Journal of Experimental Psychology: General
2010, Vol. 139, No. 4, 702–727

© 2010 American Psychological Association
0096-3445/10/$12.00 DOI: 10.1037/a0020488

Analogical and Category-Based Inference: A Theoretical Integration With Bayesian Causal Models

Keith J. Holyoak, Hee Seung Lee, and Hongjing Lu
University of California, Los Angeles

A fundamental issue for theories of human induction is to specify constraints on potential inferences. For inferences based on shared category membership, an analogy, and/or a relational schema, it appears that the basic goal of induction is to make accurate and goal-relevant inferences that are sensitive to uncertainty. People can use source information at various levels of abstraction (including both specific instances and more general categories), coupled with prior causal knowledge, to build a causal model for a target situation, which in turn constrains inferences about the target. We propose a computational theory in the framework of Bayesian inference and test its predictions (parameter-free for the cases we consider) in a series of experiments in which people were asked to assess the probabilities of various causal predictions and attributions about a target on the basis of source knowledge about generative and preventive causes. The theory proved successful in accounting for systematic patterns of judgments about interrelated types of causal inferences, including evidence that analogical inferences are partially dissociable from overall mapping quality.

Keywords: causal models, category-based inference, analogical inference, schemas, Bayesian theory

Inductive reasoning, which has been characterized as encompassing "all inferential processes that expand knowledge in the face of uncertainty" (Holland, Holyoak, Nisbett, & Thagard, 1986, p. 1), is fundamental to human cognition. The great puzzle of induction has been to understand how humans often manage to make plausible and useful inferences based on what would seem to be a bewildering array of data. The philosopher Charles Peirce (1932) argued that the key must be that people are guided by "special aptitudes for guessing right" (p. 476). Similarly, psychologists have argued that inductive reasoning must depend on some basic constraints that serve to generate preferences for those inferences that are more likely to serve the goals of the reasoner (Holland et al., 1986; Rips, 1990). But what might these constraints actually be?

Keith J. Holyoak, Hee Seung Lee, and Hongjing Lu, Department of Psychology, University of California, Los Angeles.

Hee Seung Lee is now at the Department of Psychology, Carnegie-Mellon University.

All authors contributed equally to this article; names are ordered alphabetically. Preliminary reports of part of this article were presented at the Second International Conference on Analogy (Sofia, Bulgaria, July 2009), at the Thirty-Second Annual Conference of the Cognitive Science Society (Portland, Oregon, August 2010), and in a doctoral dissertation completed by Hee Seung Lee at the University of California, Los Angeles, in 2010. The project was supported by Office of Naval Research Grant N000140810186.

We thank Patricia Cheng for helpful discussions of the derivation of theoretical predictions for causal attribution. Matlab code for the models presented here is available from the website of the UCLA Computational Vision and Learning Lab (http://cvl.psych.ucla.edu).

Correspondence concerning this article should be addressed to Keith J. Holyoak, Hee Seung Lee, or Hongjing Lu, Department of Psychology, University of California, Los Angeles, 405 Hilgard Avenue, Los Angeles, CA 90095-1563. E-mail: Keith J. Holyoak: holyoak@lifesci.ucla.edu, Hee Seung Lee: heeseungll@gmail.com, or Hongjing Lu: hongjing@ucla.edu

Inductive reasoning is often guided by some combination of prior knowledge about categories and more specific cases. Very crudely, inductive inference is based on the assumption that if two situations are alike in some respects, they may be alike in others. The problem with this characterization is that it has proved difficult to pin down what similarities between two situations are important for establishing particular inferences (Goodman, 1955; Medin, Goldstone, & Gentner, 1993). With respect to category-based inferences, a long-standing position is that induction is guided not simply by overall similarity but by people's informal theories concerning the causal processes that give rise to the properties that category members share (Murphy & Medin, 1985; also Carey, 1985; Keil, 1989; for a review see Medin & Rips, 2005). For example, people believe that biological species such as raccoons tend to inherit the properties of their parents, whereas manufactured objects such as chairs are generally constructed to fulfill certain functions important to their human creators.

Scientific hypotheses are often formed and evaluated on the basis of causal knowledge about categories or schemas for types of phenomena and/or on the basis of specific analogs that are better understood than is the target domain (Dunbar & Fugelsang, 2005; Holyoak & Thagard, 1995). When the source includes a relatively specific example, inductive reasoning is based on analogy. In some areas of science in which experimental research is impossible, such as historical ethnography, analogy may provide the only viable mechanism for evaluating hypotheses. Talalay (1987) gave the example of interpreting the function of strange clay fragments discovered in Neolithic Greek sites: individual female legs, apparently never attached to torsos, that had been manufactured in pairs and later broken apart. The best clues to their function have come from other cultures in which similar tokens are known to have served to seal contracts and provide special evidence of the identity of the bearer (in feudal China, for example, a valuable piece of

jade would be broken in two to mark a contract between a master and his vassal, with each keeping one piece so they could later be matched). Here the known function in a source domain has a causal connection to the form of relevant artifacts, and the ethnographer makes the analogical inference that a similar cause may have operated in the target domain (see Bartha, 2010).

Causal Models and Category-Based Inferences

The basic idea that causal knowledge underlies inductive inference has been formalized in terms of graphical representations termed *causal models*. A causal model is a representation of cause–effect relations, expressing such information as 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. Graphical representations of causal structures were first introduced in philosophy (Reichenbach, 1956; Salmon, 1984). In artificial intelligence, Pearl (1988) developed a detailed graphical formalism termed *causal Bayes nets* (also see Spirtes, Glymour, & Scheines, 2000). Inspired by Pearl's work, Waldmann and Holyoak (1992; Waldmann, Holyoak, & Fratianne, 1995) introduced graphical causal models as a psychological account of human causal learning. Cheng (1997) proposed the *power theory of the probabilistic contrast model*, or *power PC theory*, which provides a quantitative account of (a) how the strengths of individual links within a causal model can be estimated from noncausal contingency data and (b) how multiple causal links are combined to make inferences. Griffiths and Tenenbaum (2005) integrated causal graphs with Bayesian inference, thereby providing an explicit account of how uncertainty about causal knowledge can be represented. More recently, Lu, Yuille, Liljeholm, Cheng, and Holyoak (2008) formalized and tested multiple Bayesian models of causal learning and inference with simple graph structures, including Bayesian variants of the power PC theory.

Empirical evidence indicates that both initial formation of categories (Kemp, Goodman, & Tenenbaum, 2007; Lien & Cheng, 2000; Waldmann et al., 1995) and inferences based on learned categories (Ahn, 1999; Heit, 1998; Rehder, 2006, 2007, 2009; Rehder & Burnett, 2005; Rehder & Kim, 2006; Tenenbaum, Kemp, & Shafto, 2007) are guided by causal models (for a recent review see Lagnado, Waldmann, Hagmayer, & Sloman, 2007). For example, Rehder (2006) taught participants the causal relations that influenced the distribution of features associated with novel categories and showed that inferences about properties of category members are then guided by these causal relations, which can override the influence of overall similarity. Rehder (2009) extended the power PC theory (Cheng, 1997) to develop a formal model of category-based inferences within the framework of causal-based generalization (CBG). Rehder applied his CBG model to predict various inductive inferences that depend on integrating knowledge about the distributions of category properties and causal relations that influence (or are influenced by) these properties. The model accounted for observed influences of the frequency distribution of category properties, the direction of causal relations (cause to effect vs. effect to cause), and quantitative variations in the power of a cause to produce its effect.

The Role of Causal Relations in Inductive Inference Across Varieties of Knowledge Representations

The present article aims to explain how causal knowledge can be used to guide inductive inference given a broad range of potential sources of knowledge. To place this goal in the context of related research, Figure 1 schematizes two major dimensions along which knowledge representations appear to vary and which in turn influence the role played by causal knowledge. The *x*-axis represents variation in degree of abstraction of the knowledge representations (specific to general), and the *y*-axis represents variation in what we term *relational richness* (low to high). Abstraction ranges from the low end, at which reasoning depends heavily on specific cases, to the high end, at which it depends primarily on generalizations. Relational richness ranges from the low end, at which the representations are primarily based on simple features or properties of objects, to the high end, at which representations include many complex relations potentially instantiated with dissimilar objects.

Each corner of the quadrant in Figure 1 is labeled with a "prototypical" psychological concept related to inductive inference. The lower left corner represents inferences based on feature-defined instances. Given that causal knowledge is inherently relational and hence goes beyond simple features, its role is minimal in this quadrant, in which inferences are primarily based on featural similarity. The lower right corner corresponds to relatively simple categories based on distributions of properties over their instances, potentially accompanied by causal generalizations. This is the type of knowledge typically used in paradigms associated with category-based induction.

The top left corner focuses on relationally complex instances, the main focus of work on analogical reasoning (Gentner, 1983). Insofar as these relations include some that are causal, causal knowledge will often play an important role in analogical inference. In the absence of a well-established schema, even a single example can function as a source that supports novel inferences about a less-well-understood target situation (Gentner & Holyoak, 1997; Holyoak, 2005), especially if it is coupled with some degree of abstraction (Ahn, Brewer, & Mooney, 1992; Kemp & Jern, 2009; Mitchell, Keller, & Kedar-Cabelli, 1986). Of course, analogies may be formally complex without necessarily involving causal knowledge about the world (e.g., letter-string analogies of

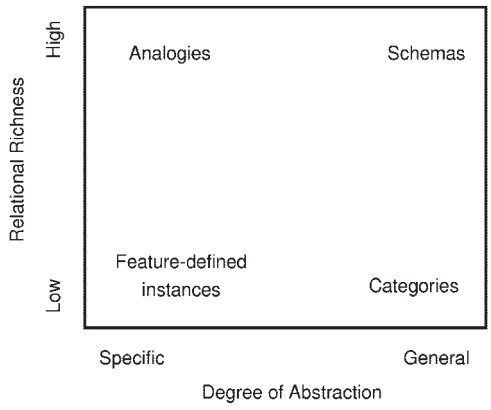

Figure 1. Schematic relationships among types of knowledge representations related to causal inference.

the sort studied by Hofstadter & Mitchell, 1994); however, in the present article we focus on inferences about empirical phenomena, for which causal knowledge is central.[1] Finally, the upper right corner focuses on abstract relational categories, often referred to as *schemas*. Comparison of multiple analogs can lead to the development of a schema representing a category of situations (Gentner, Lowenstein, & Thompson, 2003; Gick & Holyoak, 1983), and the computational processes involved in deriving inferences based on individual analogs and on generalized schemas appear to be similar (Hummel & Holyoak, 2003). Scientific hypotheses based on causal relations are especially likely to take a form consistent with the schema corner of the quadrant.

We should emphasize that the distinctions between knowledge representations depicted in Figure 1 are best viewed as fuzzy continua rather than well-defined subtypes. In actual inductive reasoning, multiple sources of knowledge at varying levels of abstraction and relational richness may be used together. In the present article, we often refer to "analogical" reasoning when in fact we mean reasoning based on some mixture of a specific relational instance and more schematic causal generalizations.

Our aim in the present article is to show in some detail how causal knowledge, represented as causal models, can be integrated with these varied types of knowledge to yield inductive inferences. The theory proposed here has much in common with previous models specifically focused on the role of causal models in making category-based inferences, notably the CBG model of Rehder (2009). However, the present theory is more general than previous models of this sort in that it can be applied to situations involving high uncertainty about causal powers (including cases in which only one or even zero specific cases are available to guide inference) and to situations involving high relational richness. We illustrate in this article how the general theory can be applied not only to category-based inferences but also to analogical and schema-based inferences, which pose the challenge of integrating causal models with more complex relational reasoning.

The Role of Causal Knowledge in Analogical Transfer

Early work in psychology on transfer of problem solutions by analogy demonstrated the close connection between causal understanding and analogical inference (Brown, 1989; Gick & Holyoak, 1980, 1983; Holyoak, Junn, & Billman, 1984; see Holyoak, 1985). Analogical transfer has been viewed as a multistate process that includes retrieval of a source analog, identifying the correspondences between the two analogs, transferring information from source to target, and evaluating the resulting inferences (e.g., Falkenhainer, Forbus, & Gentner, 1989; Holyoak, Novick, & Melz, 1994). It has been argued that all of these subprocesses of analogy are influenced by pragmatic constraints that make use of causal information (Holyoak, 1985; Holyoak & Thagard, 1995: Spellman & Holyoak, 1996).

However, the causal-model framework, which has been highly influential in the area of category-based inference, had only recently been explicitly connected to analogical transfer. Our aim in the present article is to begin to integrate analogical inference with causal models, providing a more unified account of the influence of causal knowledge on the subprocesses involved in analogical transfer. Because analogy typically involves a small number of examples (often one) as well as high relational richness, current models of category-based induction are inapplicable to the types of reasoning involved in the upper left quadrant of Figure 1. Because extending the causal-model framework to this quadrant is a key aim of the present article, we focus much of our attention on the integration of causal models with analogical transfer.

Models of analogical transfer have generally treated inference as a direct product of *mapping*—the process of determining a set of systematic correspondences between elements of the source and target. The domain-general algorithm adopted by all extant models (e.g., Falkenhainer et al., 1989; Holyoak & Thagard, 1989; Hummel & Holyoak, 2003) has been termed *copy with substitution and generation*, or CWSG (Holyoak et al., 1994). If proposition P consisting of relation r and objects a and b, notated P: $r(a, b)$, exists in the source but does not map to any proposition in the target, then a corresponding proposition in the target may be inferred by substituting mapped elements. Specifically, given the correspondences $r \rightarrow r'$, $a \rightarrow a'$, and $b \rightarrow b'$, then the inferred target proposition would be P': $r'(a', b')$. If any source element in P lacks a corresponding element in the target, then such a corresponding element may be postulated.

In order to be psychologically realistic, CWSG clearly needs to be constrained (Markman, 1997). Otherwise, for any unmapped proposition P in any possible source analog, the corresponding proposition P' could be inferred for any target, regardless of the credibility of the overall mapping between the source and target (and even if none of the elements of P map to known target elements). Algorithmic models have all hypothesized that some measure of mapping quality determines the acceptability of inferences that might be postulated on the basis of CWSG. Different models have adopted various combinations of constraints postulated to govern mapping and hence to indirectly constrain inferences generated by CWSG.

The structure-mapping theory[2] of Gentner (1983; Falkenhainer et al., 1989) emphasizes structural constraints on mapping. In particular, mappings based on structurally defined "higher order"

[1] More generally, analogical inferences can apparently be guided by various types of functional dependencies in the source analog (e.g., logical and mathematical relations) that determine which aspects of the source need to be preserved in the target in order to justify analogical inferences (Bartha, 2010). In the present article we focus on causal relations because they appear to be the major type of functional dependency underlying everyday understanding of the world and are central to scientific explanations.

[2] All theories of analogical inference that assume explicit representations of relational structure are based in part on finding a mapping between the structures of the source and target. The structure-mapping theory provides one account of how representational structures may be mapped to one another. It should be noted that our aim of integrating theories of analogy with theories of causal inference leads to an unfortunate collision between two distinct though interrelated theoretical uses of the term *structure*. In the field of analogy, *structure* refers to the systematic pattern of relations that comprise a knowledge representation, where both causal and noncausal relations may contribute to structure. In the field of causal inference, *causal structure* refers to the existence of a causal link between two variables, in contrast to *causal strength*, a continuous variable indicating the power of a cause to produce or prevent its effect. (If no causal relation holds between two variables, then there is no causal structure to which a strength value can be assigned.) In the present article we necessarily use *structure* in both senses, particularly in Experiment 4, in which we demonstrate that a preventive causal relation can play a dual role in guiding both structure mapping and causal inference.

relations (those that take one or more propositions as their arguments) are preferred over mappings based on first-order relations (which take only objects as arguments) or on simpler attributes of individual objects. Relations based on the predicate *cause* are treated as important examples of a general preference for systematic mappings that include higher order relations. For example, in mapping the structure of the solar system onto that of an atom, a key higher order relation might involve a complex proposition stating that the cause of a planet revolving around the sun is the fact that the sun is more massive than the planet it attracts. Expressed in a predicate-calculus-style formalism, this fact might be represented as:

> *cause* {*and* [*more-massive* (sun, planet), *attracts* (sun, planet)], *revolve-around* (planet, sun)}.

From the perspective of structure-mapping theory, the impact of *cause* on mapping and inference is based solely on its structural role in the static representation of each analog. It has long been recognized that this extreme structural view is implausible (Holyoak, 1985). As Rips (1990) pointed out, in the example just given one could simply replace the predicate *cause* with the operator *and*, which also can fill the syntactically defined role of a higher order relation, yielding the following:

> *and* {*and* [*more-massive* (sun, planet), *attracts* (sun, planet)], *revolve-around* (planet, sun)}.

The syntactic form of the complex proposition is unchanged; but intuitively, without the semantics associated with the relation *cause*, the basis for a credible analogical inference about the behavior of atomic particles is greatly diminished. It should be acknowledged, however, that virtually all the empirical studies supporting the importance of causal relations (not only in analogical but also in category-based inference) have used only one type of relation—a generative cause. Thus an alternative interpretation of most empirical findings is that both category-based and analogical inferences are guided by syntactically defined higher order relations, of which *cause* is simply one example. However, a study by Lassaline (1996) provides an important exception, because it demonstrated that a causal relation in fact provides stronger support for inferences than does a noncausal relation of the same syntactic form.

Other evidence indicates that causal relations exert a dynamic influence on mapping and inference. Using complex stories, Spellman and Holyoak (1996) showed that when the source–target mapping was ambiguous by structural criteria, those relations causally relevant to the reasoner's goal determined the preferred mapping, as well as guiding inferences about the target. Such evidence suggests that causal relations are perceived as especially important and hence receive greater attention during analogical reasoning (Holyoak & Thagard, 1989; Hummel & Holyoak, 1997, 2003; Winston, 1980). But regardless of whether the influence of causal relations on analogical reasoning has been treated as solely structural or also attention-based, it is the case that all algorithmic models have assumed that causal relations guide analogical inference only by influencing the preferred mapping between source and target and hence the output of the CWSG algorithm. In other words, all models have predicted that the support for analogical inferences increases monotonically with some measure of mapping quality.

In contrast, the model we propose postulates that although both structure mapping and the CWSG algorithm are necessary for transferring relational structure (including a causal model) from source to target, they are not sufficient to provide a complete account of analogical transfer. Rather, a full account requires that these processes be integrated with an explanation of how a causal model can initially be acquired for the source domain and subsequently "run" to make inferences about the target domain.

Empirical Evidence That Analogical Inference Depends on More Than Mapping Alone

A series of experiments reported by Lee and Holyoak (2008) demonstrated how causal knowledge guides analogical inference and that analogical inference is not determined solely by quality of the overall mapping between source and target. Using a common-effect structure (Waldmann & Holyoak, 1992), Lee and Holyoak manipulated structural correspondences between the source and the target as well as the causal polarity (generative or preventive) of multiple causes present in the target. In Figure 2, Panels A, B, and C show examples of causal structures used in their experiments. In the source A, three causes (two generative, G_1 and G_2, and one preventive, P) are simultaneously present, and when the influences of these three causes are combined, the effect occurs. The target analog B shares all three causal factors with the source, whereas target C shares

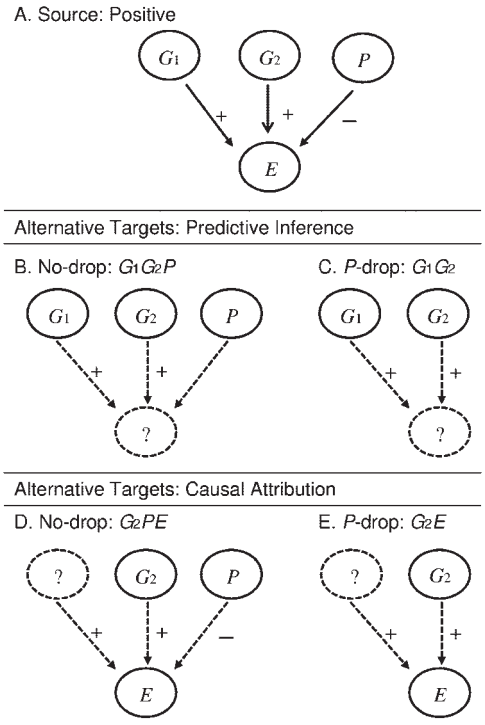

Figure 2. The use of causal models in analogical inference. *G*, *P*, and *E* represent generative causes, a preventive cause, and an effect, respectively; + and – indicate generative and preventive causes, respectively. Dotted elements are initially missing from the target and must be inferred on the basis of the source.

only the two generative factors with the source, not the preventive one. Accordingly, target B has greater semantic and structural overlap with the source than does target C. All previous computational models of analogy, which predict that the plausibility of target inferences increases monotonically with some measure of the quality of the overall mapping between the source and target analogs, therefore predict that target B is more likely to have the effect *E* than is target C.

If analogical inference is guided by causal models, however, the prediction reverses, because dropping a preventive cause, as in target C relative to target B, yields a causal model of the target in which the probability that the effect occurs will increase. Lee and Holyoak (2008) found that people in fact rated target C as more likely to exhibit the effect than target B, even though participants rated C as less similar than B to the source analog A. These findings (since replicated by Colhoun & Gentner, 2009, Study 1) suggest that understanding human use of analogy to make inferences requires deeper consideration of how causal knowledge is integrated with structural mapping.

Bayesian Theory of Inference Guided by Causal Models

In the present article we propose and test a formal model of the role of causal knowledge in making inductive inferences based on knowledge represented by categories, analogies, and/or schemas (see Figure 1). We describe a computational theory of inference that is tightly coupled with the framework provided by causal models (Cheng, 1997; Griffiths & Tenenbaum, 2005; Lu et al., 2008; Waldmann & Holyoak, 1992; Waldmann & Martignon, 1998). In order to definitively differentiate the meaning of causal relations from their syntactic form, we consider not only generative causes (those that make their effect happen) but also preventive causes (those that stop their effect from happening). Moreover, we consider both inferences from cause to effect (causal prediction) and reverse inferences from effect to cause (causal attribution; Kelley, 1973). Previous studies of category-based inferences have seldom considered inferences based on preventive causes; similarly, studies of causal attribution have never considered the impact of preventive causes. The model presented here predicts that when preventive causes are introduced, causal prediction and causal attribution will each yield a distinct inference pattern. In neither case is the predicted pattern a monotonic function of any measure of overall similarity or mapping quality.

The present theory is formalized for simple common-effect models in which one effect has multiple possible causes (Waldmann & Holyoak, 1992) and the factors are binary (each cause and its effect are present or else absent). These are the cases for which the Bayesian power PC theory of causal learning and inference (Cheng, 1997; Lu et al., 2008) has been most firmly established. Space precludes a full review of proposed models of causal judgments. By a recent count, over 40 algorithmic models of causal learning have been proposed in the literature (Hattori & Oaksford, 2007), almost all of which are nonnormative heuristics. Perales and Shanks (2007) compiled a meta-analysis of data from 114 conditions, taken from 17 experiments from 10 studies conducted in multiple labs, varying a variety of quantitative and qualitative parameters related to causal learning. Lu et al. (2008) showed that the parameter-free Bayesian power PC model provides the best quantitative fit of any model that has been applied to the data in this meta-analysis ($r = .96$). Critically for our present purposes, the Bayesian version of the power PC theory goes beyond the "classic" version (Cheng, 1997) in that it can account for the influence of sample size on estimates of causal strength and more generally is able to represent uncertainty about causal strengths. The Bayesian extension is essential for modeling analogical inference, which is often based on a single source example. In contrast, non-Bayesian models of category-based inference, even those based on the power PC theory (Rehder, 2009), are not applicable to situations in which causal powers are highly uncertain. We elaborate on this point in the General Discussion.

Causal Prediction and Causal Attribution

We focus on two general types of causal inferences in the target analog. A canonical predictive causal inference involves using a known cause to predict an effect (e.g., observing a fire being started, we may infer that smoke will be produced). However, people can also reason from effects to causes (e.g., observing smoke, we can infer that a fire may have occurred and caused the smoke). The latter type of inference is a causal attribution (closely related to abduction, diagnosis, and causal explanation). Causal attribution is more complex than predictive causal inference, as attribution requires considering possible combinations of alternative causes that may have been present, whereas prediction is based on one particular combination of observed causes (Bindra, Clarke, & Schultz, 1980; Fenker, Waldmann, & Holyoak, 2005). Causal attribution gives rise to *causal discounting*, whereby the presence of one generative cause reduces the estimated probability that some other generative cause was active (Kelley, 1973). For example, if you find wet grass in the morning, you might be tempted to suspect it rained overnight. But if you find that there was a sprinkler on, you might attribute the wet grass to the sprinkler and discount the probability that the wet grass was caused by rain (Pearl, 1988). Pearl (1988) showed that causal discounting is a normative consequence of reasoning with causal models (see also Novick & Cheng, 2004).

Previous work on causal attribution has not considered the impact of combining preventive and generative causes. It has been established that dropping a preventive cause from the target increases the strength of the predictive inference that the effect will occur (Lee & Holyoak, 2008). But intuitively, it seems that the impact of a preventive cause may reverse for causal attribution. If we again take the situation shown in Figure 2A as the source, then in target E attribution of factor G_1 as the cause of effect *E* will be discounted due to the presence of generative cause G_2. But in target D, the continued presence of preventive cause *P* seems to make it more likely that G_1 as well as G_2 played a causal role of producing *E* despite the countervailing influence of *P*. We will show that a mathematical extension of the Bayesian power PC theory to causal attribution in fact predicts that a preventive cause will decrease causal discounting and hence increase the strength of a causal attribution.

Overview of the Theoretical Framework

The two networks shown in Figure 3 schematize causal models for a source (left) and target (right) analog. The nodes represent variable causes (C) and effects (E). The superscripts (S, T) indicate the source and the target, respectively. The links represent the causal structure (only linked nodes have direct causal connections). The vectors $\boldsymbol{w}_i$ represent the causal polarity (generative or preventive) and the causal strength for links.

A key assumption is that inductive inference (including analogical transfer) uses causal knowledge of the source to develop a causal model of the target, which can in turn be used to derive a variety of inferences about the values of variables in the target. Unlike other formalisms that have been adopted by analogy models (e.g., predicate calculus), causal relations in Bayesian causal models can carry information about the existence of causal links (e.g., causal structure) and distributions of causal strength, as well as about the generating function by which multiple causes combine to influence effects. In the present theory, the first step in analogical inference is to learn a causal model of the source. The source model is then mapped to the initial (typically impoverished) representation of the target. Based on the mapping, the causal structure and strengths associated with the source are transferred to the target, creating or extending the causal model of the latter. The model of the target can then be "run," using causal reasoning to derive inferences about the values of endogenous variables in the target. Accordingly, as summarized in Figure 3, the four basic components in analogical inference are the learning of a causal model for a source (Step 1); the assessment of the analogical mapping between the source and a target (Step 2); the transfer of causal knowledge from the source to the target on the basis of the analogical mapping to construct the causal model of the target (Step 3); and inference based on the causal model of the target (Step 4).

Computational Theory

We now describe a model of transfer based on Bayesian inference, deriving predictions that can be qualitatively compared with the pattern of inference ratings obtained from human reasoners. The Bayesian model derives the probabilities of potential inferences about the target from the four computational components shown in Figure 3.

Figure 4 schematizes simple analogy problems of the sort used in the experiments reported later. Here, the source (see Panel A) has one background cause (B^{S}, assumed to be generative and constantly present in the context), two generative causes (G_1^{S} and G_2^{S}), one preventive cause (P_1^{S}), and an effect (E^{S}). The target has one background cause (B^{T}), two additional generative causes (G_1^{T} and G_2^{T}), and one preventive cause (P_1^{T}). All node variables are binary, with value 0 (absent) or 1 (present). In the predictive inference case (as shown in Figure 4B), the task is to predict the probability of the effect occurring in the target, whereas in the causal attribution case (as shown in Figure 4C), the task is to predict the probability that the cause G_1^{T} was present (value of 1) and produced the effect in the target.

A causal model includes both causal structure and strength. Causal structure is represented by directed arrows between nodes to indicate cause–effect relations, as shown in Figure 4. The causal strength associated with the link between each cause node and its effect node is denoted by $\boldsymbol{w}^{\mathrm{S}}$ for the source and $\boldsymbol{w}^{\mathrm{T}}$ for the target. Both $\boldsymbol{w}^{\mathrm{S}}$ and $\boldsymbol{w}^{\mathrm{T}}$ are vectors and convey two kinds of information: polarity of the causal power (generative or preventive) and absolute causal strength of the link. Polarity is coded as "+" when the cause is generative and "−" when the cause is preventive. Causal strength is represented as a random variable, in which higher values imply that the cause has higher power to generate an effect or else to prevent the effect from occurring.

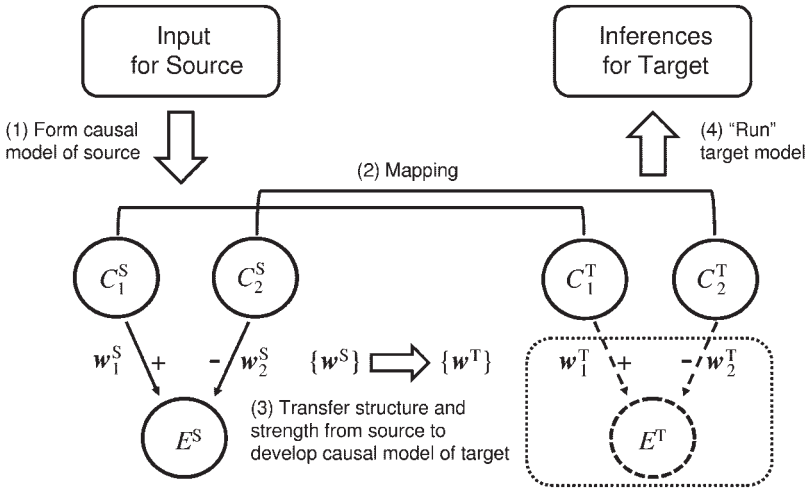

Figure 3. Framework for analogical transfer based on causal models. G, P, and E represent generative causes, a preventive cause, and an effect, respectively; w_1^{S}, w_2^{S}, w_1^{T}, and w_2^{T} each represent a distribution over causal strength for causal links in the source (S) and in the target (T), respectively. Dotted lines indicate knowledge transferred from source to target (see text).

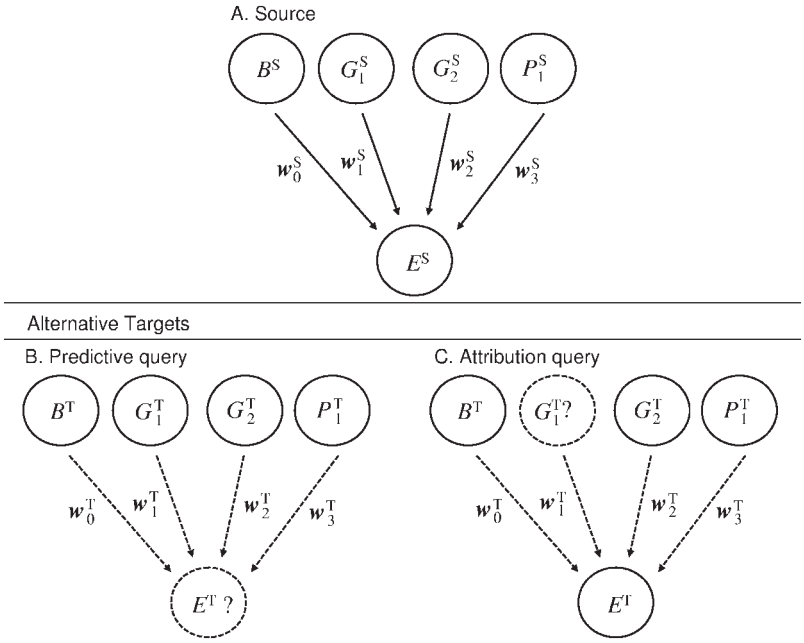

Figure 4. Causal graphs for a source (Panel A), a target requiring a predictive inference (Panel B), and a target requiring a causal attribution (Panel C). Variables *B*, *G*, *P*, and *E* represent a background cause, a generative cause, a preventive cause, and an effect, respectively. All nodes are binary variables. The vectors $\boldsymbol{w}^S$ and $\boldsymbol{w}^T$ represent the causal polarity and distribution over causal strength for causal links in the source (S) and in the target (T), respectively. Dotted elements are initially missing from the target and must be inferred on the basis of the source.

In Figure 4, the status of the effect in the target is to be inferred (predictive inference). For the first step, the model learns the causal structure of the source and estimates the strength of each causal factor. In the Bayesian framework, the inductive strength of an inference depends on two factors, priors and likelihoods of the data. Priors capture knowledge that people have about causal structure or causal strengths before they observe new data. Likelihoods of the data can be calculated by applying the power PC theory (Cheng, 1997) to assess how likely the observed data can be generated from a causal model. The probability of an effect occurring is given by a noisy-OR function when candidate causes are generative and by a noisy-AND-NOT function when a candidate cause is preventive.

In the present article we focus on situations in which it can be safely assumed that people do not have any prior preference about the values of causal strength in the source; hence, priors are assumed to be uniformly distributed causal strength over the range [0, 1]. Although alternative priors could be considered (Lu et al., 2008), uniform priors allow us to keep the model parameter-free.[3] The basic theory can readily be extended to situations in which the reasoner begins with specific priors about the source or target. We do not offer a theory of mapping in the present article, but the experiments we report involve situations in which the correspondences between the source and target elements are uncontroversial. In paradigms that involve category-based inference, if a new instance is stated to belong to Category A, it will presumably be transparent that it corresponds to previous instances of Category A. (The situation would be considerably more complex if there were uncertainty about the category membership of a new instance; see Murphy & Ross, 2010; Ross & Murphy, 1996.) Even with richer representations for which the correspondences must be established by structure mapping, the output will be determinate as long as the information provided about the source and target is sufficient to determine a specific "correct" mapping for individual causal variables. Several formal models of structure mapping are able to predict human mapping judgments for a broad range of inputs. These include the Structure Mapping Engine (SME; Falkenhainer et al., 1989), the Analogical Constraint Mapping Engine (ACME; Holyoak & Thagard, 1989), and Learning and Inference with Schemas and Analogies (LISA; Hummel & Holyoak, 1997, 2003). In the studies reported here, we took the tack of constructing experimental materials for which all three of these mapping models would yield the same set of (highly intuitive) correspondences. However, our theory is in fact sufficiently general to derive predictions even when alternative mappings are possible, as we illustrate in Experiments 3 and 4.

Derivation of predictive inferences. Figure 4 illustrates the computation for predictive inference. In Equation 1, which follows, $\boldsymbol{C}^S$ denotes the information that the source (see Panel A in Figure 4) has a background generative cause, B^S, and three addi-

[3] Lu et al. (2008) presented evidence favoring a variant of the Bayesian power PC theory that includes a generic prior indicating that causes will be relatively few in number and individually of high strength ("sparse and strong"). For analogical inferences in the experimental designs we consider, adding such a prior would not alter the qualitative predictions of the Bayesian theory; hence, we assumed uniform priors to obviate the need to introduce a free parameter.

tional causal factors, G_1^S, G_2^S, and P_1^S, that is, $\boldsymbol{C}^S = (B^S, G_1^S, G_2^S, P_1^S)$. (Vectors are indicated by bold font.) $\boldsymbol{C}^T$ provides analogous information about possible causes in the target (see Panel B). In predictive inference, the model estimates the probability of an effect occurring in the target, $E^T = 1$, on the basis of initial information about the source ($\boldsymbol{C}^S$, E^S) and the target ($\boldsymbol{C}^T$). The unknown causal strength of the target is represented by $\boldsymbol{w}^T$. The basic equation for predictive inference is

$$\begin{aligned} P(E^T \mid \boldsymbol{C}^T, E^S, \boldsymbol{C}^S) &= \sum_{\boldsymbol{w}^T} P(E^T, \boldsymbol{w}^T \mid \boldsymbol{C}^T, E^S, \boldsymbol{C}^S) \\ &= \sum_{\boldsymbol{w}^T} P(E^T \mid \boldsymbol{w}^T, \boldsymbol{C}^T) \\ &\quad \times \sum_{\boldsymbol{w}^S} [P(\boldsymbol{w}^T \mid \boldsymbol{w}^S, E^S, \boldsymbol{C}^S, \boldsymbol{C}^T) \\ &\quad \times P(\boldsymbol{w}^S \mid \boldsymbol{C}^S, E^S)], \end{aligned} \quad (1)$$

where the rightmost term on the right side of the equation, $P(\boldsymbol{w}^S \mid \boldsymbol{C}^S, E^S)$, captures the learning of a source model from observed contingency data (see Step 1 in Figure 3). Recent computational studies have developed detailed models that estimate probability distributions of causal strength by combining priors and observations (Griffiths & Tenenbaum, 2005; Lu et al., 2008).

The middle term, $P(\boldsymbol{w}^T \mid \boldsymbol{w}^S, E^S, \boldsymbol{C}^S, \boldsymbol{C}^T)$, quantifies knowledge transfer based upon analogical mapping (see Steps 2 and 3 in Figure 3). The correspondence between variables in the source and target is denoted by an assignment matrix $\mathbf{M}$ in which element M_{ij} equals 1 if the ith variable in the source maps to the jth variable in the target and $M_{ij} = 0$ otherwise. These assignment variables specify the transfer of causal structure and strength as

$$P(\boldsymbol{w}^T \mid \boldsymbol{w}^S, E^S, \boldsymbol{C}^S, \boldsymbol{C}^T) = \sum_{\mathbf{M}} P(\boldsymbol{w}^T \mid \boldsymbol{w}^S, \mathbf{M}) P(\mathbf{M} \mid E^S, \boldsymbol{C}^S, \boldsymbol{C}^T). \quad (2)$$

The first term on the right-hand side of Equation 2, $P(\boldsymbol{w}^T \mid \boldsymbol{w}^S, \mathbf{M})$, determines to what extent causal knowledge will be transferred if particular source variables match particular target variables. If an assignment variable is 1, then causal strength in the target is assumed to follow a Gaussian distribution centered at the causal strength of the matched variable in the source, with variance σ. This term can therefore be expressed as

$$P(\boldsymbol{w}^T \mid \boldsymbol{w}^S, \mathbf{M}) = \frac{\exp\left(-\sum_{i,j} M_{ij}(w_j^T - w_i^S)^2 \Big/ 2\sigma^2\right)}{\left(\sqrt{2\pi}\sigma\right)^{\sum_{i,j} M_{ij}}}. \quad (3)$$

The second term on the right-hand side of Equation 2, $P(\mathbf{M} \mid E^S, \boldsymbol{C}^S, \boldsymbol{C}^T)$, assesses the probability of possible mappings (i.e., how well variables in the source match to variables in the target). Note that the general expression in Equation 2 does not require that there be a perfect structural match between the source and target or that the match of variables between the source and target be established with probability equal to 1. Equation 2 simply specifies that the transfer of causal knowledge (including the causal link and its strength) is weighted by the probability that source variables play the same causal role as do target variables. People may well leave some parts of the source and target unmapped, especially if they are not considered causally relevant (Holyoak, 1985).

One special case concerns applying a deterministic mapping rule (such as "copy with substitution and generation"), which implies that the variance specified in Equation 2 is zero. Accordingly, in this special case the Dirac delta distribution was employed to model the probability of transfer as

$$\begin{cases} P(w_j^T = w_i^S) = 1, & \text{if the } j\text{th target variable} \\ P(w_j^T = w_i^S) = 0, & \text{matches the } i\text{th source variable and} \\ & \text{otherwise.} \end{cases} \quad (4)$$

We thus treat the matching of cause variables as a binary decision. When matching fails, Equation 4 implies that the causal model for the target will be left unchanged (i.e., neither causal structure nor causal strengths will be transferred from the source, which is simply viewed as irrelevant).

The leftmost term on the right side of Equation 1, $P(E^T \mid \boldsymbol{w}^T, \boldsymbol{C}^T)$, uses knowledge from analogical transfer and observations about the presence of causal factors in the target to estimate the probability of the effect in the target (see Step 4 in Figure 3). This probability can be directly computed with the Bayesian extension of the power PC theory (Cheng, 1997; Griffiths & Tenenbaum, 2005; Lu et al., 2008). Additional mathematical details are presented in Appendix A.

Figure 5 summarizes the conceptual decomposition of the computation in Equation 1. In summary, Equation 1 implies that the probability of an effect is computed by considering all possible values of causal strengths and structures in the target (not just a single point value, such as causal power), weighted by the probability of these values, as determined by the causal knowledge transferred from the source. The model specified by Equation 1 is thus sensitive to transferred information about causal strength, causal structure, and the uncertainty associated with this information.

Derivation of causal attributions. Several distinct queries involving causal attribution (Cheng & Novick, 2005, pp. 700–701) or diagnosis (Waldmann, Cheng, Hagmayer, & Blaisdell, 2008, pp. 464–466) can be distinguished. In the present experiments we focus on an attribution query of this form: "Given that certain causal factors are known to have occurred, and that the effect E has occurred, what is the probability that cause C (not known to have occurred) in fact occurred and produced E?" In general, such a causal attribution question requires apportioning the observed probability of an effect, $P(E^+)$, among causes of E. On the basis of

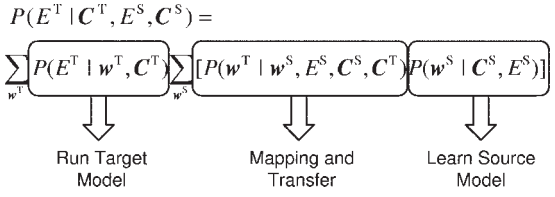

Figure 5. Major conceptual components of the computational model of predictive inference (Equation 1). $\boldsymbol{C}^S$ denotes causal information in the source; $\boldsymbol{C}^T$ provides analogous information about possible causes in the target. E^S and E^T denote the presence of the effect in the source and the target, respectively; $\boldsymbol{w}^S$ and $\boldsymbol{w}^T$ represent the unknown causal strengths associated with the source and the target, respectively.

the assumptions of the power PC theory, Cheng and Novick (2005, Equation 3, p. 700) derived the predicted probability that C is the cause of E when E occurs, namely,

$$P(C^+ \rightarrow E^+ \mid E^+) = P(C^+)q_C/P(E^+), \quad (5)$$

where $C^+ \rightarrow E^+$ denotes "C is the cause of E's occurrence" (corresponding to an unobservable state in a causal model; Cheng, 1997) and q_C denotes the generative power of C (see Lu et al., 2008, p. 978). Equation 5 yields a point estimate of causal attribution judgments. As in the case of predictive causal inference, the Bayesian extension of the power PC model derives a probability distribution for causal attribution. Here we lay out the general framework used in the present article. In Appendix B we derive predictions for two cases of causal attribution: (1) when multiple causes may be present, all generative, and (2) when a preventive cause is also present. (See Meder, Mayrhofer, & Waldmann, 2009, for a Bayesian analysis of diagnostic inference with a single generative cause.)

As shown in Figure 4, the input of the model includes the initial information in the source (see Panel A) and the target (see Panel C), ($\boldsymbol{C}^S$, E^S, $\boldsymbol{C}^T$, E^T), in which $\boldsymbol{C}^T$ denotes the known causal factors, that is, $\boldsymbol{C}^T = (B^T, G_2^T, P_1^T)$ and does not include the unknown causal factor G_1^T. The goal of the model is to predict the probability that the cause G_1^T was present and produced the effect in the target.

$$P(G_1^T = 1, G_1^T \rightarrow E^T \mid E^T, \boldsymbol{C}^T, E^S, \boldsymbol{C}^S) = \frac{P(G_1^T = 1, G_1^T \rightarrow E^T, E^T \mid \boldsymbol{C}^T, E^S, \boldsymbol{C}^S)}{P(E^T \mid \boldsymbol{C}^T, E^S, \boldsymbol{C}^S)} = \frac{P(G_1^T = 1 \mid \boldsymbol{C}^T, E^S, \boldsymbol{C}^S)P(G_1^T \rightarrow E^T, E^T \mid G_1^T = 1, \boldsymbol{C}^T, E^S, \boldsymbol{C}^S)}{P(E^T \mid \boldsymbol{C}^T, E^S, \boldsymbol{C}^S)}. \quad (6)$$

The first term in the numerator, $P(G_1^T = 1 \mid \boldsymbol{C}^T, E^S, \boldsymbol{C}^S)$, is the base rate of the cause in the target. This base rate can be estimated on the basis of the standard counting probability, using the binomial distribution. The basic computation can be described in relation to the following situation. Suppose there are two bags of marbles, and assume the probability of a marble being red is equal for the two bags. If four marbles (with replacement) are chosen from the first bag and all four are red, and then similarly three marbles are chosen from the second bag and all three are red, then one can estimate the probability of getting a red marble in the fourth draw from the second bag. Appendix B elaborates how these probabilities are estimated in different situations. The estimated base rate of the cause is determined by the number of causal factors observed in the source and the target. The qualitative result is that, after observing four causal factors to occur in the source, the probability of G_1^T occurring increases with the number of other causal factors observed to occur in the target.

The second term in the numerator, $P(G_1^T \rightarrow E^T, E^T \mid G_1^T = \boldsymbol{C}^T, E^S, \boldsymbol{C}^S)$, serves to quantify the probability of a predictive inference (i.e., how likely the effect in the target can be produced by G_1^T given the information in the source). The principle in computing this probability is the same as described in the previous section. The denominator, $P(E^T \mid \boldsymbol{C}^T, E^S, \boldsymbol{C}^S)$, is calculated by the weighted sum of the probability of the effect occurring in the presence and the absence of the cause G_1^T. The weights are determined by the estimate of the base rate of this cause.

$$P(E^T \mid \boldsymbol{C}^T, E^S, \boldsymbol{C}^S) = P(E^T \mid G_1^T = 1, \boldsymbol{C}^T, E^S, \boldsymbol{C}^S) \times P(G_1^T = 1 \mid \boldsymbol{C}^T, E^S, \boldsymbol{C}^S) + P(E^T \mid G_1^T = 0, \boldsymbol{C}^T, E^S, \boldsymbol{C}^S) \times P(G_1^T = 0 \mid \boldsymbol{C}^T, E^S, \boldsymbol{C}^S). \quad (7)$$

The mathematical derivation (see Appendix B) makes it clear that causal attribution judgments are more complex than causal predictions are (because attributions require summing over all possible combinations of unknown causal factors). In deriving predictions for the experimental conditions of concern in the present article, the Bayesian theory provides analytic solutions for both causal prediction and causal attribution.

Experiment 1

In Experiment 1 we investigated the impact of causal structure and strength in the source on predictive inferences about the target. The source consisted of an instance from the same category as the target accompanied by general information about relevant causal structure. The materials thus lie toward the low end of the continuum of relational richness (see Figure 1); the paradigm is best described as involving category-based inference coupled with a single source analog. The causal structures were similar to those used by Lee and Holyoak (2008).

The goal of Experiment 1 was to test the basic hypothesis that analogical inference is controlled in part by what has been learned about the source. Accordingly, we varied whether the effect E did or did not occur in the source analog. As we will show, the Bayesian model predicts that whether the effect occurred given the same set of causes in the source will lead to different estimations of causal strength distribution for each of the causal links. If E occurs, the distribution of the generative causal strength will be biased toward relatively strong causal power; whereas if E does not occur, then their strength distributions will be biased toward weak generative power relative to the preventive cause. In turn, these different estimates of causal strength for each causal link will influence inferences about a new target instance.

Method

Participants. Forty undergraduate students at the University of California, Los Angeles (UCLA), participated in the experiment for course credit. Twenty participants were randomly assigned to each of two conditions.

Design. A 2 × 3 mixed design was employed. The first independent variable was source outcome, positive or negative. In the positive condition, the effect was said to occur in the source, whereas in the negative condition the effect was said not to occur. Regardless of the source outcome, the source always included three causal relations (two generative causes and one preventive cause), as shown in Figure 2A. The type of source outcome was a between-subjects factor.

The second independent variable (a within-subjects factor) was argument type, defined by the presence or absence of various causes in the stated target analog. There were three possibilities: The three causes could be stated to all be present in the target (no-drop condition: G_1G_2P), the preventive cause could be stated to be absent (P-drop: G_1G_2), and a generative cause could be stated to be absent (G-drop: G_2P).

Materials and procedures. Participants read descriptions of pairs of fictional animals. Two sets of materials were employed, each using a different animal name. For our examples, we will refer to animals called "trovids." This fictional animal was described as having an abnormal characteristic (dry flaky skin) and three different gene mutations (Mutations A, B, and C). The mutations were described as tending to either produce or prevent the abnormal characteristic. It was stated that each of these gene mutations occurred randomly for unknown reasons, so any individual might have 0, 1, 2, or 3 distinct mutations. A source analog was simply referred to as "Trovid #1," and a target analog was referred to as "Trovid #2." The source analog always had three causal properties (i.e., three mutations) that were causally connected to the effect property (i.e., the abnormal characteristic). As in the Lee and Holyoak (2008) study, the phrase "tends to" was included for each causal statement to emphasize that causes might be probabilistic (Cheng, 1997). Depending on the source outcome type, the source exhibited the effect property or not. An example of the positive condition is the following:

> For Trovid #1, it happens that all three mutations have occurred.
>
> For Trovid #1, Mutation A tends to PRODUCE dry flaky skin; Mutation B tends to PRODUCE dry flaky skin; Mutation C tends to PREVENT dry flaky skin.
>
> Trovid #1 has dry flaky skin.

For the negative condition, in the last statement, *has* was simply replaced with *does NOT have*.

After reading the description of the source analog, participants were given three different judgment tasks, one for each argument type. Before making judgments, they were informed that the questions were not related to each other and that they should think about each question separately. The presence of the effect property was unknown, and the presence or absence of each of the three mutations was listed. Each target analog had two or three mutations, depending on the argument type (no-drop: G_1G_2P; *P*-drop: G_1G_2; and *G*-drop: G_2P). For example, in the no-drop (G_1G_2P) condition, all three mutations were present in the target. When a causal factor was dropped, that mutation was explicitly described as absent. In making judgments, participants were asked to suppose there were 100 animals "just like" the target animal described and to estimate how many of these 100 would have the effect property, choosing a number between 0 and 100. This procedure for eliciting causal strength ratings was introduced by Buehner, Cheng, and Clifford (2003).

Two different sets of materials were constructed (the other set based on fictitious animals called "falderols"); each participant received both sets and thus provided two judgment scores for each argument type. The order of these two sets was counterbalanced, and within each set the order of argument type was randomized for each participant.

Results and Discussion

Mean causal ratings for the two predictive conditions are shown in Figures 6A and 6B. A 2 × 3 mixed-design analysis of variance (ANOVA) was performed, in which source outcome (positive vs. negative) was a between-subjects variable and argument type (no-drop vs. *G*-drop vs. *P*-drop) was a within-subjects variable. A significant interaction between source outcome and argument type was obtained, $F(2, 76) = 16.12$, $MSE = 473.62$, $p < .001$, confirming that source outcome (whether the effect occurred in the source) influenced analogical inference in the target.

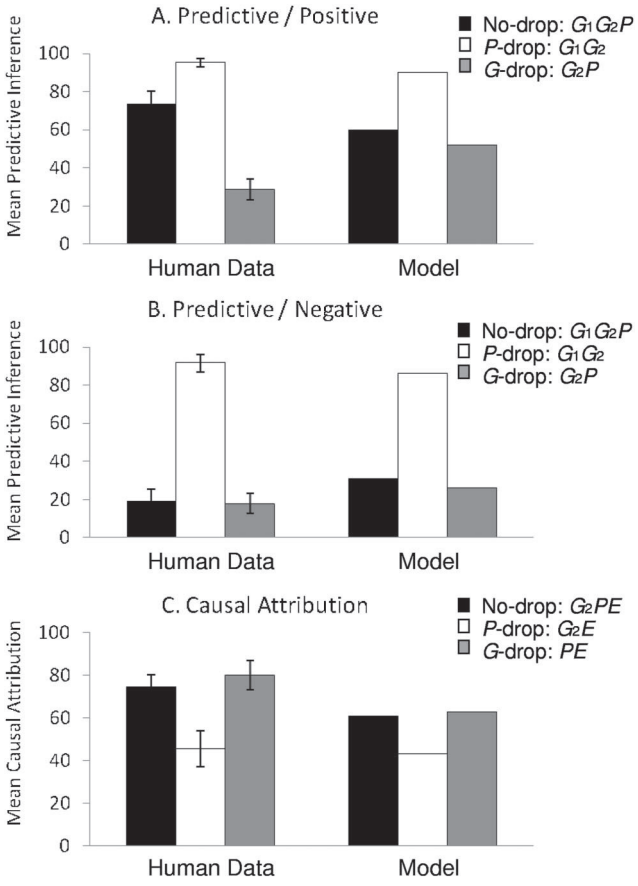

Figure 6. Mean predictive inference ratings (Experiment 1) when source outcome was positive (Panel A) or negative (Panel B) and mean causal attribution ratings (Experiment 2) for each argument type (Panel C). *G*, *P*, and *E* represent generative causes, a preventive cause, and an effect, respectively. Error bars represent 1 standard error of the mean. Predictions derived from the Bayesian model are shown in the right panel of each graph.

To examine how predictive ratings for each argument type were affected by source outcome, a separate within-subjects ANOVA was performed for each source-outcome condition (for data see Panels A and B of Figure 6). In the source-positive condition (see Figure 6A), the mean inference ratings for G_1G_2P, G_1G_2, and G_2P argument types were 73.6, 95.2, and 28.5, respectively. An ANOVA revealed a significant effect of argument type, $F(2, 38) = 43.61$, $MSE = 530.64$, $p < .001$. The argument G_1G_2 was rated as having higher inference strength than either the argument G_1G_2P, $t(19) = 3.41$, $p = .003$, or the argument G_2P, $t(19) = 10.20$, $p < .001$. The argument G_1G_2P was also rated as having higher inference strength than the argument G_2P, $t(19) = 5.16$, $p < .001$.

These results replicate the previous findings of Lee and Holyoak (2008), in that dropping a preventive cause from the target increased inductive strength whereas dropping a generative cause decreased inductive strength. When people made predictive inferences in the source-negative condition (see Figure 6B), the mean

ratings for G_1G_2P, G_1G_2, and G_2P argument types were 19.2, 92.1, and 17.8, respectively. An ANOVA revealed a significant mean difference between the argument types, $F(2, 38) = 86.74$, $MSE = 416.61$, $p < .001$. The argument G_1G_2 was rated as having significantly higher inference strength than either the argument G_1G_2P, $t(19) = 9.91$, $p < .001$, or the argument G_2P, $t(19) = 9.53$, $p < .001$; however, the G_1G_2P and G_2P types did not differ, $t(19) = 0.47$, $p = .65$.

The main difference in the pattern of ratings was that in the negative condition, participants appeared to estimate the strength of the preventive cause to be greater than the strengths of the generative causes, so that if P was present the effect was not expected to occur, regardless of whether one or both generative causes were present. The differences between the analogical inferences resulting from the two source-outcome conditions thus demonstrate that analogical transfer is sensitive to causal strength as well as structure (cf. Lu et al., 2008). Estimated probability distribution of causal strength appeared to differ depending on whether the effect occurred in the source, and these estimated strength distributions in turn were used in forming and running a causal model for a target. We report the fit of our Bayesian model to these data after presenting Experiment 2, which examined judgments of causal attribution.

Experiment 2

People can make inferences either from a cause to its effect (prediction) or from an effect to its cause (attribution). In Experiment 2, we investigated how people form and run a causal model when making a causal attribution. Previous work on causal attribution has considered the impact of multiple generative causes. The aim of Experiment 2 was to examine the impact of combining preventive with generative causes in causal attribution. The presence of a preventive cause (in contrast to a generative cause) is predicted to decrease causal discounting and hence increase the strength of a causal attribution.

Method

Twenty UCLA undergraduate students participated in the experiment for course credit. The materials and procedure were similar to those of Experiment 1. The source analog always exhibited the effect (as in the source-positive condition of Experiment 1). In the target analog, the presence of one of the mutations was described as unknown, and the presence or absence of each of the other mutations was explicitly stated. The target analog always had the effect property and one or two mutations, depending on the argument type (no-drop: G_2PE; P-drop: G_2E; and G-drop: PE). For example, in the no-drop (G_2PE) condition, two mutations and the effect property were present in the target. In making judgments, participants were to suppose there were 100 animals just like the target animal and to estimate in how many the unknown mutation (G_1) had occurred and produced the effect property, assigning a number between 0 and 100.

Results and Discussion

A different pattern was obtained in the causal attribution task (see Figure 6C) compared with predictive inference (Experiment 1; see Figures 6A and 6B). Mean causal attribution ratings for G_2PE, G_2E, and PE argument types were 74.4, 45.4, and 80.1, respectively. These means were significantly different, $F(2, 38) = 5.61$, $MSE = 1{,}237.55$, $p = .007$. The argument G_2E was rated lower than either G_2PE, $t(19) = 2.49$, $p = .02$, or PE, $t(19) = 2.55$, $p = .02$. The mean difference between the latter two argument types was not reliable, $t(19) = 0.81$, $p = .43$. In sharp contrast to the pattern observed for the corresponding predictive inference (i.e., predictive/positive condition; see Figure 6A), dropping a preventive cause decreased the rated probability of an inference about a potential generative cause. In accord with causal discounting, in the G_2E condition, because the target lacks the preventive cause and is known to have a generative cause, an additional generative cause is not as likely.

All extant models of analogical mapping would predict that mapping quality will be higher for the G_2PE condition (three factors matching between source and target) than for either the G_2E or PE conditions (two factors matching). Thus, the observed rank order of the three conditions ($PE \approx G_2PE > G_2E$) demonstrates that when a known preventive cause is involved, causal attribution yields a dissociation between mapping quality and support for an analogical inference. Moreover, this new dissociation follows a different pattern from that observed for the previous dissociation we found in the case of causal prediction.

Comparison of Human Performance to Model Predictions

Fits to Data From Experiments 1 and 2

We tested our Bayesian model of analogical inference by comparing its predictions with human performance on judgments of both causal prediction (Experiment 1) and causal attribution (Experiment 2). In Figure 6, the right side of each graph shows the model's predictions for each argument type. To allow a direct comparison with the human data, we converted the theoretical predictions for estimated probability to a 0–100 scale. Our focus is on the qualitative performance of the model. The fit of the parameter-free Bayesian model to human data across all conditions (both predictive and attribution judgments) was quite good, $r(7) = .93$, $p < .001$. Qualitatively, the model captures the fact that for predictive inferences, dropping a preventive cause from the target increases inductive strength (i.e., G_1G_2 yielded higher ratings than did G_1G_2P) and that the preventive cause is viewed as stronger (hence decreasing the probability of E more) in the source-negative than in the source-positive condition. For causal attribution, the model accounts for the fact that dropping a preventive cause increases causal discounting and hence reduces the estimated probability of the unknown cause (i.e., lower ratings for the G_2E condition than for the two conditions in which P is present). The Bayesian model thus accounts for both of the dissociations between analogical inference and overall mapping quality that we have demonstrated in Experiment 1 (causal prediction) and Experiment 2 (causal attribution).

Fits to Data From Colhoun and Gentner (2009)

Colhoun and Gentner (2009, Study 2) reported an experiment that investigated how the presence or absence of causal relations in

either the source or target analog affects the analogical transfer of causal relations. Their study also examined the impact of leaving the status of the effect in the source analog unknown, providing another source of empirical evidence that can be used to assess the predictions of our Bayesian model. Using materials similar to those employed by Lee and Holyoak (2008), Colhoun and Gentner compared four different conditions, which differed in whether causal relations were stated in the source or in the target. Participants read a description of a pair of imaginary animals representing a source and target analog, with each animal referred to simply by a letter (e.g., animal R). Simple blank properties (e.g., "blocked oil glands") were used as causal factors and the effects. Depending on the condition, the source animal might include two causal factors (one generative cause and one preventive cause), and/or causal relations might be stated (i.e., that each causal factor tended to cause or prevent the effect), and/or the effect might be stated as present. In Condition 1 (no source), people were not given a source analog, but causal relations were stated in the target. In Condition 2 (noninformative source), a source was provided with the causal factors only (no information about presence of effect), and causal relations were stated in the target as in Condition 1. In Condition 3 (schematic source), the source stated both the presence of causal factors and also the causal relations; but like in Condition 2, no information was provided about the status of the effect (i.e., the source did not constitute a fully specified instance). Finally, in Condition 4 (specific source), the source stated the presence of causal factors and the causal relations and in addition explicitly stated that the effect was present.[4]

For each of the four conditions, three different argument types were tested: *GGP*, *GP*, and *GPP*. In the argument type *GP*, two causal factors (*G* and *P*) were shared between the source and target. In the argument types *GGP* and *GPP*, one additional cause (generative in *GGP*, preventive in *GPP*) that had not appeared in the source was added to the target (in addition to the two causal factors shared by the source and target). Each participant received all of these three argument types for one of the four conditions. For each argument type, participants judged the likelihood that the effect would occur in the target animal by choosing a number between 0 and 100.

The mean ratings obtained by Colhoun and Gentner (2009) are shown in Figure 7, along with the predictions derived from the Bayesian model, on the basis of the same assumptions as used to fit data from the present Experiments 1 and 2. The human rating data can be interpreted at most as an interval scale. We did not adjust the model's predictions to match the absolute magnitudes of the human rating data, because we were only concerned with fitting the qualitative pattern of the data; hence the model's predictions are parameter-free (as was the case for the fits of data from Experiments 1 and 2). The human data reveal that people's judgments about predictive causal inferences were not influenced by whether causal relations were explicitly stated in the source or in the target (i.e., the pattern of judgments was statistically identical across Conditions 1, 2, and 3).

In modeling the findings of Colhoun and Gentner (2009), Conditions 1 and 2 (in which the causal structure is directly specified for the target) do not involve transfer of causal knowledge from the source; consequently, our model reduces to a standard Bayesian model of causal learning for the target situation (Lu et al., 2008), using the noisy-OR and noisy-AND-NOT integration rules for likelihood and a uniform prior on causal strength in the target. In Condition 3, the schematic causal structure is specified for the source; however, in the absence of information about the effect, it is not possible to update strength distributions. Given the assumption of perfect transfer from source to target (see Equation 4), the same causal model will be constructed for the target in Condition 3 as that specified directly in Conditions 1 and 2. Accordingly, the model makes identical predictions for Conditions 1–3.

Condition 4 (specific source), on the other hand, provides information in the source about the presence of causal factors and also the effect, thereby providing a specific instance that can be used to update strength distributions prior to transfer of causal knowledge to the target. Because the effect is stated to be present, the model will revise the expected strength values upward for the generative cause in the source and downward for the preventive cause. Accordingly, the effect will be more likely to be inferred in the target for all argument types, as was indeed observed in Colhoun and Gentner's (2009) experiment. In addition to making accurate ordinal predictions, the general fit of the model to human data across all 12 conditions (4 conditions $\times$ 3 argument types) was good, $r(10) = .92$, $p < .001$.

Experiment 3

Lee and Holyoak (2008, Experiment 3) demonstrated that causal models also guide predictive inferences on the basis of a cross-domain analogy in which each causal factor is itself a relation (materials that approximate the upper right quadrant in Figure 1). Experiment 3 was designed to investigate causal attribution judgments in the context of a similar cross-domain analogy, in which causal relations were structurally defined as higher-order rather than first-order relations. Attributional inferences (which require reasoning backward from an observed effect to alternative unobserved possible causes) are known to be more difficult than predictive causal judgments (Bindra et al., 1980; Fenker et al., 2005; see Appendix B); furthermore, cross-domain analogies are more difficult to process than within-domain analogies (e.g., Holyoak & Koh, 1987; Keane, 1988; see Hummel & Holyoak, 1997). Accordingly, it is possible that combining these two sources of difficulty in reasoning will interfere with people's ability to draw causal attributions.

Method

Participants. Thirty-two UCLA undergraduate students participated in the experiment for course credit.

Design, materials, and procedure. The design of Experiment 3—a within-subjects design with three argument types as conditions—was identical to that of Experiment 2. The argument types differed in whether the target analog included all three factors present in the source (no-drop: G_2PE), lacked a preventive cause (*P*-drop: G_2E), or lacked a generative cause (*G*-drop: *PE*). The source and target analogs were two stories from different domains

[4] The labels of conditions in this study have been changed to make their meanings more transparent. The corresponding condition labels used by Colhoun and Gentner (2009) were as follows: Condition 1: No Analogy; Condition 2: AN-TargetRels; Condition 3: AN-BaseRels; Condition 4: AN-BaseRels + E.

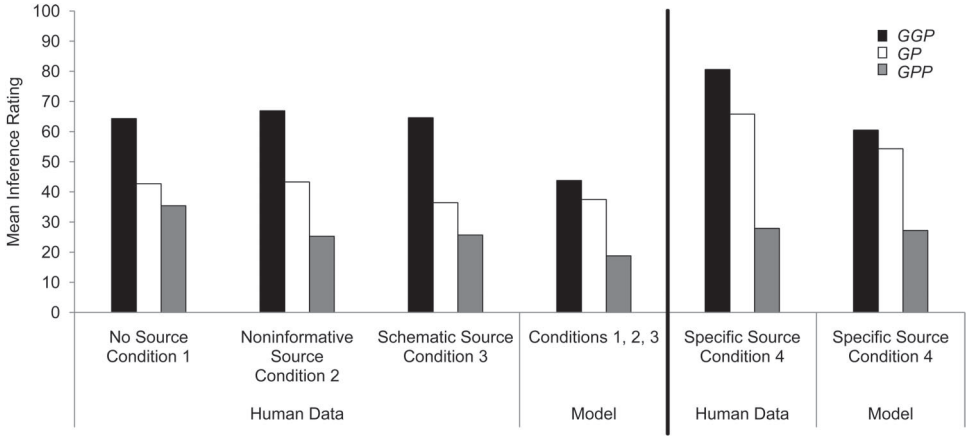

Figure 7. Mean ratings for predictive causal inferences produced by participants ("Human Data") in an experiment by Colhoun and Gentner (2009, Study 2) and predictions of the Bayesian model ("Model"). *G* and *P* represent generative causes and preventive causes, respectively. In argument type *GP*, two causal factors (*G* and *P*) were shared between the source and target. In the argument types *GGP* and *GPP*, one additional cause (generative in *GGP*, preventive in *GPP*) that had not appeared in the source was added to the target (in addition to the two causal factors shared by the source and target).

(fictitious descriptions related to chemistry and astronomy, respectively), adapted from materials developed by Lee and Holyoak (2008, Experiment 3; see description of materials on p. 1118). The source story was based on a chemist's observations about three different liquid substances, and the target story was based on an astronomer's observations about three stars. Participants first received the source story, which described how three different liquid substances could change into a solid material when they are mixed, depending on three causal factors, each based on a different semantic relation involving the three liquids. Each relation involved a comparison between two liquids with respect to one of three different characteristics: temperature, turbulence, and volume (e.g., "Denitrogel is colder than Oreor"). Two of the three relations were described as tending to produce the effect (i.e., changing the mixed liquids into a solid), and one was described as tending to prevent the effect.

Three different versions of the source story were constructed to counterbalance which semantic relation was used as the preventive factor. After listing the three relations, a schematic diagram was provided to help participants understand the causal structure of the source. This diagram resembled causal graphs of the sort commonly used in theoretical work on causal reasoning (e.g., Griffiths & Tenenbaum, 2005; Pearl, 1988; Waldmann & Holyoak, 1992). In the diagram, each causal factor was depicted as a rectangle shape, whereas the effect was depicted as an oval shape. Each rectangle shape (i.e., causal factor) was connected to the oval shape (i.e., effect) by a directed arrow. The two generative causal factors were connected to the effect by a solid line, and the preventive factor was connected by a dashed line.

After studying the source story, participants were given three different targets, each based on an astronomer's findings about three stars. In each target, participants were given exactly the same cover story except that the doctor's name and the names of the three stars were unique to each target. An example is the following:

> Recently an astronomer, Dr. Sternberg, has discovered that three stars, Acruxia, Errailel and Castoriff, located in a distant galaxy, have fused to form a superstar. He reads about the chemist's findings and thinks that this case of stellar fusion may be explained by a process similar to that identified by the chemist. The astronomer thinks the three stars may behave in a way similar to the three liquids. The theory is that gravitational attraction among all three stars could make two of the stars move closer together, so that these three stars finally fuse to form a superstar. The three stars are close to each other and no other stars have been found in that region of the galaxy.

The story went on to describe how the astronomer had identified three possible relations between the three stars that were analogous to the relations between the three liquids. Three relations involving the stars were created to have a structural mapping with the chemist's findings. To aid participants in finding the mapping, we also created semantic similarities in the corresponding relations between the chemistry and astronomy stories. The temperature relation in the source, "colder than," corresponded to "had a lower temperature than" in the target; the turbulence relation, "stirred vigorously," corresponded to "subject to more violent solar storms"; and the volume relation, "greater than," corresponded to a diameter relation, "wider than."

After the three relations that might have effects on stellar fusion were listed, participants were given a diagram completion task. This task involved deciding whether each of the relations between the three stars tended to produce or prevent the formation of the superstar on the basis of the analogical correspondences between the three stars and the three liquids. As in the case of the causal graph diagram provided with the source, each possible causal factor in the target (i.e., relation between the stars) was depicted as a rectangle shape, and the effect (i.e., formation of the superstar) was depicted as an oval shape. However, causal factors and the effect were not connected by directed arrows. Instead, participants were asked to draw lines between each of the relations and "formation of the superstar," just as with the chemist's diagram shown

earlier. This diagram completion task was included to check whether participants were able to correctly determine the analogical mapping between the source and target analogs. To prevent them from blindly copying the causal arrows on the basis of the simple presentation order of causal factors in the source and target diagrams, we randomized the presentation order of causal factors in the target diagram.

Because the basic task was causal attribution, the initial cover story made it clear that the effect (i.e., formation of the superstar) had in fact occurred. After completing the diagram completion task, the participant read the astronomer's "examination report," a table that stated whether each of the three causally relevant relations had actually occurred prior to the formation of the superstar. The presence or absence of one generative cause was always stated to be unknown (denoted by a question mark); this factor was the focus of the subsequent causal attribution judgment. The pattern of the other causal factors stated to be present or absent created the manipulation of argument type. In the no-drop condition (G_2PE), two relations (one generative factor, G_2, and one preventive factor, P) were stated to be present. In the P-drop condition (G_2E), one generative factor, G_2, was stated to be present, and the preventive factor, P, was stated to be absent. In the G-drop condition (PE), one generative factor, G_2, was stated to be absent, and the preventive factor, P, was stated to be present.

Finally, participants completed the causal attribution task. To explain why the superstar had been formed, participants had to carefully study which of the relations did or did not occur before the formation of the superstar, using the reports in the table, and then judge how likely it was that before the formation of the superstar, a specific relation (the generative cause described as unknown in the report table) was in fact present and had made the three stars fuse to form a superstar. Because it might seem odd to imagine 100 cases of apparently unusual astronomical systems, participants were simply asked to circle a number on a scale ranging from 0 (*certain false*) to 100 (*certain true*). Each participant was presented with all three conditions, and the order of the conditions was counterbalanced with three different versions. This variation in order of conditions was crossed with three versions of the source story, yielding a total of nine different counterbalancing conditions. Each counterbalancing condition was assigned randomly to different participants.

Results and Discussion

Human data. One of the participants failed to complete the diagram completion task correctly; her data were removed from analyses of causal attribution judgments. The mean causal attribution ratings for the G_2PE, G_2E, and PE conditions were 69.0, 61.9, and 70.3, respectively. This overall pattern of causal attribution ratings across the conditions was consistent with the findings of Experiment 2, in that dropping a preventive cause appeared to yield lower ratings for the unobserved generative cause; however, the differences between the three means were not reliable, $F(2, 60) = 1.19$, $MSE = 532.26$, $p = .31$.

Closer inspection of the data suggested that the overall pattern of means might be masking important individual differences in task performance. To examine whether the overall rating pattern was manifested by most participants or whether it resulted from averaging over qualitatively different response patterns, we performed a cluster analysis. The method of dividing participants into subgroups on the basis of their response profiles has been used in several previous studies of causal induction (e.g., Buehner et al., 2003; Rehder, 2006, 2009). Because we were interested in relative rather than absolute strengths of causal attribution judgments across the three argument types, Z scores were calculated for each participant on the basis of the mean rating across all three conditions for that participant. The three Z scores for each participant were entered into K-means cluster analysis using SPSS statistical software. The optimal number of clusters, which proved to be two, is that which yields the lowest value of the Schwarz Bayesian information criterion (BIC). The BIC for the two-cluster decomposition relative to the single-cluster solution decreased by 6.65 points; the BIC for the two-cluster decomposition relative to the three-cluster decomposition decreased by 9.58 points. Moreover, both the highest ratio of BIC changes and the highest ratio of distance measures were obtained for the two-cluster solution, strongly confirming the presence of two distinctive subgroups.

These two subgroups showed qualitatively different response patterns across the three argument types. In fact, as shown in Figures 8A and 8B, these two subgroups showed completely opposite response patterns. Separate within-subjects ANOVAs were performed for each subgroup. For the first subgroup ($n = 17$; see Figure 8A), the mean causal attribution ratings for G_2PE, G_2E, and PE conditions were 72.4, 52.4, and 89.4, respectively. These

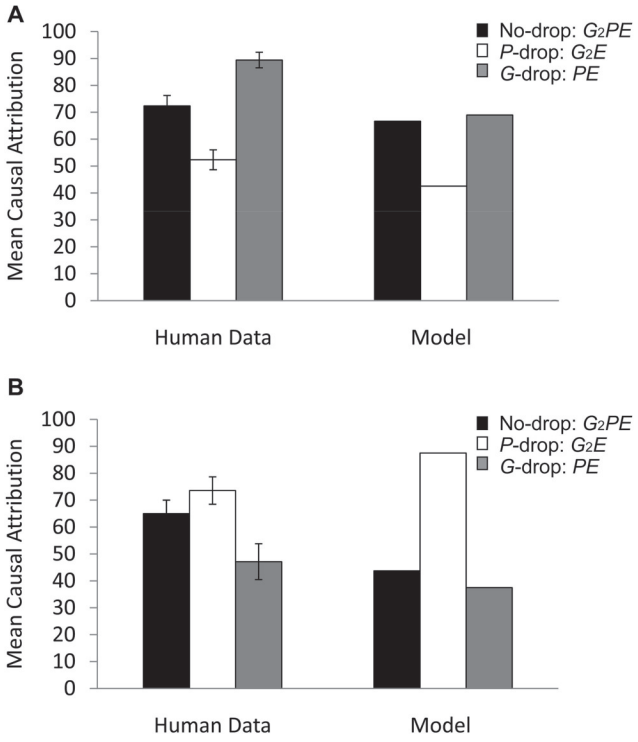

Figure 8. Mean causal attribution ratings for each subgroup in Experiment 3. Panel A shows the means for Subgroup 1 ($n = 17$); Panel B shows the means for Subgroup 2 ($n = 14$). G, P, and E represent generative causes, a preventive cause, and an effect, respectively. Error bars represent 1 standard error of the mean. Predictions derived from the Bayesian model are shown at the right of each panel.

means were reliably different, $F(2, 32) = 32.45$, $MSE = 180.27$, $p < .001$. Causal attribution ratings for argument type G_2PE were higher than those for argument type G_2E, which dropped the preventive factor, $t(16) = 3.78$, $p = .002$, but lower than for argument type PE, which dropped the generative factor, $t(16) = 4.17$, $p = .001$. This pattern is qualitatively similar to that observed in Experiment 2.

For the second subgroup ($n = 14$; see Figure 8B), the mean causal attribution ratings for G_2PE, G_2E, and PE conditions were 65.0, 73.6, and 47.1, respectively, a pattern opposite to that obtained for the first subgroup. These means were also significantly different, $F(2, 26) = 6.22$, $MSE = 409.34$, $p = .006$. Causal attribution ratings for G_2PE tended to be lower than those for G_2E, although the difference was not significant, $t(13) = 1.39$, $p = .189$, and were reliably higher than those for PE, $t(13) = 2.31$, $p = .038$.

The cluster analysis thus revealed one subgroup (consisting of over half the participants in Experiment 3) that, when making causal attribution judgments based on a cross-domain analogy, showed the same basic pattern as we observed in Experiment 2: The absence of a preventive cause decreased causal attribution ratings for an unobserved generative cause, whereas the absence of a different generative cause increased causal attribution ratings for the unobserved generative cause. The second subgroup, in contrast, showed the opposite response pattern.

Application of Bayesian model. These distinct response profiles strongly suggest that the participants in the two subgroups reasoned very differently about this cross-domain attribution problem. The first subgroup appears to have answered the attribution question on the basis of the "correct" correspondences between the source and target, just as did the participants in Experiment 2 who made simpler category-based inferences. We derived predictions for Subgroup 1 across these three conditions using our Bayesian model. The source provided information about the existence of causal links (i.e., causal structure) but not about whether the effect occurred given a specific combination of causal factors (essentially the same situation as that in the schematic source condition tested by Colhoun & Gentner, 2009). Accordingly, the theoretical derivation simply assumed uniform strength distributions, without updating. Because there are only three data points and hence the correlation between human data and model predictions had just one degree of freedom, the fit of the model, $r(1) = .92$, is simply descriptive. Nonetheless, it is apparent that the inference pattern observed for Subgroup 1 is in qualitative agreement with the predictions of our Bayesian model (see Figure 8A).

As we noted earlier, causal attribution is clearly more complex than causal prediction, and its difficulty may well be amplified when the source and target are far analogs, as was the case in Experiment 3. One explanation for the pattern of judgments produced by Subgroup 2 is that these participants in essence converted the stated attribution question into one that could be answered by causal prediction. This conversion could be accomplished by assuming that the queried cause in the target actually corresponds to the effect in the source, leading to a reversal of the causal arrow linking G_1 to E. In fact, in a debriefing session after the experiment, some participants in the second subgroup explicitly reported that the presence of a generative cause would have a positive power and the presence of a preventive cause would have a negative power with respect to the likelihood of the missing generative cause being present. Such verbal reports are consistent with the hypothesis that some participants treated the queried cause as an effect of the other factors.

The Bayesian model, when applied to the target structure with reversed causal arrows, is consistent with the qualitative pattern of differences across the three conditions for Subgroup 2, $r(1) = .82$ (see Figure 8B). For the G_2PE case, for example, the presence of the preventive causal factor would be viewed as directly decreasing the likelihood of the unknown generative cause. For the G_2E case, by contrast, the presence of one generative causal factor coupled with the absence of the preventive causal factor would increase the probability that the unknown generative causal factor had occurred.

Experiment 4

In Experiment 4 we extend our Bayesian theory to a more complex situation in which multiple source analogs are available, so that inferences about the target depend in part on which source analog is selected as the basis for analogical transfer. According to our integrated theory, a causal relation in the target potentially plays a dual role: It first may guide structure mapping between one or more source analogs and the target; then once a source is selected, the causal relation will also guide causal inference on the basis of the resulting causal model of the target. In the present study we examined analogical transfer when the structure mapping involved in selecting the relevant source was sometimes ambiguous (cf. Spellman & Holyoak, 1996).

More specifically, we examined how presence or absence of a particular causal relation (a preventive cause) in the target might increase or decrease inductive strength depending on whether the structural mapping is clear or ambiguous. The source analog included a preventive cause, which might or might not be also included in the target. When the mapping is clear, the expected effect of inclusion of the preventive cause will be to decrease inductive strength in the target, as shown in previous studies (e.g., Lee & Holyoak, 2008). However, when the mapping is ambiguous, and if the preventive cause is able to resolve the mapping ambiguity, the expected result will be reversed.

The materials in Experiment 4 were designed so that when the mapping was ambiguous, the inclusion of a preventive cause in the target provided sufficient structural information to resolve the ambiguity and hence select a particular source as the basis for transfer of causal structure to the target. Conversely, when this preventive cause was omitted from the target, the structural ambiguity would be left unresolved, thereby impairing transfer of a causal model from source to target. In such situations, our Bayesian model predicts that including the preventive cause in the target can actually increase inductive support for the occurrence of the effect that it tends to prevent. No other model of analogical transfer appears to yield such a prediction. Experiment 4 was performed to investigate this type of interactive impact of causal and structural constraints on analogical transfer.

Method

Participants. Forty-five UCLA undergraduates participated in the experiment to fulfill a course requirement. Each participant was randomly assigned to one of eight different sets of materials generated for counterbalancing purposes.

Design and materials. The source story described a biochemist's findings about an imaginary liver disease called "tibulosis," found in rats. The disease had two different subtypes, Type A and Type B, described as being caused by different factors and exhibiting quite different symptoms. The scientist had identified several factors that determine whether rats might develop either Type A or Type B tibulosis. For each type, certain hormones, enzymes, and antibodies were involved. Participants were asked to carefully study the biochemist's findings using a verbal description and diagram presented in the booklet in order to determine what characteristics are likely to produce or prevent the development of each type of the disease. Participants were then given descriptions of human patients with a liver disease and asked to apply what they had learned about tibulosis in rats to judge the probability that the human patients had tibulosis Type A or Type B.

The two disease subtypes for rats constituted two alternative source analogs for the human disease. The descriptions of the subtypes were designed to create a potential structural ambiguity. The two types had identical causal structures except for the names of causal elements but with one critical structural difference involving a preventive cause. Each source disease included two generative causes, one preventive cause, and an effect (consistent with a common-effect model; Waldmann & Holyoak, 1992). The two generative causes were certain types of hormones and enzymes, and the preventive cause was a certain type of antibody. In each case the preventive cause was narrow in scope (Carroll & Cheng, 2009), in that it served to stop the causal impact of one of the two generative causes but not the other. The description of the causal structure for Type A tibulosis was as follows:

> Hormone A tends to stimulate the production of enzyme A, and vice versa.
>
> Hormone A tends to PRODUCE Type A tibulosis.
>
> Enzyme A also tends to PRODUCE Type A tibulosis.
>
> The immune system sometimes PRODUCES antibody A in response to enzyme A, but never in response to hormone A.
>
> Antibody A tends to PREVENT enzyme A from producing Type A tibulosis. However, antibody A provides no protection against the direct effect of hormone A on Type A tibulosis.

To aid comprehension of the causal structure, we provided a schematic diagram right below the description. Figure 9 depicts the causal structure for Type A tibulosis (left), as just described, and also for Type B tibulosis (right). For Type A, hormone A and enzyme A are two generative causes that both tend to produce the effect, Type A tibulosis. Antibody A is a preventive cause with a narrow scope that prevents enzyme A (but not hormone A).

The B subtype (see Figure 9, right) was very similar to the A subtype just described, except that the effect was Type B tibulosis (rather than Type A) and the names of the hormone, enzyme, and antibody included a *B* rather than an *A*. The critical structural difference between the two sources was that in the B version, the immune system was described as producing antibody B in response to hormone B but never in response to enzyme B (opposite to the situation in the A version); furthermore, antibody B tended to prevent the effect of hormone B (not enzyme B).

In the target story, participants read reports about human patients who might have a human form of Type A or Type B tibulosis. Examination reports for seven patients were constructed. Each examination report included information about a hormone, an enzyme, and (in some versions) an antibody found in each patient. A 2×2 within-subjects design was employed, resulting in four basic versions of the target descriptions. The first independent variable was whether the target description was specific or generic. In the specific condition, specific names of the hormone, enzyme, and antibody (e.g., hormone A, enzyme A, antibody A) were explicitly stated in the description of the patient report provided in the target. Given that these names matched those for one of the two source subtypes, the mapping of the human case to Type A (or B) tibulosis was accordingly transparent.

In contrast, in the generic condition, specific names of the hormone, enzyme, and antibody were not provided. Instead, each was simply described by its general categorical description (i.e., hormone, enzyme, and antibody). Thus, in the absence of additional structural information, there was no basis for preferentially mapping the description of the factors observed in the human patient onto those related to Type A versus Type B tibulosis in rats.

This manipulation of the target description was crossed with a second independent variable: presence or absence of the preventive cause (antibody) in the description of the human patient. As previously explained, the critical structural difference between Type A and Type B tibulosis was that for Type A, the *enzyme* produced the antibody, which then acted to block the *enzyme's*

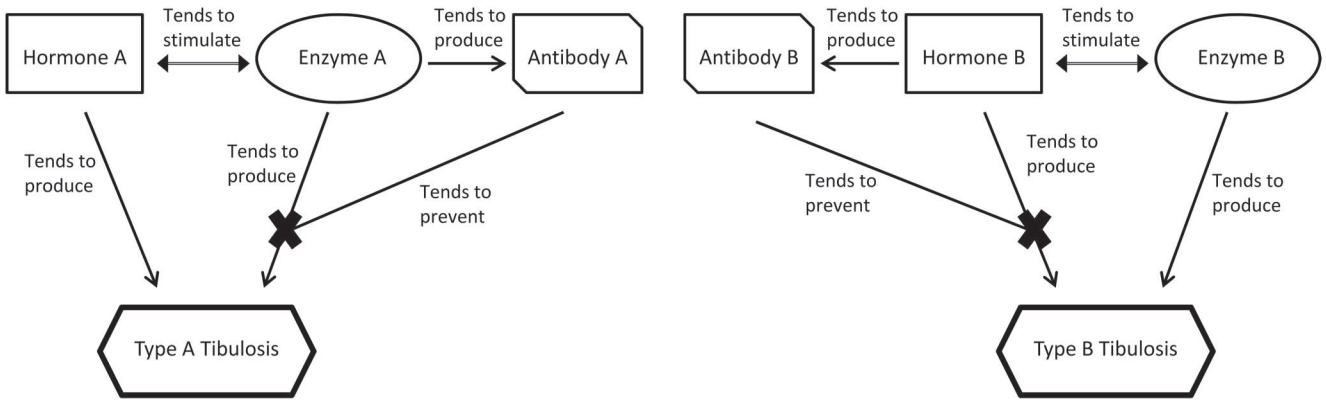

Figure 9. Example of causal structure for the two disease subtypes used as source analogs in Experiment 4. Left: Type A tibulosis; right: Type B tibulosis.

impact; whereas for Type B, it was the *hormone* that produced the antibody, which then acted to block the *hormone's* impact. In the *P*-present condition (in which *P* represents a preventive cause), the target description included analogous information about the human case. For example, in the specific, *P*-present condition, the description might state the following:

> Hormone A and enzyme A are present, and each stimulates production of the other.
>
> The immune system produced antibody A in response to the enzyme (but not the hormone).

More critically, in the generic, *P*-present condition, the description stated this:

> A hormone and an enzyme are present, and each stimulates production of the other.
>
> The immune system produced an antibody in response to the enzyme (but not the hormone).

Note that even though no specific names are provided, this generic, *P*-present description (on the basis of the second statement in the description) provides structural information sufficient to disambiguate the mapping between the human case in the target and the two disease descriptions for rats. That is, only Type A tibulosis involves an antibody produced in response to an enzyme, which then blocked the enzyme's effect. Any of the major models of structure mapping (e.g., Falkenhainer et al., 1989; Holyoak & Thagard, 1989; Hummel & Holyoak, 1997) would be able to use the structural information provided in the generic, *P*-present condition to resolve the potential ambiguity and identify a determinate mapping between the disease described in the target and one of the two sources in the description of disease subtypes found in rats.

In the *P*-absent versions (both specific and generic), the second statement in the relevant description was simply replaced with "no antibody is present." Critically, in the generic, *P*-absent condition, no information was provided that could possibly serve to resolve the structural ambiguity inherent in the mapping; hence, the target case could be mapped to either Type A or Type B tibulosis as the source. If a preventive cause plays a dual role in analogical transfer, as the integrated theory postulates, then in this experiment its inclusion will have a paradoxical influence on the judged probability of an effect in the target. Specifically, given a specific description of the target, inclusion of the preventive cause will decrease the judged probability of the effect (by acting as a preventer within the causal model of the target); but given a generic description of the target, its presence will increase the judged probability of the same effect (by serving to disambiguate the mapping so that a particular source is selected and hence the corresponding causal model of the target can in fact be constructed).

For each condition except the generic, *P*-absent condition, two patient reports were constructed, resulting in seven patient reports in total. For each of the first three conditions, one of the two patient reports supported mapping to Type A, and the other supported mapping to Type B. Because the generic, *P*-absent condition did not support mapping to one type over the other, only one version of this patient report could be constructed. Two different sets of materials were constructed by counterbalancing whether the hormone or the enzyme produced an antibody in Type A and in Type B. Within each set, four different orders of targets were constructed, resulting in eight versions of materials in total.

Procedure. Participants were given a booklet that included the story about the two alternative source diseases, the target story, and a series of inference tasks. First, participants read the story about a biochemist's findings about a new liver disease found in rats and studied what factors were likely to produce or prevent the development of two types of the disease on the basis of the verbal descriptions and diagrams. After reading the descriptions of the two types of the disease, participants were asked to briefly describe a major difference between the two types of tibulosis. This task was intended to call attention to structural information that might later serve to disambiguate the mapping of the target to one of the two alternative sources.

In the generic conditions (but not in the specific conditions), a mapping task was included before the inference task to check whether the potential mapping ambiguity was resolved. This task required identifying the generic hormone as "hormone A," "hormone B," or "can't tell." The analogous question was also asked about the generic enzyme. Regardless of the answers the participants gave on this mapping task, they then completed the task of analogical inference.

For the analogical inference task, participants were given the examination reports for seven different patients. For each patient, participants were asked to judge how likely it was that the patient had each type of the disease. To answer each question, they were to imagine there were 100 cases with the same known characteristics as for the specific case and judge how many of these 100 cases would be expected to have each type of the disease. A number between 0 and 100 was elicited for each type of the disease.

Results and Discussion

Human data. On the mapping task, 33 of the 45 participants reported the structurally justified mappings for the hormone and enzyme in the generic, *P*-present condition. The other 12 participants gave a variety of responses in this critical condition. Some of them chose "can't tell" for both mapping questions, some chose the correct mapping for one question but "can't tell" for the other, and some gave structurally incorrect mappings. In reporting the results for the analogical inference task, we first report the results based on data from all of the participants ($N = 45$) and then the results for those who gave completely correct mappings in the generic, *P*-present condition ($n = 33$). In the generic, *P*-absent condition, in which structural information to resolve the ambiguity was lacking, all but one participant chose "can't tell" for both mapping questions. (The single exception was one of those who also gave a structurally incorrect response in the generic, *P*-present condition.)

For each patient case, participants estimated both the probability that the patient had Type A of the disease and the probability that the patient had Type B. The format encouraged participants to treat the two types as mutually exclusive, and assignments of Type A versus Type B were fully counterbalanced across conditions. To code the responses on the inference task, we defined the "correct" disease type as that supported by the preferred mapping in the three unambiguous conditions (specific, *P*-present; specific, *P*-absent; and generic, *P*-present). For comparison, this same disease

type (either A or B) was defined as "correct" in the matched generic, *P*-absent condition (in which the selection of a source was structurally ambiguous).

The mean rated probability of the correct effect for each of the four conditions is shown in Figure 10. These data were analyzed with a 2 × 2 ANOVA in which both target description (specific vs. generic) and presence of the preventive cause (*P*-present vs. *P*-absent) were within-subjects variables. A significant main effect of specificity of the target description was obtained, $F(1, 44) = 123.09$, $MSE = 302.52$, $p < .001$, in that inference strength was significantly higher when the description was specific ($M = 83.0$, $SD = 16.09$) than when it was generic ($M = 54.3$, $SD = 17.70$). The main effect of presence of the preventive cause was not significant ($F < 1$). Most importantly, a significant interaction was obtained between target specificity and presence of preventive cause, $F(1, 44) = 79.7$, $MSE = 281.49$, $p < .001$, implying that the presence of a preventive cause had a different impact on analogical inference depending on the specificity of the target description. When the description of the target was specific so that the mapping to one of the disease types in the source was transparent, participants gave significantly higher estimates of the probability of the correct effect in the *P*-absent condition ($M = 92.4$, $SD = 13.99$) than in the *P*-present condition ($M = 73.6$, $SD = 28.52$), $t(44) = 4.02$, $p = .001$. This result replicates the present Experiment 1 and previous findings (Lee & Holyoak, 2008), in that dropping a preventive cause from the target increased the strength of a predictive inference. In contrast, when the target description was generic, the effect of including the preventive cause was reversed. The estimated probability of the correct effect was now higher in the *P*-present condition ($M = 67.2$, $SD = 29.6$), in which the preventive cause served to disambiguate the source selection, than in the *P*-absent condition ($M = 41.3$, $SD = 23.89$), in which the selection of a source was structurally indeterminate, $t(44) = 4.28$, $p < .001$.

As mentioned earlier, 12 participants failed to solve the mapping task correctly in the generic, *P*-present condition. Because the inference task critically depended on the mapping, incorrect mappings might lead to an erroneous prediction about the target. We therefore performed a second set of analyses using only the data from the 33 participants who solved the mapping task correctly. A significant main effect of specificity of the target description was obtained, $F(1, 32) = 101.28$, $MSE = 185.56$, $p < .001$, with higher ratings of the correct effect when the target description was specific ($M = 81.8$, $SD = 16.46$) than when it was generic ($M = 57.9$, $SD = 16.93$). The main effect of presence of a preventive cause was not reliable ($F < 1$). A significant interaction between target description specificity and presence of a preventive cause was obtained, $F(1, 32) = 129.67$, $MSE = 168.45$, $p < .001$. When the description was specific, participants gave significantly higher inference ratings in the *P*-absent condition ($M = 92.5$, $SD = 13.57$) than in the *P*-present condition ($M = 71.1$, $SD = 29.58$), $t(32) = 3.83$, $p = .001$. In contrast, when the description was generic, the *P*-present condition ($M = 72.9$, $SD = 28.77$) yielded significantly higher inference strength than did the *P*-absent condition ($M = 42.9$, $SD = 22.01$), $t(32) = 4.49$, $p < .001$. The pattern for the subgroup of participants that fully solved the mapping task was thus statistically identical to that found for the entire set of participants.

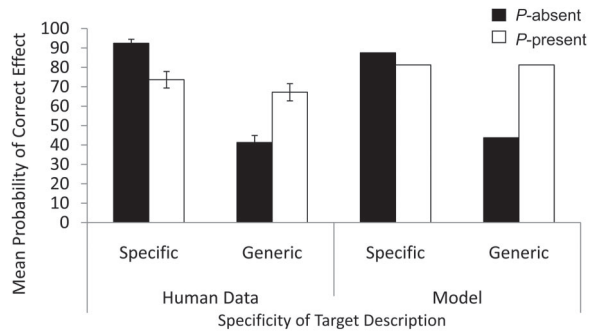

Figure 10. Mean probability of the correct effect in each condition of Experiment 4. In the *P*-present condition the preventive cause *P* was present in the target; in the *P*-absent condition this preventive cause was absent. Left: human data; right: predictions derived from the Bayesian model. Error bars represent 1 standard error of the mean. Note that the selection of the source analog is ambiguous for the generic, *P*-absent condition.

Application of the Bayesian model. Our Bayesian model can be extended to deal with the situation created by the design of Experiment 4, in which any one of a set of alternative source analogs might provide a basis for analogical inference about a target. As shown in Equation 8, the probability that an effect occurs in the target is the sum of its probabilities based on each possible source, weighted by the probability of the mapping between the target and each source. That is,

$$P(E^{\mathrm{T}} \mid \boldsymbol{C}^{\mathrm{T}}) = \sum_{\mathbf{S}} P(E^{\mathrm{T}} \mid \boldsymbol{C}^{\mathrm{T}}, E^{\mathbf{S}}, \boldsymbol{C}^{\mathbf{S}}) P(E^{\mathbf{S}}, \boldsymbol{C}^{\mathbf{S}} \mid \boldsymbol{C}^{\mathrm{T}}) \qquad (8)$$

where $\mathbf{S}$ denotes a set of possible sources, each consisting of a causal model with variables ($E^{\mathbf{S}}$, $C^{\mathbf{S}}$). In calculating the first term, we assumed that participants had no prior knowledge about causal structure or strength of the source (as in the derivations used in modeling the previous experiments); hence, the stated causal relations were assigned a uniform strength distribution ranging between 0 and 1. As was the case for the design used in Experiment 3, because no further information about causal strengths was provided in the source, these distributions remained uniform (no updating on the basis of examples); thus, in effect only causal structure (not strength) was available to be transferred to the target. On the basis of Equation 4, causal links with uniform strength distributions were directly transferred from the source to the target analog when the mapping was determinate. In the derivations, the functional form of the preventive cause (a noisy-AND-NOT function) was applied in a manner that reflected the appropriate narrow scope of the preventer (Carroll & Cheng, 2009). The influences of the causes were integrated sequentially. After applying a noisy-AND-NOT function to integrate the influence of the preventer with that of its related generative cause, a noisy-OR function was applied to combine this intermediate result with the influence of the other generative cause and an assumed background cause.

The second term in Equation 8 assesses how likely the target can map to a specific source. In Experiment 4, this probability was either 1 or 0 for the three unambiguous conditions (specific, *P*-present; specific, *P*-absent; and generic, *P*-present), so the causal model for the correct effect was always transferred to the target. However, in the ambiguous condition (generic, *P*-absent), this

mapping probability was equal (.50) for the two sources due to the structural ambiguity.

Figure 10 depicts the parameter-free predictions of the Bayesian model. The model captured the qualitative pattern of the human data, $r(2) = .93$. When data from just those participants who solved the mapping task correctly were modeled, the fit improved slightly, $r(2) = .94$. The model captures the trade-off that arises in the generic, *P*-present condition, in which the presence of the preventer exerts a positive influence on analogical transfer by guiding the mapping to one particular source but then reduces transfer somewhat by acting to prevent the effect within the causal model created for the target.

General Discussion

Summary

We have presented a Bayesian theory of inductive inference using the framework provided by representations based on causal models (Cheng, 1997; Griffiths & Tenenbaum, 2005; Lu et al., 2008; Pearl, 1988; Waldmann & Holyoak, 1992). In doing so, we extended similar models that have addressed the role of causal knowledge in making category-based inferences (e.g., Rehder, 2009) to situations involving very small numbers of specific cases (one or even zero) and relatively high levels of relational richness. Our Bayesian model of causal inference is integrated with structure mapping, making it possible to derive predictions regarding transfer on the basis of relationally complex analogies and schemas, including situations in which structure mapping and causal inference interact (Experiment 4). By integrating analogical inference with causal models, we were able to provide a more unified account of the influence of causal knowledge on the subprocesses involved in analogical transfer, ranging from initial acquisition of causal knowledge about the source to evaluation of potential inferences about the target.

Our parameter-free Bayesian theory was able to account for a range of human inductive inferences involving causal prediction (cause to effect; Experiments 1 and 4) and causal attribution (effect to cause; Experiments 2 and 3) when the situations involved a mix of both generative and preventive causes. The theory accounts for the dissociation between overall mapping quality and the strength of a causal inference about the effect, which is obtained when a generative causal factor in the source is omitted from the target. This manipulation reduces mapping quality yet increases inferences that the effect will occur in the target (Experiment 1; also Colhoun & Gentner, 2009, Study 1a; Lee & Holyoak, 2008). In addition, the theory accounts for the pattern of predictive causal judgments created by varying causal knowledge about the source, which depends on whether the effect occurs in the source, fails to occur (Experiment 1), or is of unknown status (Experiments 3 and 4; Colhoun & Gentner, 2009, Study 2).

The present article is the first to address the impact of preventive causes on causal attribution (see Appendix B). For causal attribution, the theory accounts for the reversal of the impact of dropping a preventive causal factor (as derived in Appendix B and demonstrated for the first time in the present Experiments 2 and 3). That is, the same manipulation (dropping a preventive factor from the target) decreases inferences that an unobserved generative cause produced the observed effect (or to state the same phenomenon in a different way, keeping a preventive cause in the target diminishes causal discounting). The overall pattern across the three attribution conditions tested in Experiments 2 and 3 thus showed a novel dissociation between overall mapping quality and support for inductive inferences (involving an inference about a generative cause rather than an effect).

In addition, Experiment 3 revealed that when confronted with an attribution question requiring cross-domain analogical transfer, two subgroups of participants reasoned very differently. One subgroup answered as would be expected given the apparent correspondences between target and source. However, a second subgroup appeared to convert the attribution question into a simpler predictive question by assuming that the queried target variable was an effect rather than a cause. If we assume the underlying mapping differed in this manner, the Bayesian model is able to account for the opposite pattern of judgments produced by the two subgroups. Finally, the data from Experiment 4 confirmed another novel prediction of our integrated theory of transfer: When multiple source analogs compete, including a preventive cause in the target can either decrease or increase the estimated probability that an effect will occur, depending on whether a structural ambiguity influences the selection of a source analog. This interaction confirms that causal relations in the target can play a dual role in transfer, influencing structure mapping between the source and target as well as the operation of the causal model in the target.

Comparison to Previous Approaches

Our approach is broadly similar to Bayesian theories developed to account for other varieties of inductive inference (e.g., Anderson, 1990; Griffiths & Tenenbaum, 2005; Hahn & Oaksford, 2007; Kemp et al., 2007; Kemp & Jern, 2009; Kemp & Tenenbaum, 2009; Lu et al., 2008; Oaksford & Chater, 2007). The present model differs from previous approaches to analogical inference in several interrelated ways. First, the theory we propose integrates analogical reasoning with probabilistic inference, which takes into consideration the uncertainty inherent in analogical inferences. Following Holyoak (1985) and Lee and Holyoak (2008), we argue that the core computational goal in reasoning with analogies is to arrive at accurate and useful inferences about the target. To do so, the model incorporates an explicit theory of how causal knowledge (including uncertainty) about the source is learned and represented. A causal relation, rather than being viewed as a static relational description, is represented as a probability distribution over values of the causal factor's power to dynamically influence the state of the effect variable. Analogical transfer (defined as the generation of new inferences about the target on the basis of knowledge of the source) is explicitly decomposed into development of the target model by transferring causal structure and strength, followed by "running" it. Importantly, inferences about the values of variables determined endogenously by the causal model of the target follow from the final step rather than being imported directly from the source.

The present proposal is closely related to approaches to category-based inference that have adopted the framework of causal models. Like the model developed by Rehder (2009), the present model incorporates the basic assumptions of the power PC theory (Cheng, 1997). The model presented here is based on a Bayesian formulation (Griffiths & Tenenbaum, 2005; Lu et al.,

2008). The Bayesian framework treats causal strength as a distribution rather than a point estimate, which is critical for dealing with the uncertainty inherent in strength estimates based on a single source analog. The great advantage of a Bayesian formulation of reasoning with causal models is that it makes it possible to derive inferences when causal strengths are highly uncertain. In particular, previous work has shown that the Bayesian version of the power PC theory accounts for the impact of sample size on patterns of causal judgments. In contrast, classical causal power as defined by Cheng (1997) is independent of sample size and thus fails to capture the difference in certainty provided by a large sample relative to a single case.

The Bayesian power PC theory can readily account for data concerning category-based inferences based on large samples or when causal powers can be specified with certainty. The general relationship between the Bayesian and classical formulations of the power PC theory is that derivations based on the former converge on derivations based on the latter as sample size grows large (or more generally, as the power value is established with high certainty). In particular, the present proposal subsumes Rehder's (2009) CBG model for the class of situations in which both are applicable (i.e., common-effect structures based on binary variables).

This is illustrated by one experiment in a series reported by Rehder (2009), in which he investigated how people generalize new properties to an entire class of category members. The CBG model predicts that increasing the strength of a causal link between a known feature associated with a category and a new feature increases the judged prevalence of a new feature when the latter is an effect, whereas decreasing causal strength increases the judged prevalence of the new feature when the latter is a cause. To test this prediction of the CBG model, Rehder (in his Experiment 2) employed a 2 × 2 within-subjects design to assess the interaction between causal strength and causal direction. Participants first received summary information about the distribution of various known features associated with a category. Each known feature was described as occurring in 67% of category members (i.e., base rate of a known feature was always .67).

After participants learned about the distribution of known features, a causal law linking an existing feature with a new feature was stated. In this causal law, two variables were manipulated. First, causal strength was described as either low or high. In the low-strength condition, participants were told that whenever a category member had a certain feature, it would cause the member to have another feature 67% of the time. In the high-strength condition, 67% was replaced with 100%. Participants were also told that there were no other causes of the effects (i.e., there was no background cause), implying that generative causal power was proportional to the stated frequency of the effect given the cause (i.e., .67 or else 1). Second, causal directionality was manipulated, with the new feature described as being either a cause or an effect of a known feature. Finally, participants were asked to judge what proportion of category members would have the new feature (by positioning a slider on a scale from *none* to *all*).

The results supported the predictions of the CBG model. When the new feature was described as an effect, it was judged to be more prevalent when causal strength was high rather than low. In contrast, when the new feature was described as a cause, the pattern was reversed: The new feature was judged to be more prevalent when causal strength was low rather than high. Qualitatively, this pattern is consistent with the basic prediction that causes need to be more prevalent to generate the observed effect when the causal link is weak than when it is strong. Quantitative predictions can be derived using the assumptions of the power PC theory.

In modeling the data from this experiment, our Bayesian theory is essentially identical to the CBG model. Because Rehder (2009) directly taught his participants that the causal power was a specific point value, uncertainty associated with the strength distribution would be minimal. A Bayesian model can learn a close approximation to a specified point value by being presented with a large sample of "imaginary cases" that follow the appropriate contingency table for occurrences of cause and effect. To derive predictions for Rehder's Experiment 2, the Bayesian model was first presented with 200 cases that instantiated the appropriate contingency table for the manipulation of causal power, which was either .67 or 1. The resulting distributions of causal strength (sharply peaked at the value of power) were then transferred to the target, and the causal model for the target was used to infer the frequency of the novel feature. When the new feature was described as an effect, the standard equation for predictive causal inference was used. When the new feature was described as a cause, the model summed over the predicted probabilities of the new feature, given presence versus absence of the effect. That is,

$$P(C = 1) = P(C = 1 \mid E = 1)P(E = 1) + P(C = 1 \mid E = 0)P(E = 0), \quad (9)$$

where C is the causal (new) feature and E is the effect (known) feature, and the values 1 and 0 represent their presence versus absence. Our parameter-free Bayesian model yields quantitative predictions virtually identical to those based on the CBG model, and the fit to the human data is excellent, $r(2) = .996$, $p = .004$.

Whereas the Bayesian theory can thus account for data from category-based inference that supports the CBG model (the former in essence subsuming the latter), the CBG model is simply inapplicable to the kinds of inference tasks on which we focused in the present Experiments 1–4. As we noted in the introduction, the fact that our Bayesian theory can cope with small numbers of examples is critical to the success of its extension to analogical and schema-based reasoning, in which the number of specific instances provided in the source is typically just one or even zero (when the causal structure is conveyed without any specific case). The basic limitation of the classical definition of causal power is that its value (a point estimate rather than a distribution) is essentially constant over sample size, thus failing to capture the empirical form of learning curves. Of greater present concern (given the goal of constructing a theory of analogical inference), assessing classical causal power requires computing the contrast between the probability of the effect in the presence versus the absence of a cause. This computation necessarily requires an absolute minimum of two observed cases (at least one observation of what happens when the cause is present and at least one observation of what happens when it is absent). Thus, in a typical case of analogical inference from a single instance in which various causes are present and the effect occurs, classical causal power is left undefined. Accordingly, models of category-based inference based on classical causal power, such as the CGB model of Rehder (2009),

are inherently inapplicable to typical paradigms involving analogical inference.

Broader Implications of the Bayesian Framework for Analogical Transfer

Our Bayesian theory clarifies what might be termed the "value added" by a source analog. Although models of analogy typically represent causal knowledge about a source analog by simply stating cause relations as part of its predicate-calculus-style description, this is clearly an oversimplification. Causal relations will seldom if ever be solely based on a single example; rather, they reflect more general prior information derived from multiple examples and/or more direct instruction. In this sense, categorical knowledge and analogical inference are closely interconnected, as suggested by Figure 1.

A source analog can, however, provide additional information beyond prior knowledge about individual causal links. In the experiments on which we have focused in the present article, the source (when it includes a specific instance) provides information about the joint impact of a set of causal factors that do or do not yield the effect, thus allowing revision of strength distributions to reflect the relative strengths of generative and preventive causes. More generally, the source may also provide information about (a) a sequence of causal events that leads (or fails to lead) to goal attainment or (b) side effects generated in the course of attaining a goal. This type of detailed information about a specific pattern of causal events will often go well beyond prior knowledge about individual causal relations, enabling the source to provide critical guidance in making inferences about a complex and poorly understood target analog.

As a computational level theory, our proposal in no way denies that multiple component mechanisms are involved in the process of analogical inference. Indeed, recent neuroimaging studies have linked relational integration (a reasoning process closely tied to comparison of role-governed relations; Waltz et al., 1999) to a specific brain region, the rostrolateral prefrontal cortex (Bunge, Wendelken, Badre, & Wagner, 2005; Cho et al., 2010; Christoff et al., 2001; Green, Kraemer, Fugelsang, Gray & Dunbar, 2010; Kroger et al., 2002; for a review see Knowlton & Holyoak, 2009). However, the fact that relational comparison is a key component of analogical inference does not obviate the need for an integrated theory. The predictive power of the present Bayesian theory of analogical inference is evidenced by its initial success in providing a unified account of how causal knowledge about the source is acquired, how this knowledge is linked to the target, and how the derived representation of the target is used to generate systematic causal inferences.

Future Directions

The Bayesian theory of analogical inference we have presented in the present article is best viewed as a promissory note for future theoretical developments. Although the present theory can in principle be extended to more complex causal structures, the initial applications we have considered involve simple networks (a small number of causal factors and a single effect). Much work remains to account for inferences based on analogies involving complex causal chains with intermediate nodes and multiple effects. We conjecture that the role of analogy in guiding causal inference will in fact increase dramatically as the size of the search space of potential causal factors and effects in the target increases. A source analog with a well-established causal model can serve to focus attention on portions of a complex target in which analogous causal relations are particularly likely to be found, obviating the need for unconstrained search through a prohibitively large space of possibilities.

The central principle that guides Bayesian analyses of perception and cognition for core inductive tasks—including categorization, causal learning, and perhaps analogical inference—is that these core processes serve the computational goal of yielding accurate knowledge that will help achieve the reasoner's goals (Anderson, 1990). Nonetheless, actual human induction is clearly also constrained by limitations in attention and working memory, as more recent algorithmic models of analogy have stressed (Halford, Wilson, & Phillips, 1998; Hummel & Holyoak, 1997, 2003). In extending the approach to more complex analogies, it will be necessary to confront the processing limitations that constrain human relational reasoning. We already found evidence of such processing limitations in Experiment 3, in which a substantial minority of participants apparently evaded the task of drawing causal attributions on the basis of a cross-domain analogy, instead assuming the queried variable was actually an effect, not a cause, and then reasoning forward from known factors to the one that was unobserved. The joint demands of processing a cross-domain analogy and making the "reverse" inferences required for causal attribution (reasoning from an observed effect to a possible unobserved cause) may have exceeded the processing capacity of some participants.

Finally, we would like to emphasize that the integration of models of analogical inference with models of causal learning and inference can continue to be of mutual benefit. We have aimed in the present article to illustrate how the past two decades of work on causal models can be used to further the development of theories of analogical inference. Future advances in understanding human causal reasoning will thus have immediate implications for work on analogy.

Equally promising, however, is the potential for transfer across these two research areas in the reverse direction. As Lien and Cheng (2000, p. 98) noted, "New causal relations are sometimes learned by analogy to a known causal relation." We have already mentioned the potential usefulness of a source analog in guiding search for causal relations within a complex target. An even more basic theoretical link involves the identification of relevant data for making causal inferences. In the extensive literature on the acquisition of causal knowledge from observations of empirical contingency data, the issue of what observations count as the "same" has generally been glossed over. For example, to decide whether a new disease is communicable across people, it is typical to assume that we can readily identify occasions on which a healthy person does or does not come in contact with a person afflicted with the disease, and then observe whether the previously healthy person does or does not contract the disease. But suppose we now find that a similar disease in pigs can spread from one animal to another. On the face of it, observations of pigs cannot alter contingencies related to human disease transmission. Yet intuitively, such knowledge about a different domain in fact increases the estimated probability that the human disease is also communicable. Why? The answer, which calls for advances in the theoretical integration of causal models with analogical inference, may provide a major clue as to why the inductive power of human reasoning exceeds

that of any other form of biological or artificial intelligence (Holland et al., 1986; Penn, Holyoak, & Povinelli, 2008).

References

Ahn, W. (1999). Effect of causal structure on category construction. *Memory & Cognition, 27,* 1008–1023.

Ahn, W., Brewer, W. F., & Mooney, R. J. (1992). Schema acquisition from a single example. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 391–412.

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

Bartha, P. (2010). *By parallel reasoning: The construction and evaluation of analogical arguments.* Oxford, England: Oxford University Press.

Bindra, D., Clarke, K. A., & Shultz, T. R. (1980). Understanding predictive relations of necessity and sufficiency in formally equivalent "causal" and "logical" problems. *Journal of Experimental Psychology: General, 109,* 422–443.

Brown, A. L. (1989). Analogical learning and transfer: What develops? In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 369–412). New York, NY: Cambridge University Press.

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

Bunge, S. A., Wendelken, C., Badre, D., & Wagner, A. D. (2005). Analogical reasoning and prefrontal cortex: Evidence for separable retrieval and integration mechanisms. *Cerebral Cortex, 15,* 239–249.

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

Carroll, C. D., & Cheng, P. W. (2009). Preventative scope in causal inference. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 833–838). Austin, TX: Cognitive Science Society.

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

Cheng, P. W., & Novick, L. R. (2005). Constraints and nonconstraints in causal learning: Reply to White (2005) and Luhmann and Ahn (2005). *Psychological Review, 112,* 694–707.

Cho, S., Moody, T. D., Fernandino, L., Mumford, J. A., Poldrack, R. A., Cannon, T. D., . . . Holyoak, K. J. (2010). Common and dissociable prefrontal loci associated with component mechanisms of analogical reasoning. *Cerebral Cortex, 20,* 524–533.

Christoff, K., Prabhakaran, V., Dorfman, J., Zhao, Z., Kroger, J. K., Holyoak, K. J., & Gabrieli, J. D. (2001). Rostrolateral prefrontal cortex involvement in relational integration during reasoning. *NeuroImage, 14,* 1136–1149.

Colhoun, J., & Gentner, D. (2009). Inference processes in causal analogies. In B. Kokinov, K. J. Holyoak, & D. Gentner (Eds.), *New frontiers in analogy research: Proceedings of the Second International Conference on Analogy* (pp. 82–91). Sofia, Bulgaria: New Bulgarian University Press.

Dunbar, K., & Fugelsang, J. (2005). Scientific thinking and reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 705–725). Cambridge, England: Cambridge University Press.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure mapping engine: Algorithm and examples. *Artificial Intelligence, 41,* 1–63.

Fenker, D. B., Waldmann, M. R., & Holyoak, K. J. (2005). Accessing causal relations in semantic memory. *Memory & Cognition, 33,* 1036–1046.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7,* 155–170.

Gentner, D., & Holyoak, K. J. (1997). Reasoning and learning by analogy: Introduction. *American Psychologist, 52,* 32–34.

Gentner, D., Loewenstein, J., & Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology, 95,* 393–408.

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

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15,* 1–38.

Goodman, N. (1955). *Fact, fiction, and forecast.* Cambridge, MA: Harvard University Press.

Green, A., Kraemer, D., Fugelsang, J., Gray, J., & Dunbar, K. (2010). Connecting long distance: Semantic distance in analogical reasoning modulates frontopolar cortex activity. *Cerebral Cortex, 20,* 70–76.

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

Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychological Review, 114,* 704–732.

Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences, 21,* 803–864.

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

Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 248–274). Oxford, England: Oxford University Press.

Hofstadter, D. R., & Mitchell, M. (1994). The Copycat Project: A model of mental fluidity and analogy-making. In K. J. Holyoak & J. A. Barnden (Eds.), *Advances in Connectionist and Neural Computation Theory: Vol. 2. Analogical connections* (pp. 31–112). Norwood, NJ: Ablex.

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

Holyoak, K. J. (1985). The pragmatics of analogical transfer. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 19, pp. 59–87). New York, NY: Academic Press.

Holyoak, K. J. (2005). Analogy. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 117–142). Cambridge, England: Cambridge University Press.

Holyoak, K. J., Junn, E. N., & Billman, D. O. (1984). Development of analogical problem-solving skill. *Child Development, 55,* 2042–2055.

Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory & Cognition, 15,* 332–340.

Holyoak, K. J., Novick, L. R., & Melz, E. R. (1994). Component processes in analogical transfer: Mapping, pattern completion, and adaptation. In K. J. Holyoak & J. A. Barnden (Eds.), *Advances in Connectionist and Neural Computation Theory: Vol. 2. Analogical connections* (pp. 113–180). Norwood, NJ: Ablex.

Holyoak, K. J., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science, 13,* 295–355.

Holyoak, K. J., & Thagard, P. (1995). *Mental leaps: Analogy in creative thought.* Cambridge, MA: MIT Press.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review, 104,* 427–466.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review, 110,* 220–264.

Keane, M. T. (1988). *Analogical problem solving.* Chichester, England: Ellis Horwood.

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

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

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2007). Learning causal schemata. In D. S. McNamara & G. Trafton (Eds.), *Proceedings of the*

Twenty-ninth Annual Conference of the Cognitive Science Society (pp. 389–394). Austin, TX: Cognitive Science Society.

Kemp, C., & Jern, A. (2009). Abstraction and relational learning. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems* (Vol. 22, pp. 934–942). Cambridge, MA: MIT Press.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review, 116,* 20–58.

Knowlton, B. J., & Holyoak, K. J. (2009). Prefrontal substrate of human relational reasoning. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences IV* (pp. 1005–1018). Cambridge, MA: MIT Press.

Kroger, J. K., Saab, F. W., Fales, C. L., Bookheimer, S. Y., Cohen, M. S., & Holyoak, K. J. (2002). Recruitment of anterior dorsolateral prefrontal cortex in human reasoning: A parametric study of relational complexity. *Cerebral Cortex, 12,* 477–485.

Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., & Sloman, S. A. (2007). Beyond covariation: Cues to causal structure. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 154–172). Oxford, England: Oxford University Press.

Lassaline, M. E. (1996). Structural alignment in induction and similarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 754–770.

Lee, H. S., & Holyoak, K. J. (2008). The role of causal models in analogical inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 1111–1122.

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

Lu, H., Yuille, A., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review, 115,* 955–982.

Markman, A. B. (1997). Constraints on analogical inference. *Cognitive Science, 21,* 373–418.

Meder, B., Mayrhofer, R., & Waldmann, M. R. (2009). A rational model of elementary diagnostic inference. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2176–2181). Austin, TX: Cognitive Science Society.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review, 100,* 254–278.

Medin, D. L., & Rips, L. J. (2005). Concepts and categories: Memory, meaning, and metaphysics. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 37–72). Cambridge, England: Cambridge University Press.

Mitchell, T. M., Keller, R. M., & Kedar-Cabelli, S. T. (1986). Explanation-based generalization. *Machine Learning, 1,* 47–80.

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

Murphy, G. L., & Ross, B. H. (2010). Uncertainty in category-based induction: When do people integrate across categories? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 263–276.

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

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning.* Oxford, England: Oxford University Press.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* San Mateo, CA: Morgan Kaufmann.

Peirce, C. S. (1932). *The collected papers of Charles Sanders Peirce: Vol. 2. Elements of logic* (C. Hartshorne & P. Weiss, Eds.). Cambridge, MA: Harvard University Press.

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences, 31,* 109–178.

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

Rehder, B. (2006). When similarity and causality compete in category-based property generalization. *Memory & Cognition, 34,* 3–16.

Rehder, B. (2007). Property generalization as causal reasoning. In A. Feeney & E. Heith (Eds.), *Inductive reasoning: Experimental, developmental, and computational approaches* (pp. 81–113). Cambridge, England: Cambridge University Press.

Rehder, B. (2009). Causal-based property generalization. *Cognitive Science, 33,* 301–343.

Rehder, B., & Burnett, R. C. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology, 50,* 264–314.

Rehder, B., & Kim, S. (2006). How causal knowledge affects classification: A generative theory of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 659–683.

Reichenbach, H. (1956). *The direction of time.* Berkeley: University of California Press.

Rips, L. J. (1990). Reasoning. *Annual Review of Psychology, 41,* 321–353.

Ross, B. H., & Murphy, G. L. (1996). Category-based predictions: Influence of uncertainty and feature associations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 736–753.

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

Spellman, B. A., & Holyoak, K. J. (1996). Pragmatics in analogical mapping. *Cognitive Psychology, 31,* 307–346.

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

Talalay, L. E. (1987). Rethinking the function of clay figurine legs from Neolithic Greece: An argument by analogy. *American Journal of Archaeology, 91,* 161–169.

Tenenbaum, J. B., Kemp, C., & Shafto, P. (2007). Theory-based Bayesian models of inductive reasoning. In A. Feeney & E. Heit (Eds.), *Inductive reasoning* (pp. 167–204). Cambridge, England: Cambridge University Press.

Waldmann, M. R., Cheng, P. W., Hagmayer, Y., & Blaisdell, A. P. (2008). Causal learning in rats and humans: A minimal rational model. In N. Chater & M. Oaksford (Eds.), *Rational models of cognition* (pp. 453–484). Oxford, England: Oxford University Press.

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

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

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

Waltz, J. A., Knowlton, B. J., Holyoak, K. J., Boone, K. B., Mishkin, F. S., de Menezes Santos, M., . . . Miller, B. L. (1999). A system for relational reasoning in human prefrontal cortex. *Psychological Science, 10,* 119–125.

Winston, P. (1980). Learning and reasoning by analogy. *Communications of the ACM, 23,* 689–703.

Appendix A

Derivation of Bayesian Model for Predictive Inference

We use Figure 4 as an example to illustrate the computation for predictive inference. $\boldsymbol{C}^{\mathrm{S}}$ denotes the information that the source (see Panel A) has a background generative cause, B^{S}, and three additional causal factors, G_1^{S}, G_2^{S}, and P_1^{S}, that is: $\boldsymbol{C}^{\mathrm{S}} = (B^{\mathrm{S}}, G_1^{\mathrm{S}}, G_2^{\mathrm{S}}, P_1^{\mathrm{S}})$. $\boldsymbol{C}^{\mathrm{T}}$ provides analogous information about possible causes in the target. Specific causal factors included in $\boldsymbol{C}^{\mathrm{T}}$ change depending on different experiment conditions. We will address each condition separately.

As shown in Equation A1 (the same as Equation 1 in the main article), in predictive inference, the model estimates the probability of an effect occurring in the target, $E^{\mathrm{T}} = 1$, on the basis of initial information about the source, $(\boldsymbol{C}^{\mathrm{S}}, E^{\mathrm{S}})$, and the target, $\boldsymbol{C}^{\mathrm{T}}$. The unknown causal strengths for the source and target are represented by $\boldsymbol{w}^{\mathrm{S}}$ and $\boldsymbol{w}^{\mathrm{T}}$, respectively. The basic equation for predictive inference includes three basic components, as elaborated in the following equation:

$$\begin{aligned} P(E^{\mathrm{T}} \mid \boldsymbol{C}^{\mathrm{T}}, E^{\mathrm{S}}, \boldsymbol{C}^{\mathrm{S}}) &= \sum_{\boldsymbol{w}^{\mathrm{T}}} P(E^{\mathrm{T}}, \boldsymbol{w}^{\mathrm{T}} \mid \boldsymbol{C}^{\mathrm{T}}, E^{\mathrm{S}}, \boldsymbol{C}^{\mathrm{S}}) \\ &= \sum_{\boldsymbol{w}^{\mathrm{T}}} P(E^{\mathrm{T}} \mid \boldsymbol{w}^{\mathrm{T}}, \boldsymbol{C}^{\mathrm{T}}) \\ &\quad \times \sum_{\boldsymbol{w}^{\mathrm{S}}} [P(\boldsymbol{w}^{\mathrm{T}} \mid \boldsymbol{w}^{\mathrm{S}}, E^{\mathrm{S}}, \boldsymbol{C}^{\mathrm{S}}, \boldsymbol{C}^{\mathrm{T}}) \\ &\quad \times P(\boldsymbol{w}^{\mathrm{S}} \mid \boldsymbol{C}^{\mathrm{S}}, E^{\mathrm{S}})]. \end{aligned} \quad \text{(A1)}$$

1. $P(\boldsymbol{w}^{\mathrm{S}} \mid \boldsymbol{C}^{\mathrm{S}}, E^{\mathrm{S}})$ captures the learning of a source model from observed contingency data (see Step 1 in Figure 3). Recent computational studies have developed detailed models that estimate distributions of causal strength by combining priors and observations (Griffiths & Tenenbaum, 2005; Lu et al., 2008). Using Figure 4 as an example, this term can be computed by combining likelihoods and priors using Bayes rule as follows:

$$P(\boldsymbol{w}^{\mathrm{S}} \mid \boldsymbol{C}^{\mathrm{S}}, E^{\mathrm{S}}) = \frac{P(E^{\mathrm{S}} \mid \boldsymbol{C}^{\mathrm{S}}, \boldsymbol{w}^{\mathrm{S}}) P(\boldsymbol{w}^{\mathrm{S}})}{P(E^{\mathrm{S}} \mid \boldsymbol{C}^{\mathrm{S}})}. \quad \text{(A2)}$$

We adopt the likelihood calculation based on noisy-OR and noisy-AND-NOT functions as specified by the power PC theory (Cheng, 1997) as follows:

$$P(E^{\mathrm{S}} \mid \boldsymbol{C}^{\mathrm{S}}, \boldsymbol{w}^{\mathrm{S}}) = (1 - (1 - \boldsymbol{w}_0^{\mathrm{S}})(1 - \boldsymbol{w}_1^{\mathrm{S}})(1 - \boldsymbol{w}_2^{\mathrm{S}}))(1 - \boldsymbol{w}_3^{\mathrm{S}}), \quad \text{(A3)}$$

where $\boldsymbol{w}_0$ indicates the causal strength of background cause, $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ are the strengths associated with the two generative causes, and $\boldsymbol{w}_3$ indicates the strength of the preventive cause. Equation A3 applies in the case in which the source analog consists of a single instance. In the general case (Lu et al., 2008), where the source is based on N independent instances, the right-hand side of Equation A3 is raised to the power of N. The prior of $P(\boldsymbol{w}^{\mathrm{S}})$ before observing any data assumes a uniform distribution. The denominator is a normalization term to ensure that the calculated probability is bounded within the range of 0 and 1. This term is calculated by summing over all possible values $\boldsymbol{w}^{\mathrm{S}}$ for the numerator.

2. $P(\boldsymbol{w}^{\mathrm{T}} \mid \boldsymbol{w}^{\mathrm{S}}, E^{\mathrm{S}}, \boldsymbol{C}^{\mathrm{S}}, \boldsymbol{C}^{\mathrm{T}})$ quantifies knowledge transfer on the basis of analogical mapping (see Steps 2 and 3 in Figure 3). As described in the article, we model the probability of transfer as

$$\begin{cases} P(\boldsymbol{w}_j^{\mathrm{T}} = \boldsymbol{w}_i^{\mathrm{S}}) = 1, & \text{if the } j\text{th target variable} \\ P(\boldsymbol{w}_j^{\mathrm{T}} = \boldsymbol{w}_i^{\mathrm{S}}) = 0, & \text{matches the } i\text{th source variable and otherwise.} \end{cases} \quad \text{(A4)}$$

3. $P(E^{\mathrm{T}} \mid \boldsymbol{w}^{\mathrm{T}}, \boldsymbol{C}^{\mathrm{T}})$ uses knowledge derived from analogical transfer and observations about the presence of causal factors in the target to estimate the probability of the effect in the target (see Step 4 in Figure 3). This term varies depending on the different information about causal factors provided in the target, following the basic principles of the Bayesian version of the power PC theory (Cheng, 1997; Lu et al., 2008). Here we provide the derivation for each of the three experimental conditions tested in Experiment 1. All derivations are analytic, calculated using Matlab programs (Lu et al., 2008).

3a. *Case 1: Target G_1G_2P, depicted in Figure 4B.* The probability of the effect in the target can be computed as

$$P(E^{\mathrm{T}} \mid \boldsymbol{w}^{\mathrm{T}}, \boldsymbol{C}^{\mathrm{T}}) = [1 - (1 - \boldsymbol{w}_0^{\mathrm{T}})(1 - \boldsymbol{w}_1^{\mathrm{T}})(1 - \boldsymbol{w}_2^{\mathrm{T}})](1 - \boldsymbol{w}_3^{\mathrm{T}}). \quad \text{(A5)}$$

3b. *Case 2: Target G_1G_2, which drops the preventer in the target.* Given that the background cause always occurs, the probability of the effect in the target is calculated as

$$P(E^{\mathrm{T}} \mid \boldsymbol{w}^{\mathrm{T}}, \boldsymbol{C}^{\mathrm{T}}) = 1 - (1 - \boldsymbol{w}_0^{\mathrm{T}})(1 - \boldsymbol{w}_1^{\mathrm{T}})(1 - \boldsymbol{w}_2^{\mathrm{T}}). \quad \text{(A6)}$$

3c. *Case 3: Target G_2P, which drops a generative cause in the target.* We compute the effect probability as

$$P(E^{\mathrm{T}} \mid \boldsymbol{w}^{\mathrm{T}}, \boldsymbol{C}^{\mathrm{T}}) = [1 - (1 - \boldsymbol{w}_0^{\mathrm{T}})(1 - \boldsymbol{w}_2^{\mathrm{T}})](1 - \boldsymbol{w}_3^{\mathrm{T}}). \quad \text{(A7)}$$

The probability of the effect in the target G_1G_2P can be estimated by substituting Equations A3, A4, and A5 into Equation A1. Similar calculations can be made to estimate the probability of the effect in the target when it includes different information (e.g., targets G_1G_2 and G_2P).

(*Appendices continue*)

Appendix B

Derivation of Bayesian Model for Causal Attribution

Here we illustrate the derivation for causal attribution using the problem shown in Figure 4 as an example. The input to the model includes the initial information in the source (see Panel A) and the target (see Panel C), ($\boldsymbol{C}^{\mathrm{S}}$, E^{S}, $\boldsymbol{C}^{\mathrm{T}}$, E^{T}), in which $\boldsymbol{C}^{\mathrm{T}}$ denotes the known causal factors in the target, that is, $\boldsymbol{C}^{\mathrm{T}} = (B^{\mathrm{T}}, G_2^{\mathrm{T}}, P_1^{\mathrm{T}})$, and does not include the unknown causal factor G_1^{T}. The goal of the model is to predict the probability that the cause G_1^{T} was present and produced the effect in the target, which is given by Equation B1 (the same as Equation 6 in the main article):

$$P(G_1^{\mathrm{T}} = 1,\, G_1^{\mathrm{T}} \rightarrow E^{\mathrm{T}} \mid E^{\mathrm{T}}, \boldsymbol{C}^{\mathrm{T}}, E^{\mathrm{S}}, \boldsymbol{C}^{\mathrm{S}}) = \frac{P(G_1^{\mathrm{T}} = 1,\, G_1^{\mathrm{T}} \rightarrow E^{\mathrm{T}}, E^{\mathrm{T}} \mid \boldsymbol{C}^{\mathrm{T}}, E^{\mathrm{S}}, \boldsymbol{C}^{\mathrm{S}})}{P(E^{\mathrm{T}} \mid \boldsymbol{C}^{\mathrm{T}}, E^{\mathrm{S}}, \boldsymbol{C}^{\mathrm{S}})}$$

$$= \frac{P(G_1^{\mathrm{T}} = 1 \mid \boldsymbol{C}^{\mathrm{T}}, E^{\mathrm{S}}, \boldsymbol{C}^{\mathrm{S}}) P(G_1^{\mathrm{T}} \rightarrow E^{\mathrm{T}}, E^{\mathrm{T}} \mid G_1^{\mathrm{T}} = 1, \boldsymbol{C}^{\mathrm{T}}, E^{\mathrm{S}}, \boldsymbol{C}^{\mathrm{S}})}{P(E^{\mathrm{T}} \mid \boldsymbol{C}^{\mathrm{T}}, E^{\mathrm{S}}, \boldsymbol{C}^{\mathrm{S}})}. \quad \text{(B1)}$$

We now provide a detailed explanation of the computation for each term in the Equation B1. The attribution derivations are analytic, calculated using Matlab programs.

1. The first term in the numerator, $P(G_1^{\mathrm{T}} = 1 \mid \boldsymbol{C}^{\mathrm{T}}, E^{\mathrm{S}}, \boldsymbol{C}^{\mathrm{S}})$, is the base rate of the causal factor in the target (i.e., an estimate of the probability that the causal factor would occur in the target environment). This base rate can be estimated on the basis of standard counting probability, using the binomial distribution. The qualitative result is that, after observing four causal factors to occur in the source, the probability of G_1^{T} occurring increases with the number of other causal factors observed to occur in the target. As a result, the estimated base rate of the cause is determined by the number of causes observed in the source and the target.

We assume in the following equation that all the causes share an equal probability for their occurrence but that this probability, denoted as θ, is unknown:

$$P(G_1^{\mathrm{T}} = 1 \mid \boldsymbol{C}^{\mathrm{T}}, E^{\mathrm{S}}, \boldsymbol{C}^{\mathrm{S}}) = \int P(G_1^{\mathrm{T}} = 1 \mid \theta) P(\theta \mid \boldsymbol{C}^{\mathrm{T}}, \boldsymbol{C}^{\mathrm{S}}) d\theta$$

$$= \int P(G_1^{\mathrm{T}} = 1 \mid \theta) \frac{P(\boldsymbol{C}^{\mathrm{T}} \mid \theta) P(\theta \mid \boldsymbol{C}^{\mathrm{S}})}{Z} d\theta, \quad \text{(B2)}$$

where $P(G_1^{\mathrm{T}} = 1 \mid \theta) = \theta$ and $P(\boldsymbol{C}^{\mathrm{T}} \mid \theta) = \theta^{n_{\mathrm{T}}}$ (n_{T} indicates the number of known causes in the target) following the binomial distribution for a binary variable and with the occurrence of causes assumed to be independent. $P(\theta \mid \boldsymbol{C}^{\mathrm{S}})$ is the learned base rate of the causal factor based on information provided in the source. Using the Bayes rule, this probability can be evaluated as $P(\theta \mid \boldsymbol{C}^{\mathrm{S}}) = (n_{\mathrm{S}} + 1)\theta^{n_s}$ by assuming a uniform distribution on θ as the prior (n_s indicates the number of known causes in the source). Z is the normalization constant to ensure the calculated probability is bounded within 0 and 1.

Following the calculation just laid out, the estimated base rate of the cause for the G_2PE condition, in which there are three causes (G_1, G_2, and P) in the source and two causes (G_2, and P) in the target, is 6/7. The base rate of the cause for both the G_2E and PE conditions is 5/7.

2. The second term in the numerator of Equation B1, $P(G_1^{\mathrm{T}} \rightarrow E^{\mathrm{T}}, E^{\mathrm{T}} \mid G_1^{\mathrm{T}} = 1, \boldsymbol{C}^{\mathrm{T}}, E^{\mathrm{S}}, \boldsymbol{C}^{\mathrm{S}})$, serves to quantify the probability of a predictive inference (i.e., how likely the effect in the target can be produced by G_1^{T} given the information in the source). This is a predictive inference problem, so the principle used to compute this probability is the same as described in Appendix A, that is,

$$P(G_1^{\mathrm{T}} \rightarrow E^{\mathrm{T}}, E^{\mathrm{T}} \mid G_1^{\mathrm{T}} = 1, \boldsymbol{C}^{\mathrm{T}}, E^{\mathrm{S}}, \boldsymbol{C}^{\mathrm{S}}) = \sum_{\boldsymbol{w}_1^{\mathrm{T}}} P(E^{\mathrm{T}} \mid \boldsymbol{w}_1^{\mathrm{T}}, \boldsymbol{C}_1^{\mathrm{T}}) \sum_{\boldsymbol{w}_1^{\mathrm{S}}} [P(\boldsymbol{w}_1^{\mathrm{T}} \mid \boldsymbol{w}_1^{\mathrm{S}}, \boldsymbol{C}_1^{\mathrm{S}}, \boldsymbol{C}_1^{\mathrm{T}}) P(\boldsymbol{w}_1^{\mathrm{S}} \mid \boldsymbol{C}^{\mathrm{S}}, E^{\mathrm{S}})]. \quad \text{(B3)}$$

3. The denominator in Equation B1, $P(E^{\mathrm{T}} \mid \boldsymbol{C}^{\mathrm{T}}, E^{\mathrm{S}}, \boldsymbol{C}^{\mathrm{S}})$, is calculated by the weighted sum of the probability of the effect occurring in the presence and the absence of the cause G_1^{T}. The weights are determined by the estimate of the base rate of this causal factor, as shown in Equation B2. Specifically,

$$P(E^{\mathrm{T}} \mid \boldsymbol{C}^{\mathrm{T}}, E^{\mathrm{S}}, \boldsymbol{C}^{\mathrm{S}}) = \sum_{G_1^{\mathrm{T}}} P(E^{\mathrm{T}}, G_1^{\mathrm{T}} \mid \boldsymbol{C}^{\mathrm{T}}, E^{\mathrm{S}}, \boldsymbol{C}^{\mathrm{S}})$$

$$= P(E^{\mathrm{T}} \mid G_1^{\mathrm{T}} = 1, \boldsymbol{C}^{\mathrm{T}}, E^{\mathrm{S}}, \boldsymbol{C}^{\mathrm{S}}) P(G_1^{\mathrm{T}} = 1 \mid \boldsymbol{C}^{\mathrm{T}}, E^{\mathrm{S}}, \boldsymbol{C}^{\mathrm{S}})$$

$$+ P(E^{\mathrm{T}} \mid G_1^{\mathrm{T}} = 0, \boldsymbol{C}^{\mathrm{T}}, E^{\mathrm{S}}, \boldsymbol{C}^{\mathrm{S}}) P(G_1^{\mathrm{T}} = 0 \mid \boldsymbol{C}^{\mathrm{T}}, E^{\mathrm{S}}, \boldsymbol{C}^{\mathrm{S}}). \quad \text{(B4)}$$

The calculation of this term makes it clear that causal attribution judgments are more computationally demanding than are causal predictions, because of the requirement to estimate the effect probability in two alternative situations (i.e., the presence and the absence of the unknown causal factor).

3a. *Case 1: Generative causes only.* This case includes the G_2E condition in Experiment 2. Case 1 yields causal discounting, as the causal attribution to the unknown causal factor is diminished if another generative causal factor is known to have occurred. The calculation of the denominator term is straightforward. We can follow Appendix A to estimate the probabilities of the target effect in two situations, when (a) both G_1 and G_2 causes occur and (b) only G_2 occurs. Then we calculate the weighted sum of the two predictive probabilities, weighted by the estimated base rate of the target cause given by Equation B2.

(*Appendices continue*)

3b. *Case 2: Generative causes combined with a preventive cause.* This case includes the G_2PE and PE conditions in Experiment 2. Case 2 is essentially the same as Case 1, except that it is necessary to take account of how the influence of a preventive cause is combined with those of generative causes. Whereas the noisy-OR operator for generative causes is invariant across the order in which causes are combined (i.e., satisfies the property of associativity), the noisy-AND-NOT operator for preventive causes is not. The net outcome produced by a set of generative causes accompanied by a preventive cause therefore depends on the order in which causes are combined. Novick and Cheng (2004) distinguished between sequential and parallel integration in certain situations involving interactive causes, arguing that sequential integration is generally the default. By analogy, we derive causal attribution for Case 2 under the assumption that integration of causes is performed sequentially (Carroll & Cheng, 2009). Sequential processing was encouraged in our experiments by the form of the question, which first stated that certain causal factors (sometimes including a preventive factor) had occurred along with the effect and then queried a further generative causal factor of unknown status. We assume that, following the same sequential order, the preventive noisy-AND-NOT operator is first applied to the net generative influence of those causal factors stated to be present; the net output resulting from this operation is then in turn combined with the influence of the generative cause being queried. This sequential procedure implies that the estimated impact of known generative factors is diminished by the preventive cause, whereas the estimated impact of the queried generative factor is not. This asymmetry in the impact of the preventive cause will diminish causal discounting. That is, given that one generative cause is known to be present, causal attribution to a further generative cause of unknown status will be greater if a preventive cause is present than if it is not.

We use the PE condition as an example to illustrate the operation of this sequential procedure. The calculation of the numerator in Equation B1 is the same as described earlier and is not affected by the sequential procedure. However, the sequential procedure will play a role in assessing the predictive probabilities of the effect in the denominator of Equation B1. The model needs to assess the probability of the effect in two situations: (a) when G_1 and P occur together with the background cause and (b) when only P occurs with the background cause. The sequential process affects the calculation in the first case, because it determines the combination order. The noisy-AND-NOT function is used to integrate the expected influence of the background cause B and preventer P, after which this output is integrated with the influence of G_1 using the noisy-OR function to assess the probability of the effect in the target on the basis of transferred causal knowledge (see Step 4 in Figure 3), that is,

$$P(E^{\mathrm{T}} \mid \boldsymbol{w}^{\mathrm{T}}, \boldsymbol{C}^{\mathrm{T}}) = 1 - (1 - \boldsymbol{w}_0^{\mathrm{T}}(1 - \boldsymbol{w}_3^{\mathrm{T}}))(1 - \boldsymbol{w}_1^{\mathrm{T}}). \quad \text{(B5)}$$

The same principle applies for the G_2PE condition and all other situations involving causal attribution judgments about an unknown generative cause in the presence of a known preventer.

Received November 9, 2009
Revision received May 31, 2010
Accepted June 1, 2010 ■

Showcase your work in APA's newest database.

PsycTESTS®

Make your tests available to other researchers and students; get wider recognition for your work.

"PsycTESTS is going to be an outstanding resource for psychology," said Ronald F. Levant, PhD. "I was among the first to provide some of my tests and was happy to do so. They will be available for others to use—and will relieve me of the administrative tasks of providing them to individuals."

Visit http://www.apa.org/pubs/databases/psyctests/call-for-tests.aspx
to learn more about PsycTESTS and how you can participate.

Questions? Call 1-800-374-2722 or write to tests@apa.org.

Not since PsycARTICLES has a database been so eagerly anticipated!