# Causal Learning and Inference as a Rational Process: The New Synthesis

Keith J. Holyoak and Patricia W. Cheng

Department of Psychology, University of California, Los Angeles, Los Angeles, California 90095-1563; email: holyoak@lifesci.ucla.edu, cheng@lifesci.ucla.edu

Annu. Rev. Psychol. 2011. 62:135–63

The *Annual Review of Psychology* is online at psych.annualreviews.org

This article's doi: 10.1146/annurev.psych.121208.131634



0066-4308/11/0110-0135$20.00

## Key Words

causal cognition, Bayesian inference, dynamic models, attribution, diagnosis, reasoning, neuroimaging

## Abstract

Over the past decade, an active line of research within the field of human causal learning and inference has converged on a general representational framework: causal models integrated with Bayesian probabilistic inference. We describe this new synthesis, which views causal learning and inference as a fundamentally rational process, and review a sample of the empirical findings that support the causal framework over associative alternatives. Causal events, like all events in the distal world as opposed to our proximal perceptual input, are inherently unobservable. A central assumption of the causal approach is that humans (and potentially nonhuman animals) have been designed in such a way as to infer the most invariant causal relations for achieving their goals based on observed events. In contrast, the associative approach assumes that learners only acquire associations among important observed events, omitting the representation of the distal relations. By incorporating Bayesian inference over distributions of causal strength and causal structures, along with noisy-logical (i.e., causal) functions for integrating the influences of multiple causes on a single effect, human judgments about causal strength and structure can be predicted accurately for relatively simple causal structures. Dynamic models of learning based on the causal framework can explain patterns of acquisition observed with serial presentation of contingency data and are consistent with available neuroimaging data. The approach has been extended to a diverse range of inductive tasks, including category-based and analogical inferences.

## Contents



## INTRODUCTION

The ability to learn and reason about the causal structure of the world has obvious adaptive significance for bringing about desired outcomes. Humans have learned how to diagnose and treat diseases, build machines that fly, and predict the hypothetical consequences of mounting carbon emissions. These and many other cognitive achievements, including scientific reasoning (Dunbar & Fugelsang 2005), depend on the ability to learn and reason about cause-effect relations. Yet 25 years ago, work on thinking and reasoning within the field of cognitive psychology paid almost no attention to causal induction (i.e., the acquisition of causal knowledge from empirical observations). This situation has changed dramatically, and in the years since the turn of the millennium (the period on which we focus in this review), new developments have made causal induction a central focus of current research in cognitive science. A sense of the recent pace of change is given by an informal inspection of the tables of contents for two sets of *Proceedings of the Conference of the Cognitive Science Society* (Gleitman & Joshi 2000, Taatgen & van Rijn 2009), separated by about a decade. Our rough count of the number of presentations related to causality yields just 4 in the former set but 25 in the latter.

Our capability for causal induction poses a key question: Given that causal events, like all events in the distal world, are inherently unobservable, what minimal set of processes and assumptions must an intelligent cognitive system be endowed with so that it would be able to infer causal relations based on observed events and bring about desired outcomes, as humans do? This is the central question we address in the present review. Our aim is not to survey the full range of recent work on causal cognition, but rather to examine one major strand of empirical and theoretical development related to the rationality, and hence adaptiveness, of causal inference. A major "sea change" in the field over the past decade has been the emergence of a general framework for understanding human representations of causal knowledge. Earlier work on causal inference had primarily adopted either a heuristic approach related to Tversky & Kahneman's (1973, Kahneman et al. 1982) work on decision making (e.g., Schustack & Sternberg 1981, White 1998) or an associative approach inspired by David Hume (1739/1987) that reflects the predominant use of associative statistics by scientists (e.g., Dickinson et al. 1984, Shanks & Dickinson

1987). Both of these approaches forego rational causal inference as a feasible goal.

Three recent developments, however, have converged to show that the human causal process is surprisingly rational, in terms of accuracy, flexibility, and coherence (i.e., logical consistency and simplicity). The first is the development of rational Bayesian network models (Pearl 1988, 2000; Spirtes et al. 1993/2000). The second is the introduction of a Kantian a priori causal framework (Kant 1781/1965) that builds on but goes beyond Hume's legacy (e.g., Cheng 1997, Novick & Cheng 2004, Waldmann & Holyoak 1992; for a nontechnical exposition, see Sloman 2005). And the third is the introduction of probabilistic Bayesian mathematics as a modeling tool (Griffiths & Tenenbaum 2005, 2009; Kemp & Tenenbaum 2009; Lu et al. 2008b; Tenenbaum & Griffiths 2001), bringing greater power to the analysis of rationality and providing a language that allows more precise formulations of issues regarding the representation of causal knowledge.

This causal framework represents a new synthesis to which many investigators have contributed. We hasten to add that a synthesis does not imply a consensus; this active research area continues to be enlivened by vigorous debates. Indeed, it might prove challenging to find two researchers in the area who are in full agreement about the nature of causal models used by human (and perhaps nonhuman) reasoners! In this review, we discuss a number of the issues that continue to be discussed, offering our own perspective on the interpretation of the available evidence while acknowledging alternative perspectives. Our focus is on adult human causal induction (for a recent review covering causal cognition in nonhuman animals, see Penn & Povinelli 2007; for a review of evidence showing that even young children are rational causal reasoners, see Gopnik 2009; also see Gopnik & Schulz 2007).

In reviewing the emergence of the causal approach, we discuss the evidence that supports this framework over associationist accounts of causal induction. As an example to illustrate the alternative views, an associative account would treat learning facts such as that cigarette smoking, or having yellow fingers, covaries with certain forms of cancer—based on the proximal contingency data available to the learner—as the core goal of inductive learning. The causal framework, by contrast, assumes that people use such proximal data as evidence concerning distal causal relations in the real world—thereby potentially inferring that smoking, but not having yellow fingers, is a cause of cancer. Although someone who has been a smoker, or someone who has yellow fingers, is more likely than someone who does not have either characteristic to have cancer, only an intervention to reduce the prevalence of smoking may reduce the cancer rate—an intervention to reduce yellow fingers (e.g., by wearing gloves while smoking) will not (Spirtes et al. 1993/2000).

## THE EMERGENCE OF THE CAUSAL APPROACH

We begin with some background. In the early- and mid-twentieth century, interest in the psychology of causal understanding was almost exclusively confined to developmental psychologists working in the Piagetian tradition (Piaget 1930) and to social psychologists guided by attribution theory (Kelley 1967), which dealt specifically with inferences about the causes of behavior by participants in social interactions. About two decades ago, however, the nature and mechanisms of domain-general causal learning began to attract the attention of researchers in cognitive psychology. In an ironic twist, modern work on the topic was initially stimulated by researchers working in the associationist tradition then dominant in the field of animal learning and conditioning (Dickinson et al. 1984, Shanks & Dickinson 1987). Working in the empiricist tradition of David Hume (1739/1987), their goal was not to elucidate the cognitive representations underlying causal understanding, but rather to reduce human causal learning to the acquisition of associative links of the sort that might underlie conditioning in nonhuman animals.

In the first flush of enthusiasm for connectionist models based on parallel distributed processing (Rumelhart et al. 1986), various proposals emerged to account for human causal learning using variants of Rescorla & Wagner's (1972) learning rule for classical conditioning (e.g., Gluck & Bower 1988, Shanks 1991); these efforts continued into the current decade (e.g., McClelland & Thompson 2007, Stout & Miller 2007). Although the Rescorla-Wagner model predated modern connectionism, it readily lent itself to connectionist implementations. The general notion was that cause-effect relations can be represented as if they were simply cue-outcome associations, where cause and effect factors correspond to neuron-like units, and the "strength" of a cause-effect relation corresponds to a numerical weight on a synapse-like link between the input and output units. The magnitude of the learned weight controls the degree to which presentation of the cause evokes its associated effect.

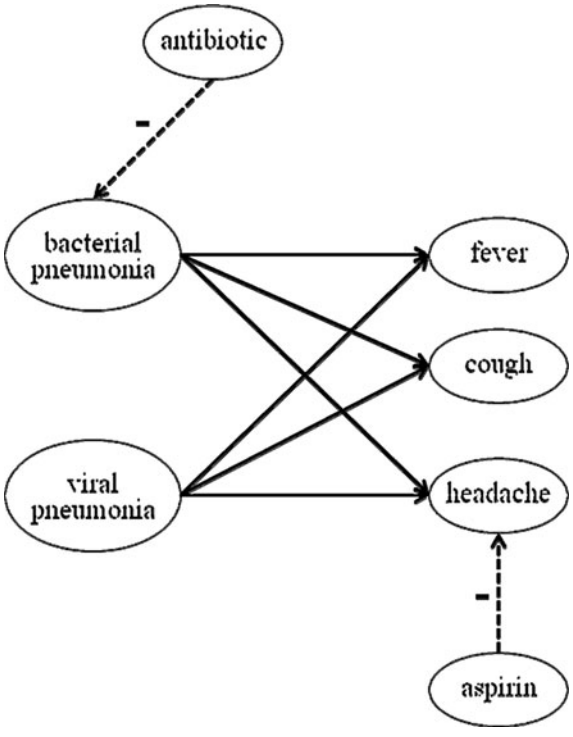


**Figure 1**

An example of a causal model, including both generative (*solid lines*) and preventive (*dashed lines*) causal links.

## Networks of Causal Relations

An alternative to this associationist view was that human causal knowledge is not reducible to associative links that exist only in the mind, but rather is based on mental representations of cause-effect relations assumed to be in the external world (Gallistel 1990, Waldmann & Holyoak 1992). Networks of cause-effect relations lend themselves to graphical representations, which were introduced earlier in philosophy (Reichenbach 1956, Salmon 1984). In artificial intelligence, Pearl (1988, 2000) and Spirtes et al. (1993/2000) developed a rich graphical formalism termed "Bayes nets." Although Bayes nets have primarily been developed as a practical tool for automated inference and data mining (e.g., Pourret et al. 2008), many of the principles that underlie computational Bayes nets have guided psychological work on causal models.

Waldmann & Holyoak (1992; Waldmann et al. 1995), influenced by the work of Pearl (1988), hypothesized that human causal learning results in explicit cause-effect (rather than merely associative) representations organized into causal models—networks of interconnected relationships. At their most basic level, causal models represent the direction of the causal arrow, the polarity of causal links (generative causes make things happen, preventive causes stop things from happening), the strength of individual causal links, and the manner in which the influences of multiple causes combine to determine their joint influence on an effect. **Figure 1** schematizes a simple causal model of a fragment of medical knowledge. The key assumption is that people preferentially acquire knowledge about asymmetrical causal links, each directed from a variable representing a causal factor to a variable representing an effect. In many situations, the variables are assumed to be binary (the factor is either present or absent). As we will see, this type of variable is particularly informative in

distinguishing between the causal and associative approaches.

A causal model provides a compact representation of the statistical associations among the factors. Rather than requiring an explicit representation of the joint distribution of all variables (the size of which scales exponentially with the number of nodes in the network), only direct causal connections, those that cannot be further decomposed into finer-grained causal relations, are explicitly represented. Based on this core knowledge, other relationships among factors can be derived. Fundamental types of questions that can be answered based on a causal approach include, "What will happen if a cause is either observed or made to occur?" (e.g., if viral influenza is present, what is the probability of getting a fever?), "Why did this effect occur?" (e.g., given that fever is present, what is the probability that viral influenza was present and caused it?), and "What should I do to influence an effect?" (e.g., if the goal is to prevent a headache, what intervention might succeed?). As these questions illustrate, not only is a causal representation compact, but it also effectively supports predictions about the consequences of actions (Spirtes et al. 1993/2000).

The "cause" and "effect" roles are fundamentally asymmetric: causes produce (or prevent) their effects, but not vice versa. Because a cause must be present in order to act, it follows that a cause is understood to occur temporally prior to its effect (even for situations in which there is no perceptible temporal gap between the occurrence of the cause and of the effect). Fenker and colleagues (2005) found that college students are faster to verify the existence of a causal relation between two lexically expressed concepts when the words are presented in the temporal order corresponding to cause-effect order (e.g., *spark* prior to *fire*) rather than in the reverse temporal order (*fire* prior to *spark*). No such asymmetry was observed when participants were asked to verify whether the two concepts were "associated." The greater psychological naturalness of the cause-effect temporal order supports the basic assumption that causal models preferentially encode links directed from cause to effect (cf. Tversky & Kahneman 1982).

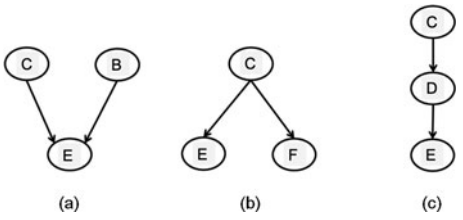


**Figure 2**

Three basic causal structures: (*a*) common effect, (*b*) common cause, and (*c*) chain. Letters are arbitrary labels for variables.

Early work pitting causal-model theory against associative accounts of human causal learning made use of paradigms adapted from work in animal learning. These paradigms involve situations in which multiple cues may co-occur. The general approach was to hold cue-outcome contingencies constant while using different cover stories to vary the causal model that learners would use to guide their learning. For example, Waldmann (2000) presented two groups of participants with cover stories that specified either a common-effect model (**Figure 2*a***) or a common-cause model (**Figure 2*b***). In Phase 1, participants in both groups learned that a cue predicts an outcome (P+). In Phase 2, this cue was constantly paired with a second, redundant light (i.e., PR+) followed by the outcome. As shown in **Figure 3**, for participants presented with the common-effect model, the R cue showed a strong "blocking" effect (i.e., lower mean "predictiveness" rating relative to the P cue; cf. Kamin 1969). In contrast, for participants presented with the common-cause model, no blocking was observed—cues P and R were judged to be equally predictive. In the past few years, many studies have yielded similar asymmetries in contingency learning (Booth & Buehner 2007, López et al. 2005, Waldmann 2001). Such demonstrations that the psychological representation of cause-effect relations can be dissociated from the overt temporal order of presented cues and their outcomes

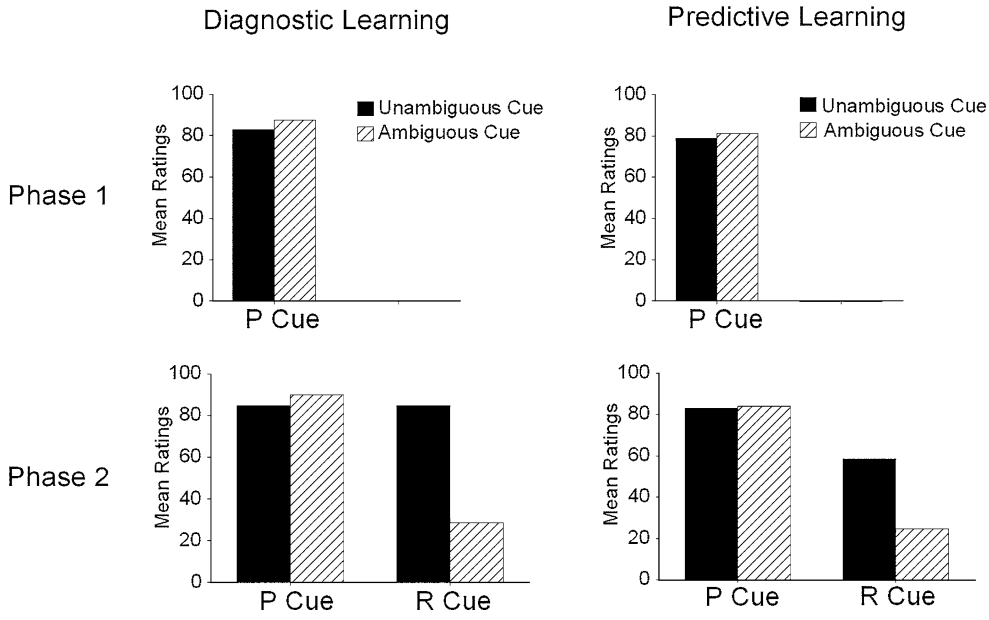


**Figure 3**

Mean predictiveness ratings for diagnostic (common cause) and predictive (common effect) conditions in phases 1 and 2 of experiment 3a for the predictive cue (P cue) and redundant cue (R cue). Text describes unambiguous cues; ambiguous cues provide a control. (From Waldmann 2000, reprinted by permission.)

support the position that human causal learning makes use of causal representations.

## The Origin of Causal Knowledge

To achieve or avoid an outcome, one may want to predict with what probability an effect will occur given that a certain cause of the effect occurs. Such probabilities might be acquired by direct instruction. But the issue remains of how such knowledge is initially formed. In order to acquire a causal model that can yield an accurate answer to such questions, there must be some process that takes as its input noncausal empirical observations of cues and outcomes and that yields as its output values of causal "strength" associated with the individual links in a causal model. Since causes of an effect other than the target potential cause may well be unobserved or unknown, it is impossible to rule out their occurrence. The absence of the effect in the absence of the target cause need not imply that no alternative causes occurred, as it is possible for background causes to be preventive. A theory of strength learning therefore necessitates assumptions about alternative causes of the effect. These assumptions imply the form of the integration function[1] that specifies how multiple causes jointly determine the effect. For example, if we know that smoking causes cancer with some probability less than 1, and so does exposure to asbestos, what would we predict to be the probability of cancer occurring in a smoker exposed to asbestos?

Researchers in the associative tradition proposed various learning rules (the Rescorla-Wagner model and several variants of it) that update cue-outcome strength values based on contingency data (for a recent review, see López & Shanks 2008). The classic "statistical" measure of the strength of a contingency has been

[1] In Bayesian models, the integration function is commonly referred to as the "generating function." The same basic concept has also been termed the "parameterization" for combining multiple link strengths (Griffiths & Tenenbaum 2005) and the "functional form" (Griffiths & Tenenbaum 2009).

termed $\Delta P$ (Jenkins & Ward 1965), which in causal terms is simply the difference between the probability of the effect $e$ in the presence versus absence of the candidate cause $c$, i.e.,

$$\Delta P = P(e^+ \mid c^+) - P(e^+ \mid c^-). \qquad (1)$$

where $e$ and $c$are binary variables, and +/− indicate the value of a binary variable to be 1 (present) versus 0 (absent). Under certain common conditions, the asymptotic strength computed by the Rescorla-Wagner model is equivalent to $\Delta P$ (Danks 2003). For both $\Delta P$ and the Rescorla-Wagner model, the integration function according to which the influences of multiple causes combine to influence the probability of the effect is additive, with a correction to bound the predicted probability of the effect between 0 and 1 (see Griffiths & Tenenbaum 2005).

However, in some situations in which multiple causal factors co-occur, studies of both human causal judgments and of animal conditioning have identified systematic deviations from the predictions of the Rescorla-Wagner model and the $\Delta P$ rule (see Cheng & Holyoak 1995), suggesting that the integration function is not always additive and that an alternative approach to causal induction is required. A natural interpretation of the strength of a causal link is that it represents the power of the cause operating in the external world to produce (or prevent) the corresponding effect. Generative causal power, for example, would correspond to the probability that the target cause, if it were to act alone (i.e., in the absence of other causal factors), would produce the effect. Cheng (1997) proposed a normative theory of how a reasoner could estimate causal power from noncausal contingency data by adopting a set of a priori causal assumptions (cf. Kant 1781/1965). This theory gives a causal explanation of the earlier probabilistic contrast model (Cheng & Novick 1992) and hence was termed the power PC theory (causal power theory of the probabilistic contrast model).

We can state the key psychological claims of the power PC theory in relation to the simple common-effect model in **Figure 2*a***. The model represents the partitioning of all causes of an effect $E$ into candidate cause $C$ and the rest of the causes, represented by $B$, an amalgam of observed and unobserved background causes and enabling conditions that occur with unknown frequencies that may vary from situation to situation. $C$, $B$, and $E$ are binary variables with a "present" and an "absent" value. The model is a general default structure that maps onto all learning situations. Causal power is represented by the weights on each causal link. The focus is typically on $w_1$, a random variable representing the strength of the candidate cause $C$ to influence effect $E$.

The power PC theory postulates that people approach causal learning with four general prior beliefs:

1) $B$ and $C$ influence effect $E$ independently,
2) $B$ could produce $E$ but not prevent it,
3) causal powers are independent of the frequency of occurrences of the causes (e.g., the causal power of $C$ is independent of the frequency of occurrence of $C$), and
4) $E$ does not occur unless it is caused.

Assumptions 1 and 2 serve as default hypotheses for the reasoner, adopted unless evidence discredits them (in which case alternative models apply, see Novick & Cheng 2004; for implications of the relaxation of these assumptions, see Cheng 2000). Assumptions 3 and 4 are viewed as essential to causal inference. Assumption 4 is supported by research showing that adults (Kushnir et al. 2003), preschool children (Gelman & Kremer 1991, Schulz & Sommerville 2006), and even infants (Saxe et al. 2005) interpret events as having causes, even when the causes are unobservable.

This set of assumptions, which is stronger than that assumed in standard Bayes nets, requires less processing capacity. It allows causal relations to be learned one at a time, when there is information on only two variables, a single candidate cause and an effect. Without these assumptions, it is impossible to distinguish between causation and mere association in this simple situation. These assumptions of the power PC theory imply a specific

integration function for contingency data (Cheng 1997, Glymour 2001), different from the additive function assumed by associative models. For the situation in which a potentially generative candidate cause $C$ occurs independently of other causes, the probability of observing the effect $E$ is given by a noisy-OR function,

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

where $b, c \in \{0, 1\}$ denotes the absence and the presence of the causes $B$ and $C$. Variables $w_0$ and $w_1$ are causal strengths of the background cause $B$ and the candidate cause $C$, respectively. In the preventive case, the same assumptions are made except that $C$ is potentially preventive. The resulting noisy-AND-NOT integration function for preventive causes is

$$P(e^+ \mid b, c; w_0, w_1) = w_0 b(1 - w_1 c). \quad (3)$$

Using these "noisy-logical" integration functions (terminology from Yuille & Lu 2008), Cheng (1997) derived normative quantitative predictions for judgments of causal strength (Equations 4 and 5). Causal power, $q$, is a maximum likelihood point estimate of $w_1$, the causal power of the candidate cause (see Griffiths & Tenenbaum 2005). The causal power for a generative cause $c$ with respect to effect $e$ is estimated by

$$q_G = \frac{\Delta P}{1 - P(e^+ \mid c^-)}, \quad (4)$$

and the power for a preventive cause $c$ is estimated by

$$q_P = \frac{-\Delta P}{P(e^+ \mid c^-)}, \quad (5)$$

where $\Delta P$ is the difference between the probability of the effect $e$ in the presence versus absence of the candidate cause $c$ (Equation 1).

The term $P(e^+ \mid c^-)$ in the denominator of Equations 4 and 5 is often termed the "base rate of the effect," as it gives the prevalence of $e$ in the absence of $c$. A key qualitative implication of these equations is that learning of generative and preventive causes will be asymmetrical with respect to the base rate of the effect. Holding $\Delta P$ constant, Equation 4 implies that generative power increases with the base rate of the effect, whereas Equation 5 implies that preventive power decreases with the base rate. In contrast, the additive function underlying associative models predicts no such asymmetry due to causal polarity. Numerous studies have shown that the impact of the base rate on human judgments of the causal strength of generative versus preventive factors is in fact asymmetrical, as the power PC theory predicts (Buehner et al. 2003, Novick & Cheng 2004, Wu & Cheng 1999; see meta-analysis reported by Perales & Shanks 2007). To take an extreme case, Equation 4 predicts that when the base rate is 1 ($e$ always occurs in absence of $c$), the values of the generative power of the candidate will be indeterminate. This situation corresponds to the familiar concept of a "ceiling effect" in experimental design, which makes it impossible to assess whether an independent variable increases the probability of an outcome. Conversely, Equation 5 predicts that when the base rate is 0 ($e$ never occurs in absence of $c$, which is the preventive analog of a ceiling effect), the values of the preventive power of the candidate will be indeterminate. Note that in both these extreme situations in which the power PC theory predicts asymmetrical uncertainty, the value of $\Delta P$ is precisely 0; hence this measure, and associative models in general, predict anomalously that when a ceiling effect (or its preventive analog) arises, people will be certain that the candidate is noncausal.

## INTEGRATING CAUSAL REPRESENTATION WITH BAYESIAN INFERENCE

### Representing Uncertainty

The most important methodological advance in the past decade in psychological work on causal learning has been the introduction of Bayesian inference to causal inference. This began with the work of Griffiths & Tenenbaum

(2005, 2009; Tenenbaum & Griffiths 2001; see also Waldmann & Martignon 1998). Although the power PC theory (Cheng 1997) had been proposed as a rational model of causal induction, as initially formulated it did not provide any general account of how uncertainty impacts causal judgments. In particular, just like the $\Delta P$ rule, the point estimate of causal power (Equations 4 and 5) is insensitive to sample size. For example, if the base rate of the effect is 0, then both power and $\Delta P$ will have the value 0.5 regardless of whether the effect occurs in the presence of the candidate cause in 1 out of 2 cases or in 50 out of 100 cases. But intuitively, the learner's degree of uncertainty should be lower in the latter situation, when the sample size is large (as reflected in standard statistical measures of independence, such as the $\chi^2$ statistic). The lack of an account of uncertainty in early models of human causal learning played a role in prolonging the debate between proponents of associationist treatments and of the power PC theory. For some data sets (e.g., Lober & Shanks 2000), human causal judgments for some conditions were found to lie intermediate between the values predicted by causal power versus $\Delta P$, perhaps reflecting uncertainty outside the scope of either model.

Even though it had been argued that causal induction is fundamentally rational (Cheng 1997), and causal models had been formalized as "causal Bayes nets," until recently causal induction had not been treated as Bayesian inference. In fact, as we will see, adopting Bayesian inference is entirely orthogonal to the longstanding debate between causal and associationist approaches. Bayesian inference can be either causal or associative, depending on whether causal assumptions (e.g., those in the power PC theory) are made, leading to different causal-judgment predictions. The term "Bayes net" was initially introduced by Pearl (1988) to highlight the role of Bayesian inference in deriving rational inferences from a known causal network. However, work on learning within Bayes nets (e.g., Spirtes et al. 1993/2000) generally emphasized non-Bayesian algorithms.

The heart of Bayesian inference is Bayes rule,

$$P(H \mid D) = \frac{P(D \mid H)P(H)}{P(D)}, \qquad (6)$$

where $H$ denotes a hypothesized state of the world, and $D$ denotes observed data. Conceptually, Bayes rule provides a mathematical tool to calculate the posterior probability of a hypothesis, $P(H \mid D)$, from prior belief about the probability of the hypothesis, $P(H)$, coupled with the likelihood of the new data in view of the hypothesis, $P(D \mid H)$. Assuming the hypothesis is causal (i.e., it can be represented as a link in a directed graph of the sort shown in **Figure 1**), developing a Bayesian model of causal learning further requires specification of relevant prior beliefs and of the function linking causal hypotheses to data.

## Causal Support Model

Griffiths & Tenenbaum (2005) introduced the Bayesian analysis of causal learning in the context of what they termed the "causal support model." This model focused on a different causal query than the "strength" judgments that had been emphasized in most empirical studies. We first discuss the particular model proposed by Griffiths and Tenenbaum and then consider its limitations.

A strength judgment concerns the weight on a causal link, which in essence aims to answer the query, "What is the probability with which a cause produces (alternatively, prevents) an effect?" (e.g., for Graph 1 in **Figure 4**, this probability is the weight $w_1$ on the link from $C$ to effect $E$; note that Graph 1 carries no implication that $w_1$ must be greater than 0). Within a Bayesian framework, strength judgments pose a problem of parameter estimation. Griffiths & Tenenbaum (2005) focused on a different causal query, termed a "structure" judgment, which aims to answer, "How likely is it that a causal link exists between these two variables?" Within a Bayesian framework, structure judgments pose a problem of model selection (see Mackay 2003). The causal support model

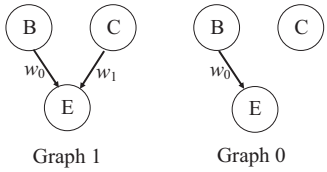


**Figure 4**

Graphs contrasting hypotheses that the candidate cause, *C*, causes the effect, *E* (Graph 1) or does not (Graph 0). *B*, *C*, and *E* respectively denote the background cause, the candidate cause and the effect. *B*, *C*, and *E* are binary variables that represent the absence and presence of the cause and the effect. Weights $w_0$ and $w_1$ indicate causal strength of the background cause (*B*) and the candidate cause (*C*), respectively.

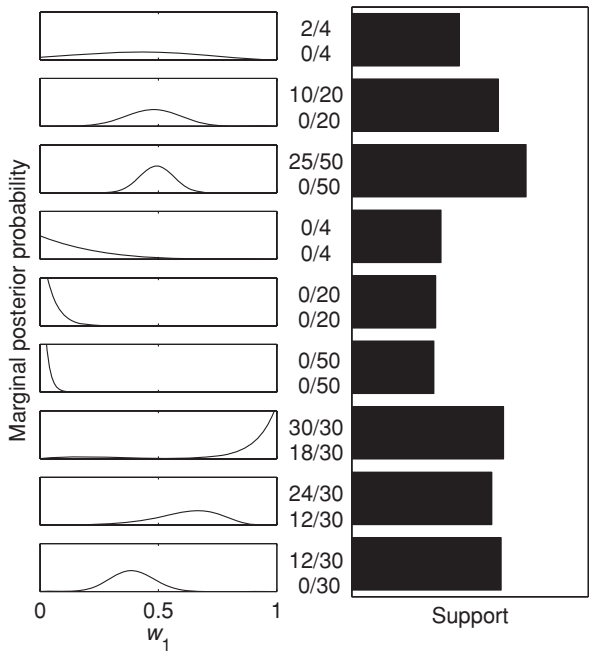


**Figure 5**

Marginal posterior distributions on $w_1$ and values of causal support for three different sets of contingencies. Each contingency is expressed in terms of two fractions, respectively: The relative frequency with which the effect *e* occurs in the presence of the candidate cause *c* (e.g., "2/4," denoting that *e* occurs in two out of four trials in which *c* is present) and that with which *e* occurs in the absence of *c* (e.g., "0/4"). (From Griffiths & Tenenbaum 2005; reprinted with permission.)

aims to account for judgments as to whether a set of observations (*D*) was generated by a causal graphical structure in which a link may exist between candidate cause *C* and effect *E* (Graph 1 in **Figure 4**) or by a causal structure in which no link exists between *C* and *E* (Graph 0).

Griffiths & Tenenbaum (2005) defined "causal support" as

$$\text{support} = \log \frac{P(D \mid Graph1)}{P(D \mid Graph0)}. \qquad (7)$$

Griffiths & Tenenbaum (2009) define a variant of causal support in terms of $P(Graph1 \mid D)$, the posterior probability of Graph 1. The likelihoods on graphs are computed by averaging over the unknown parameter values, causal strengths $w_0$ and $w_1$, which lie in the range [0, 1] and are associated with causes *B* and *C*, respectively. Stated formally, $P(D \mid w_0, w_1, Graph1)$ and $P(D \mid w_0, Graph0)$ are the likelihoods of the observed data given specified causal strengths and structures, and $P(w_0, w_1 \mid Graph1)$ and $P(w_0 \mid Graph0)$ are prior probability distributions that model the learner's beliefs about the distributions of causal strengths given a specific causal structure (assumed to be uniform, reflecting complete ignorance about the parameter values).

One way of viewing the relationship between causal support (Griffiths & Tenenbaum 2005) and causal power (Cheng 1997) is that the latter provides the basis (Equations 4 and 5) for calculating the likelihoods used in computing causal support. From here on, unless the linear generating function is specified, the causal (rather than associative) variant of the causal support model is assumed. Griffiths & Tenenbaum (2005) noted that, "Speaking loosely, causal support is the Bayesian hypothesis test for which causal power is an effect size measure: it evaluates whether causal power is significantly different from zero" (p. 359). They are answers to different questions. As illustrated on the left side of **Figure 5**, the Bayesian analysis yields a posterior distribution of causal strength ($w_1$), in terms of which causal power is the

point estimate corresponding to the maximum likelihood (i.e., the peak value of the marginal posterior distribution). The top three contingencies in **Figure 5** illustrate conditions that yield different posterior distributions of values of $w_1$ and different values of causal support, but the distributions have the same peak, corresponding to both $\Delta P$ and causal power. The second set of three contingencies illustrates that an increase in sample size can result in a slight decrease in causal support, indicating greater certainty that $w_1$ is zero. The third set of three contingencies illustrates how causal support can change in a nonmonotonic fashion while the peak of the distribution over $w_1$ decreases.

As noted above, a Bayesian analysis can be applied to both strength and structure judgments (along with other interrelated types of queries, such as causal attribution). For strength judgments, a natural Bayesian extension of the power PC theory would base predictions on some function of the posterior distribution of $w_1$ (e.g., its mean). As is apparent from the example distributions shown on the left in **Figure 5**, the posterior distribution is sensitive to sample size, and its mean can differ from its peak.

Rather than testing a direct Bayesian measure of causal strength judgments, Griffiths & Tenenbaum (2005) argued that people may often make support judgments when nominally asked to make strength judgments. However, since these investigators did not elicit both strength and structure judgments from participants, their study provided no evidence to support the assumption that people are unable to distinguish the two queries. Griffiths and Tenenbaum supported their argument with data from Buehner et al.'s (2003) experiment 1, in which subjects were presented with an ambiguous question intended to elicit strength judgments. However, in Buehner et al.'s experiment 2, subjects were presented with a less ambiguous strength question, and the resulting estimates of causal strength were in close accord with causal power, contradicting Griffiths and Tenenbaum's argument. Griffiths and Tenenbaum did report three experiments designed to elicit structure judgments; however, the structure question posed to the participants was ambiguous (see discussion in Buehner et al. 2003), and sample size was not systematically manipulated. More recently, some experimental results have revealed ordinal violations of the support model as an account of human judgments about causal structure (Lu et al. 2008b). Relative to the support model, human reasoners appear to place greater emphasis on power and the base rate of the effect, and less emphasis on sample size. In addition, some contingency conditions yield specific differences due to causal polarity (generative versus preventive), which are not captured by the support model.

## Bayesian Extensions of the Power PC Theory

Lu et al. (2008b) developed and compared several variants of Bayesian models as accounts of human judgments about both strength and structure. As Griffiths & Tenenbaum (2005) had noted, a Bayesian model can incorporate either the noisy-logical integration functions derived from the power PC theory or the linear function underlying the Rescorla-Wagner model and the $\Delta P$ rule, resulting in a causal and an associative variant, respectively. In addition to directly comparing predictions based on these alternatives, Lu et al. (2008b) considered two different sets of priors on causal strength. One possible prior is simply a uniform distribution, as assumed in the causal support model. Since the power PC theory makes no assumptions about priors, a uniform distribution is its natural default. The alternative "generic" (i.e., domain-general) prior tested by Lu et al. (2008b) is based on the assumption that people prefer simple causal models (Chater & Vitányi 2003, Lombrozo 2007, Novick & Cheng 2004). Sparse and strong (SS) priors imply that people prefer causal models that minimize the number of causes of a particular polarity (generative or preventive) while maximizing the strength of each individual cause that is in fact potent (i.e., of nonzero strength).

Lu et al. (2008b) performed an experiment to assess the predictive power of the four Bayesian models defined by the factorial combination of the two types of integration functions (noisy-logical or additive) and the two priors (uniform or SS). The design, in which the single candidate cause could be either potentially generative or preventive, included a range of contingencies and sample sizes. In order to minimize memory issues and other possible performance limitations, their study employed a procedure developed by Buehner et al. (2003; also Liljeholm & Cheng 2009), in which individual trials are presented simultaneously in a single organized display. Such presentations provide a vivid display of individual cases, making salient the frequencies of the various types of cases while minimizing memory demands. To clearly assess causal strength rather than causal structure, participants were asked to make a judgment of the frequency (out of 100) with which cases with the cause would be expected to show the effect, when none of the 100 showed the effect without the cause.

For all four Bayesian models, Lu et al. (2008b) compared the average observed human strength rating for a given contingency condition with the mean of $w_1$ computed using the posterior distribution. In contrast to the causal support model, which explicitly compares the probabilities of two graphs, the Bayesian models tested by Lu et al. (2008b) assume that only a single graph structure (Graph 1 in **Figure 4**) is required to make basic strength judgments. Model fits revealed that the two causal variants based on the noisy-logical integration function were much more successful overall than the associative variants. The assumption of SS priors was able to explain subtle asymmetries in causal judgments across generative versus preventive causes, primarily attributable to conditions with extreme base rates of the effect (base rates near 1 for generative causes, near 0 for preventive causes), for which preventive strength was judged as exceeding generative strength for matched contingencies. In addition, and without any further parameter-fitting, the two causal Bayesian variants proved very successful in fitting datasets from a meta-analysis based on 17 experiments selected from 10 studies in the literature (Perales & Shanks 2007; see also Hattori & Oaksford 2007), achieving an overall correlation as high as $r = 0.96$. (See Griffiths & Tenenbaum 2009 for a similar fit to this set of data using a related Bayesian model.) Indeed, the causal Bayesian models (with one or zero free parameters) performed at least as well as the most successful nonnormative model of causal learning (with four free parameters) and much better than the Rescorla-Wagner model. Thus, although both causal and associative approaches can be given a Bayesian formulation, empirical tests of human causal learning reported by Lu et al. (2008b) favor the integration of the causal approach with Bayesian inference, providing further evidence for the rationality of human causal inference. As their results illustrate, Bayesian models, like all other models, are only as rational as their assumptions (see Liljeholm & Cheng 2007).

Lu et al. (2008b) also evaluated structure analogs of the two causal variants of Bayesian strength models to account for observed structure judgments from additional experiments in which participants were explicitly asked to judge whether or not the candidate was indeed a cause. One analog, assuming uniform priors, was simply Griffiths & Tenenbaum's (2005) causal support model. The alternative model was based on the assumption that the priors should support Graph 1 only if $C$ is a strong cause, thereby justifying its addition to the set of accepted causes of $E$. Lu et al. (2008b) therefore assumed that structure judgments reflect an additional preference that $C$ (in Graph 1) be a strong cause of $E$. For two experiments, the latter Bayesian extension of the power PC theory provided a better quantitative account of human structure judgments than did the support model. **Figure 6** shows the model fits for one experiment (Lu et al. 2008b, experiment 4). This design (for generative conditions only) compared judgments for two contingencies close to the generative peak favored by priors for a strong cause but with a small sample size (8) to two contingencies far from the

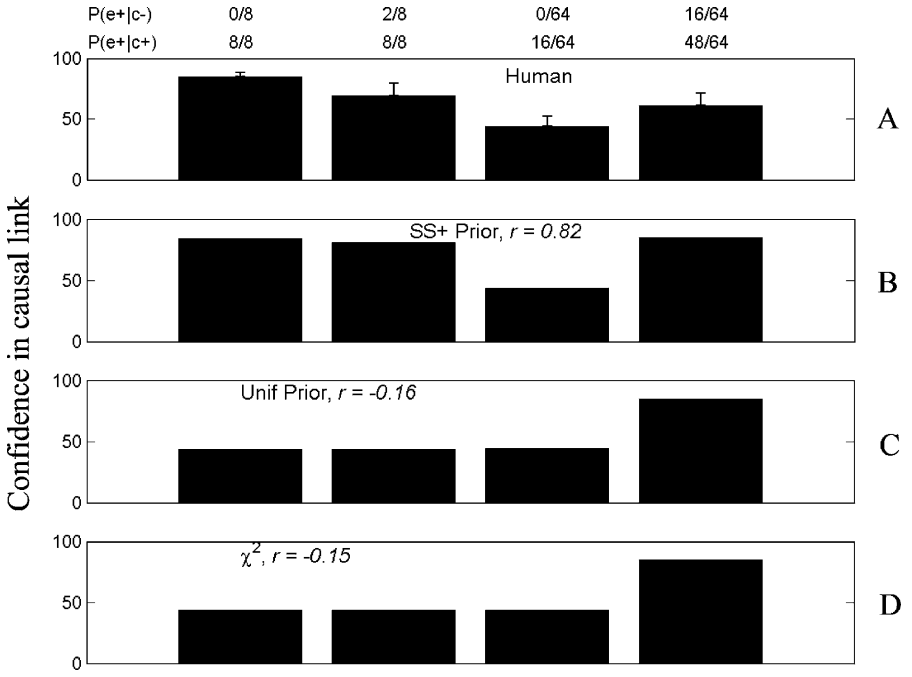


**Figure 6**

Confidence in a causal link (Lu et al. 2008b, experiment 4). (*A*) Mean human confidence judgments (error bars indicate one standard error), (*B*) predictions of Bayesian model with priors favoring a strong cause, (*C*) predictions of the causal variant of the causal support model with uniform (Unif) priors, and (*D*) predictions of $\chi^2$ statistic. (From Lu et al. 2008b; reprinted with permission.)

favored peak but with a substantially larger sample size (64; see **Figure 6*A***). The model incorporating priors for a strong cause (**Figure 6*B***) yielded a high positive correlation across the four conditions, whereas the correlations were actually in the wrong direction for both the causal support model (**Figure 6*C***) and $\chi^2$ (**Figure 6*D***).

The causal support model provides an important demonstration of how Bayesian analysis can be applied to causal learning; however, in its original formulation it has turned out to be empirically inadequate as an account of human causal judgments. When questions are clearly worded, human reasoners can evaluate either causal strength or causal structure, providing distinct patterns of responses to each type of query. Moreover, both types of judgments (but especially judgments of causal structure) are guided by generic priors that favor strong causes. Nonetheless, Griffiths & Tenenbaum's (2005) contribution in integrating the Bayesian framework with causal representation has proven to be extremely fruitful (see Chater et al. 2006, Goodman et al. 2007, Griffiths et al. 2008, Griffiths & Tenenbaum 2009, Kemp & Tenenbaum 2009).

The work we have described focuses on predictive inferences, from knowledge of causes to the status of an effect. However, one of the strengths of the causal approach is that it allows the derivation of many different types of inferences (Cheng & Novick 2005). Other major types of inference include causal attribution (given that an effect has occurred, inferring the probability that some particular factor caused it) and diagnostic inference (Waldmann et al. 2008) (given that an effect occurred, inferring the probability that some other factor was present). Recent studies suggest that at least for fairly simple causal networks, people are able to reason from effects to possible causes in accord

with the pattern predicted by Bayesian extensions of the power PC theory (Holyoak et al. 2010, Meder et al. 2009).

### Integrating Prior Causal Knowledge with Current Data

Once some knowledge about distal causal relations has been acquired, the issue arises: How does one make use of this prior knowledge and new observations to make further causal inference? This question invites a natural answer based on Bayesian inference. Previous work has demonstrated the important influence of prior abstract causal knowledge in everyday causal inference (e.g., Ahn & Kalish 2000, Ahn et al. 1995). Griffiths & Tenenbaum (2009) presented a formalization of this influence in terms of a hierarchical Bayesian model. The model represents three key aspects of prior abstract causal knowledge: (*a*) an ontology that organizes a domain in terms of entities, their properties, and relations between their properties, (*b*) the plausibility of specific causal relations in the domain, and (*c*) the functional form defining how the causal influences of different entities combine. Following standard Bayesian methods, the model enumerates, or randomly samples from, the set of all plausible causal structures given the ontology, assigns prior probabilities to the alternative structures as well as to the values of the strength of causal relations, and computes answers to various questions by applying Bayes theorem. The hierarchical nature of the model enables simultaneous learning at multiple levels of abstraction. An interesting illustration is the application of the hierarchical model to Gopnik et al.'s (2001, experiment 1) "two-causes" condition. The model can learn that a machine operated probabilistically rather than deterministically at the same time that it learns which entity activates the machine and which does not (see also Kemp & Tenenbaum 2009). This recent work illustrates that Bayesian mathematics provides a powerful tool for representing the role of prior knowledge across diverse domains.

## DYNAMIC MODELS OF SEQUENTIAL CAUSAL LEARNING

Models such as the power PC theory and the $\Delta P$ rule address causal induction at the computational level. Because they focus on asymptotic knowledge, computational-level models of this sort do not address phenomena related to factors such as sample size, trial order, and forgetting, all of which clearly impact human causal learning. We have seen that by integrating Bayesian inference with the power PC theory, the influence of sample size can be addressed (Griffiths & Tenenbaum 2005, Lu et al. 2008b). However, many issues remain in developing a learning model that can operate when the observations from which causal relations might be induced are distributed over time rather than presented in summary form. Clearly, much causal learning by humans (and perhaps all causal learning by nonhuman animals) depends on sequential accumulation of evidence. An attraction of the Rescorla-Wagner model and similar associative accounts of causal learning has been that they are inherently sequential in nature and hence can potentially account for a broader range of acquisition phenomena. But as reviewed above, these models incorrectly predict that the asymptotic value of causal strength will approach $\Delta P$ (due to the assumption of an additive integration function), whereas in fact it approaches the value of causal power (Buehner et al. 2003, Lu et al. 2008b). Other approaches to sequential learning are required, which can incorporate alternative integration functions as well as updating of distributions of causal strength rather than simply of point estimates.

### Computational Models of Sequential Learning

Recent years have in fact seen a surge of developments related to dynamic models of causal learning, most of which involve various techniques for making use of Bayesian inference. As Danks and colleagues (2003) observed, any model of causal learning from summary data

can be applied to sequential learning simply by keeping a running tally of the observations required to assess contingencies, applying the model after accumulating $n$ observations, and repeating as $n$ increases. These investigators were the first to develop a sequential model incorporating the noisy-logical integration function.

This type of "rational" model of sequential learning suffices to account not only for the basic negatively accelerating acquisition curve, but also for a phenomenon involving zero contingencies that had previously been viewed as nonrational. When the probability of the effect is the same regardless of whether the candidate cause occurs, the asymptotic causal rating approaches zero; however, after a limited number of trials, it has often been observed that a zero contingency with a higher effect probability initially receives more positive strength ratings than does a zero contingency with a lower effect probability (e.g., Allan & Jenkins 1983, Shanks 1987, White 2004). For example, if the effect occurs with probability 0.75 in both the presence and absence of the candidate cause, after a few trials the cause will be judged to be moderately generative; if instead the effect occurs with probability 0.25 in both the presence and absence of the cause, it will initially be judged to be weaker. As the number of observations increases, the judged strength of the candidate cause in all such noncontingent conditions will approach 0.

It turns out that this acquisition phenomenon can be explained by the inherent asymmetry between the noisy-OR integration function for generative causes and the noisy-AND-NOT function for preventive causes when applied to stochastic observations (Buehner et al. 2003). The net result will be an initial bias toward more positive strength estimates when the effect probability is relatively high.

Ultimately, a full treatment of sequential learning requires moving beyond models that assume complete memory for all observations. Phenomena involving trial order imply that causal learning depends not simply on what observations have been encountered, but also on their temporal order. One class of such phenomena involves retrospective reevaluation of causal relations. For example, "blocking" (Kamin 1969) involves pairing a redundant cue (R) with another cue (P) that has been shown to be individually predictive. In standard forward blocking, the P cue is first presented alone with the effect, followed by the paired presentations of P and R (i.e., P+, PR+). Cue R is said to be blocked by P because it is subsequently rated as less predictive of the effect.

It turns out blocking is also observed, but reduced in magnitude, when the order of presentation is reversed (i.e., PR+, P+) (e.g., Beckers & de Houwer 2005). In backward blocking, it is as if learners retroactively downgrade their assessment of the causal strength of R after learning that P alone has high strength. Such retrospective effects are problematic for models such as Rescorla-Wagner, which provide no mechanism for revising the strengths of absent cues (but see Dickinson & Burke 1996, Stout & Miller 2007, and Van Hamme & Wasserman 1994 for alternative associationist models that can account for retrospective reevaluation). In contrast, Bayesian models based on strength distributions have no problem explaining the existence of backward as well as forward blocking; however, models that lack some way of responding differentially depending on trial order fail to explain why backward blocking (and similar retrospective effects) are typically weaker than their forward counterparts.

A number of dynamic models have been developed that can account both for the fact that backward blocking is obtained (as are similar types of retrospective reevaluation) and the equally important fact that its magnitude is reduced relative to forward blocking (Daw et al. 2007, Kruschke 2006, Lu et al. 2008a). All of these models introduce the basic assumption that the learner uses each successive observation to update strength distributions and then discards each individual observation after it has been so used. Bayesian revision of strength distributions has the effect of revising causal knowledge about cues that are absent on a given

trial, as well as cues that are present, yielding retrospective reevaluation. Because uncertainty associated with strength distributions tends to decrease with the number of observations, observations presented early in the sequence generally have greater impact than do those later in the sequence; hence trial order matters, and retrospective effects are typically weaker than their forward counterparts.

Although they have important commonalities, the recent dynamic models also differ in important ways, especially in regard to their assumptions about integration functions. The model developed by Daw et al. (2007) is essentially a Bayesian extension of the Rescorla-Wagner model. Its additive integration function, although likely appropriate for some situations involving continuous variables, is unsuitable for causal learning with binary variables, as we have seen. Kruschke (2006) developed a model that assumes learning is based on a sequence of component modules, each equivalent to a layer in a neural network. Parameter updating within each layer is treated as locally Bayesian, although the behavior of the system as a whole is not globally Bayesian. The approach is implemented as an associative learning model that maps inputs filtered by attention to outputs. The model can be specified using either an additive integration function or the causal noisy-logical function. The Bayesian updating allows the model to exhibit retrospective revaluation effects such as backward blocking and unovershadowing.

Lu et al. (2008a) developed a dynamic Bayesian learning model that incorporates multiple integration functions and then uses model selection to choose the particular function most appropriate for a given set of observations. They applied their model to an intriguing set of findings reported by Beckers & de Houwer (2005). These investigators first trained participants with certain cue-outcome pairs, such as bacon (cue G) and eggs (cue H), each paired with a moderate allergic reaction. The combination of the two cues, bacon and eggs (cue GH) was paired with either a moderate or a strong allergic reaction. The participants were then transferred to a classic forward blocking paradigm with unrelated cues, such as cheese (cue A) paired with moderate reaction, and cheese and nuts (cue AX) also paired with moderate reaction. Finally, participants were tested on how likely nuts alone (cue X) was to cause allergy. Human participants provided different ratings on the transfer test for cue X depending on whether cue combination GH had been paired with moderate or strong allergic reaction during the pretraining. In particular, less blocking was observed for cue X if the compound GH had been paired with a moderate reaction (suggesting that the joint influence of the two cues was subadditive) than if GH had been paired with a strong reaction (suggesting that the two cues in combination had a greater impact than either cue alone).

None of the models we have reviewed so far learn about the background cause. Luhmann & Ahn (2007) proposed a sequential learning algorithm for learning the strength of the background cause. Their BUCKLE (bidirectional unobserved cause learning) model, which adopts the noisy-OR and noisy-AND-NOT integration functions, iteratively (*a*) estimates the probability that the background cause occurs based on the status of the observed variables (namely, the candidate cause and the effect), and (*b*) updates the estimates of causal strength for both the candidate and background causes.

In summary, recent theoretical work is opening up prospects for developing more detailed, algorithmic accounts of dynamic learning within causal models. This line of research can incorporate all the core insights concerning integration functions and the role of prior knowledge that have been generated by previous work as well as show how causal learning can take place while operating under realistic processing limitations, even when observations are noisy, distributed over time, and subject to forgetting.

## Causal Reasoning in the Brain

Another development over the past decade, closely related to dynamic causal processing,

has been the exploration of the neural basis for causal learning and inference. A number of investigators have used functional magnetic resonance imaging (fMRI) to detect brain areas that appear sensitive to causal processing (for a review, see Fugelsang & Dunbar 2009). Some studies have adapted the paradigm introduced by Fenker et al. (2005), which involves semantic judgments of whether or not verbally presented concepts (e.g., *virus* and *epidemic*) are causally related (Fenker et al. 2010, Satpute et al. 2005). Others studies have investigated causal learning using standard paradigms based on sequential presentation of possible causes and their associated outcomes (Corlett et al. 2004, Fletcher et al. 2001, Turner et al. 2004). Although the brain areas related to causal tasks have differed across experiments using different materials and paradigms, the general pattern of findings implicates a broad network including frontal areas related to working memory and reasoning, as well as areas related to prediction and error detection, such as the striatum and substantia nigra.

Imaging studies of causal learning have not directly attempted to distinguish among alternative integration functions; however, some findings involving retrospective reevaluation favor the causal noisy-logical function over the additive function posited by associative theories. Corlett et al. (2004) first presented pairs of foods coupled with an allergic reaction (AB+). In a subsequent phase, one of the foods presented alone led to the allergy (A+, which yields backward blocking of cue B) or else did not lead to the allergy (A-, which yields unovershadowing of cue B). Finally, the remaining food from the original pair (B) was presented. Backward blocking is known to result in reduced expectation of the allergy given B, whereas unovershadowing will result in increased expectation of the allergy given B. The Rescorla-Wagner model is unable to explain such retrospective reevaluation effects. Van Hamme & Wasserman's (1994) modification of the Rescorla-Wagner model predicts that the two conditions will have symmetrical effects on expectations about cue B. The power PC theory predicts that the effects will be asymmetrical, as backward unblocking will leave the causal power of B uncertain, whereas unovershadowing implies B is a strong generative cause. The findings concerning surprise reported by Corlett et al. (2004) supported the prediction of the power PC theory: Absence of the allergy given B was more surprising in the unovershadowing condition than was presence of the allergy given B in the backward blocking condition. Dickinson & Burke's (1996) modified SOP model is also able to account for the findings.

## LEARNING AND INFERENCE WITH DIFFERENT TYPES OF CUES AND CAUSAL STRUCTURES

The research we have reviewed above has generally employed fairly simple common-effect structures, with a small number of potential causes, where both causes and their effect are binary variables. Of course, causal relations in the world often form much more diverse and complex networks. Here we review recent work that has addressed some of the issues that arise in attempting to generalize the causal approach to a broader range of structures.

### Covariation and Beyond: Alternative Cues to Causal Structure

In the studies reviewed above, learners were typically informed about the direction of possible causal links (i.e., which factors might be causes and which the effect). In an unfamiliar domain in which uncertainty may exist even about the direction of the causal arrow, what cues might help in inducing causal structure? Beginning with the work of the philosopher Hume (1739/1987), several basic cues have been proposed as input to the causal induction process. An excellent summary of recent work on this topic has been provided by Lagnado et al. (2007). These researchers highlight four types of potential cues: statistical relations (i.e., patterns of covariation among events), temporal order (i.e., the order of occurrence of events in the world), intervention (i.e., observing the

consequences of one's own actions), and prior knowledge. Unlike Hume, these cognitive scientists are concerned with causal learning in everyday life rather than with epistemology.

Covariation is of course a basic cue to causal relationships, but in the absence of temporal cues (an effect never occurs prior to its causes) or prior knowledge, covariation between two variables alone does not determine causal direction. Note, however, that the issue of when and how people make use of previously acquired causal knowledge should not be confused with the issue of when and how such knowledge is acquired by bottom-up mechanisms. No proponent of bottom-up processes would claim that people do not use prior causal knowledge; in fact, the assumed goal of such processes is to acquire causal knowledge that can be applied, for example, to guide subsequent inference and learning. Bottom-up models and models that assume prior knowledge address different issues and are complementary rather than competing.

Researchers in statistics and artificial intelligence have developed constraint-based (CB) algorithms to extract causal structures formalized as Bayes nets solely from covariational data (Pearl 2000, Spirtes et al. 1993/2000; also see Sloman, 2005). CB algorithms operate on Bayes nets but (despite the terminological confusion) do not employ Bayesian inference; rather, they rely on current-data-driven hypothesis testing. The basic CB algorithm involves generating conditional dependency and independence relations associated with every potential causal network for a given set of variables based on certain formal assumptions, notably the causal Markov condition, and performing a search through the space of potential causal networks to identify which set of networks is consistent with the pattern of observed independence relations. The causal Markov condition states that any variable is independent of all variables that are not its (direct or indirect) effects, conditional on knowing its direct causes. For example, people in southern California often set their lawn sprinklers to go on automatically, regardless of whether it rains (a rare occurrence). In this situation, the sprinkler going on and rain occurring represent statistically independent events. However, if the lawn is found to be wet one morning (a common effect that might be produced by either rain or sprinklers), then the two alternative causes become negatively correlated (corresponding to the phenomenon of causal discounting; Kelley 1973).

In knowledge-engineering applications, Bayes nets using CB algorithms offer valuable supplements to human observers. But despite some recent claims (Glymour 2001, Gopnik et al. 2004), there is no evidence that these data-intensive algorithms, which require considerable processing capacity, are used in human causal induction (see critical discussions by Griffiths & Tenenbaum 2009, Lu et al. 2008b). Human learners often make inferences that appear to violate the causal Markov condition (Mayrhofer et al. 2008, Rehder & Burnett 2005, Walsh & Sloman, 2008). Rather than treating their causal knowledge as fixed, humans may sometimes interpret the nonoccurrence of an expected event as a trigger to revise their causal model (falsifying the assumption under which the causal Markov condition should hold), thereby altering their expectations about other effects. Moreover, people have great difficulty extracting cause-effect relations in the absence of critical cues provided by perceived temporal order and their own interventions (Lagnado & Sloman 2004, 2006; Steyvers et al. 2003; see Lagnado et al. 2007), reflecting their processing-capacity limitations. It has also been noted that when such additional cues and/or prior knowledge are made available, human learners (even young children) acquire causal knowledge from much smaller data sets than is possible for Bayes nets using CB algorithms (Gopnik et al. 2001, Sobel & Kirkham 2007) (for Bayesian models of causal learning based on sparse data, see Griffiths & Tenenbaum 2007, 2009; Lu et al. 2008b). There is no conceptual mystery as to why learning based on sparse data is possible (previously collected information is added to the database); the appropriate application of Bayes' theorem for this purpose nonetheless closes a gap in modeling.

Other work has confirmed that human causal learning is heavily influenced by other cues besides "raw" covariation. Temporal order is important (as the associative and causal approaches agree), as it reduces the search space of candidate causes. People expect causes to precede (or at least occur simultaneously with) their effects (Buehner & May 2003; Greville & Buehner 2007; Lagnado & Sloman 2004, 2006; Waldmann & Hagmayer 2005). A surprising finding is that people's perception of time can be warped by their experience of causality (Buehner & Humphreys 2009). In a stimulus-anticipation task, participants' response (a key press) reflected a shortened experience of time when the participants' response caused a target stimulus than when it did not cause it. The received view dating from David Hume is that temporal information, which is observable, serves as input to the causal learning process. Buehner & Humphrey's (2009) finding suggests that temporal information needs to be differentiated by stage: Early sensory information serves as input to the causal inference process, and later (probably perceptual) information may be influenced by top-down knowledge.

People are also sensitive to the fact that their own actions are often causal. In general, knowledge acquired through active intervention is a more reliable guide to causal relations than is sheer observation (Hagmayer et al. 2007, Sloman & Lagnado 2005, Steyvers et al. 2003). People are able to reason suppositionally or counterfactually about what would be expected to happen if some intervention were made.

Although there is general agreement that interventions are a powerful source of causal knowledge, there has been some argument as to whether they actually have some special psychological status beyond their effectiveness in controlling for confounding variables. What scientists and laypeople call interventions, or experimental manipulations, differ from interventions as defined in Bayes nets. In artificial intelligence, Pearl (2000) posited a formal operation on causal Bayes nets called "do," which has the effect of setting the value of a variable directly and hence cutting it off from its usual causal influences ("graph surgery"). Pearl showed that various causal inferences can be derived by applying the "do" operator. Some psychologists have argued that human reasoning about the causal impact of actual actions in the world can be modeled by the "do" operator in a Bayes net (Sloman & Lagnado 2005). Others (Waldmann et al. 2008) have noted that actions (in their typical everyday and scientific sense) may be especially informative for a more general reason, as they constitute one basic way to control for possible alternative causes (the justification for scientific experimentation). In addition, intervention ordinarily conveys information about temporal order of events, which is itself a critical cue to causality (Lagnado & Sloman 2006). But real actions, unlike the formal "do" in Bayes nets, may sometimes prove misleading due to unanticipated or unintended co-occurrences (e.g., an educational intervention may enhance learning not because of its substantive content, but because the intervention attracts attention, which increases the learners' motivation). The placebo effect captures this well-known caveat: Interventions do not necessarily warrant causal conclusions. Under the latter view, one's own actions are often especially informative cues to causal relations, but intervention is not an infallible guide. (For a discussion of models of realistic interventions, see Meder et al. 2010.) Moreover, even for interventions as defined in Bayes nets, it has been argued that when the intervention involves a probabilistic causal relation, a successful intervention does not result in graph surgery (Waldmann et al. 2008).

Finally, it is clear that once causal knowledge at various degrees of generality is acquired, it influences the acquisition of representations of higher-order causal relations (Griffiths & Tenenbaum 2009, Waldmann 1996). Lien & Cheng (2000) argued that people use information at higher levels of generality to distinguish genuine from spurious causal relations at a more specific level. The rooster may invariably crow before the sun rises, but people believe that the actions of small terrestrial animals are simply

not the kind of thing that can move astronomical bodies. Lien & Cheng (2000) showed experimentally that people acquire general causal knowledge about types of entities that cause types of effects and then apply this knowledge to help decide whether novel covariations at a more specific level are in fact causal. They found that the level of generality of the "cause" categories acquired by their subjects depended on how accurately candidate causes at various levels of generality predicted the effect, supporting the view that category formation and causal learning serve the same goal: to represent the world in such a way as to best predict the consequences of actions. Empirical categories are what obey causal laws (Lewis 1929).

Bayes nets designed to use CB algorithms (Pearl 1988, 2000; Spirtes et al. 1993/2000) lack the representational power to acquire knowledge about causes at multiple levels of abstraction (see Schulz et al. 2008). However, approaches to causal learning based on Bayesian inference are able to incorporate hierarchical priors that can capture the influence of known causal regularities at a general level on the acquisition of a novel specific regularity (Kemp et al. 2007). This approach also allows explicit specification of the role of prior causal theories in the acquisition of novel causal relations (Griffiths & Tenenbaum 2009) and allows a natural solution to learning at multiple levels of abstraction at the same time. It provides a language for representing prior knowledge (causal or otherwise), enabling psychological models to better capture the power and flexibility of human causal inference.

## Causal Interactions and Alternative Integration Functions

Though the causal relations in the world may be subtle and complex, there is reason to suppose that humans have a strong preference for simplicity, both in evaluating causal explanations (Lombrozo 2007) and in learning systems of causal relations (Lu et al. 2008b). One aspect of causal simplicity, emphasized in the power PC theory (Cheng 1997), is that people operate under the default assumption that multiple causes exert independent influences on their common effect. That is, the impact of each individual cause on the effect is the same when other causes are also present as it would have been if other causes had been absent. The notion of independent causal influence in fact provides the basis for deriving the noisy-OR and noisy-AND-NOT integration functions for binary variables.

Importantly, however, the assumption of independent causal influence can be overturned by evidence. In many experimental designs analogous to those employed in classical conditioning, combinations of causal cues influence the effect in ways that deviate from what would be expected given the independence assumption. In "negative patterning," for example, cues A and B are each followed by the effect when presented separately, but the AB compound is never followed by the effect. Both rats and humans are capable of learning to respond appropriately to such apparent causal interactions. Novick & Cheng (2004) showed how such interactions can be understood in terms of conjunctive causes (unitized representations of a cue combination) that follow the same integration function, as do simple causes. The assumption of independent causal influence is in fact critical for learning about causal interactions. By definition, two causes interact (to produce an effect or to prevent it) in the manner and to the extent that their joint impact deviates from what would be predicted by the independence assumption (Novick & Cheng 2004).

Moreover, the assumption of independent causal influence also guides inferences about whether or not a causal model should be revised. Liljeholm & Cheng (2007) showed college students an initial set of contingency data that established both the background and a single specific cue A as generative causes. In one condition, the data shown in Phase 1 established the power of cue A as 0.75 (assuming the noisy-OR integration function), whereas normative associative measures show an associative strength of 0.25. In Phase 2, additional events were shown

that occurred in a new background context. The new background had a different causal power; moreover, cues A and B were always paired. Under the assumption of independent causal influence, the pattern in Phase 2 was consistent with the causal power of cue A remaining at 0.75 (even though associative strength had increased from 0.25 to 0.75). The participants were then asked to give their "best bet" regarding whether cue B was causal. The majority indicated that cue B was not causal, as would be expected if people have a default preference to maintain a simpler causal network (with fewer causes), indicating that causal rather than associative strength is what defines "sameness of change" due to a cause.

In a second condition, by contrast, the apparent causal power of cue A changed from Phase 1 to Phase 2. If reasoners are sensitive to independent causal influence, they would tacitly "explain" the apparent change in Phase 2 by assuming that cue B was also causal, rather than that cue A's causal power had actually changed. And indeed, the majority of participants agreed that cue B was in fact causal. These findings confirm that people's readiness to accept a more complex causal network (in the sense of adding an additional cause) depends on whether or not they detect an apparent violation of independent causal influence in a simpler network. The associative measure, $\Delta P$ (to which the asymptotic value computed by the Rescorla-Wagner rule is equivalent), could not explain the observed pattern of judgments about the causal power of B in the Liljeholm & Cheng (2007) study and, more generally, does not provide a coherent (i.e., logically consistent) definition of independent causal influence. Thus, coherence appears to be an a priori assumption, as might be expected because past experiences would be useless otherwise. Anything would follow from anything else if logical consistency is not required.

Recent work has begun to explore additional integration functions that may be evoked by different content. The noisy-OR and noisy-AND-NOT functions apply in the case of binary variables; other integration functions may be evoked when the cause and/or effect variables are viewed as continuous in magnitude (Lu et al. 2008a). Waldmann (2007) demonstrated that people apply radically different integration functions for alternative types of variables. In particular, causes that involve intensive quantities (e.g., taste) or preferences (e.g., liking) bias people toward averaging the causal influences, whereas extensive quantities (e.g., strength of a drug) lead to a tendency to add. However, the knowledge underlying these processes is often fallible and unstable. People are easily influenced by additional task-related cues, including the way data are presented, the difficulty of the inference task, and transfer from previous tasks. Understanding the nature of the integration functions that define independent causal influence will be essential in extending the causal approach to more complex and diverse situations (Lucas & Griffiths 2010).

## Causal Inference Based on Categories and Analogies

Causal knowledge plays an important role in learning and reasoning based on both general categories (Kemp et al. 2007, Lien & Cheng 2000, Marsh & Ahn 2009, Waldmann & Hagmayer 2006) and on small numbers of specific examples that lend themselves to reasoning by analogy (Lee & Holyoak 2008). Numerous studies suggest that the strength of inferences depends on causal models of the interconnections between members of various categories. For example, if salmon are known to have a disease, it may then seem more likely that bears will have it, presumably because people believe diseases can sometimes be transmitted from prey to predator; in contrast, knowing that bears have a disease would constitute weaker evidence that salmon will have it (e.g., Bailenson et al. 2002, Medin et al. 2003, Shafto & Coley 2003). Rehder (2006) found that similarity-based influences on inferences were almost entirely eliminated when a generalization could be based on causal relations instead. Causal knowledge appears to play a major role in expert

reasoning about complex categories such as those involved in clinical diagnosis (Ahn et al. 2009).

The precise manner in which causal knowledge impacts category inferences has been debated (see discussion in Waldmann & Hagmayer 2010). Some investigators have claimed that causes (especially those earlier in a causal chain) are inherently more central than effects in supporting inferences (Ahn et al. 2000; see also Hadjichristidis et al. 2004). Rehder and his colleagues (Rehder 2009, Rehder & Burnett 2005, Rehder & Kim 2006) have argued that a more general framework is required to understand how people use knowledge of categories to make causal-based generalizations (CBGs), which can involve effects as well as causes. This approach has been formalized in the CBG model (Rehder 2009), which incorporates the causal assumptions in the power PC theory. The basic hypothesis is that people expect exemplars of categories to be causally coherent, in that an instance should exhibit the features that would be expected based on a causal model of the category. For example, the CBG model predicts that, given the same set of observed events, increasing the strength of a causal link between a known feature associated with a category and a new unobserved feature will increase the judged prevalence of the new feature when the latter is an effect, whereas decreasing that causal strength will increase the judged prevalence of the new feature when the latter is a cause. The experiments reported by Rehder (2009) provide support for this and other predictions derived from the CBG model.

Causal models also play an important role in guiding inferences based on specific examples in combination with more general causal regularities. Lee & Holyoak (2008) demonstrated how causal knowledge guides analogical inference, showing that analogical inference is not solely determined by quality of the overall similarity of the source and target analogs. Using a common-effect structure, Lee and Holyoak manipulated structural correspondences between an initial source analog and a novel target analog, as well as the causal polarity (generative or preventive) of multiple causes present in the target. The source always showed that the effect occurred given the combination of two generative causes and one preventive cause. Then if the target analog dropped the preventive cause, people rated the target as more likely to exhibit the effect than if the preventive cause was present, even though dropping the preventer reduced overall similarity between the analogs. Holyoak et al. (2010) identified an additional dissociation between inference strength and similarity of analogs using queries that required causal attribution. These investigators showed that a Bayesian extension of the power PC theory (Lu et al. 2008b) could explain the impact of causal models on analogical inference, even when causal knowledge is based on a single example and hence is highly uncertain.

## CONCLUSIONS AND CONTINUING CONTROVERSIES

Let us return to our key question: Given that our understanding of the causal world is entirely our mental reconstruction from noncausal input, with the goal of identifying invariant empirical relations that support predictions regarding the consequences of actions, what has the field learned about this reconstruction process? After two decades of vigorous debate and active empirical research, recent developments have firmly established that humans learn networks of explicit cause-effect relations rather than associations, and use the resulting causal models to predict future effects and make attributions about past causes. Perhaps the most significant conclusion from the research of the past decade is that human causal reasoning (for relatively simple situations that do not exceed available working memory resources) is both robust and rational. This conclusion implies that, for binary variables, people have the defeasible default assumption that multiple causes of an effect influence it independently, as reflected in noisy-logical

integration functions. The domain-general causal assumptions underlying the integration functions bootstrap the rational process of human causal learning. Different types of causal variables evoke alternative default integration functions, probably due to different definitions of independence logically implied by the variable type. Augmented by the machinery of Bayesian inference to handle the inherent uncertainty of induction and the inclusion of prior causal knowledge, and adopting a preference for the simplest explanation and an assumption that the world is logically consistent, the causal approach can explain a wealth of data involving judgments of causal strength, causal structure, attribution, and diagnosis. A causal framework, logical consistency, and simplicity are essential a priori assumptions that enable the human causal-learning process to reconstruct order underlying seeming chaos for the purpose of achieving understanding and planning actions.

Many important questions remain, and others have been recently raised. One important question follows from the rationality and success of the causal approach. If humans have evolved to approach causal learning with a set of a priori causal assumptions, should scientists (who at present predominantly use associative statistics in their research) consider using a similar set of assumptions if they test causal hypotheses? Another important question for further research involves the role of the hypothesis space for causal learning (an issue highlighted by the Bayesian approach, in which the hypothesis space is made explicit). Relative to the obvious space of possible causal structures defined on pre-existing variables (of the sort illustrated in **Figures 1**, **2**, and **4**), do humans evaluate possible states of the world that are not only more fine-grained (e.g., distributions of causal strength), but that emerge from a broader hypothesis space (e.g., alternative definitions of cause and effect categories, the variables linked by causal relations)? If so, given that enumerating and evaluating possibilities in a larger hypothesis space must require greater processing capacity, what are the criteria for hypothesis revision (in particular, for enlarging one's hypothesis space)?

Much more work will be required to investigate how people learn and reason with more complex causal networks that tax the limits of their working memory capacity (Waldmann & Walker 2005). In addition, further work on the neural basis of causal learning and inference may help to refine our understanding of the nature of causal models. The limited available evidence has not established whether causal relations involve a distinct neural substrate or whether they are processed using the same neural machinery as other core conceptual relations (e.g., category membership) that enable predictions.

The findings over the past decade regarding the use of a causal framework in humans have raised another natural question: Do nonhuman species also adopt a similar framework? Because many comparative psychologists view the cognitive capacities of humans as continuous with those of nonhuman animals, the evidence that humans use causal models has encouraged research on whether nonhuman animals do so as well (e.g., Blaisdell et al. 2006, Call 2004). Penn & Povinelli (2007) agree that nonhuman causal cognition is significantly more sophisticated than can be accounted for by traditional associationist theories. In particular, nonhuman animals do not simply learn about observable contingencies; they appear to be sensitive to the unobservable constraints specific to causal inference. However, Penn & Povinelli (2007) argue there is no compelling evidence that nonhuman animals are capable of reasoning about higher-order causal relations or abstract causal principles (also Penn et al. 2008). What are the similarities and differences between human and nonhuman causal inference? Among other possibilities, might the differences between humans and nonhumans reflect differences in the space of causal hypotheses available to different types of reasoners?

## DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

## ACKNOWLEDGMENTS

Preparation of this review was supported by grant FA9550-08-1-0489 from the Air Force Office of Scientific Research. We thank Hee Seung Lee, Hongjing Lu, Derek Penn, Robert Sternberg, Michael Vendetti, Michael Waldmann, and Alan Yuille for helpful comments, discussions, and other assistance.

## LITERATURE CITED

Ahn W-K, Kim NS, Lassaline ME, Dennis MJ. 2000. Causal status as a determinant of feature centrality. *Cogn. Psychol.* 41:361–416

Ahn W-K, Proctor CC, Flanagan EH. 2009. Mental health clinicians' beliefs about the biological, psychological, and environmental bases of mental disorders. *Cogn. Sci.* 33:147–82

Ahn W-K, Kalish CW. 2000. The role of covariation versus mechanism information in causal attribution. In *Cognition and Explanation*, ed. R Wilson, F Keil, pp. 227–53. Cambridge, MA: MIT Press

Ahn W-K, Kalish CW, Medin DL, Gelman SA. 1995. The role of covariation versus mechanism information in causal attribution. *Cognition* 54:299–352

Allan LG, Jenkins HM. 1983. The effect of representations of binary variables on judgments of influence. *Learn. Motiv.* 14:381–405

Bailenson JN, Shum MS, Atran S, Medin DL, Coley JD. 2002. A bird's eye view: biological categorization and reasoning within and across cultures. *Cognition* 84:1–53

Beckers T, de Houwer J. 2005. Outcome additivity and outcome maximality influence cue competition in human causal learning. *J. Exp. Psychol.: Learn. Mem. Cogn.* 11:238–49

Blaisdell AP, Sawa K, Leising KJ, Waldmann MR. 2006. Causal reasoning in rats. *Science* 311:1020–22

Booth SL, Buehner MJ. 2007. Asymmetries in cue competition in forward and backward blocking designs: further evidence for causal model theory. *Q. J. Exp. Psychol.* 60:387–99

Buehner MJ, Cheng PW, Clifford D. 2003. From covariation to causation: a test of the assumption of causal power. *J. Exp. Psychol.: Learn. Mem. Cogn.* 29:1119–40

Buehner MJ, Humphreys GR. 2009. Causal binding of actions to their effects. *Psychol. Sci.* 20:1221–28

Buehner MJ, May J. 2003. Rethinking temporal contiguity and the judgment of causality: effects of prior knowledge, experience, and reinforcement procedure. *Q. J. Exp. Psychol.* 56A:865–90

Call J. 2004. Inferences about the location of food in the great apes (*Pan paniscus*, *Pan troglodytes*, *Gorilla gorilla*, and *Pongo pygmaeus*). *J. Comp. Psychol.* 118:232–41

Chater N, Tenenbaum JB, Yuille A. 2006. Probabilistic models of cognition: conceptual foundations. *Trends Cogn. Sci.* 10:287–91

Chater N, Vitányi P. 2003. Simplicity: a unifying principle in cognitive science? *Trends Cogn. Sci.* 7:19–22

Cheng PW. 1997. From covariation to causation: a causal power theory. *Psychol. Rev.* 104:367–405

Cheng PW. 2000. Causality in the mind: estimating contextual and conjunctive causal power. In *Explanation and Cognition*, ed. F Keil, R Wilson, pp. 227–53. Cambridge, MA: MIT Press

Cheng PW, Holyoak KJ. 1995. Complex adaptive systems as intuitive statisticians: causality, contingency, and prediction. In *Comparative Approaches to Cognitive Science*, ed. HL Roitblat, J-A Meyer, pp. 271–302. Cambridge, MA: MIT Press

Cheng PW, Novick LR. 1992. Covariation in natural causal induction. *Psychol. Rev.* 99:365–82

Cheng PW, Novick LR. 2005. Constraints and nonconstraints in causal learning: reply to White (2005) and to Luhmann and Ahn (2005). *Psychol. Rev.* 112:694–706

Cheng PW, Novick LR, Liljeholm M, Ford C. 2007. Explaining four psychological asymmetries in causal reasoning: implications of causal assumptions for coherence. In *Topics in Contemporary Philosophy, Vol. 4: Explanation and Causation*, ed. M O'Rourke, pp. 1–32. Cambridge, MA: MIT Press

Corlett PR, Aitken MRF, Dickinson A, Shanks DR, Honey GD, et al. 2004. Prediction error during retrospective revaluation of causal associations in humans: fMRI evidence in favor of an associative model of learning. *Neuron* 44:877–88

Danks D. 2003. Equilibria of the Rescorla-Wagner model. *J. Math. Psychol.* 47:109–121

Danks D, Griffiths TL, Tenenbaum JB. 2003. Dynamical causal learning. In *Advances in Neural Information Processing Systems*, ed. S Becker, S Thrun, K Obermayer, Vol. 15, pp. 67–74. Cambridge, MA: MIT Press

Daw N, Courville AC, Dayan P. 2007. Semi-rational models of conditioning: the case of trial order. In *The Probabilistic Mind: Prospects for Rational Models of Cognition*, ed. M Oaksford, N Chater, pp. 431–52. New York: Oxford Univ. Press

Dickinson A, Burke J. 1996. Within-compound associations mediate the retrospective revaluation of causality judgements. *Q. J. Exp. Psychol.* 37B:397–416

Dickinson A, Shanks DR, Evenden JL. 1984. Judgment of act-outcome contingency: the role of selective attribution. *Q. J. Exp. Psychol.* 36A:29–50

Dunbar K, Fugelsang J. 2005. Causal thinking in science: how scientists and students interpret the unexpected. In *Scientific and Technical Thinking*, ed. ME Gorman, RD Tweney, D Gooding, A Kincannon, pp. 57–80. Mahwah, NJ: Erlbaum

Fenker DB, Schoenfeld MA, Waldmann MR, Schuetze H, Heinze H-J, Duezel E. 2010. "Virus and epidemic": Causal knowledge activates prediction error circuitry. *J. Cogn. Neurosci.* 22:2151–63

Fenker DB, Waldmann MR, Holyoak KJ. 2005. Accessing causal relations in semantic memory. *Mem. Cogn.* 33:1036–46

Fletcher PC, Anderson JM, Shanks DR, Honey RAE, Carpenter TA, et al. 2001. Responses of human frontal cortex to surprising events are predicted by formal associative learning theory. *Nat. Neurosci.* 4:1043–48

Fugelsang J, Dunbar K. 2009. Brain-based mechanisms underlying causal reasoning. In *Neural Correlates of Thinking*, ed. E Kraft, pp. 269–79. Berlin: Springer

Gallistel CR. 1990. *The Organization of Learning*. Cambridge, MA: MIT Press

Gelman S, Kremer KE. 1991. Understanding natural cause: children's explanations of how objects and their properties originate. *Child Dev.* 62:396–414

Gleitman LR, Joshi AK, eds. 2000. *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum

Gluck MA, Bower GH. 1988. Evaluating an adaptive network model of human learning. *J. Mem. Lang.* 27:50–55

Glymour C. 2001. *The Mind's Arrows: Bayes Nets and Graphical Causal Models in Psychology*. Cambridge, MA: MIT Press

Goodman ND, Mansinghka VK, Tenenbaum JP. 2007. Learning grounded causal models. In *Proceedings of the Twenty-Ninth Annual Conference of the Cognitive Science Society*, ed. DS McNamara, G Trafton, pp. 305–10. Austin, TX: Cogn. Sci. Soc.

Gopnik A. 2009. *The Philosophical Baby: What Children's Minds Tell Us About Truth, Love, and the Meaning of Life*. New York: Farrar, Straus & Giroux

Gopnik A, Schulz L. 2007. *Causal Learning: Psychology, Philosophy, and Computation*. New York: Oxford Univ. Press

Gopnik A, Glymour C, Sobel DM, Schulz LE, Kushnir T, Danks D. 2004. A theory of causal learning in children: causal maps and Bayes nets. *Psychol. Rev.* 111:1–30

Gopnik A, Sobel DM, Schulz LE, 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. *Dev. Psychol.* 37:620–29

Greville WJ, Buehner MJ. 2007. The influence of temporal distributions on causal induction from tabular data. *Mem. Cogn.* 35:444–53

Griffiths TL, Kemp C, Tenenbaum JB. 2008. Bayesian models of cognition. In *Cambridge Handbook of Computational Psychology*, ed. R Sun, pp. 59–100. London: Cambridge Univ. Press

Griffiths TL, Tenenbaum JB. 2005. Structure and strength in causal induction. *Cogn. Psychol.* 51:354–84

Griffiths TL, Tenenbaum JB. 2007. Two proposals for causal grammars. In *Causal Learning: Psychology, Philosophy, and Computation*, ed. A Gopnik, L Schulz, pp. 323–45. London: Oxford Univ. Press

Griffiths TL, Tenenbaum JB. 2009. Theory-based causal induction. *Psychol. Rev.* 116:661–716
Hadjichristidis C, Sloman SA, Stevenson R, Over D. 2004. Feature centrality and property induction. *Cogn. Sci.* 28:45–74
Hagmayer Y, Sloman SA, Lagnado DA, Waldmann MR. 2007. Causal reasoning through intervention. In *Causal Learning: Psychology, Philosophy, and Computation*, ed. A Gopnik, L Schulz, pp. 86–100. London: Oxford Univ. Press
Hattori M, Oaksford M. 2007. Adaptive noninterventional heuristics for covariation detection in causal induction: model comparison and rational analysis. *Cogn. Sci.* 31:765–814
Holyoak KJ, Lee HS, Lu H. 2010. Analogical and category-based inference: a theoretical integration with Bayesian causal models. *J. Exp. Psychol.: Gen.* 139: In press
Hume D. 1739/1987. *A Treatise of Human Nature*. Oxford, UK: Clarendon. 2nd ed.
Jenkins HM, Ward WC. 1965. Judgment of contingency between responses and outcomes. *Psychol. Monogr.: Gen. Appl.* 79(1, Whole No. 594)
Kahneman D, Slovic P, Tversky A. 1982. *Judgment Under Uncertainty: Heuristics and Biases*. London: Cambridge Univ. Press
Kamin LJ. 1969. Predictability, surprise, attention, and conditioning. In *Punishment and Aversive Behavior*, ed. BA Campbell, RM Church, pp. 276–96. New York: Appleton-Century-Crofts
Kant I. 1781/1965. *Critique of Pure Reason*. London: Macmillan
Kelley HH. 1967. Attribution theory in social psychology. In *Nebraska Symposium on Motivation*, ed. D Levine, Vol. 15, pp. 192–238. Lincoln: Univ. Nebraska Press
Kelley HH. 1973. The process of causal attribution. *Am. Psychol.* 28:107–28
Kemp C, Goodman ND, Tenenbaum JB. 2007. Learning causal schemata. In *Proceedings of the Twenty-Ninth Annual Conference of the Cognitive Science Society*, ed. DS McNamara, G Trafton, pp. 389–94. Austin, TX: Cogn. Sci. Soc.
Kemp C, Tenenbaum JB. 2009. Structured statistical models of inductive reasoning. *Psychol. Rev.* 116:20–58
Kruschke JK. 2006. Locally Bayesian learning with applications to retrospective revaluation and highlighting. *Psychol. Rev.* 113:677–99
Kushnir T, Gopnik A, Schulz LE, Danks D. 2003. Inferring hidden causes. In *Proceedings of the Twenty-Fifth Annual Meeting of the Cognitive Science Society*, ed. R Alterman, D Kirsh, pp. 699–703. Mahwah, NJ: Erlbaum
Lagnado DA, Sloman SA. 2004. The advantage of timely intervention. *J. Exp. Psychol.: Learn. Mem. Cogn.* 30:856–76
Lagnado DA, Sloman SA. 2006. Time as a guide to cause. *J. Exp. Psychol.: Learn. Mem. Cogn.* 32:451–60
Lagnado DA, Waldmann MR, Hagmayer Y, Sloman SA. 2007. Beyond covariation: cues to causal structure. In *Causal Learning: Psychology, Philosophy, and Computation*, ed. A Gopnik, L Schulz, pp. 154–72. London: Oxford Univ. Press
Lee HS, Holyoak KJ. 2008. The role of causal models in analogical inference. *J. Exp. Psychol.: Learn. Mem. Cogn.* 34:1111–22
Lewis CI. 1929. *Mind and the World Order*. New York: Scribner
Lien Y, Cheng PW. 2000. Distinguishing genuine from spurious causes: a coherence hypothesis. *Cogn. Psychol.* 40:87–137
Liljeholm M, Cheng PW. 2007. When is a cause the "same"? Coherent generalization across contexts. *Psychol. Sci.* 18:1014–21
Liljeholm M, Cheng PW. 2009. The influence of virtual sample size on confidence and causal strength judgments. *J. Exp. Psychol.: Learn. Mem. Cogn.* 35:157–72
Lober K, Shanks D. 2000. Is causal induction based on causal power? Critique of Cheng (1997). *Psychol. Rev.* 107:195–212
Lombrozo T. 2007. Simplicity and probability in causal explanation. *Cogn. Psychol.* 55:232–57
López FJ, Cobos PL, Caño A. 2005. Associative and causal reasoning accounts of causal induction: symmetries and asymmetries in predictive and diagnostic inferences. *Mem. Cogn.* 33:1388–98
López FJ, Shanks DR. 2008. Models of animal learning and their relations to human learning. In *Cambridge Handbook of Computational Psychology*, ed. R Sun, pp. 589–611. London: Cambridge Univ. Press

Lu H, Rojas RR, Beckers T, Yuille AL. 2008a. Sequential causal learning in humans and rats. In *Proceedings of the Thirtieth Annual Conference of the Cognitive Science Society*, ed. BC Love, K McRae, VM Sloutsky, pp. 195–88. Austin, TX: Cogn. Sci. Soc.

Lu H, Yuille AL, Liljeholm M, Cheng PW, Holyoak KJ. 2008b. Bayesian generic priors for causal learning. *Psychol. Rev.* 115:955–82

Lucas CG, Griffiths TL. 2010. Learning the form of causal relationships using hierarchical Bayesian models. *Cogn. Sci.* 34:113–47

Luhmann CC, Ahn W-K. 2007. BUCKLE: a model of unobserved cause learning. *Psychol. Rev.* 114:657–77

Mackay D. 2003. *Information Theory, Inference and Learning Algorithms*. London: Cambridge Univ. Press

Marsh JK, Ahn W-K. 2009. Spontaneous assimilation of continuous values and temporal information in causal induction. *J. Exp. Psychol.: Learn. Mem. Cogn.* 35:334–52

Mayrhofer R, Goodman ND, Waldmann MR, Tenenbaum JB. 2008. Structured correlation from the causal background. In *Proceedings of the Thirtieth Annual Conference of the Cognitive Science Society*, ed. BC Love, K McRae, VM Sloutsky, pp. 303–8. Austin, TX: Cogn. Sci. Soc.

McClelland JL, Thompson RM. 2007. Using domain-general principles to explain children's causal reasoning abilities. *Dev. Sci.* 10:333–56

Meder B, Gerstenberg T, Hagmayer Y, Waldmann MR. 2010. Observing and intervening: rational and heuristic models of causal decision making. *Open Psychol. J.* 3:119–35

Meder B, Mayrhofer R, Waldmann MR. 2009. A rational model of elementary diagnostic inference. In *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society*, ed. N Taatgen, H van Rijn, pp. 2176–81. Austin, TX: Cogn. Sci. Soc.

Medin DL, Coley JD, Storms G, Hayes BK. 2003. A relevance theory of induction. *Psychon. Bull. Rev.* 10:517–32

Novick LR, Cheng PW. 2004. Assessing interactive causal influence. *Psychol. Rev.* 111:455–85

Pearl J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann

Pearl J. 2000. *Causality*. London: Cambridge Univ. Press

Penn DC, Holyoak KJ, Povinelli DJ. 2008. Darwin's mistake: explaining the discontinuity between human and nonhuman minds. *Behav. Brain Sci.* 31:109–78

Penn DC, Povinelli DJ. 2007. Causal cognition in human and nonhuman animals: a comparative, critical review. *Annu. Rev. Psychol.* 58:97–118

Perales JC, Shanks DR. 2007. Models of covariation-based causal judgment: a review and synthesis. *Psychon. Bull. Rev.* 14:577–96

Piaget J. 1930. *The Child's Conception of Physical Causality*. London: Kegan Paul, Trench, Trubner

Pourret O, Naïm P, Marcot B, eds. 2008. *Bayesian Networks: A Practical Guide to Applications*. New York: Wiley

Rehder B. 2006. When similarity and causality compete in category-based property generalization. *Mem. Cogn.* 34:3–16

Rehder B. 2009. Causal-based property generalization. *Cogn. Sci.* 33:301–43

Rehder B, Burnett R. 2005. Feature inference and the causal structure of categories. *Cogn. Psychol.* 50:264–314

Rehder B, Kim S. 2006. How causal knowledge affects classification: a generative theory of categorization. *J. Exp. Psychol.: Learn. Mem. Cogn.* 32:659–83

Reichenbach H. 1956. *The Direction of Time*. Berkeley: Univ. Calif. Press

Rescorla RA, Wagner AR. 1972. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In *Classical Conditioning II: Current Theory and Research*, ed. AH Black, WF Prokasy, pp. 64–99. New York: Appleton-Century-Crofts

Rumelhart DE, McClelland JL, PDP Res. Group. 1986. *Parallel Distributed Processing: Explorations in the Microstructures of Cognition. Vol. 1: Foundations*. Cambridge, MA: MIT Press

Salmon WC. 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton, NJ: Princeton Univ. Press

Satpute AB, Fenker DB, Waldmann MR, Tabibnia G, Holyoak KJ, Lieberman MD. 2005. An fMRI study of causal judgments. *Eur. J. Neurosci.* 22:1233–38

Saxe R, Tenenbaum JB, Carey S. 2005. Secret agents: inferences about hidden causes by 10- and 12-month infants. *Psychol. Sci.* 16:995–1001

Schulz LE, Goodman ND, Tenenbaum JB, Jenkins AC. 2008. Going beyond the evidence: preschoolers' inferences about abstract laws and anomalous data. *Cognition* 109:211–23

Schulz LE, Sommerville J. 2006. God does not play dice: causal determinism and preschoolers' causal inferences. *Child Dev.* 77:427–42

Schustack MW, Sternberg RJ. 1981. Evaluation of evidence in causal inference. *J. Exp. Psychol.: Gen.* 110:101–20

Shafto P, Coley JD. 2003. Development of categorization and reasoning in the natural world: novices to experts, naive similarity to ecological knowledge. *J. Exp. Psychol.: Learn. Mem. Cogn.* 29:641–49

Shanks DR. 1987. Acquisition functions in contingency judgment. *Learn. Motiv.* 18:147–66

Shanks DR. 1991. Categorization by a connectionist network. *J. Exp. Psychol.: Learn. Mem. Cogn.* 17:433–43

Shanks DR, Dickinson A. 1987. Associative accounts of causality judgment. In *The Psychology of Learning and Motivation*, ed. GH Bower, Vol. 21, pp. 229–61. San Diego, CA: Academic

Shanks DR, Holyoak KJ, Medin DL, eds. 1996. *The Psychology of Learning and Motivation, Vol. 34: Causal Learning.* San Diego, CA: Academic

Sloman SA, Lagnado DA. 2005. Do we "do"? *Cogn. Sci.* 29:5–39

Sobel DM, Kirkham NZ. 2007. Bayes nets and babies: infants' developing statistical reasoning abilities and their representation of causal knowledge. *Dev. Sci.* 10:298–306

Spirtes P, Glymour C, Scheines R. 1993/2000. *Causation, Prediction, and Search (Springer Lecture Notes in Statistics)*. Cambridge, MA: MIT Press. 2nd ed., rev.

Steyvers M, Tenenbaum JB, Wagenmakers EJ, Blum B. 2003. Inferring causal networks from observations and interventions. *Cogn. Sci.* 27:453–89

Stout SC, Miller RR. 2007. Sometimes-competing retrieval (SOCR): a formalization of the comparator hypothesis. *Psychol. Rev.* 114:759–83

Taatgen N, van Rijn H, eds. 2009. *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society*. Austin, TX: Cogn. Sci. Soc.

Tenenbaum JB, Griffiths TL. 2001. Structure learning in human causal induction. In *Advances in Neural Information Processing Systems*, ed. TK Leen, TG Dietterich, V Tresp, Vol. 13, pp. 59–65. Cambridge, MA: MIT Press

Turner DC, Aitken MRF, Shanks DR, Sahakian BJ, Robbins TW, et al. 2004. The role of the lateral frontal cortex in causal associative learning: exploring preventative and superlearning. *Cereb. Cortex* 14:872–80

Tversky A, Kahneman D. 1973. Judgment under uncertainty: heuristics and biases. *Science* 185:1124–31

Tversky A, Kahneman D. 1982. Causal schemata in judgements under uncertainty. In *Judgment Under Uncertainty: Heuristics and Biases*, ed. D Kahneman, P Slovic, A Tversky, pp. 117–28. London: Cambridge Univ. Press

Van Hamme LJ, Wasserman EA. 1994. Cue competition in causality judgments: the role of nonpresentation of compound stimulus elements. *Learn. Motiv.* 25:127–51

Waldmann MR. 1996. Knowledge-based causal induction. In *The Psychology of Learning and Motivation, Vol. 34: Causal Learning*, ed. DR Shanks, KJ Holyoak, DL Medin, pp. 47–88. San Diego, CA: Academic

Waldmann MR. 2000. Competition among causes but not effects in predictive and diagnostic learning. *J. Exp. Psychol.: Learn. Mem. Cogn.* 26:53–76

Waldmann MR. 2001. Predictive versus diagnostic causal learning: evidence from an overshadowing paradigm. *Psychon. Bull. Rev.* 8:600–8

Waldmann MR. 2007. Combining versus analyzing multiple causes: how domain assumptions and task context affect integration rules. *Cogn. Sci.* 31:233–56

Waldmann MR, Cheng PW, Hagmayer Y, Blaisdell AP. 2008. Causal learning in rats and humans: a minimal rational model. In *Rational Models of Cognition*, ed. N Chater, M Oaksford, pp. 453–84. London: Oxford Univ. Press

Waldmann MR, Hagmayer Y. 2005. Seeing versus doing: two modes of accessing causal knowledge. *J. Exp. Psychol.: Learn. Mem. Cogn.* 31:216–27

Waldmann MR, Hagmayer Y. 2006. Categories and causality: the neglected direction. *Cogn. Psychol.* 53:27–58

Waldmann MR, Hagmayer Y. 2012. Causal reasoning. In *Oxford Handbook of Cognitive Psychology*, ed. D Reisberg. New York: Oxford Univ. Press. In press

Waldmann MR, Holyoak KJ. 1992. Predictive and diagnostic learning within causal models: asymmetries in cue competition. *J. Exp. Psychol.: Gen.* 121:222–36

Waldmann MR, Holyoak KJ, Fratianne A. 1995. Causal models and the acquisition of category structure. *J. Exp. Psychol.: Gen.* 124:181–206

Waldmann MR, Martignon L. 1998. A Bayesian network model of causal learning. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, ed. MA Gernsbacher, SJ Derry, pp. 1102–7. Mahwah, NJ: Erlbaum

Waldmann MR, Walker JM. 2005. Competence and performance in causal learning. *Learn. Behav.* 33:211–29

Walsh CR, Sloman SA. 2008. Updating beliefs with causal models: violations of screening off. In *Memory and Mind: A Festschrift for Gordon H. Bower*, ed. MA Gluck, JR Anderson, SM Kosslyn, pp. 345–58. New York: Psychol. Press

White PA. 1998. Causal judgment: use of different types of contingency information as confirmatory and disconfirmatory. *Eur. J. Cogn. Psychol.* 10:131–70

White PA. 2004. Causal judgment from contingency information: a systematic test of the *p*CI rule. *Mem. Cogn.* 32:353–68

White PA. 2008. Accounting for occurrences: a new view of the use of contingency information in causal judgment. *J. Exp. Psychol.: Learn. Mem. Cogn.* 34:204–18

Wu M, Cheng PW. 1999. Why causation need not follow from statistical association: boundary conditions for the evaluation of generative and preventive causal powers. *Psychol. Sci.* 10:92–97

Yuille AL, Lu H. 2008. The noisy-logical distribution and its application to causal inference. In *Advances in Neural Information Processing Systems*, ed. JC Platt, D Koller, Y Singer, S Roweis, Vol. 20, pp. 1673–80. Cambridge, MA: MIT Press