1102–1107). Mahwah, NJ: Erlbaum.

Wasserman, E. A., Elek, S. M., Chatlosh, D. C., & Baker, A. G. (1993). Rating causal relations: Role of probability in judgments of response-outcome contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 174–188.

Wellman, H. M. (1990). *The child's theory of mind.* Cambridge, MA: MIT Press.

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology, 43,* 337–375.

Westfall, R. S. (1977). *The construction of modern science: Mechanisms and mechanics.* Cambridge, England: Cambridge University Press.

White, P. A. (1990). Ideas about causation in philosophy and psychology. *Psychological Bulletin, 108,* 3–18.

White, P. A. (1995). *The understanding of causation and the production of action: From infancy to adulthood.* Hillsdale, NJ: Erlbaum.

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

Zelazo, P. R., & Shultz, T. R. (1989). Concepts of potency and resistance in causal prediction. *Child Development, 60,* 1307–1315.

Appendix A

Materials for Experiment on Manipulation of Functional Form

Generative Condition

Imagine that you are working in a laboratory and you want to find out whether certain chemicals cause certain genes to be expressed in mice. Below, you can see laboratory records for a number of studies. In each study, a random sample of mice were injected with a certain chemical and later examined to see if they expressed a particular gene. Each study investigated the effects of a *different* chemical on a *different* gene, so the results from different studies bear no relation to each other.

Of course, these genes may sometimes be expressed in animals not injected with a chemical substance. Thus, a random sample of mice who were not injected with any chemical were also checked to see if they expressed the same genes as the injected mice. For each study below, you can see how many of the injected mice were found to have expressed the gene, as well as how many of the uninjected mice were found to have expressed the same gene. What you must decide is whether a given chemical increases the likelihood of expressing a given gene.

Preventive Condition

Imagine that you are working in a laboratory and you want to find out whether certain chemicals prevent viruses in mice. Below, you can see laboratory records for a number of studies. In each study, a random sample of mice were injected with a certain chemical and later examined to see if they caught a particular virus. Each study investigated the effects of a *different* chemical on a *different* virus, so the results from different studies bear no relation to each other.

Of course, these viruses vary in how often they occur. Thus, a random sample of mice who were not injected with any chemical were also checked to see if they caught the same viruses as the injected mice. For each study below, you can see how many of the injected mice were found to have the virus, as well as how many of the uninjected mice were found to have the same virus. What you must decide is whether a given chemical decreases the likelihood of catching a given virus.

Difference Condition

Imagine that you are working in a laboratory and you want to find out about the genetic properties of different kinds of mice. Below, you can see records for a number of studies conducted by the laboratory. In each study, a random sample of mice from two different species were examined to see if they expressed a particular gene. Each study investigated the expression of a *different* gene in *different* species of mice, so the results from different studies bear no relation to each other.

For each study below, you can see how many of the mice were found to have expressed the gene for each of the two species. The mice in *GROUP 1* were all from one species, while the mice in *GROUP 2* were all from a second species. The number of mice expressing the gene provides you with information about the two species, but these numbers are also affected by chance. What you must decide is whether there is a real difference between the two species in their tendency to express the gene.

Appendix B

Bidirectional Causal Relationships

The hypothesis spaces generated by the theory of stick-ball machines can include structures in which balls can cause one another to move, such as that shown in Figure B1a. Cyclic relationships are usually prohibited in graphical models (Pearl, 1988), although some kinds of cycles can be dealt with in causal graphical models (e.g., Pearl, 2000). One common approach to dealing with cycles is to impose temporal structure on a set of events, unrolling the cycle into a set of dependencies that hold between two variables in successive time slices. In this Appendix, we outline how this approach can be used to deal with cyclic causal relationships in stick-ball machines.

Given N_B balls, with B_i indicating **Moves($\mathbf{b}_i$, T)** and N_H hidden causes, with H_j indicating **Active($\mathbf{h}_j$, T)**, the following procedure can be used to generate values of B_i and H_j on any given trial:

1. Determine which hidden causes are active by sampling the values of H_j.

2. Determine which balls are moved by the hidden causes by sampling the values of B_i, conditioned just on H_j. If $H_j \rightarrow B_i$ and $h_j = 1$, then $b_i = 1$ with probability ω.

3. Determine which balls move other balls. Every ball that moves has one opportunity to move the balls to which it has causal connections. If $B_i \rightarrow B_j$, b_j is currently 0, and $\mathbf{b}_i$ has not previously attempted to move $\mathbf{b}_j$, then $b_j = 1$ with probability ω.

4. Repeat Step 3 with the balls that were just moved by other balls. This procedure continues until all balls that have moved have had one opportunity to influence each of the balls to which they are connected.

This procedure implicitly defines a temporal succession of events, with hidden causes becoming active, then moving a subset of the balls, each of which moves some subset of the remaining balls, each of which moves further balls, and so forth. This temporal succession removes the cycles in the underlying causal graphical model, allowing events to unroll through time.

The generative procedure described in the previous paragraph can be expressed as a recipe for constructing an "unrolled" graphical model from the basic graphical model. The unrolled model can be used to compute the probability of events, and is constructed as follows:

1. Create $N_B + 1$ copies of B_i, numbered from 0 to N_B, indicating successive points in time within the trial. We use $B_i^{(t)}$ to refer to the copy of B_i at time t.

2. $B_i^{(t-1)} \rightarrow B_i^{(t)}$ for all $t \geq 1$. If $H_j \rightarrow B_i$ in the basic model, then $H_j \rightarrow B_i^{(1)}$ in the unrolled model. If $B_i \rightarrow B_j$ in the basic model, then $B_i^{(t-1)} \rightarrow B_j^{(t)}$ and $B_i^{(t-2)} \rightarrow B_j^{(t)}$ for all $t \geq 2$ in the unrolled model.

3. Set $b_i^{(0)} = 0$. Parameterize $B_i^{(1)}$ as a noisy-OR of H_j. For $t \geq 2$, parameterize $B_i^{(t)}$ as

$$P(b_i^{(t)} = 1 | b_\cdot^{(t-1)}, b_\cdot^{(t-2)}) = 1 - (1 - b_i^{(t-1)})(1 - \omega)^{\sum_{j \neq i} b_j^{(t-1)}(1 - b_j^{(t-2)})},$$

where $b_\cdot^{(t-1)}$ denotes the value of all b_j at $t - 1$. Under this parameterization $b_i^{(t)} = 1$ if $b_i^{(t-1)} = 1$, and is otherwise a noisy-OR of all B_j that changed between $t - 2$ and $t - 1$.

The probability of a set of observed values for B_i in the basic model is the probability of $B_i^{(N_B)}$ taking those values in the unrolled model, summing over all latent variables. Figure B1b shows the unrolled model for the basic model shown in Figure B1a.

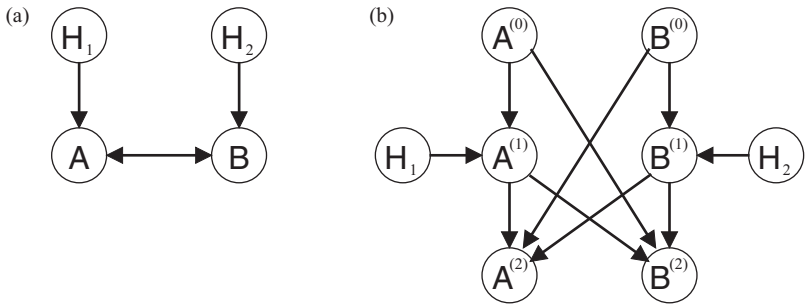

Figure B1. Dealing with bidirectional causal relationships in stick-ball machines. (a) A causal graphical model generated by the theory of stick-ball machines. H_1 and H_2 are hidden causes, with A and B the motion of two balls. (b) The same model "unrolled" through time, removing the cyclic causal relationship. Superscripts indicate a sequence of points in time, with each variable now being represented distinctly at each point.

(*Appendixes continue*)

Appendix C
Materials for Cancer Cluster Experiment

Instructions for the cancer cluster experiment were as follows:

> Researchers are investigating the effects of environmental contaminants on the spatial distribution of rare forms of cancer. They are studying 12 different rare cancers, trying to establish which cancers show evidence of being caused at least in part by some localized environmental factor, such as living near to a toxic waste dump or a chemical leak, in addition to random genetic mutations that are thought to be the usual causes of these diseases in the general population.
>
> Each of the images below shows the places of residence ("x" marks) for every individual who developed a particular rare form of cancer within a one square mile area of a major city in a single year. (Approximately 20,000 people live within this area.) For each image, please rate HOW LIKELY you think it is that there is an underlying environmental cause that makes people living near some particular location more likely than the general population to develop that particular disease. Use a scale from 1 to 10, where 1 means "very UNLIKELY to have a localized environmental cause," and 10 means "very LIKELY to have a localized environmental cause."

The stimuli consisted of three sets of 12 images. Each set was generated by the same stochastic process—a mixture of a uniform and a Gaussian distribution—but the parameters of the mixture were varied as shown in Table C1. Since there was some randomness in the stimulus generation program, three different sets were used to make sure the effects found were not due to some idiosyncrasy in the particular stimuli used. Each participant saw one of these three sets.

Results with each of the three sets were virtually identical and they were averaged together for all analyses presented in the article. Shown next to each of the 12 types of stimuli are the parameters of the random process used to generate them. In addition, the stimulus generating program filtered out any stimuli where events occurred too close to the boundary, too close to each other, or without sufficient separation between the events generated by the Gaussian cluster and those drawn uniformly over the whole rectangle.

Table C1
Parameters Used to Generate Stimuli for Cancer Cluster Experiment

Image	Number of points	Number in cluster	Variance of cluster
1	3	2	.2
2	6	4	.2
3	12	8	.2
4	8	2	.2
5	8	3	.2
6	8	4	.2
7	8	5	.2
8	6	4	.8
9	6	4	.05
10	3	0	.2
11	8	0	.2
12	12	0	.2

Note. All points were plotted in a square 10 units in width and length, with the locations of points outside the cluster and the mean of the cluster all being chosen uniformly in that space. The variance of the cluster refers to the diagonal entries of the covariance matrix Σ of the two-dimensional Gaussian from which the points in the cluster were generated.

Received November 20, 2007
Revision received July 21, 2009
Accepted July 23, 2009 ■