CHAPTER 12

Causal Learning

Patricia W. Cheng *and* Marc J. Buehner

Abstract

This chapter is an introduction to the psychology of causal inference using a computational perspective, with the focus on causal discovery. It explains the nature of the problem of causal discovery and illustrates the goal of the process with everyday and hypothetical examples. It reviews psychological research under two approaches to causal discovery, an associative approach and a causal approach that incorporates causal assumptions in the inference process. The latter approach provides a framework within which to answer different questions regarding causal inference coherently. The chapter ends with a consideration of causality as unfolding over time. We conclude with a sketch of future directions for the field.

Key Words: causal learning, rationality, associative models, causal models, causal Bayes nets, Bayesian causal models, temporal contiguity, causal invariance, intervention, empirical knowledge

Why Causality?

The central question of this chapter is: "How can any intelligent system put on Planet Earth, if given cognitive resources and types of information similar to those available to us, discover how the world works so that it can best achieve its goals?" Before we attempt to answer this question, let us imagine that our cognition were different in various respects. First, suppose we were unable to learn associations between events (i.e., detect statistical regularity in the occurrence of events). We would be unable to predict any events, causal or otherwise. For example, we would be unable to predict that if the traffic light turns red, we should stop or an accident is likely to happen, or that if we hear a knock on our door, someone will be on the other side when we open the door. Nor would we be able to predict the weather, even imperfectly. We would be unable to learn the sequences of sound in language or music, or the meaning of words. We would behave as if we had prosopagnosia, unable to relate our parents' faces, or the face of the person we have been dating for the past month, to their past history. A nonassociative world would be grim.

Now, imagine a world where we were able to learn associations but unable to reason about causes and effects. What would it be like? In that world, for a child growing up on a farm who has always experienced sunrise after the rooster crows, if the rooster is sick one morning and does not crow, she would predict that the sun would not rise until the rooster crows. Similarly, if the rooster has been deceived into crowing, say, by artificial lighting in the middle of the night, the child would predict that the sun would rise soon after. Notice that under normal conditions noncausal associations do enable one to reliably predict a subsequent event from an observation (e.g., sunrise from a rooster's crowing, a storm soon to come from a drop in the barometric reading, or from ants migrating uphill). They do not, however, support predictions about the event (e.g., sunrise) when the observation (crowing or no

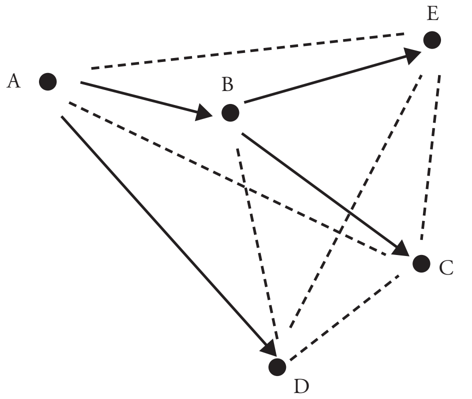


Fig. 12.1 A causal tree and the implied associative links. Nodes represent variables, labeled by letters. Arrows represent direct causal links. Dotted lines represent implied associations, which include indirect causal links as well as associations between direct and indirect effects of a common cause.

crowing) is produced by an action or an extraneous cause (respectively, artificial light and the rooster's sickness). An associative world without causation would be exasperating.

Consider how often we would be wrong, and how inefficient we would be, were we to store all associations, both causal and noncausal. We illustrate the problem with the causal tree in Figure 12.1. In the figure, each node represents a variable, and each arrow represents a causal relation. There are four causal links, but six additional associations (the dotted lines).[1] These additional associations can be inferred from the causal links, and thus are redundant. In an associative world, if information on any of the six extra associations is salient (e.g., as information on a rooster's crowing and sunrise might be), they would be indistinguishable from the causal links. Thus, not only would the extra associations be inefficient to store, many would yield erroneous predictions based on actions. For example, node D in the figure is linked by a single arrow, but by three additional associations, to nodes B, C, and E; manipulating any of these variables would not lead to D, the "desired" outcome predicted by these three associations.

Finally, not only would we be unable to achieve our goals, but we would be unable to structure an otherwise chaotic flux of events into meaningful episodes. We explain events by causation. Returning to our storm example, whereas we might explain that a car skidded and rolled down the mountainside because of the rainstorm, it would be odd to explain that the car skidded because of the low barometric reading. Causal explanations are universal, as anthropologists who study everyday narratives across cultures have observed; they serve to imbue life events with an orderliness, to demystify unexpected events, and establish coherence (Ochs & Capps, 2001).

Predicting the Consequences of Actions to Achieve Goals: A Framework for Causal Learning, Category Formation, and Hypothesis Revision

For the just mentioned reasons, it is easy to see why it is important to distinguish between causation and mere association. A less obvious reason, one that has implications for the formulation of the problem to be solved, is that whereas associations are observable, causal relations are inherently unobservable and can only be inferred. For example, one can observe the number of lung cancer patients among cigarette smokers and among nonsmokers and see the association between cigarette smoking and lung cancer, but the association can be causal or noncausal. It may, for example, be due to confounding by the higher incidence of radon in the smokers' dwellings, and the exposure to radon is what caused the smokers' lung cancer. The challenge is how to encapsulate causal relations, even though they are unobservable, so that the causal knowledge can be applied to best achieve desired outcomes.

We have so far implicitly assumed that cause-and-effect variables, such as "sunrise," "drop in barometric reading," and "storm," are predefined, given to the causal learner, and only the relations between them are to be discovered. A more realistic description of the situation is: In order for our causal knowledge to be generalizable from the learning context (e.g., prior experience, whether one's own or that of others, shows that icy roads cause skidding) to the application context (I don't want my car to skid, so I will wait until the ice has melted before I drive), we construct a representation of the world in which cause-and-effect variables are so defined that they enable ideally invariant causal relations to be constructed. As the philosopher C. I. Lewis (1929) observed, "Categories are what obey laws." Defining the objects, events, and categories linked by causal relations is part of the problem of causal discovery. Fortunately for cognitive psychologists studying human causal discovery, some of the work defining variables is already taken care of by evolution, prewired into our perceptual system and emotions. For example, we see a rooster as figure against the ground of the farm landscape. Strong gusts of wind alarm us, and getting wet in the rain is unpleasant. But there is definitely remaining work; for example, what determines that "ants migrating uphill" should be defined as a variable?

Whenever we apply causal knowledge to achieve a goal, we are assuming that the causal relations in question remain invariant from the learning context to the application context. Because of our limited causal knowledge, however, a causal relation that we assume to be invariant would no doubt in fact often change (e.g., a scientist might hypothesize "vitamin E has antioxidant effects" but find instead that whereas natural foods rich in vitamin E have antioxidant effects, vitamin E pills do not). The assumption of causal invariance in our everyday application of causal knowledge might seem too strong. Although simplistic as a static hypothesis, however, this assumption is rational as a defeasible default within the dynamic process of hypothesis testing and hypothesis revision. Given our limited access to information at any particular moment, the criterion of causal invariance serves as a compass aimed at formulating the simplest explanation of a phenomenon that allows invariance to obtain (e.g., the scientist might search for other substances in natural foods that in conjunction with vitamin E consistently produce the effects). Observed deviation from the default indicates a need for hypothesis revision, a change in direction aimed at capturing causal invariance (Carroll & Cheng, 2010).

Cheng (2000) showed that an alternative assumption that would also justify generalization of a causal relation regarding an outcome to a new context is that enabling conditions (causal factors in the contextual background interacting with the hypothesized cause) and preventive causes that occur in the background (all of which are often unobserved) occur just as frequently in the generalization context as in the learning context. She also showed that with respect to the accuracy of generalization to new contexts, the two assumptions are equivalent. In the rest of our chapter, we use causal invariance (which we term "independent causal influence" when we define it mathematically) because it is the simpler of the two equivalent conceptions.

Now that we have considered some goals and constraints of causal inference, let us rephrase the question of causal learning with respect to those goals and constraints: How can any intelligent agent given the information and resources available to humans discover ideally invariant causal relations that support generalization from the learning context to an application context? In particular, would it suffice to have a powerful statistical process that detects regularities among events but lacks any a priori assumptions about how the world works? Because humans have limited access to information, an accompanying question is, What hypothesis testing and revision process would allow the ideally invariant causal relations to be constructed?

By posing our question in terms of discovery, we by no means rule out the possibility that there exist some innate domain-specific biases. Classic studies by Garcia, McGowan, Ervin, and Koelling (1968) demonstrated two such biases. In each of four groups of rats, one of two cues, either a novel size or a novel flavor of food pellets, was conditionally paired with either gastrointestinal malaise induced by X-ray or with pain induced by electrical shock. The combination of flavor and illness produced a conditioned decrement in the amount consumed but that of the size of the pellet and illness did not. Conversely, the combination of size and pain produced hesitation before eating, but flavor and pain did not. Apparently, the novelty had to be of the right kind for effective causal learning regarding the malaise and shock to occur.

Most of what we do know about the world, however, must have been acquired due to experience. How else could we have come to know that exposure to the sun causes tanning in skin but causes bleaching in fabrics? Or come to know that billiard ball A in motion hitting billiard ball B at rest would not jump over B, rebound leaving B still, or explode (Hume, 1739/1888)? Notice that it is not necessary for the causal learner to know *how* sunlight causes tanning in skin or bleaching in fabrics to discover that it does. Neither was it necessary, for that matter, for the rats to know *how* the X-ray or electricity caused their respective effects for their learning to occur. We will return to the issue of adding intervening nodes in a causal network to explain how an outcome is achieved via a causal mechanism.

Proposed Solutions: Two Dominant Approaches

We have only gone so far as posing the problem to be solved. Hopefully, posing the problem clearly will mean much of the work has been done. In the rest of this chapter, we review proposed solutions according to two dominant approaches: the associative approach, including its statistical variants, and the causal approach. We follow each of these accounts with a review of main empirical tests of the approach. (For a discussion of how the perceptual view [Michotte, 1946/1963], the mechanism view [Ahn, Kalish, Medin, & Gelman, 1995], and the coherence view [Thagard, 2000] relate to these approaches, see Buehner and Cheng [2005].) We then broaden our scope to consider the role of temporal information in causal learning. We end the chapter with a sketch of future research directions.

The Associative Approach

An intuitive approach that has dominated psychological research on learning is the associative approach (e.g., Allan & Jenkins, 1980; Jenkins & Ward, 1965; Rescorla & Wagner, 1972), which traces its roots to the philosopher David Hume (1739/1888). Hume made a distinction between analytic and empirical knowledge, and argued that causal knowledge is empirical. Only experience tells us what effect a cause has. The strong conviction of causality linking two constituent events is but a mental construct. The problem of causal learning posed by Hume radically shaped subsequent research on the topic and set the agenda for the study of causal learning from a cognitive science perspective. Both the associative and causal approaches are predicated on his posing of the problem.

To Hume, the relevant observed aspects of experience that give rise to the mentally constructed causal relations were the repeated association between the observed states of a cause and its effect, their temporal order and contiguity, and spatial proximity. Our examples have illustrated that one can predict a future event from a *covariation*—the concerted variation among events—provided that causes of that event remain unperturbed. Predictions of this kind are clearly useful; we appreciate weather reports, for example. To early associative theorists, causality is nothing more than a fictional epiphenomenon floating unnecessarily on the surface of indisputable facts.[2] After all, causal relations are unobservable. In fact, Karl Pearson, one of the fathers of modern statistics, subscribed to a positivist view and concluded that calculating correlations is the ultimate and only meaningful transformation of evidence at our disposal: “Beyond such discarded fundamentals as ‘matter’ and ‘force’ lies still another fetish amidst the inscrutable arcana of modern science, namely, the category of cause and effect” (Pearson, 1911, p. iv).

But, as we saw earlier, mere associations are inadequate for predicting the consequences of actions and would also be inefficient to store. Thus, in addition to dissecting the traditional associative view to understand its shortcomings, we will also consider a more viable augmented variant of the associative view, one similar to how scientists infer causation. The augmented view assumes that rational causal learning requires not only a sophisticated detector of covariations among events but also the use of actions as a causality marker: When the observed states of events are obtained by an action, by oneself or others, intervening in the normal course of events, the observed associations are causal; otherwise, they are noncausal. After all, one can observe that actions are what they are; there is therefore no deviation from Hume’s constraint that causal discovery begins with observable events as input. In entertaining this variant, we are taking the perspective of the design of an intelligent causal learner on our planet, rather than adhering to how the associative view has been traditionally interpreted. This more viable variant of the associative view implicitly underlies the use of associative statistics in typical tests of causal hypotheses in medicine, business, and other fields. It retains the strong appeal of the associative approach, namely, its objectivity. Other things being equal, positing unobservable events, as the causal view does, seems objectionable.

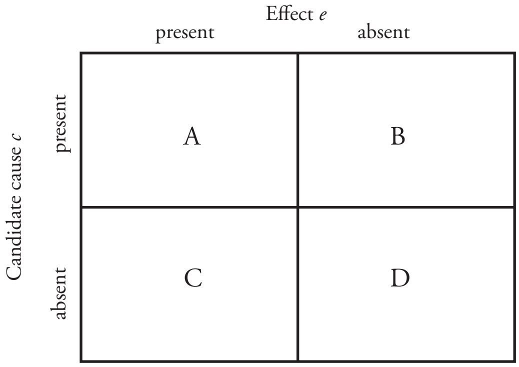
A growing body of research is dedicated to the role of intervention in causal learning, discovery, and reasoning (e.g., Blaisdell, Sawa, Leising, & Waldmann, 2006; Gopnik et al., 2004; Lagnado & Sloman, 2004; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003). Indeed, the general pattern reported is that observations based on intervention allow causal inferences that are not possible based on mere observations.

A STATISTICAL MODEL

For situations involving only one varying candidate cause, an influential decision rule for more than four decades has been the ΔP rule:

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

according to which the strength of the relation between binary causes *c* and effects *e* is determined by their *contingency* or *probabilistic contrast*—the difference between the probabilities of *e* in the presence and absence of *c* (see, e.g., Allan & Jenkins, 1980; Jenkins & Ward, 1965). ΔP is estimated by relative frequencies. In our equations, we denote the



		Effect *e* present	Effect *e* absent
Candidate cause *c*	present	A	B
	absent	C	D

Fig. 12.2 A standard 2 x 2 contingency table; *a* through *d* are labels for event types resulting from factorial combination of the presence and absence of cause *c* and effect *e*.

"presence" value of a binary variable by a "+" superscript and the "absence" value by a "–" superscript (e.g., c+ denotes the presence of c). Figure 12.2 displays a standard *contingency table* where cells *A* and *B respectively,* represent the frequencies of the occurrence, and nonoccurrence, of *e* in the presence of *c;* cells *C* and *D* represent, respectively, the frequencies the occurrence, and of nonoccurrence, of *e* in the absence of *c*.

If ΔP is noticeably positive, then *c* is thought to produce *e*; if it is noticeably negative, then *c* is thought to prevent *e*; and if ΔP is not noticeably different from zero, then *c* and *e* are thought not to be causally related to each other. Several modifications of the ΔP rule to include various parameters have been proposed (e.g., Anderson & Sheu, 1995; Perales & Shanks, 2007; Schustack & Sternberg, 1981; White, 2002). By allowing extra degrees of freedom, these modified models fit certain aspects of human judgment data better than the original rule. Another type of modification is to compute the ΔP value of a candidate cause conditioned on constant values of alternative causes (Cheng & Holyoak, 1995). This modification allows the model to better account for the influence of alternative causes (as illustrated later). Like all other psychological models of causal learning, all variants of the ΔP model assume that the candidate causes are perceived to occur before the effect in question.

AN ASSOCIATIONIST MODEL

In the domain of animal learning, an organism's capacity to track contingencies in its environment has long been of central interest, and apparent parallels between conditioning and causal learning have led many researchers (see Shanks & Dickinson, 1987; for a review see De Houwer & Beckers, 2002) to search for explanations of human causal learning in neural-network models that specify the algorithm of learning. The most influential associationist theory, the Rescorla-Wagner (RW) model (Rescorla & Wagner, 1972), and all its later variants, is based on an algorithm of error correction driven by a discrepancy between the expected and actual outcomes. For each learning trial where a cue was presented the model specifies

$$\Delta V_{CS} = \alpha_{CS}\, \beta_{US}\, (\lambda - \Sigma V) \tag{2}$$

where ΔV is the change in the strength of a given CS-US association on a given trial (CS stands for conditioned stimulus, e.g., a tone; US stands for unconditioned stimulus, e.g., a foot-shock), α and β represent learning rate parameters reflecting the saliencies of the CS and US, respectively, λ stands for the actual outcome of each trial (usually 1.0 if it is present and 0 if it is absent), and ΣV is the expected outcome defined as the sum of all associative strengths of all CSs present on that trial.

For situations involving only one varying cue, its mean weight at equilibrium according to the RW algorithm has been shown to equal ΔP if the value of β remains the same when the US is present and when it is absent for the λ values just mentioned (Chapman & Robins, 1990; Danks, 2003). In other words, this simple and intuitive algorithm elegantly explains why causal learning is a function of contingency. It also explains a range of results for designs involving multiple cues, such as *blocking* (see section on "Blocking" to follow), *conditioned inhibition*, *overshadowing*, and *cue validity* (Miller, Barnet, & Grahame, 1995).

Blocking: Illustrating an Associationist Explanation

"Blocking" (Kamin, 1969) occurs after a cue (A) is established as a perfect predictor (A+, with "+" representing the occurrence of the outcome), followed by exposure to a compound consisting of A and a new, redundant, cue B. If AB is also always followed by the outcome (AB+), cue B receives very little conditioning; its conditioning is *blocked* by cue A. According to RW, A initially acquires the maximum associative strength supported by the stimulus. Because the association between A and the outcome is already at asymptote when B is introduced, there is no error left for B to explain, hence the lack of conditioning to B. What RW computes is the ΔP for B conditioned on the constant presence of A. Shanks (1985) replicated the same finding in a causal reasoning experiment with human participants, although the human responses seemed to reflect uncertainty of the causal status of cue B rather than certainty that it is noncausal (e.g., Waldmann & Holyoak, 1992).

Failure of the RW Algorithm to Track Covariation When a Cue Is Absent

However, Shanks' (1985) results also revealed evidence for *backward* blocking; in fact, there is evidence for backward blocking even in young children (Gopnik et al., 2004). In this procedure, the order of learning phases is simply reversed; participants first learn about the perfect relation between AB and the outcome (AB+), and subsequently learn that A by itself is also a perfect predictor (A+). Conceptually, forward and backward blocking are identical, at least from a causal perspective. A causal explanation might go: If one knows that A and B together always produce an effect, and one also knows that A by itself also always produces the effect, one can infer that A is a strong cause. B, however, might be a cause, even a strong one, or noncausal; its causal status is unclear. Typically, participants express such uncertainty with low to medium ratings relative to ratings for control cues that have been paired with the effect an equal number of times (see Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008, for a review).

Beyond increasing susceptibility to attention and memory biases (primacy and recency; see, e.g., Dennis & Ahn, 2001), there is no reason why the temporal order in which knowledge about AB and A is acquired should play a role from a rational standpoint. This is not so for the RW model, however. The model assumes that the strength of a cue can only be updated when that cue is present. In the backward blocking paradigm, however, participants *retrospectively* alter their estimate of B on the A+ trials in phase 2. In other words, the ΔP of B, conditioned on the presence of A, decreases over a course of trials in which B is actually absent, and the algorithm therefore fails to track its covariation.

Several modifications of RW have been proposed to allow the strengths of absent cues to be changed, for instance, by setting the learning parameter α negative on trials where the cue is absent: Van Hamme and Wasserman's (1994) modified RW model, Dickinson and Burke's modified sometimes-opponent-process model (1996), and the comparator hypothesis (Denniston, Savastano, & Miller, 2001; Miller & Matzel, 1988; Stout & Miller, 2007). Such modifications can explain backward blocking and some other findings showing retrospective revaluation (for an extensive review of modifications to associative learning models applicable to human learning see De Houwer & Beckers, 2002). But these modifications also oddly predict that one will have difficulty learning that there are multiple sufficient causes of an effect. For example, if one drinks tea by itself and finds it quenching, but one sometimes drinks both tea and lemonade, then learning subsequently that lemonade alone can quench thirst will cause one to unlearn that tea can quench thirst. Carroll, Cheng, and Lu (2010) found that in such situations human subjects do not revise causal relations for which they have unambiguous evidence (e.g., that tea is quenching).

Causal Inference: Empirical Findings on Humans and Rats

Association does not equal causation, as we illustrated earlier and as every introductory statistics text warns. We now review how humans and rats reason causally rather than merely associatively.

THE DIRECTION OF CAUSALITY

The concept of causality is fundamentally directional (Reichenbach, 1956) in that causes produce effects, but effects cannot produce causes. Thus, whereas we might say that, given the angle of the sun at a certain time of the day, the height of a flagpole explains the length of its shadow on the ground, it would be odd to say the reverse.[3] A straightforward demonstration that humans make use of the direction of the causal arrow was provided by Waldmann and Holyoak (1992), who reasoned that only causes, but not effects, should "compete" for explanatory power. If P is a perfect cause of an outcome A, and R, a redundant cue, is only presented preceding A in conjunction with P, one has no basis of knowing to what extent, if at all, R actually produces A. Consequently, the predictiveness of R should be depressed relative to P in a predictive situation. But if P is instead a consistent effect of A, there is no reason why R cannot also be an equally consistent effect of A. Alternative causes need to be kept constant to allow causal inference, but alternative effects do not. Consequently, the predictiveness of R should not be depressed in a diagnostic situation.

This asymmetry prediction was tested with the blocking design, using scenarios to manipulate whether a variable is interpreted as a candidate cause or as an effect. Participants in Waldmann and Holyoak's (1992) Experiment 3 had to learn the relation between several light buttons and the state of an alarm system. The instructions introduced the buttons as causes for the alarm in the *predictive* condition, but as potential consequences of the state of the alarm system in the *diagnostic* condition.

Waldmann and Holyoak found exactly the pattern of results they predicted: There was blocking in the predictive condition, but not the diagnostic condition. These results reveal that humans make use of the direction of the causal arrow. Follow-up work from Waldmann's lab (Waldmann & Holyoak, 1997; Waldmann, 2000, 2001) as well as others (Booth & Buehner, 2007; López, Cobos, & Caño, 2005) has demonstrated that the asymmetry in cue competition is indeed a robust finding.

CEILING EFFECTS

One might think that augmenting statistical models with intervention would solve the problem of the directionality of causation. But although intervention generally allows causal inference, it does not guarantee it. Consider a food allergy test that introduces samples of food into the body by needle punctures on the skin. The patient may react with hives on all punctured spots, and yet one may not know whether the patient is allergic to any of the foods. Suppose her skin is allergic to needle punctures, so that hives appear also on punctured spots without food. In this example, there is an intervention, but no causal inference regarding food allergy seems warranted (Cheng, 1997). More generally, interventions are subject to the problem of the well-known *placebo effect*, in which the intended intervention is accompanied by a concurrent intervention (as adding allergens into the bloodstream is accompanied by the puncturing the skin), resulting in confounding. Our example illustrates that intervention does not guarantee causal inference. Not only is intervention insufficient for differentiating causation from association, it is also unnecessary. Mariners since ancient times have known that the position and phase of the moon is associated with the rising and falling of the tides (Salmon, 1989). Notably, they did not consider the association causal, and they had no explanation for the ebb and flow of the tides, until Newton proposed his law of universal gravitation. No intervention on the moon and the tides is possible, but there was nonetheless a dramatic change in causal assessment.

A revealing case of the distinction between covariation and causation that does not involve confounding has to do with what is known in experimental design as a *ceiling effect*. We illustrate this effect with the preventive version of it (a principle never covered in courses on experimental design); the underlying intuition is so powerful it needs no instruction. Imagine that a scientist conducts an experiment to find out whether a new allergy drug relieves migraine as a side effect. She follows the usual procedure and administers the medicine to an experimental group of patients, while an equivalent control group receives a placebo. At the end of the study, the scientist discovers that none of the patients in the experimental group but also none of the patients in the control group suffered from migraine. The effect never occurred, regardless of the intervention. If we enter this information into the ΔP rule, we see that $P(e^{+}|c^{+}) = 0$ and $P(e^{+}|c^{-}) = 0$, yielding $\Delta P = 0$. According to the ΔP rule and RW, this would indicate that there is no causal relation, that is, the drug does not relieve migraine. Would the scientist really conclude that? No, the scientist would instead recognize that she has conducted a poor experiment and hence withhold judgment on whether the drug relieve migraine. If the effect never occurs in the first place, how can a preventive intervention be expected to prove its effectiveness?

Even rats seem to appreciate this argument (Zimmer-Hart & Rescorla, 1974). For associative models, however, when an inhibitory cue (i.e., one with negative associative strength) is repeatedly presented without the outcome, so that the actual outcome is 0 whereas the expected outcome is negative, the prediction is that the cue reduces its strength toward 0. That is, in a noncausal world, we would unlearn our preventive causes whenever they are not accompanied by a generative cause. For example, if we inoculate child after child with polio vaccine in a country, and there is no occurrence of polio in that country, we would come to believe that the polio vaccine does not function anymore, rather than merely that it is not needed. To the contrary, even for rats, the inhibitory cue retains its negative strength (Zimmer-Hart & Rescorla, 1974). In other words, when an outcome in question never occurs, either when a conditioned inhibitory cue is present or when it is not, rats apparently treat the zero ΔP value as uninformative, retaining the inhibitory status of the cue. In this case, in spite of a discrepancy between the expected and actual outcomes, there is no revision of causal strength.

Table 12.1. Relative Frequencies of Headache and Model Values for Each Hypothetical Study and Each Condition in Liljeholm and Cheng (2007, Experiment 2)

	Study 1		Study 2		Study 3	
	e|no A	e|A	e|no A	e|A	e|no A	e|A
Varying-base rate	16/24	22/24	8/24	20/24	0/24	18/24
Constant base rate	0/24	6/24	0/24	12/24	0/24	18/24

Notice that given the aforementioned hypothetical migraine-relief experiment, from the same exact data, showing that migraine never occurs one can conclude that the drug *does not cause* migraine rather than withhold judgment. Thus, given the exact same covariation, the causal learner can simultaneously have two conclusions depending on the direction of influence under evaluation (generative vs. preventive). Wu and Cheng (1999) conducted an experiment that showed that beginning college students, just like experienced scientists, do and do not refrain from making causal inferences in the generative and preventive ceiling effects situations depending (in opposite ways) on the direction of influence to be evaluated. We are not aware of any convincing modification of associationist models that can accommodate the finding.

DEFINITION OF CAUSAL INVARIANCE: BEYOND AUGMENTATION OF ASSOCIATIONS WITH INTERVENTION AND OTHER PRINCIPLES OF EXPERIMENTAL DESIGN

The same problem that leads to the ceiling effect—namely, the lack of representation of causal relations—manifests itself even when *all* the principles of experimental design are obeyed. Even in that case, the associative view makes anomalous predictions. Liljeholm and Cheng (2007, Experiment 2) presented college students with a scenario involving three studies of a single specific cue A (Medicine A, an allergy medicine) as a potential cause of an outcome (headache as a potential side-effect of the allergy medicine). In the scenario, allergy patients in the studies were randomly assigned to an experimental group that received Medicine A and a control group that received a placebo. In the three studies, the probability of the outcome was higher by, respectively, ¼, ½, and ¾ in the experimental group than in the control group (i.e., ΔP = ¼, ½, and ¾; see Table 12.1). In a varying-base-rate condition, the base rate of headache differed across the three studies. In a constant-base-rate condition, the base rate of the effect remained constant: Headache never occurred without the medicine. The students were asked to assess whether the medicine interacted with unobserved causes in the background across the studies or influenced headache the same way across them. As intuition suggests, more students in the constant-base-rate condition than in the varying-base-rate condition (13 out of 15, vs. 5 out of 15, respectively) judged the medicine to interact with the background causes.

Because the changes in covariation, as measured by associative models such as Δ*P* (Jenkins & Ward, 1965) or RW (Rescorla & Wagner, 1972), were the same across conditions, these associative models could not explain the observed pattern of judgments. Thus, even when there is an effective intervention and there is no violation of the principles of experimental design, a statistical account will not suffice. We return to discuss the implications of these results later.

INTERVENTION VERSUS OBSERVATION

Following analogous work on humans (Waldmann & Hagmayer, 2005), Blaisdell et al. (2006) reported a result that challenges associative models: Rats are capable of distinguishing between observations and interventions. In Experiment 1, during a Pavlovian learning phase rats were trained on two interspersed pairs of associations: A light cue (L) is repeatedly followed by either a tone (T) or food (F). If the rats learned that L is a common cause of T and F (see Fig. 12.5a), then in the test phase, observing T should lead them to infer that L must have occurred (because L was the only cause of T), which should in turn lead them to predict F (because L causes F). The number of nose pokes into the food bin measures prediction of F. Consider an alternative condition in which during a test phase the rats learn that pressing on a newly introduced lever turns on T. Because generating T by means of an alternative cause does not influence its cause (L), turning T on by pressing a lever should not lead the rats to predict F. After the learning phase, rats were allocated to either the observation or the intervention condition. The occurrences of T in the test phase were equated across the two conditions by yoking the observation rats to the intervention rats, so that when a rat in the intervention condition pressed the lever and T followed, a rat in the observation condition heard T at the same time, independently of their lever pressing. L and F never occurred during the test phase. Remarkably, the observation rats nose-poked more often than the intervention rats in the interval following T, even though during the learning phase, T and F never occurred simultaneously on the same trial.

Because all occurrences of L, T, and F were identical across the observation and intervention groups, associations alone cannot explain the difference between observing T and intervening to obtain T. Even if augmented with the assumption that interventions have special status, so that the pairing between lever pressing and T, for example, is learned at a much faster rate than purely observed pairings, there would still be no explanation for why the intervention rats apparently associate T with L less than did the observation rats. We will return to discuss a causal account of the observed difference.

A Causal Approach

A solution to the puzzles posed by the distinction between covariation and causation is to have a leap of faith that causal relations exist, even though they are unobservable (Kant, 1781/1965). This leap of faith distinguishes the diverse variants of the causal approach from all variants of the associative approach. Some psychologists have proposed that human causal learning involves positing candidate causal relations and using deductive propositional reasoning to arrive at possible explanations of observed data (De Houwer, Beckers, & Vandorpe, 2005; Lovibond, Been, Mitchell, Bouton, & Frohardt, 2003; Mitchell, De Houwer, & Lovibond, 2009). Others (Gopnik et al., 2004) have proposed that human causal learning is described by *causal Bayes nets*, a formal framework in which causal structures are represented as directed acyclic graphs (Pearl, 2000; Spirtes, Glymour, & Scheines, 1993/2000; see Sloman, 2005, for a more accessible exposition). The graphs consist of arrows connecting some nodes to other nodes, where the nodes represent variables and each arrow represents a direct causal relation between two variables; "acyclic" refers to the constraint that the paths formed by the arrows are never loops. Others have proposed a variant of causal Bayes nets that makes stronger causal assumptions; for example, assume as a defeasible default that causes do not interact, and revise that assumption only when there is evidence against it. The stronger assumptions enable the learner to construct causal knowledge incrementally (Buehner, Cheng, & Clifford, 2003; Cheng, 1997, 2000; Griffiths & Tenenbaum, 2005; Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008; Waldmann, Cheng, Hagmayer, & Blaisdell, 2008).

These variants of the causal view, in addition to their explicit representation of causal relations, share a rational perspective (see Chater & Oaksford, Chapter 2). Thus, they all have a goal of inferring causal relations that best explain observed data. They all make use of deductive inference (for examples of the role of analytic reasoning in empirical learning, see Mermin, 2005; Shepard, 2008). It may be said that they all assume that the causal learner deduces when to induce! Our focus in this chapter is on explaining basic ways in which the causal approach provides a solution to what appears to be impasses from an associative perspective.

A THEORY OF CAUSAL INDUCTION

We use Cheng (1997)'s *causal power* theory (also called the power PC theory, short for "a causal power theory of the probabilistic contrast model") to illustrate how a causal theory explains many of the puzzles mentioned earlier. This theory starts with the Humean constraint that causality can only be inferred, using observable evidence (e.g., covariations, temporal ordering, and spatial information) as input to the reasoning process. It combines that constraint with Kant's (1781/1965) postulate that reasoners have a priori notions that types of causal relations exist in the universe.

This unification can best be illustrated with an analogy. The relation between a causal relation and a covariation is like the relation between a scientific theory and a model. Scientists postulate theories (involving unobservable entities) to explain models (i.e., observed regularities or laws); the kinetic theory of gases, for instance, is used to explain Boyle's law. Boyle's law describes an observable phenomenon, namely that pressure × volume = constant (under certain boundary conditions), and the kinetic theory of gases explains in terms of unobservable entities why Boyle's law holds (gases consist of small particles moving at a speed proportional to their temperature, and pressure is generated by the particles colliding with the walls of the container). Likewise, a causal relation is the unobservable entity that reasoners strive to infer in order to explain observable regularities between events. This distinction between a causal relation as an unobserved, distal, postulated entity and covariation as an observable, proximal stimulus property implies that there can be situations where evidence is observable, but inference is not licensed, and the goal of causal inference thus cannot be met. Specifically, this means that the desired distal unknown, such as causal strength, is represented as a variable (cf. Doumas & Hummel, Chapter 4; Holyoak & Hummel, 2000), separately from covariation, allowing situations where covariation has a definite value (e.g., 0, as in the ceiling effect), but the causal variable has no value.

How, then, does the causal power theory (Cheng, 1997) go beyond the proximal stimulus and explain the various ways in which covariation does not imply causation? The path through the derivation of the estimation of causal strength reveals the answers. For inferring simple (i.e., elemental) causal relations, the theory partitions all causes of effect *e* into the candidate cause in question, *c*, and *a*, a composite of all (observed and unobserved) alternative causes of *e*. "Alternative causes" of *e* include all and only those causes of *e* that are not on the same causal path to *e* as *c*. Thus, *c* can be thought of as a composite that includes all causes on the same causal path as *c* preceding *e*. This partitioning is a general structure that maps onto all learning situations involving candidate causes and effects that are binary variables with a "present" and an "absent" value. We focus on this type of variables because they best reveal how the associative and causal views differ.

The unobservable probability with which *c produces e* (i.e., the probability that *e* occurs *as a result of c* occurring), termed the *generative* causal power of *c* with respect to *e*, is represented by a variable, q_c. When $\Delta P \geq 0$, q_c is the desired unknown. Likewise, when $\Delta P \leq 0$, the *preventive* causal power of *c*, denoted by p_c, is the desired unknown. Two other relevant theoretical unknowns are q_a, the probability with which *a* produces *e* when it occurs, and P(*a*), the probability with which *a* occurs. The composite *a* may include unknown or unobservable causes. Because any causal power variable may have a value of 0, or an unknown or undefined value, these variables are merely hypotheses—they do not presuppose that *c* and *a* indeed have causal influence on *e*. The idea of a cause producing an effect and of a cause preventing an effect are primitives in the theory (see Goodman, Ullman, & Tenenbaum, 2011, and Tenenbaum, Kemp, Griffiths & Goodman, 2011, for an alternative view).

The theory assumes four general simplifying beliefs (Cheng, 1997; Novick & Cheng, 2004):

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

Assumption 1 is a leap of faith inherent to this incremental learning variant of causal discovery. This is the defeasible default assumption we termed "causal invariance" earlier. Two causes influencing effect *e* "independently" means that the influence of each on *e* remains unchanged regardless of whether *e* is influenced by the other cause. Assumption 2 is likewise a default hypothesis, adopted unless evidence discredits it. (Alternative models apply if assumption 1 or 2 is discredited; see Novick & Cheng 2004; see Cheng, 2000, for implications of the relaxation of these assumptions.) This set of assumptions, which is stronger than that assumed in standard Bayes nets, enables causal relations to be learned one at a time, when there is information on only the occurrences of two variables, a single candidate cause and an effect. The type of learning described by the theory therefore requires less processing capacity. It is assumed that, as in associative models, when there is information on which variable is an effect, the causal learner iterates through candidate causes of the effect, grouping all potential causes other than the candidate in question as the composite alternative cause. Otherwise, the causal learner iterates through all possible variable pairs of candidate causes and effects.

These assumptions imply a specific function for integrating the influences of multiple causes (Cheng, 1997; Glymour, 2001), different from the additive function assumed by associative models. For the situation in which a potentially generative candidate cause *c* occurs independently of other causes, the probability of observing the effect *e* is given by a noisy-OR function,

$$P(e^+ \mid c; w_a, q_c) = q_c \cdot c + w_a - q_c \cdot c \cdot w_a \qquad (3)$$

where $c \in \{0,1\}$ denotes the absence and the presence of the candidate cause *c*. Recall that in our equations we denote the "presence" value of a binary variable by a "+" superscript and the "absence" value by a "–" superscript. In contrast, variables have no superscripts. As just mentioned, variable q_c represents the generative power of the candidate cause c. Because it is not possible to estimate the causal power of unobserved causes, variable w_a represents $P(a^+) \cdot q_a$. In the preventive case, the same assumptions are made except that c is potentially preventive. The resulting noisy-AND-NOT integration function for preventive causes is

$$P(e^+ \mid c; w^{a}{}_{,}p_c) = w_a(1 - p_c \cdot c)\ , \tag{4}$$

where p_c is the preventive causal power of c.

Using these "noisy-logical" integration functions (terminology due to Yuille & Lu 2008), Cheng (1997) derived normative quantitative predictions for judgments of causal strength. Under the aforementioned set of assumptions, the causal power theory explains the two conditional probabilities defining ΔP as follows:

$$P(e^+ \mid c^+) = q_c + P(a^+ \mid c^+) \cdot q_a - q_c \cdot P(a^+ \mid c^+) \cdot q_a \tag{5}$$

$$P(e^+ \mid c^-) = P(a^+ \mid c^-) \cdot q_a \tag{6}$$

Equation 5 "explains" that given that c has occurred, e is produced by c or by the composite a, nonexclusively (e is jointly produced by both with a probability that follows from the independent influence of c and a on e). Equation 6 "explains" that given that c did not occur, e is produced by a alone.

Explaining the Role of "No Confounding" and Why Manipulation Encourages Causal Inference But Does Not Guarantee Success

It follows from Equations 3 and 4 that

$$\begin{aligned}\Delta P_c = q_c + P(a^+ \mid c^+) \cdot q_a - q_c \cdot P(a^+ \mid c^+) \cdot q_a \\ - P(a^+ \mid c^-) \cdot q_a\end{aligned} \tag{7}$$

From Equation 7, it can be seen that there are four unknowns: q_c, q_a, $P(a^+|c^+)$, and $P(a^+|c^-)$! It follows that in general, despite ΔP having a definite value, there is no unique solution for q_c. This failure to solve for q_c corresponds to our intuition that covariation need not imply causation.

When there is no confounding. Now, in the special case in which *a occurs* independently of c (e.g., when alternative causes are held constant), $P(a^+ \mid c^+) = P(a^+ \mid c^-)$. If one is willing to assume "no confounding," then making use of Equation 6, Equation 7 simplifies to Equation 8,

$$q_c = \frac{\Delta P}{1 - P(e^+ \mid c^-)} \tag{8}$$

in which all variables besides q_c are observable. In this case, q_c can be solved. Being able to solve for q_c only under the condition of *independent occurrence* explains why manipulation by free will encourages causal inference in everyday reasoning—alternative causes are unlikely to covary with one's decision to manipulate. For the same reason, it explains the role of the principle of *control* in experimental design.

At the same time, the necessity of the "*no confounding*" condition explains why causal inferences resulting from interventions are not always correct; although alternative causes are unlikely to covary with one's decision to manipulate, they still may do so, as our needle-puncture allergy example illustrates. Note that the "no confounding" condition is a *result* in this theory, rather than an unexplained axiomatic assumption, as it is in current scientific methodology (also see Dunbar & Klahr, Chapter 35).

An analogous explanation yields p_c, the power of c to prevent e

$$p_c = \frac{-\Delta P}{P(e^+ \mid c^-)} \tag{9}$$

Griffiths and Tenenbaum (2005) showed that if one represents uncertainty about the estimates of causal power by a distribution of the likelihood of each possible strength given the data, then Equation 8 and 9, respectively, are maximum likelihood point estimates of the generative and preventive powers of the candidate cause; that is, they are the peak of the posterior likelihood distributions.

Explaining Two Ceiling Effects

Equations 8 and 9 explain why ceiling effects block causal inference and do so under different conditions for evaluating generative and preventive causal influence. When the outcome does not occur at either a ceiling (i.e., extreme) level, both equations yield either causal power of 0 when the occurrence of c makes no difference to the occurrence of e ($\Delta P = 0$). When e always occurs (i.e., $P(e^+|c^+) = P(e^+|c^-) = 1$) regardless of the manipulation, however, q_c in Equation 8 (the generative case) is left with an undefined value. In contrast, in the preventive case, when *e* never occurs (i.e., $P(e^+|c^+) = P(e^+|c^-) = 0$), again regardless of the manipulation, p_c in Equation 9 is left with an undefined value.[4]

As we mentioned, most causes are complex, involving not just a single factor but a conjunction of factors operating in concert, and the assumption that *c* and *a* influence *e* independently may be false most of the time. When this assumption is violated, if an alternative cause (part of *a*) is observable, the independent influence assumption can be given up for the observable alternative cause, and progressively more complex causes can be evaluated using the same distal approach that represents causal powers (see Novick & Cheng, 2004, for an extension of this approach to evaluate conjunctive causes involving two factors). Even if alternative causes are unknown, however, Cheng (2000) showed that as long as they occur with about the same probability in the learning context as in the generalization context, predictions according to simple causal power involving a single factor will hold.

Some have claimed that the causal power approach cannot account for reasoning that combines observations with interventions. As just shown, however, this approach explains the role of interventions in causal learning and how it differs from observation. Likewise indicating that this approach readily accommodates the combination, Waldmann et al. (2008) derived an equation under the causal power assumptions that explains Blaisdell et al.'s results regarding the distinction between observations and interventions in diagnostic reasoning.

Experimental Tests of a Causal Approach

We examine three findings in support of the causal approach. None of these findings can be explained by the associative view, even when augmented with the assumption that only interventions enable causal inference. The first two findings test the two leaps of faith: that causal relations exist and that they are invariant across contexts. The first finding concerns the independent causal influence assumption as manifested in a qualitative pattern of the influence of $P(e^+|c^+)$, the base rate of *e*, for candidate causes with the same ΔP. The second illustrates the parsimony of a causal explanation that assumes independent causal influence across different types of "effect" variables, specifically, dichotomous and continuous (Beckers, De Houwer, Pineno, & Miller, 2005; Beckers, Miller, De Houwer, & Urushihara, 2006; Lovibond et al., 2003). The third concerns the test reviewed earlier of the distinction between observation and intervention ("seeing" vs. "doing") in diagnostic causal inference (Blaisdell et al., 2006; Waldmann & Hagmayer, 2005). We consider explanations of this distinction as an illustration of the compositionality of the causal view and of the role of deductive reasoning in causal inference.

The Independent Causal Influence Assumption Affects Causal Judgments: Base- Rate Influence on Conditions with Identical ΔP

As we noted, a major purpose of causal discovery is to apply the acquired causal knowledge to achieve goals, and that the independent causal influence assumption is a leap of faith that justifies generalization from the learning context to the application context. Here, we see that the assumption leads to causal judgments that differ from those predicted by associative models, even those augmented with a privileged status for interventions and other principles of experimental design. In other words, this assumption not only affects the application of causal knowledge, it affects the very discovery of that knowledge itself.

Do people have this leap of faith? Let us examine the predictions based on the causal power assumptions in greater detail. If we consider Equation 8, for any constant positive ΔP, generative causal ratings should *increase* as $P(e^+|c^-)$ increases. Conversely, according to Equation 9, for any constant negative ΔP, preventive causal ratings should *decrease* as $P(e^+|c^-)$ increases. On the other hand, according to both equations, zero contingencies should be judged as noncausal regardless of the base rate of *e* except when the base rate is at the respective ceiling levels.

No associative model of causal inference, descriptive or prescriptive, predicts this qualitative pattern of the influence of the base rate of *e*. Normative models are symmetric around the probability of .5 and therefore do not predict an asymmetric pattern either for generative causes alone or for preventive causes alone. Although some psychological associative learning models can explain one or another part of this pattern given felicitous parameter values, the same parameter values will predict notable deviations from the rest of the pattern. For example, in the RW, if $\beta_{US} > \beta_{\neg US}$ causal ratings for generative and preventive causes will *both* increase as base-rate increases , whereas they will *both* decrease as base-rate

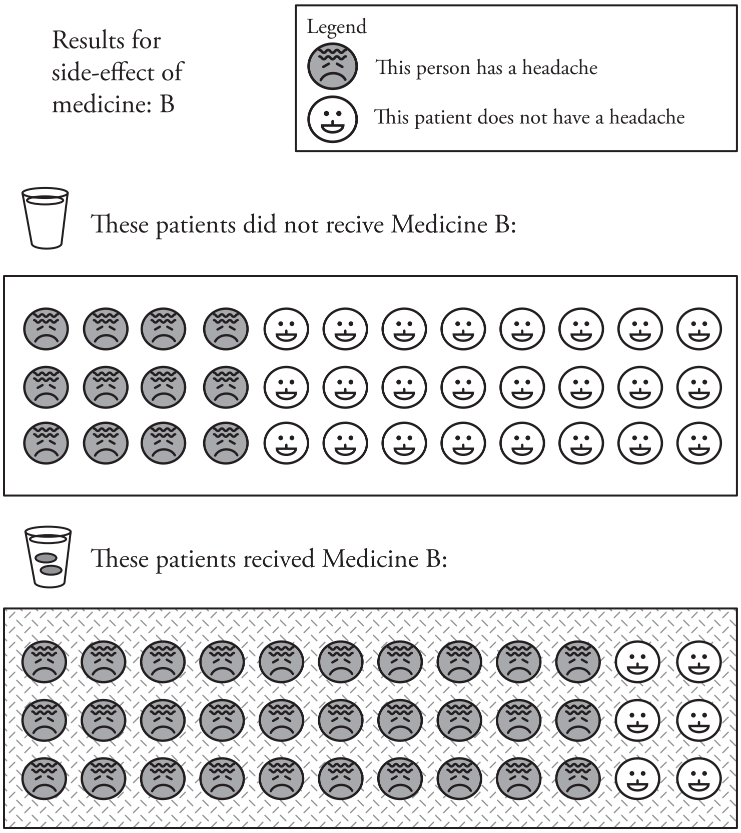

Results for side-effect of medicine: B
Legend
This person has a headache
This patient does not have a headache
These patients did not recive Medicine B:
These patients recived Medicine B:

Fig. 12.3 Example stimulus materials from a condition in Buehner et al. (2003).

increases if the parameter ordering was reversed. No consistent parameter setting will predict opposite trends for generative as for preventive causes for the same change in the base rate of *e*. Another prominent associative learning model, Pearce's (1987) model of stimulus generalization, can account for opposite base rate influences in positive and negative contingencies if the parameters are set accordingly, but this model would then predict a base-rate influence on noncontingent conditions.

Figure 12.3 illustrates the intuitiveness of a deviation from ΔP. The reasoning is counterfactual. $P(e^+|c^-)$, 1/3 in the figure, estimates the "expected" probability of *e* in the *presence* of *c*, if *c* had been absent so that only causes other than *c* exerted an influence on *e*. A deviation from this counterfactual probability is evidence for *c* being a simple cause of *e*. Under the assumption that the patients represented in the figure were randomly assigned to the two groups, one that received the drug and another that did not, one would reason that about 1/3 of the patients in the "drug" group would be expected to have headache if they had not received the drug. For the remaining patients—the 2/3 who did not have already have headaches caused by other factors—the drug would be the sole cause of headaches. In this subgroup, headache occurred in 3/4 of the patients. One might therefore reason, one's best guess for the probability of the drug producing headache is 3/4. If one assumes that for every patient in the control group, *regardless* of whether the patient had a headache, the drug causes headache with a probability of 3/4, this estimate would result. Among those who already had a headache produced by alternative causes, headache due to the drug is not observable.

In contrast, consider what estimate would result if one assumes instead that the drug causes headache with a probability of 1/2, the estimated causal strength according to associative models such as the ΔP model. Applying that probability to every patient, one's best guess would be that 2/3 of the patients would have headaches after receiving the medicine, rather than the 5/6 shown in Figure 12.3. As should be clear, associative models give estimates that are inconsistent with the assumption that the causes involved influenced headache independently, even though the additivity in those models is generally assumed to represent independence, and thus to justify generalization to new contexts. This inconsistency, due to the outcome variable being dichotomous, leads to irrational applications of causal knowledge to achieve desired outcomes.

Are people rational or irrational in their estimation of causal strength? To discriminate between the causal power theory and the associative approach, Buehner, Cheng, and Clifford (2003, Experiment 2) made use of the pattern of causal-strength predictions according to Equations 8 and 9 just discussed. They gave subjects a task of assessing whether various allergy medicines have a side effect on headaches, potentially causing headaches or preventing them, when presented with fictitious results of studies on allergy patients (see Fig. 12.3 for an example) in which the patients were randomly assigned to two groups, one receiving the medicine and the other not. The subjects were also asked to rate the causal strengths of each candidate after viewing the results for each fictitious study, using a frequentist counterfactual causal question that specified a novel transfer context: "Imagine 100 patients who do *not* suffer from headaches. How many would have headaches if given the medicine?" The novel context for assessing generative causal power, as just illustrated, is one in which there are no alternative generative causes of headache. By varying the base rate of the target effect, for both generative and preventive causes, the experiment manipulated *(1)* causal power while keeping ΔP constant, *(2)* ΔP while keeping causal power constant. The experiment also manipulated the base rate of *e* for noncontingent candidate causes. Their results clearly indicate that people make the leaps of faith assumed by the causal power theory, contrary to the predictions of all associative models, including normative associative models.

Integrating Causal Representation with Bayesian Inference: Representing Uncertainty and Evaluating Causal Structure

The reader might have noticed that, just like the ΔP rule, the point estimate of causal power in the causal power theory (Equations 8 and 9) is insensitive to sample size. As initially formulated, the theory did not provide any general account of how uncertainty impacts causal judgments. The point estimates are the most likely strength of the causal link that would have produced the observed data, but causal links with other strength values, although less likely to have produced the data, could well have also, for smaller sample sizes more so than for larger sizes. The lack of an account of uncertainty in early models of human causal learning, together with methodological problems in some initial experiments testing the causal power theory (see Buehner et al., 2003), contributed to prolonging the debate between proponents of associationist treatments and of the causal power theory. For some data sets (e.g., Buehner & Cheng, 1997; Lober & Shanks, 2000), human causal-strength judgments for some conditions were found to lie intermediate between the values predicted by causal power versus ΔP. This pattern was especially salient for studies in which the causal question, which was ambiguously worded, could be interpreted to concern confidence in the existence of a causal link. Intriguingly, a subtle, statistically insignificant, but consistent trend toward this pattern seemed to occur even for the disambiguated counterfactual question illustrated earlier. These deviations from the predictions of the causal power theory perhaps reflect the role of uncertainty, which is outside the scope of the theory.

An important methodological advance in the past decade is to apply powerful Bayesian probabilistic inference to causal graphs to explain psychological results (e.g., Griffiths & Tenenbaum, 2005; Lu et al., 2008; Tenenbaum et al., 2011; see Griffiths, Tenenbaum, & Kemp, Chapter 3; for a review of recent work, see Holyoak & Cheng, 2011). This new tool enables rationality in causal inference to be addressed more fully. For example, it enables a rich representation of uncertainty and a formulation of qualitative queries regarding causal structure.

Griffiths and Tenenbaum (2005; Tenenbaum & Griffiths, 2001) proposed the *causal support* model, a Bayesian model that addresses the causal query, termed a "structure" judgment, which aims to answer, "How likely is it that a causal link exists between these two variables?" This is in contrast to the causal query regarding causal strength that has been emphasized in previous psychological research on causal learning. Strength judgment concerns the weight on a causal link, which aims to answer the query, "What is the probability with which a cause produces (alternatively, prevents) an effect?"

In terms of the graphs in Figure 12.4, the causal support model aims to account for judgments as to whether a set of observations (D) was generated by Graph 1, a causal structure in which a link may exist between candidate cause C and effect E or by a causal structure in which no link exists between C and E

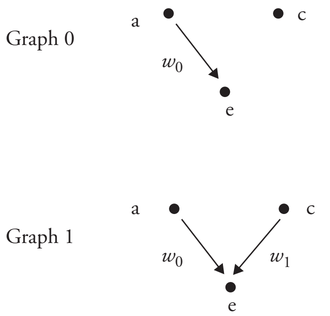


Fig. 12.4 Candidate causal structures varying in whether *c* causes *e*.

(Graph 0). Causal-strength models, by contrast, aim to account for people's best estimates of the weight w_1 on the link from *C* to effect *E* in Graph 1 that generated *D*, with w_1 ranging from 0 to 1.

In the causal support model, the decision variable is based on the posterior probability ratio of Graphs 1 and 0 by applying Bayes' rule. Support is defined as:

$$\text{support} = \log \frac{P(D \mid Graph1)}{P(D \mid Graph0)}. \qquad (10)$$

associative versus causal bayesian models: uniform versus sparse and strong priors

Note that adopting Bayesian inference is entirely orthogonal to the longstanding debate between causal and associationist approaches. Because mathematics is a tool rather than an empirical theory, the Bayesian approach can be causal or associative depending on whether causal assumptions are made, even while they are applied to supposedly causal graphs. As Griffiths and Tenenbaum (2005) had noted, a Bayesian model can incorporate either the noisy-logical integration functions derived from the causal power theory or the linear function underlying the Rescorla-Wagner model and the ΔP rule. In addition, a Bayesian analysis can be applied to both strength and structure judgments, as well as to other types of causal queries, such as causal attribution. For strength judgments, rather than basing predictions on the peak of the posterior distribution of w_1 in Graph 1, which corresponds to causal power and ΔP according to the respective models, a natural Bayesian extension of the causal power theory would base predictions on other functions of the posterior distribution of w_1, such as its mean. Thus, for the fictitious data regarding the side effect of an allergy medicine in Figure 12.3, rather than estimating that the medicine causes headache with a probability of 3/4 or 1/2, as predicted by the causal and associative causal-strength models, respectively, the estimate would fall slightly below 3/4, in between those estimates.

Lu et al. (2008) developed and compared several variants of Bayesian models as accounts of human judgments about both strength and structure. In addition to directly comparing predictions based on these alternatives, Lu et al. considered two different sets of priors on causal strength. One possible prior is simply a uniform distribution, as assumed in the causal support model. The alternative "generic" (i.e., domain-general) prior tested by Lu et al. is based on the assumption that people prefer parsimonious causal models (Chater & Vitányi, 2003; Lombrozo, 2007; Novick & Cheng, 2004). Sparse and strong (SS) priors imply that people prefer causal models that minimize the number of causes of a particular polarity (generative or preventive) while maximizing the strength of each individual cause that is in fact potent (i.e., of nonzero strength).

The sparse and strong priors, although admitted post hoc, point to the role of parsimony in explanation, an interesting issue for future research. When one is presented with a Necker cube, for example, one perceives two possible orientations. The human perceptual system has implicitly screened out the infinitely many other possible non-cube-shaped objects that would project the same eight corners onto our retina. The visual system makes a parsimony assumption: It favors the simplest "explanations" of the input on our retina. The human causal learning appears to similarly favor parsimonious causal explanations.

For all four Bayesian models, Lu et al. (2008) compared the average observed human strength rating for a given contingency condition with the mean of w_1 computed using the posterior distribution. Model fits revealed that the two causal variants based on the noisy-logical integration function were much more successful overall than the associative variants. For datasets from a meta-analysis based on 17 experiments selected from 10 studies in the literature (Perales & Shanks, 2007; see also Hattori & Oaksford, 2007), the causal Bayesian models (with one or zero free parameters) performed at least as well as the most successful nonnormative model of causal learning (with four free parameters) and much better than the Rescorla-Wagner model. Thus, although both causal and associative approaches can be given a Bayesian formulation, the empirical tests of human causal learning reported by Lu et al. favor the causal Bayesian formulation, providing further evidence for the rationality of human causal inference.

Lu et al. (2008) also evaluated structure analogs of the two causal variants of Bayesian strength models as accounts for observed structure judgments from experiments in which participants were explicitly asked to judge whether the candidate was indeed a cause. Relative to the support model, human reasoners appear to place greater emphasis on causal power and the base rate of the effect, and less emphasis on sample size.

The Independent Causal Influence Assumption for Dichotomous and Continuous Outcome Variables

Across multiple studies on humans (Beckers, de Houwer, Pineno, & Miller, 2005; De Houwer, Beckers, & Glautier, 2002; Lovibond et al., 2003) and even rats (Beckers, Miller, De Houwer, & Urushihara, 2006), an intriguing set of findings has emerged, showing that information regarding the additivity of the causal influences of two causes and the range of magnitudes of the outcome both influence judgments regarding unrelated candidate causes of that outcome. We illustrate the finding with parts of a broader study by Lovibond et al. (2003). In a backward blocking design, cues A and B (two food items) in combination were paired with an outcome (an allergic reaction); in a second phase, cue B alone was paired with the outcome. Thus, target cue A made no difference to the occurrence of the outcome (holding B constant, there was always an allergic reaction regardless of whether A was there). The critical manipulation in Lovibond et al. was a "pretraining compound" phase during which one group of subjects, the *ceiling* group, saw that a combination of two allergens produced an outcome at the same level ("an allergic reaction") as a single allergen (i.e., the ceiling level). In contrast, the *nonceiling* group saw that a combination of two allergens produced a stronger reaction ("a STRONG allergic reaction") than a single allergen ("an allergic reaction"). Following this pretraining phase, all subjects were presented with information regarding novel cues in the main training phase. Critically, the outcome in this training phase always only occurred at the intermediate level ("an allergic reaction"), both for subjects in the *ceiling* and *nonceiling* groups.

As a result of pretraining, however, subjects' perception of the nature of the outcome in this phase would be expected to differ. For the exact same outcome, "an allergic reaction," the only form of the outcome in that phase, whereas the *ceiling* group would perceive it to occur at the ceiling level, the *nonceiling* group would perceive it to occur at an intermediate level. As mentioned, for both groups, cue A made no difference to the occurrence of the outcome. Because the causal view represents causal relations separately from covariation, it explains why when the outcome occurs at a ceiling level, the generative effect of a cause has no observable manifestation. At a nonceiling level, causal and associative accounts coincide: The most parsimonious explanation for no observable difference is noncausality. However, at the ceiling level, observing no difference does not allow causal inference, as explained by the causal power theory. In support of this interpretation, the mean causal rating for cue A was reliably lower for the *nonceiling* group than for the *ceiling* group.

Beckers et al. (2005) manipulated pretraining on possible levels of the outcome and on the additivity of the influences of the cues separately and found that each type of pretraining had an enormous effect on the amount of blocking. Beckers et al. (2006) obtained similar results in rats. These and other researchers have explained these results in terms of the use of propositional reasoning to draw conclusions regarding the target cue (Beckers et al., 2005, 2006; Lovibond et al., 2003; Mitchell et al., 2009). For example, a subject might reason: "If A and B are each a cause of an outcome, the outcome should occur with a greater magnitude when both A and B are present than when either occurs by itself. The outcome in fact was *not* stronger when A and B were both present as when B occurred alone. Therefore, A must not be a cause of the outcome." These researchers have explained the impact of the pretraining in terms of learning the appropriate function for integrating the influences of multiple causes in the experimental materials (e.g., additivity vs. subadditivity) from experiences during the pretraining phase, in line with proposals by Griffiths and Tenenbaum (2005) and Lucas and Griffiths (2010).

It is important to distinguish between domain-specific integrating functions that are the outputs of causal learning, and domain-independent integrating functions that enable an output, in view of the essential role they play in the inference process. In the causal power theory, the latter are the noisy-logicals: functions representing independent causal influence. As seen in the preceding section, whether independent causal influence is assumed in the inference process leads to different causal judgments. Moreover, independent causal influence enables compositionality. Even if we were to disregard the role of that assumption in the inference process, without it generalization of the acquired causal knowledge to new contexts would be problematic: If integrating functions were purely empirically learned, every new combination of causes, such as the combination of a target cause with unobserved causes in a new context, would require new learning (i.e., causal inference would not be compositional).

An alternative interpretation of Lovibond et al.'s (2003) results is that for all types of outcome variables, independent causal influence is always the default assumed in the causal discovery process, but the mathematical function defining independent influence differs for different types of outcome variables. For continuous outcome variables, independent causal influence is represented by additivity, as is generally known; for dichotomous outcome variables, independent causal influence is represented by the noisy-logicals, as explained earlier (see pp. 219–220 & 222–223). The unifying underlying concept is the superposition of the influences, a concept borrowed from physics. Under this interpretation, the pretraining conveys information on the nature of the outcome variable: continuous or dichotomous. Thus, subjects in their ceiling group, who received pretraining showing that two food items in combination produced an "allergic reaction" just as each item alone did, learned that the outcome is dichotomous. But subjects in their nonceiling group, who received pretraining showing that two food items in combination produced a stronger allergic reaction than each item alone, learned that the outcome is continuous.

INTERVENTION VERSUS OBSERVATION AND DIAGNOSTIC CAUSAL INFERENCE

A hallmark of a rational causal reasoner is the ability to formulate flexible and coherent answers to different causal queries. A goal of accounts of causal inference is to explain that ability. We have illustrated the causal view's answers to queries regarding causal strength and structure. (For formulations of answers to questions regarding causal attribution [how much an target outcome is due to certain causes], see Cheng & Novick, 2005; for answers to questions regarding enabling conditions, see Cheng & Novick, 1992.) Let us consider here answers to queries involving *diagnostic* causal inference, inference from the occurrence of the effect to the occurrence of its causes. Recall Blaisdell et al's (2006) finding regarding rats' ability to distinguish between an event that is merely observed and one that follows an intervention. When a tone that occurred only when a light occurred during the learning phase was merely observed in the test phase, the rats in the experiment (the Observe group) nose-poked into the food bin more often than when the tone occurred immediately after that rats pressed a lever newly inserted in the test phase (the Intervene group). The Observe rats apparently diagnosed that the light must have occurred, whereas the Intervene rats diagnosed that it need not have occurred; light was never followed by food in the test phase.

Blaisdell et al.'s (2006) results were initially interpreted as support for causal Bayes nets. Note that the different diagnostic inferences in the two groups are consistent with simple deductive inference. For the Observe group, because the light was the only cause of the tone, when the tone occurred, the light must have occurred. For the Intervene group, because both the lever press and the light caused the tone, the tone occurring need not imply that the light occurred. Because causal Bayes nets (Pearl, 2000; Spirtes et al., 1993/2000) and the causal power approach (Waldmann et al., 2008) both make use of deductive inference, it is not surprising that they can also explain diagnostic reasoning.

The graphs in causal Bayes nets are assumed to satisfy the Markov condition, which states that for any variable X in the graph, conditional on its parents (i.e., the set of variables that are direct causes of X), X is independent of all variables in the graph except its descendents (i.e., its direct or indirect effects). A direct effect of X is a variable that has an arrow directly from X pointing into it, and an indirect effect of X is a variable that has a pathway of arrows originating from X pointing into it. Candidate causal networks are evaluated by assessing patterns of conditional independence and dependence entailed by the networks using the Markov and other assumptions. Candidate causal networks that are inconsistent with the observed pattern of conditional independence are eliminated, and the remaining candidate causal networks form the basis of causal judgments.

The causal Bayes nets approach explains Blaisdell et al.'s results by a distinction it makes between

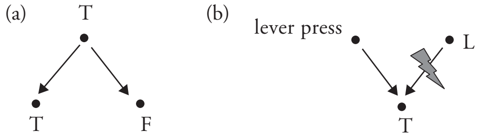


Fig. 12.5 (*a*) L (Light) causes T (tone) and F (food). (*b*) Lever press and L each causes T.

intervening to set a variable at a specific value and merely observing that value. As illustrated in Figure 12.5a, observing T allows diagnostic inference regarding L because of the arrow from L to T. But intervening to produce T severs all other incoming arrows into T, a result called graph surgery, so that the resulting causal network no longer has the arrow from L to T (see Fig. 12.5b).

Although this approach explains the results in the test phase if one assumes that the rats inferred the causal structure intended by the researchers, namely, that L is the common cause of T and F (see Fig. 12.5a), the perfect negative correlation between T and F conditional on L during the learning phase in fact violates the Markov assumption applied to this causal structure (see Rehder & Burnett, 2005; Steyvers et al., 2003 for human results indicating violations of the Markov assumption). Causal Bayes nets therefore predict from the learning phase data that there is some inhibitory connection between T and F, and that both the Intervention and Observation rats should equally avoid going to the food bin when T occurred, contrary to the responses observed.

An alternative solution, one that causal psychological theories (e.g., Cheng, 1997, 2000; Waldmann et al., 2008) inherited from traditional associative accounts (e.g., Rescorla & Wagner, 1972) is that people (and perhaps other species) incrementally construct causal networks by evaluating one (possibly conjunctive or otherwise complex) causal relation involving a single target effect at a time, while taking into consideration other causes of the effect. Motivated by consideration of limited processing capacity and of limited access to information at any one time, the incremental feature is shared by associative theorists (e.g., Jenkins & Ward, 1965 Rescorla & Wagner, 1972). Notably, whereas standard Bayes nets fail to explain Blaisdell et al.'s results, the incremental approach fully explains them. One difference is that the Markov assumption plays a different role in the latter approach: It is the consequence of the causal power assumptions (specifically, the independence assumptions), rather than a constraint used for generating the inferences. Thus, noticing the negative correlation takes effort and thus need not occur until there is sufficient training, as is consistent with the findings in rats (Savastano & Miller, 1998; Yin, Barnet, & Miller, 1994).

In summary, the three lines of evidence just discussed all lie beyond even the augmented associative view. They converge in their support for the two leaps of faith underlying the causal view, as well as for the conviction that the causal world is logically consistent.

TIME AND CAUSALITY: MUTUAL CONSTRAINTS

We have concentrated on theoretical approaches that specify how humans take the mental leap from covariation to causation. Irrespective of any differences in theoretical perspective, all these approaches have in common that they assume that covariation can be readily assessed. This assumption is reflected in the experimental paradigms most commonly used; typically, participants are presented with evidence structured in the form of discrete, simultaneous or sequential learning trials in which each trial contains observations on whether the cause occurred and whether the effect occurred. In other words, in these tasks it is always perfectly clear whether a cause is followed by an effect on a given occasion. Such tasks grossly oversimplify the complexities of causal induction in some situations outside experimental laboratories: Some events have immediate outcomes, and others do not reveal their consequences until much later. Before an organism can evaluate whether a specific covariation licenses causal conjecture, the covariation needs to be detected and parsed in the first place.

Although the problem had been neglected for many years, the last decade has seen interesting and important developments. It has long been documented that cause-effect contiguity (one of Hume's cues toward causality) appears to be essential for causal discovery. Shanks, Pearson, and Dickinson (1989), for example, reported that in an impoverished computerized instrumental learning task, people failed to discriminate between conditions where they had strong control over an outcome ($\Delta P = .75$) and noncontingent control conditions, when their actions and the associated outcomes were separated by more than 2 seconds. In a completely different domain, Michotte (1946/1963) found that impressions of causal "launching" only occur when the collision of the launcher with the launchee is followed immediately by motion onset in the launchee: Temporal gaps of 150 ms or more destroy the impression.

From a computational perspective, it is easy to see why delays would produce decrements in causal reasoning performance. Contiguous event pairings are less demanding on attention and memory. They are also much easier to parse. When there is a temporal delay, and there are no constraints on how the potential causes and effects are bundled, as in Shanks et al. (1989), the basic question on which contingency depends no longer has a clear answer: Should this particular instance of *e* be classified as occurring in the presence of *c* or in its absence? Each possible value of temporal lag results in a different value of contingency. The problem is analogous to that of the possible levels of abstractions of the candidate causes and the effects at which to evaluate contingency (and may have an analogous solution). Moreover, for a given *e*, when alternative intervening events occur, the number of hypotheses to be considered multiplies. The result is a harder, more complex inferential problem, one with a larger search space.

Buehner and May (2002, 2003, 2004) have demonstrated that prior knowledge about delayed time frames constrains the search process, such that noncontiguous relations are judged to be just as causal as contiguous ones. Buehner and McGregor (2006) have further shown that when prior assumptions about delays are sufficiently salient, contiguous relations are perceived as less causal than delayed ones—an apparent contradiction to Hume's tenets.

If causal learning operates according to the principles of Bayesian evidence integration, then these results on contiguous and delayed causation make sense: Reasoners may focus on the expected delay for a type of causal relation and evaluate observations with respect to it. In Bayesian terms, they evaluate likelihoods, the probability of the observations resulting from a hypothesis. In the earlier demonstrations of detrimental effects of delay (Michotte, 1946/1963; Shanks et al., 1989), the prior assumption would have been that there is no delay: Michotte's stimuli were simulations of well-known physical interactions (collisions), while Shanks et al. used computers, which (even in those days!) were expected to operate fast. Once these prior assumptions are modified via instructions (Buehner & May, 2002, 2003, 2004), or via constraints in the environment (Buehner & McGregor, 2006), then delayed relations pose no problem.

More recent work has found that prior expectations about time frames are relevant not only for the *extent* of delays but also with respect to their *variability*. Consider two hypothetical treatments against headache. Drug A always provides relief 2 hours after ingestion, while drug B sometimes starts working after just 1 hour, while other times it can take 3 hours to kick in. Which would we deem as a more effective drug? The answer to that question depends on how exactly temporal extent is interpreted when drawing causal conclusions. One possibility would be that causal attribution decays over time, similarly to discounting functions found in intertemporal choice (for an overview, see Green & Myerson, 2004). Under such an account, the appeal of a causal relation would decay over time according to a hyperbolic function.

One consequence of hyperbolic discounting is that variable relations may appear more attractive than stable ones, even when they are equated for mean expected delay. This conjecture is rooted in the diminishing sensitivity to delay: Variable relations accrue more net strength than constant relations matched for mean delay. And indeed, Cicerone (1976) has found that pigeons preferred variable over constant delays of reinforcement. Thus, if human causal learners approach time in a similar manner (and apply well-established principles of discounting as regards to intertemporal choice), we would expect drug B to emerge as the favorite. Interestingly, the opposite is the case: Greville and Buehner (2010) found that causal reasoners consistently prefer stable, predictable relations. Presumably we have strong a priori expectations that (most) causal relations are associated with a particular, relatively constant time frame. Where such expectations are violated, less learning takes place.

Cause-effect timing not only impacts assessments of causal strength but critically also constrains our ability to infer structure. Pace Hume, causes must occur before their effects, even though the intervening interval may extremely closely approximate 0 (e.g., the interval between a fist's contact with a pillow and the pillow's indentation, the interval between a cat walking into the sun and its shadow appearing on the ground). While such considerations are relatively trivial when there are only two variables involved, finding structure in multivariable causal systems gets increasingly difficult as the size of the system grows. Moreover, many structures are Markov-equivalent (Pearl, 2000), meaning that they cannot be distinguished by mere observation of the statistical patterns they produce. Lagnado and Sloman (2004, 2006) have shown that in such situations, people rely on temporal ordering to infer causal structure. More specifically, temporal order constrains structure inference to a greater extent than the observed patterns of statistical dependencies.

As we highlighted earlier, cognitive science approaches to causality are rooted in the Humean conjecture that causality is a mental construct, inferred from hard, observable facts. Recent evidence suggests that Hume's route from sensory experience to causal knowledge is not a one-way street but in fact goes in both directions. Not only does our sensory experience determine our causal knowledge, but causal knowledge also determines our sensory experience. The latter direction of influence was first documented by Haggard, Clark, and Kalogeras (2002), who showed that sensory awareness of actions and resultant consequences are shifted in time, such that actions are perceived as later, and consequences as earlier (with reference to a baseline judgment error). Causes and effects thus mutually attract each other in our subjective experience. Originally, the effect was thought to be specific to motor action and intentional action control (Wohlschläger, Haggard, Gesierich, & Prinz, 2003). Buehner and Humphreys (2009), however, have shown that causality is the critical component of temporal binding: Intentional actions without a clear causal relation do not afford attraction to subsequent (uncaused) events. Moreover, Humphreys and Buehner (2009, 2010) have shown that the causal binding effect exists over time frames much longer than originally reported, and outside the range of motor adaptation (Stetson, Cui, Montague, & Eagleman, 2006), as would be required for action-control based approaches. Buehner and Humphreys (2010) have furthermore demonstrated a binding effect in spatial perception using Michottean stimuli—a finding that is completely outside the scope of motor-specific accounts of binding. It appears as if our perception of time and space, and our understanding of causality, mutually constrain each other to afford a maximally coherent and parsimonious experience.

Our chapter has reviewed multiple lines of evidence showing a strong preference for parsimonious causal explanations. This preference holds for scientific as well as everyday explanations. Among the many alternative representations of the world that may support predictions equally well, we select the most parsimonious. Hawking and Mlodinow (2010) note that, although people often say that Copernicus's sun-centered model of the cosmos proved Ptoloemy's earth-centered model wrong, that is not true; one can explain observations of the heavens assuming either Earth or the sun to be at rest. Likewise, although the city council of Monza, Italy, barred pet owners from keeping goldfish in curved fishbowls—on the grounds that it is cruel to give the fish a distorted view of reality through the curved sides of the bowl—the goldfish could potentially formulate a model of the motion of objects outside the bowl no less valid as ours. The laws of motion in our frame are simpler than the fish's, but theirs are potentially just as coherent and useful for prediction. But the members of the city council of Monza, like the rest of us, have such an overpowering preference for the more parsimonious model of the world that they perceive it as "truth."

Conclusions and Future Directions

Our chapter began with the question: With what cognitive assets would we endow an intelligent agent—one that has processing and informational resources similar to humans—so that the agent would be able to achieve its goals? We have taken the perspective that generalization from the learning context to the application context is central to the achievement of its goals. From this perspective, we first examined the crippling inadequacies of the associative view, which attempts to maintain objectivity by restricting its inference process to computations on observable events only. We considered variants of the associative view augmented with a special status for interventions and other principles of experimental design, in line with typical scientific causal inference.

We then considered the causal view, which resolves major apparent impasses by endowing the agent with two leaps of faith, that *(1)* the world is causal even though causal relations are never observable, and *(2)* causal laws are uniform and compositional. These empirical leaps are grounded in the conviction that existence is logically consistent. They enable the agent to incrementally construct an understanding of how the world works and coherently generalize its acquired causal knowledge. Analysis in cognitive research shows that the common belief that justifies the augmented associative view—that assumptions about independent causal influence justify the application of causal knowledge to new contexts but do not influence the output of statistical analyses—is mistaken. Likewise, the common belief that assumptions about estimations of causal strength are secondary, and do not affect judgments regarding causal structure, is mistaken.

Remarkably, observed causal judgments reveal that humans make those leaps of faith, and that their causal judgments are based on a definition of independent causal influence that is logically consistent across the learning and application contexts. The use of the sharper tool of Bayesian mathematics shows even more unequivocal support for the causal view. This tool also extends the capability to formulate answers to different kinds of causal queries.

The potential to discover how the world works must of course be accompanied by the requisite computational capabilities. We have identified three intertwined capabilities so far. The agent must be able to *(1)* make deductive logical inferences, *(2)* compute statistical regularities, and *(3)* represent uncertainty. The last two allow the agent to make progress in the face of errors in even its best hypotheses. The first is an essential component of the parsimony assumption and of coherent and flexible reasoning.

Three outstanding issues seem especially pertinent to us in view of our analysis and review. For each issue, a rational analysis in tandem with empirical work differentiating between alternative plausible explanations would deepen our understanding of causal learning.

1. Hypothesis revision: If causal learning is incremental, by what criteria do causal learners revise their hypotheses, and what do their criteria and revisions reveal about the intended destination of the revision process? Recent research found that for preventers with a narrow scope, which violate the independent influence assumption, people are more likely to posit a hidden cause to explain and remove the violation (Carroll & Cheng, 2010). Standard causal Bayes nets would not interpret the violation to signal a need for representation revision.
2. Category formation and causal learning: We have taken the perspective that causal discovery is the driving force underlying our mental representation of the world, not just in the sense that we need to know how things influence each other but also in the sense that causal relations define what should be considered *things* in our mental universe (Lewis, 1929). Are causal learning and category formation two aspects of the same challenge, as the goal of generalization of causal beliefs to application contexts would suggest? How do people arrive at their partitioning of the continuous stream of events into candidate causes and effects? Likewise, how do people arrive at their partitioning of events into candidate causes and effects at particular levels of abstraction? In Lien and Cheng's (2000) experiments, human subjects were presented with causal events involving visual stimuli for which candidate-cause categories were undefined; there was no specification of either the potential critical features or the relevant level of abstraction of the features. It was found that subjects seemed to form candidate-cause categories that maximized ΔP, perhaps in an attempt to maximize the necessity and sufficiency of the cause to produce the effect in question. The topic awaits better formulations of explanations as well as additional empirical work.
3. Parsimony in causal explanations: We have encountered the critical role of parsimony in causal explanations multiple times in our chapter. Although models of parsimony (e.g., Chater & Vitányi, 2003; Lombrozo, 2007) are consistent with the psychological findings, they do not predict them. Better integration of theories of simplicity with theories and findings in causal learning would be a major advance.

Acknowledgments

Preparation of this chapter was supported by AFOSR FA 9950-08-1-0489. We thank Jessica Walker and Keith Holyoak for comments on an earlier draft.

Notes

[1] More generally, for a causal tree with n nodes, the number of direct causal links would be $n - 1$ (because every node other than the root node has one and only one arrow going into it). But the number of associations between nodes (including causal ones) would be $n(n - 1)/2$, because every node in the tree is linked by an arrow to at least one other node, so that there is an non-zero association between every pair of nodes.

[2] Ulrike Hahn provided this interpretation.

[3] The example was provided by Sylvain Bromberger.

[4] Although the theory obtains different equations for estimating generative and preventive causal powers, the choice between the two equations does not constitute a free parameter. Which of the two equations applies follows from the value of ΔP. On occasions where ΔP = 0, both equations apply and make the same prediction, namely, that causal power should be 0, except in ceiling-effect situations. Here, the reasoner does have to make a pragmatic decision on whether she is evaluating the evidence to assess a preventive or generative relation, and whether the evidence at hand is meaningful for that purpose.

References

Ahn, W-K., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation vs. mechanism information in causal attribution. *Cognition*, *54*, 299–352.

Allan, L. G., & Jenkins, H. M. (1980). The judgment of contingency and the nature of response alternatives. *Canadian Journal of Psychology*, *34*(1), 1–11.

Anderson, J. R., & Sheu, C. F. (1995). Causal inferences as perceptual judgments. *Memory and Cognition*, *23*(4), 510–524.

Beckers, T., De Houwer, J., Pineno, O., & Miller, R. R. (2005). Outcome additivity and outcome maximality influence cue competition in human causal learning. *Journal of Experimental Psychology Learning Memory and Cognition*, *31*(2), 238–249.

Beckers, T., Miller, R. R., De Houwer, J., & Urushihara, K. (2006). Reasoning rats: Forward blocking in Pavlovian animal conditioning is sensitive to constraints of causal inference. *Journal of Experimental Psychology: General*, *135*, 92–102.

Blaisdell, A. P., Sawa, K., Leising K. J., & Waldmann, M. R. (2006). Causal reasoning in rats. *Science*, *311*, 1020–1022.

Booth, S. L., & Buehner, M. J. (2007). Asymmetries in cue competition in forward and backward blocking designs: Further evidence for causal model theory. *Quarterly Journal of Experimental Psychology*, *60*, 387–399.

Buehner, M. J., & Cheng, P. W. (1997). Causal induction: The power PC theory versus the Rescorla-Wagner model. In M. G. Shafto & P. Langley (Eds.), *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society* (pp. 55–60). Hillsdale, NJ: Erlbaum.

Buehner, M. J., & Cheng, P. W. (2005). Causal learning. In K. J. Holyoak & R. Morrison (Eds.), *Handbook of thinking and reasoning* (pp. 143–168). Cambridge, England: 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*(6),1119–11140.

Buehner, M. J., & Humphreys, G. R. (2009). Causal binding of actions to their effects. *Psychological Science*, *20*(1), 1221–1228.

Buehner, M. J., & Humphreys, G. R. (2010). Causal contraction: Spatial binding in the perception of collision events. *Psychological Science*, *21*(1), 44–48.

Buehner, M. J., & May, J. (2002). Knowledge mediates the timeframe of covariation assessment in human causal induction. *Thinking and Reasoning*, *8*(4), 269–295.

Buehner, M. J., & May, J. (2003). Rethinking temporal contiguity and the judgment of causality: Effects of prior knowledge, experience, and reinforcement procedure. *Quarterly Journal of Experimental Psychology Section A-Human Experimental Psychology*, *56A*(5), 865–890.

Buehner, M. J., & May, J. (2004). Abolishing the effect of reinforcement delay on human causal learning. *Quarterly Journal of Experimental Psychology*, *57B*(2), 179–191.

Buehner, M. J., & McGregor, S. (2006). Temporal delays can facilitate causal attribution: Towards a general timeframe bias in causal induction. *Thinking and Reasoning*, *12*(4), 353–378.

Carroll, C. D., & Cheng, P. W. (2010). The induction of hidden causes: Causal mediation and violations of independent causal influence. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 913–918). Portland, OR: Cognitive Science Society.

Carroll, C. D., Cheng, P. W., & Lu, H. (2010). Uncertainty and dependency in causal inference. In Catrambone, R. & Ohlsson, S. (Eds.*)*, *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1076–1081). Portland, OR: Cognitive Science Society.

Chapman, G. B., & Robbins, S. J. (1990). Cue interaction in human contingency judgment. *Memory and Cognition*, *18*(5), 537–545.

Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Science*, *7*, 19–22.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367–405.

Cheng, P. W. (2000). Causality in the mind: Estimating contextual and conjunctive causal power. In F. Keil & R. Wilson (Eds.), *Cognition and explanation* (pp. 227–253). Cambridge, MA: MIT Press.

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

Cheng, P., & Novick, L. (2005). Constraints and nonconstraints in causal learning: Reply to White (2005) and to Luhmann and Ahn (2005). *Psychological Review*, *112*(3), 694–707.

Cicerone, R. A. (1976). Preference for mixed versus constant delay of reinforcement. *Journal of the Experimental Analysis of Behavior*, *25*, 257–261.

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

De Houwer, J., & Beckers, T. (2002). A review of recent developments in research and theories on human contingency learning. *Quarterly Journal of Experimental Psychology: Comparative and Physiological Psychology*, *55B*(4), 289–310.

De Houwer, J., Beckers,T., & Glautier, S. (2002). Outcome and cue properties modulate blocking. *Quarterly Journal of Experimental Psychology*, *55A*, 965–985.

De Houwer, J., Beckers, T., & Vandorpe, S. (2005). Evidence for the role of higher-order reasoning processes in cue competition and other learning phenomena. *Learning and Behavior*, *33*, 239–249.

Dennis, M. J., & Ahn, W-K. (2001). Primacy in causal strength judgments: The effect of initial evidence for generative versus inhibitory relationships. *Memory and Cognition*, *29*(1), 152–164.

Denniston, J.C., Savastano, H. I., & Miller, R. R. (2001). The extended comparator hypothesis: Learning by contiguity, responding by relative strength. In R. R. Mowrer & S. B. Klein (Eds.), *Handbook of contemporary learning theories* (pp. 65–117). Mahwah, NJ: Erlbaum.

Dickinson, A., & Burke, J. (1996). Within-compound associations mediate the retrospective revaluation of causality judgements. *Quarterly Journal of Experimental Psychology: Comparative and Physiological Psychology*, *49B*(1), 60–80.

Garcia, J., McGowan, B. K., Ervin, F. R., & Koelling, R. A. (1968). Cues: Their relative effectiveness as a function of the reinforcer. *Science*, *160*(3829), 794–795.

Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review*, *118*(1), 110–119.

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

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, *111*(1), 3–32.

Green, L., & Myerson, J. (2004). A discounting framework for choice with delayed and probabilistic rewards. *Psychological Bulletin*, *130*(5), 769–792.

Greville, W. J., & Buehner, M. J. (2010). Temporal predictability facilitates causal learning. *Journal of Experimental Psychology: General*, *139*(4), 756–771.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51*(4), 285–386.

Haggard, P., Clark, S., & Kalogeras, J. (2002). Voluntary action and conscious awareness. *Nature Neuroscience*, *5*(4), 382–385.

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

Hawking, S., & Mlodinow, L. (2010). *The grand design*. New York: Bantam Books.

Holyoak, K. J., & Cheng, P. W. (2011). Causal learning and inference as a rational process: The new synthesis. *Annual Review of Psychology*, *62*, 135–163.

Holyoak, K. J., & Hummel, J. E. (2000). The proper treatment of symbols in a connectionist architecture. In E. Dietrich & A. Markman (Eds.), *Cognitive dynamics: Conceptual change in humans and machines* (pp. 229–263). Mahwah, NJ: Erlbaum.

Hume, D. (1739/1888). A treatise of human nature. Oxford, England: Clarendon Press.

Humphreys, G. R., & Buehner, M. J. (2009). Magnitude estimation reveals temporal binding at super-second intervals. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(5), 1542–1549.

Humphreys, G. R., & Buehner, M. J. (2010). Temporal binding of action and effect in interval reproduction. *Experimental Brain Research*, *203*(2), 465–470.

Jenkins, H., & Ward, W. (1965). Judgment of contingencies between responses and outcomes. *Psychological Monographs*, *7*, 1–17.

Kamin, L. J. (1969). Predictability, surprise, attention and conditioning. In B. A. Campbell & R. M. Church (Eds.), *Punishment and aversive behavior* (pp. 279–296). New York: Appleton Century Crofts.

Kant, I. (1781/1965). *Critique of pure reason*. London: Macmillan.

Lagnado, D. A., & Sloman, S. A. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(4), 856–876.

Lagnado, D. A., & Sloman, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology-Learning Memory and Cognition*, *32*(3), 451–460.

Lewis, C. I. (1929). *Mind and the world order*. New York: Scribner.

Lien, Y. W., & Cheng, P. W. (2000). Distinguishing genuine from spurious causes: A coherence hypothesis. *Cognitive Psychology*, *40*(2), 87–137.

Liljeholm, M., & Cheng, P. W. (2007). When is a cause the "same?" Coherent generalization across contexts. *Psychological Science*, *18*, *1014–1021*.

Lober, K., & Shanks, D. R. (2000). Is causal induction based on causal power? Critique of Cheng (1997). *Psychological Review*, *107*(1), 195–212.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, *55*, 232–257.

López, F. J., Cobos, P. L., & Caño, A. (2005). Associative and causal reasoning accounts of causal induction: Symmetries and asymmetries in predictive and diagnostic inferences. *Memory and Cognition*, *33*, 1388–1398.

Lovibond, P. F., Been, S. L., Mitchell, C. J., Bouton, M. E., & Frohardt, R. (2003). Forward and backward blocking of causal judgment is enhanced by additivity of effect magnitude. *Memory and Cognition*, *31*(1), 133–142.

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

Lucas, C. G., & Griffiths, T. L. (2010). Learning the form of causal relationships using hierarchical Bayesian models. *Cognitive Science*, *34*,113–147.

Mermin, N. D. (2005). *It's about time: Understanding Einstein's relativity*. Princeton, NJ: Princeton University Press.

Michotte, A. E. (1946/1963). *The perception of causality* (T. R. Miles, Trans.). London: Methuen & Co.

Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla-Wagner model. *Psychological Bulletin*, *117*, 363–386.

Miller, R. R., & Matzel, L. D. (1988). The comparator hypothesis: A response rule for the expression of associations. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 22, pp. 51–92). San Diego, CA: Academic Press.

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences*, *32*, 183–198.

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

Ochs, E., & Capps, L. (2001). *Living narrative: Creating lives in everyday storytelling*. Cambridge, MA: Harvard University Press.

Pearce, J. M. (1987). A model for stimulus generalization in Pavlovian conditioning. *Psychological Review*, *94*(1), 61–73.

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

Pearson, K. (1911). *The grammar of science*. (3rd ed.). New York: Meridian Books.

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

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

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

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

Salmon, W. C. (1989). Four decades of scientific explanation. In P. Kitcher & W. C. Salmon (Eds.), *Minnesota studies in the philosophy of science. Vol. 13: Scientific explanation* (pp. 3–219). Minneapolis: University of Minnesota Press.

Savastano, H. I., & Miller, R. R. (1998). Time as content in Pavlovian conditioning. *Behavioural Processes*, *44*(2), 147–162.

Schustack, M. W., & Sternberg, R. J. (1981). Evaluation of evidence in causal inference. *Journal of Experimental Psychology: General*, *110*, 101–120.

Shanks, D. R. (1985). Forward and backward blocking in human contingency judgement. *Quarterly Journal of Experimental Psychology: Comparative and Physiological Psychology*, *37B*(1), 1–21.

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

Shanks, D. R., Pearson, S. M., & Dickinson, A. (1989). Temporal contiguity and the judgment of causality by human subjects. *Quarterly Journal of Experimental Psychology Section B- Comparative and Physiological Psychology*, *41*(2), 139–159.

Shepard, R. N. (2008). The step to rationality: the efficacy of thought experiments in science, ethics, and free will. *Cognitive Science*, *32*, 3–35.

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

Spirtes, P., Glymour, C., & Scheines, R. (1993/2000). *Causation, prediction and search* (2nd ed.). Boston, MA: MIT Press.

Stetson, C., Cui, X., Montague, P. R., & Eagleman, D. M. (2006). Motor-sensory recalibration leads to an illusory reversal of action and sensation. *Neuron*, *51*(5), 651–659.

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

Stout, S. C., & Miller, R. R. (2007). Sometimes-competing retrieval (SOCR): A formalization of the comparator hypothesis. *Psychological Review*, *114*(3), 759–783.

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

Tenenbaum, J. B., Kemp, C., & Griffiths, T., & Goodman, N. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*, 1279–1285.

Thagard, P. (2000). Explaining disease: Correlations, causes, and mechanisms. In F. Keil & R. Wilson (Eds.), *Cognition and explanation* (pp. 227–253). Cambridge, MA: MIT Press.

Van Hamme, L. J., & Wasserman, E. A. (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning and Motivation*, *25*(2), 127–151.

Waldmann, M. R. (2000). Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 53–76.

Waldmann, M. R. (2001). Predictive versus diagnostic causal learning: Evidence from an overshadowing paradigm. *Psychonomic Bulletin and Review*, *8*, 600–608.

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., & Hagmayer, Y. (2005). Seeing versus doing: Two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 216–227.

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General*, *121*(2), 222–236.

Waldmann, M. R., & Holyoak, K. J. (1997). Determining whether causal order affects cue selection in human contingency learning: Comments on Shanks and Lopez (1996). *Memory and Cognition*, *25*(1), 125–134.

White, P. A. (2002). Causal attribution from covariation information: The evidential evaluation model. *European Journal of Social Psychology*, *32*(5), 667–684.

Wohlschläger, A., Haggard, P., Gesierich, B., & Prinz, W. (2003). The perceived onset time of self- and other-generated actions. *Psychological science*, *14*(6), 586–591.

Wu, M., & Cheng, P. W. (1999). Why causation need not follow from statistical association: Boundary conditions for the evaluation of generative and preventive causal powers. *Psychological Science*, *10*(2), 92–97.

Yin, H., Barnet, R. C., & Miller, R. R. (1994). Second-order conditioning and Pavlovian conditioned inhibition: Operational similarities and differences. *Journal of Experimental Psychology: Animal Behavior Processes*, *20*, 419–428.

Yuille, A. L., & Lu, H. (2008). The noisy-logical distribution and its application to causal inference. In J. C. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in neural information processing systems* (Vol. 20, pp. 1673–1680). Cambridge, MA: MIT Press.

Zimmer-Hart, C. L., & Rescorla, R. A. (1974). Extinction of Pavlovian conditioned inhibition. *Journal of Comparative and Physiological Psychology*, *86*, 837–845.