Causal Learning from Observations and Manipulations

David Danks

Department of Philosophy, Carnegie Mellon University; and
Institute for Human & Machine Cognition

Address correspondence to:
David Danks
Department of Philosophy
135 Baker Hall
Carnegie Mellon University
Pittsburgh, PA 15213
ddanks@cmu.edu
(412) 268-8047 (phone)
(412) 268-1440 (fax)

## Introduction

Data from the world only have value if we can use them in some way. For many uses, such as predicting the weather, we only need to understand the correlational structure among the various features. For other purposes, though, we must know something about the causal structure of the environment, including other people: for example, to make accurate decisions, we must know the likely effects of our actions; to explain events in the world, we need to know what could have caused them; to predict other people's actions, we must know how their beliefs, desires, etc. cause them to act in particular ways; and so on. We extract causal beliefs from the patterns we see in the world, even though we never directly observe causal influence (Hume, 1748), and then use those beliefs systematically in our cognizing (Sloman, 2005). We are thus faced with a fundamental psychological problem: how do we in fact learn causal relationships in the world, which we then use in myriad ways to adjust or control our environment?

In particular, we will focus here on the problem of learning novel causal relationships given only data from the world, and where our prior knowledge does not significantly aid our causal learning.[1] The paradigmatic learning situation involves trying to determine the causal relationships among a set of variables (e.g., hormone levels and a disease, or fertilizers and plant growth) given a series of cases, where they are observed one at a time, or perhaps in a summary. Most of the theories we will discuss further assume that the variables are binary (typically present/absent), and that we can use prior knowledge (including temporal information) to label the variables as either potential causes or the effect. Thus, we have a quite well-defined situation

[1] Alternately, one could ask how we exploit prior causal knowledge to infer novel causal relationships. That is, how do we leverage our prior knowledge in novel situations and environments to draw interesting causal conclusions? This question has been studied extensively by Woo-Kyoung Ahn and her colleagues under the heading of "mechanism theory" (Ahn & Bailenson, 1996; Ahn, Kalish, Medin, & Gelman, 1995; see also Lien & Cheng, 2000). Although not addressed here, there are potentially interesting connections with the work discussed here (Danks, 2005; Glymour, 1998), and it is important to recognize that data-driven causal learning is not the only kind.

and experimental design: given observations of a series or summary of cases of binary potential causes and effect, determine the causal "influence" of each potential cause on the effect (where "influence" is deliberately being left vague).

A number of different theories have been proposed to explain just how people solve this problem, and though there are some known theoretical results connecting pairs of theories, they are widely scattered. In addition, recent surveys of the literature (e.g., De Houwer & Beckers, 2002; Shanks, 2004) have primarily focused on comparisons of the theories to empirical work, rather than the interesting connections among the theories themselves. Understanding these connections is particularly important from an experimental design perspective, since that enables us to determine the class of problems on which the various theories make different predictions. This chapter is aimed at providing just such a unification of the theory space, rather than an answer to the question of which theory best fits the empirical data. The latter task is particularly challenging given the growing evidence that a range of learning strategies occur in experimental populations (Lober & Shanks, 2000; White, 2000).

The central problem identified above is underspecified in certain ways. In particular, the temporal relationships of the variables seem to be relevant—positively and negatively—for people's ability to learn causal relationships (Buehner & May, 2002, 2003, 2004; Hagmayer & Waldmann, 2002; Lagnado & Sloman, 2004). For example, if no prior knowledge is provided and the potential causes occur significantly prior to the effect, people will tend not to infer a causal relationship. Also, people's understanding of causal relationships seems to be partially task-dependent, as their experimental responses depend systematically on the probe question, particularly counterfactual vs. "influence" terminology (Collins & Shanks, 2006). These complications and subtleties in causal inference are clearly relevant from a theoretical point of

view, but they have not been a central focus of theoretical work to this point, and so we set them aside for the remainder of this chapter.

The study of data-driven induction of causal beliefs has recently grown substantially in both experiments and theories: at least twelve substantively different theories have been independently proposed just in the last ten years. Roughly speaking, there are two major dimensions on which theories of human causal learning vary: whether they describe dynamic or long-run learning; and whether they describe learning of causal parameters or of causal structure. The first dimension is quite natural and obvious: (a) "dynamic," if the theory describes changes in causal belief after each observed case; or (b) "long-run," if the theory describes the causal beliefs that the individual will hold after observing a "sufficiently long" sequence (i.e., when the causal beliefs have stabilized).[2] The second dimension—parameter vs. structure inference—is roughly the distinction between learning "*C* causes *E*" and learning "how strongly *C* causes *E*." Unfortunately, this characterization is not quite right, since parameter learning is a kind of structure learning: learning that *C* has a non-zero (or zero) causal strength implies having learned that *C* causes (or does not cause) *E*.

To get a more accurate picture of this second distinction, we will need to make a brief excursion in Section 3 into causal Bayesian networks (or simply, causal Bayes nets), a mathematical framework from computer science, statistics, and philosophy that has emerged in the past twenty years as a normative framework for modeling causal systems. We will return to the parameter/structure distinction in more detail in Section 3.2. Before that, however, Section 2 will survey a variety of dynamical and long-run causal learning theories proposed in the

[2] Some recent summaries (e.g., De Houwer & Beckers, 2002) seem to use "associative" vs. "non-associative" (or "probabilistic") where I use "dynamic" vs. "long-run." However, that work interprets the two classes of theories as competing, whereas I will argue that they are complementary. In addition, they rarely provide an explicit characterization of "associative," and so classification of new theories is not always obvious.

psychological literature. Section 3 will then describe the framework of causal Bayes nets, and detail more of the substantial relationships between the various theories. For all of these theories, we will focus on inference from observational data. Section 4 will turn to focus on the problem of inference from our manipulations and actions in the world around us. This shift in data source will reveal further interesting connections among the various theoretical accounts of human causal learning.

## A Menagerie of Models

Many psychological theories of human causal learning can be placed into four families of theories, centered on: the well-known Rescorla-Wagner model, a configural cue version of Rescorla-Wagner, Cheng's causal power approach, and hypothesis confirmation testing. In this section, we focus on characterizing these families in terms of the within-family relationships: connecting dynamical and long-run theories that share certain crucial features. In Section 3, we will return to the cross-family comparisons, as well as describing the causal Bayes net psychological theories. One historical note: although many of these theories are presented here as dynamical/long-run versions of each other, these connections were almost all established only after each individual theory had independently been proposed. This section is geared towards theoretical clarity, not historical accuracy. A summary of the theoretical relationships is provided in Table 1 in Section 3.2.

### Rescorla-Wagner, Its Descendants, and Probabilistic Contrasts

The Rescorla-Wagner (1972) model (henceforth, the R-W model) is the paradigmatic instance of an associationist learning theory: people have beliefs about the associative strengths between cues and an outcome, and given a novel case (i.e., a set of cues and an outcome), they change their beliefs based on the error in the prediction made by their current beliefs. More

precisely, if $V_i^t$ is the associative strength of cue *i* after case *t*, the R-W model predicts that $V_i$ after the next case will be given by $V_i^{t+1} = V_i^t + \Delta V_i^{t+1}$, where the latter term is:

$$\Delta V_i^{t+1} = \begin{cases} \alpha_i\beta_1\left(\lambda - \sum_{\text{cue } j \text{ appears}} V_j^t\right), & \text{if } V_i \text{ and the outcome both occur} \\ \alpha_i\beta_2\left(0 - \sum_{\text{cue } j \text{ appears}} V_j^t\right), & \text{if } V_i \text{ occurs and the outcome does not} \\ 0, & \text{if } V_i \text{ does not occur} \end{cases} \qquad (2.1.1)$$

where $\lambda$ is the maximum association supported by the outcome (usually assumed to be 1); $\alpha_i$ is the salience of cue *i*; and $\beta_1$ and $\beta_2$ are the learning rates when the outcome is present and absent, respectively (typically with $\beta_1 \geq \beta_2$). In the R-W model, associative strengths only change when their corresponding cues occur, and the change is proportional to the "error" between the actual occurrence (or absence) of the outcome and the predicted value of the outcome (given by the linear sum of associative strengths).

The R-W model has proven to be a very good model of many animal associative learning phenomena (Miller, Barnet, & Grahame, 1995), and has been proposed as a model of human causal induction by reinterpreting the associative strengths as perceived causal strengths, in which case "cues" are potential causes (Baker, Mercier, Vallee-Tourangeau, Frank, & Pan, 1993; Lober & Shanks, 2000; Shanks, 1995; Shanks & Dickinson, 1987).[3] The R-W model cannot be the correct model of human causal learning, however, because it mistakenly predicts that people will not update their causal beliefs about a potential cause after a case in which it is absent. Retrospective revaluation seemingly only occurs in animals in special conditions (e.g., Blaisdell,

[3] There is significant debate about whether this causal interpretation is a legitimate use of the R-W model. Michael Waldmann and his colleagues (Waldmann, 2000, 2001; Waldmann & Holyoak, 1992; Waldmann, Holyoak, & Fratianne, 1995) have argued that cues in the R-W model must be the variables learned about first, and so are not mapped onto potential causes in cases of diagnostic learning (that is, reasoning from effects to causes). On this more narrow view of the R-W model, the dynamic theory I discuss here corresponds to an analogue of the R-W model, not to the model itself.

Denniston, & Miller, 2001; Blaisdell & Miller, 2001), but has been found multiple times in human causal learning (e.g., Chapman, 1991; Van Hamme & Wasserman, 1994; Wasserman, Kao, Van Hamme, Katagiri, & Young, 1996). Van Hamme & Wasserman (1994; henceforth, VHW) and Tassoni (1995) have proposed modified R-W models in which associative strengths can change even when a cue is not presented. Although these theories differ in exact details, the basic intuition is the same: given information about the occurrence or non-occurrence of other cues and the outcome, the *absence* of a cue can be informative, and so we may need to "error correct" a cue's associative strength even when it does not occur. The short-run behavior of these descendants of the R-W model have only been partially explored.

There has been a long history of research into the behavior of the R-W model in the long run (a very limited sample of the research is: Chapman & Robbins, 1990; Cheng, 1997; Danks, 2003; Gluck & Bower, 1988; Sutton & Barto, 1981). For many interesting cases, the R-W model does not have well-defined asymptotic behavior.[4] Instead, we can only talk of equilibrium states, which are themselves sometimes quite difficult to calculate; Danks (2003) provides their most general characterization, as well as a general algorithm for determining them. With experimental designs and parameter values that are typical of standard practice, the R-W model ends up living in the neighborhood of the equilibrium value in the long-run. And for a large class of problems described below, the R-W model's equilibrium state turns out to be the conditional probabilistic contrasts for each variable, which were *independently* proposed as the basis for a long-run theory of human causal learning (Cheng & Novick, 1990, 1992; Spellman, 1996).

The conditional probabilistic contrast, often called conditional $\Delta P$, is essentially a measure of conditional association, and the conditional $\Delta P$ theory proposes that people's

[4] The R-W model only has well-defined asymptotes for systems that are (i) deterministic; and (ii) have a perfect equilibrium (Danks, 2003).

judgments of causal influence for each variable will be proportional to that variable's conditional $\Delta P$ (Cheng & Novick, 1990, 1992; Spellman, 1996). Suppose $Q$ is some specification of cue values for every cue *except* cue $i$: for example, cue 1 is present, cue 2 is absent, and so on. Since we have $n$ cues, there are $2^{n-1}$ different $Q$'s. Given some particular $Q$, the conditional probabilistic contrast for cue $i$ is:

$$\Delta P_{i.Q} = P(E \mid i \,\&\, Q) - P(E \mid \neg i \,\&\, Q). \qquad (2.1.2)$$

In other words, the conditional probabilistic contrast is the change in the outcome probability between the $i$-present and $i$-absent cases, conditional on $Q$.

The various conditional probabilistic contrasts for a particular cue need not all be equal, but might differ depending on the particular variable values in the conditioning set. In these cases, the conditional $\Delta P$ model does not make a determinate prediction. However, if the conditional probabilistic contrasts for each variable are defined and equal, then the R-W model's equilibrium state for that problem is exactly conditional $\Delta P$ (Danks, 2003). That is, whenever the conditional $\Delta P$ theory is well-defined, the R-W model makes the same long-run prediction. Moreover, in these cases, the VHW and Tassoni variations have the same equilibrium states as the R-W model (Danks, 2003); they all make the same long-run predictions as the conditional $\Delta P$ theory whenever the latter theory makes any prediction at all. We can thus naturally think of the R-W model and its descendants as dynamical implementations of the conditional $\Delta P$ theory.

Pearce and Configural Cues

In the R-W model, each cue has its own associative strength, and the associative strength of compound cues (e.g., $A$ and $B$ both occur) is just the sum of the individual associative strengths. Pearce's (1987) theory of associative learning reverses this picture: each compound cue has its own associative strength, and the associative strength of individual cues is derived

from those compound cue strengths. This theory was originally proposed in the animal learning literature, but it has since been proposed as a model of human causal learning (Lopez, Shanks, Almaraz, & Fernandez, 1998; Perales & Shanks, 2003). More formally, using notation slightly different from Pearce, we define $S(Q \rightarrow R)$ to be the extent of generalization from compound $Q$'s strength to compound $R$'s strength. So, for example, if $V_{XC}$ is the associative strength for the compound cue $XC$ (and this is the only compound cue in which $C$ occurs), then the associative strength for the individual cue $C$ (due to $XC$'s associative strength) is given by $S(XC \rightarrow C) \times V_{XC}$. Pearce (1987) assumed the generalization parameters were symmetric; Perales and Shanks (2003) remove that assumption.

The associative strength of a compound cue changes only when that compound is presented. If we let $\delta(E) = 1$ if the outcome occurs and 0 otherwise, then given the presentation of compound cue $Q$ on trial $t$+1, the change in the strength of $Q^t$ is given by:

$$\Delta V_Q^{t+1} = \beta\left(\lambda\delta(E) - \left(V_Q^t + \sum_{R \in Compounds} S(Q \rightarrow R)V_R^t\right)\right) \qquad (2.2.1)$$

Given some set of updated compound cue associative strengths, the associative strength for some individual cue $C$ is the weighted (by the generalization parameters) sum of the associative strengths of all compound cues containing $C$. As with the R-W model, this model only rarely has true asymptotics, but has well-defined equilibrium states. The equilibrium states for one individual cue $C$ and a constant background $X$ were given in Perales & Shanks (2003), and in our notation, they are:

$$V_C = S(CX \rightarrow C) \times \frac{\lambda\left(P(E \mid C) - S(CX \rightarrow X)P(E \mid \neg C)\right)}{1 - S(CX \rightarrow X)S(X \rightarrow CX)} \qquad (2.2.2)$$

$$V_X = \frac{\lambda\left(P(E \mid \neg C) - S(X \rightarrow CX)P(E \mid C)\right)}{1 - S(CX \rightarrow X)S(X \rightarrow CX)} \qquad (2.2.3)$$

The equilibrium states for more complicated situations involving multiple individual cues have not been determined, but can easily be calculated using the matrix method of Danks (2003), since we will again have *n* equations in *n* unknowns (though here the *n* unknowns are the compound cue strengths, rather than the individual ones). There are currently no known equivalencies between the equilibrium states of Pearce's model and any other independently proposed long-run theories of causal or associative learning (except the connection with causal Bayes net parameter estimation provided in Section 3.2).

Causal Power Estimation

The R-W, Δ*P*, Pearce, and other models all essentially try to model the observed statistics. No particular metaphysics is proposed to explain the occurrence of those statistics; they are simply learned. Patricia Cheng's power PC theory incorporates a quite different picture of human causal learning: it posits that humans assume (or operate as if making the assumption) that the influence of a cause on its effect cannot be directly observed, and so the task of causal learning is to determine the strength of that unobserved influence (Buehner & Cheng, 1997; Buehner, Cheng, & Clifford, 2003; Cheng, 1997; Novick & Cheng, 2004). Focusing on generative causes (i.e., those that cause the effect to occur, rather than prevent it), each is presumed to have some capacity—in the sense of Cartwright (1989)—to bring about the effect. Moreover, the presence of the cause is necessary, but not sufficient for the operation of the capacity; the cause's capacities might sometimes fail to operate.[5] We also suppose that there is some always-present, generative background cause (whose capacity to produce the effect also operates only probabilistically). If the occurrence of *C* and the operation of its capacity are independent of the operation of the background cause's capacity, then $p_C$, the probability that *C*'s

[5] For those who favor a purely deterministic metaphysics, the theory works out exactly the same if we suppose that each generative cause always brings about the effect, unless there is a cause-specific, unobserved, not-always-present preventive cause that disables (in some way) the generative cause's operation.

capacity operates, can be estimated from purely observational data using the following equation derived in Cheng (1997):

$$p_C = \frac{\Delta P_C}{1 - P(E \mid \neg C)} \tag{2.3.1}$$

$p_C$ is a corrected $\Delta P$, where the correction factor accounts for the fact that some instances of $C$'s capacity operating will also be instances in which the background cause's capacity was operating. That is, sometimes $E$ will be doubly-produced, and so any estimate of $C$'s causal influence must include some information about the likelihood of $E$'s being doubly-caused. A slightly different equation is used to estimate preventive causal power.

In the power PC theory, the estimation in equation (2.3.1) does not necessarily occur over all cases, but only over those in a "focal set": a set of cases in which the reasoner judges the cause's occurrence and operation to be independent of the background's operation. The focal set also supports the extension of the power PC theory to multiple potential causes: for each variable, causal power is estimated for a set of cases in which the reasoner judges the various causes' occurrences and operation to be independent of each other. Typical examples of focal sets are: all cases; all cases in which one potential cause is always present; all cases in which a potential cause is always absent; and so on. No fully-specified dynamical theory for focal set selection has been proposed.

The power PC theory is an asymptotic theory: it predicts people's beliefs in the long-run, when those beliefs have stabilized. Equation (2.3.2) gives a dynamical theory whose equilibrium states are the power PC predictions, where the $V_k$'s are generative causes, and the $V_j$'s are preventive causes (Danks, Griffiths, & Tenenbaum, 2003).

$$\Delta V_i = \alpha_i \beta_{\delta(E)} \left( \lambda \delta(E) - \prod_{\delta(V_k)=1} (1 - V_k) \left[ 1 - \prod_{\delta(V_j)=1} (1 - V_j) \right] \right) \tag{2.3.2}$$

This dynamical theory is analogous to the VHW and Tassoni variations on the R-W model, except that a different prediction function is used. Rather than simply taking the sum of the present potential causes' strengths, we integrate the present potential causes according to the underlying metaphysics of the power PC theory.

Dis/Confirming Evidence

A quite different way of approaching the problem of causal inference is to suppose that people are explicitly testing the hypothesis that *C* causes *E*. The most direct way to do so is to track one's evidence for and against the hypothesis (Catena, Maldonado, & Candido, 1998; White, 2003a, 2003c). The cases that confirm the hypothesis are those in which both *C* and *E* are either present or absent, and cases that disconfirm it are those in which *C* and *E* differ on presence/absence. More specifically, White's proportion of Confirming Instances (pCI) theory predicts that judgments of *C*'s causal strength will be proportional to:

$$\text{pCI} = [P(C \,\&\, E) + P(\neg C \,\&\, \neg E)] - [P(C \,\&\, \neg E) + P(\neg C \,\&\, E)] \qquad (2.4.1)$$

That is, causal judgments are predicted to be proportional to the difference between the relative frequencies of confirming and disconfirming instances. (The theory's name is thus a little misleading.) We can also consider a version of pCI in which the various probabilities are differentially weighted (White, 2003c) to reflect the possibility that some kinds of evidence might be more important than others, perhaps because of rarity or some other asymmetry (McKenzie & Mikkelsen, 2000; see also White, 2003b). These weights can easily be incorporated into equation (2.4.2) below, and do not make a substantial theoretical difference, and so we ignore them for the present purposes.

In Catena, *et al.*'s (1998) belief adjustment model, the judgment of *C*'s causal strength is given by an updating equation: $J_i = J_{i-1} + \gamma \times (NewEvidence - J_{i-1})$, where $\gamma$ is a learning rate

parameter (called $\beta$ by Catena, *et al*.), and *NewEvidence* is just pCI (or possibly weighted pCI) for the cases seen since $J_{i-1}$. That is, the belief adjustment model says that people do not track causal strength based on a whole sequence of cases, but rather update their beliefs based on the difference between their last judgment and the evidence seen since that judgment. When people only make one judgment for a whole series, then the belief adjustment model makes the same predictions as the (weighted) pCI model. If people make multiple judgments during observation of a series, then $\gamma$ controls the importance of recent cases: if $\gamma = 0$, then no learning occurs from one judgment to the next; if $\gamma = 1$, then only cases observed since the last judgment matter; intermediate values correspond to various weightings of recent versus past data. If the inter-judgment observation distribution is stationary (i.e., the number of cases of each type is the same in every between-judgment interval), then $\gamma$ indicates how rapidly the belief adjustment model converges to the (weighted) pCI value. The belief adjustment model obviously depends crucially on people's judgment frequencies, but no non-experimental account of judgments has been provided; that is, Catena, *et al*. (1998) do not say when multiple judgments occur in the real world. We thus focus on the pCI theory since it simultaneously functions as a critical part of the belief adjustment model, and is fully-specified for more realistic learning situations.

A shortcoming shared by both pCI and the belief adjustment model is that neither has been extended to more complicated situations in which there are multiple potential causes. This extension is critical, given the range of experimental phenomena that require multiple causes (such as blocking phenomena). One natural extension of the theories would be to continue using the pCI equation in its current form, where the various probabilities now must be computed by summing over the different possible states of the other potential causes. This extension determines a value we can call the "unconditional pCI" for each potential cause. We could also

consider extensions to "conditional pCI": the unconditional pCI given a fixed specification of the other variable values. As with conditional $\Delta P$, there will be multiple conditional pCIs for a single variable. When that value is well-defined, the behavior of the conditional pCI theory is easily determinable. Its behavior in other conditions will depend substantially on the way the theory is cashed out. To our knowledge, none of these extensions has been endorsed or tested by proponents of the pCI or belief adjustment theories.

Notwithstanding the above concerns, we can naturally inquire about the existence of a dynamical theory for estimating pCI.[6] As with the dynamical theories for conditional $\Delta P$ (i.e., R-W and variants) and power PC, we require an updating equation for the strength estimate for the potential cause given the observation of some case. For the pCI theories, we update $C_j$'s strength estimate with the following equation:

$$\Delta V_j^{i+1} = \beta\left((-1)^{|\delta(E)-\delta(C_j)|}\lambda - V_j^i\right) \qquad (2.4.2)$$

That is, we update the strength estimate based on the difference between the current estimate for $C_j$ and either (i) $\lambda$, if $C_j$ and $E$'s presence/absence is the same; or (ii) $-\lambda$, if their presence/absence is different. Note that the presence or absence of other potential causes is completely irrelevant to the estimate for $C_j$. The equilibrium states for this updating function are the unconditional pCI values for the particular causal learning situation.

Unfortunately, the pCI theory simply cannot work in the real world as it is currently stated, for the same reason that Hempel's (1965) theory of scientific confirmation failed: for some "$C$ causes $E$" claim, there will be many $\neg C$, $\neg E$ things, but very few of those instances will

[6] To my knowledge, this is the first appearance of a dynamical theory for pCI or the between-judgment periods in the belief adjustment model.

actually provide any meaningful confirmation of the causal hypothesis.[7] As an extreme example, consider the causal claim that dropping objects on the moon ($C$) causes them to turn into wedges of green cheese ($E$). Every experience that I have had in my lifetime is an instance of $\neg C, \neg E$ with respect to this causal claim. Therefore, since the evidential weight of $\neg C, \neg E$ cases is supposed to be at least the same order of magnitude as for $C, E$ cases in the pCI theory, I should think it highly likely that the causal claim is true. But clearly these $\neg C, \neg E$ observations in fact give me no real information at all about the causal claim. In order to save the pCI theory, we need to provide some account of which $\neg C, \neg E$ instances count as relevant for judgments about "$C$ causes $E$"; this is essentially another version of the "frame problem" (McCarthy & Hayes, 1969). Of course, there is no ambiguity in experimental settings about which cases are relevant, but the real world is substantially more complicated. One natural move would be to use only pragmatically selected cases; unfortunately, no well-specified theory of the relevant pragmatics has been offered.

Moreover, even if we restrict our attention to artificial situations in which the pCI theory is closer to being well-defined (e.g., an experiment), it still has a strange implication: there must be a deep flaw in one (or both) of (i) our psychological experiments, or (ii) our actual causal cognition. Specifically, suppose that we have equal weights on the four terms in the pCI sum (unequal weights are discussed in fn. 9). Except in very particular circumstances (specifically, $P(C) = 0.5$), if $C$ and $E$ are statistically independent, then pCI $\neq 0$; and if they are associated, then

[7] A similar problem arises for the $\Delta P$ theory. Since almost all real-world cases will be $\neg C$ (and many will be $\neg E$), an unrestricted application of the $\Delta P$ theory should lead to $P(E \mid \neg C) \approx P(E)$, which will be approximately zero for many $E$. But this is clearly not the intended application of the $\Delta P$ theory (Shanks, 1993). The power PC theory and probabilistic contrast models avoid this problem by appealing to "pragmatically determined focal sets." The various associative theories either (i) ignore cases in which the cause is absent (e.g., R-W and Pearce); or (ii) only update on unexplained, "salient" absences of the cause (e.g., VHW and Tassoni).

there is a range of cases for which pCI = 0.[8] For the highly specific situations tested in an experiment—only two, temporally ordered, variables, no unobserved common causes, no other anomalous circumstances—every plausible theory of actual causal influence says that *C* causes *E* if and only if *C* and *E* are associated. Thus, the pCI theory says: when $P(C) \neq 0.5$, (a) if *C* does not cause *E*, then people will conclude *C does* cause *E*, since pCI ≠ 0; and (b) sometimes when *C* does cause *E*, people will conclude the opposite, since pCI = 0. So even if the pCI theory is the best explanation of people's responses in certain experiments (and there is reason to doubt this; see Griffiths & Tenenbaum, 2005), either our experiments are not measuring what we think, or else people are systematically wrong in their causal attributions.[9] The latter possibility, systematic error, would be quite surprising in light of our success in moving through our world.

## Causal Learning with Bayes Nets

The previous section focused on a variety of psychological theories of human causal learning. We now turn to consider a normative framework for representing causal structures: causal Bayes nets. The causal Bayes net framework originally emerged from a mixture of statistics, computer science, and philosophy, and has successfully been applied in a variety of contexts (examples from a wide range of fields include: Bessler, 2003; Conati, Gertner, Van Lehn, & Druzdzel, 1997; Cooper, Fine, Gadd, Obrosky, & Yealy, 2000; Lerner, Moses, Scott, McIlraith, & Koller, 2002; Ramsey, Gazis, Roush, Spirtes, & Glymour, 2002; Shipley, 2000; Smith, Jarvis, & Hartemink, 2002; Waldemark & Norqvist, 1999). In this section, I first outline

[8] The derivation is straightforward: pCI and $\Delta P$ make the same prediction if and only if $P(C) = 0.5$. Since $\Delta P = 0$ if and only if *C* and *E* are statistically independent, we can conclude: if $P(C) \neq 0.5$, then if pCI = 0, then *C* and *E* are associated. The converse does not hold, since there are many ways for pCI to differ from a non-zero $\Delta P$, but still be non-zero itself. Nevertheless, there are some cases in which pCI ≠ 0, but *C* and *E* are independent.

[9] One might hope to save pCI by using unequal weights. While this move helps somewhat for the $P(C) \neq 0.5$ situation, it actually harms the theory for the $P(C) = 0.5$ situation. Specifically, regardless of $P(C)$, we have: (i) For all situations and a measure one set of weights, [statistical independence ⇒ pCI ≠ 0]; and (ii) there are situations and a measure zero set of weights such that [statistical association ⇒ pCI = 0].

the causal Bayes net framework, and several psychological theories of human causal learning that have been based on it. This discussion is not intended as a formal introduction to causal Bayes nets; many other, more comprehensive introductions are available elsewhere (e.g., Glymour & Cooper, 1999; Pearl, 2000; Spirtes, Glymour, & Scheines, 1993). I then show how Bayes nets provide a powerful *lingua franca* in which to express almost all of the other extant psychological theories. Also, one of the real strengths of causal Bayes nets is their ability to model manipulations in the causal system. We will return to that issue in more detail in Section 4. One final note before continuing: the term "Bayesian" has become loaded with substantial theoretical baggage, and it is important to realize that the word "Bayes" in the framework name is due only to historical accident[10]; there is nothing intrinsically Bayesian about causal Bayes nets.

An Introduction, and Applications to Human Causal Learning

Suppose we have a set of variables.[11] In this setting, the variables might be the various cues and outcomes, and the possible values for each variable would be present or absent; more complicated sets of variables are also possible. Additionally, if we have time series data, the variables might be time-indexed. A causal Bayes net is composed of two related pieces: (i) a directed acyclic graph (DAG) over the variables; and (ii) a probability distribution over the variables. In the DAG, there is a node for each variable, and $X \rightarrow Y$ means "*X* causes *Y*," though no particularly strong metaphysical theory of causation is required; Woodward (2003) carefully explores the metaphysical commitments of causal Bayes nets. The probability distribution is a specification of the probability of all possible combinations of variable values. These two components are connected by two assumptions:

[10] They were originally used to improve performance on Bayesian updating, principally in medical diagnosis.
[11] I will assume throughout that the variables are discrete. Structural equation models are the continuous-variable analogues of causal Bayes nets, and every claim in this chapter also holds for them.

**Causal Markov Assumption:** Every variable is independent of its non-effects conditional on its direct causes.

**Causal Faithfulness Assumption:** The only probabilistic independencies are those entailed by the causal Markov assumption.

These two assumptions are essentially claims about the ways in which causal structure reveals itself in observable statistics or probabilities. They could be empirically false in certain domains, but there are reasons to think that they hold of many systems (Glymour, 1999). Moreover, there is growing evidence that people naturally make these assumptions, particularly the causal Markov assumption (Gopnik, Glymour, Sobel, Schulz, Kushnir, & Danks, 2004).

The two components of a causal Bayes net describe different kinds of causal information: the DAG encodes the qualitative causal structure; and the probability distribution encodes the quantitative types and strengths of (i.e., the parameters for) the various causal influences. Thinking about causal Bayes nets in this way helps illuminate the two assumptions: the causal Markov assumption says (roughly) that we don't need to determine the parameters for variables not connected by an edge since there is no direct causal influence; the causal Faithfulness assumption says (roughly) that the parameter values do not obscure the causal structure, for example by multiple causal pathways exactly offsetting each other. In this framework, the distinction drawn at the end of Section 1 between structure and parameter learning corresponds to learning about the DAG and probability distribution, respectively.

In addition to modeling known causal structures, a causal Bayes net can be learned from purely observational data given the causal Markov and Faithfulness assumptions (experimental and mixed data are discussed in Section 4). That is, given purely observational data and these two assumptions about the way in which causation reveals itself in associations, we can often

recover substantial parts of the actual causal structure. This result is perhaps surprising, given the oft-repeated mantra that "correlation does not imply causation." While this statement is true for single pairs of variables, it is not true for *patterns* of correlations (given these two assumptions). A simple example might help to show why. First, define a "causal connection" between *X* and *Y* to be one or more of: (i) *X* causes *Y*; (ii) *Y* causes *X*; or (iii) there is an unobserved common cause of *X* and *Y*. Now suppose that we have three variables, *A*, *B*, and *C*, and that the only independence among these variables is between *A* and *C*, unconditionally. From these data, we can conclude that (i) there is a causal connection between *A* and *B*, as well as *B* and *C*; but (ii) *B* does not cause either *A* or *C*; and (iii) there is no causal connection (direct or indirect) between *A* and *C*. To illustrate just one of these conclusions, consider (ii), and suppose instead that *B* causes *C*. That implies that there must be an indirect causal connection between *A* and *C*, where the exact nature of the connection depends on the causal connection between *A* and *B*. But an indirect causal connection implies unconditional association (because of the causal Faithfulness assumption), which contradicts the known data. Hence, we can conclude that *B* does not cause *C*.

A variety of learning algorithms have been developed within the machine learning community over the past fifteen years that exploit this fact that we can infer some causal structure from patterns of correlations (Chickering, 2002; Heckerman, 1998; Spirtes, *et al.*, 1993). Although the algorithms differ in important details, they all infer possible causal structures from patterns of conditional and unconditional associations and independencies. That is, the algorithms determine (in varying ways) the set of causal structures that could have, or were likely to have, produced data such as that actually observed. Roughly speaking, these algorithms divide into two camps: constraint-based algorithms determine the full set of possible causal Bayes nets (including those with unobserved common causes) from the pattern of

statistically significant independencies and associations; Bayesian and score-based algorithms search through the space of graphs, typically in a heuristic manner, to find causal structures that are highly probable, given the observed data. Recovering causal connections from observation correlations using these algorithms is more than a theoretical possibility: causal Bayes nets have been applied in a wide range of domains for both causal discovery and various types of inference (see references in the introduction to this section). That being said, they are not an ideal representation for some domains, such as feedback or epidemiological models. Perhaps more importantly for cases of human causal learning, causal Bayes nets do not currently provide good models of continuous time phenomena, though continuous time Bayes nets are the subject of ongoing research (Nodelman, Shelton, & Koller, 2002, 2003).

For other situations, causal Bayes nets provide an excellent representational framework for causal relationships; the causal learning situations modeled by psychological theories form one such class of suitable situations. Thus, a natural strategy would be to test whether people represent and learn these causal structures as though they were causal Bayes nets. A number of different researchers have pursued this line of thinking, which has resulted in essentially two different kinds of causal Bayes net-based psychological theories. One approach has been to use causal Bayes net learning algorithms to provide a computational-level account (i.e., a rational analysis) of causal learning (Gopnik & Glymour, 2002; Gopnik, *et al.*, 2004; Gopnik, Sobel, Schulz, & Glymour, 2001; Griffiths & Tenenbaum, 2005; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003; Tenenbaum & Griffiths, 2001, 2003).[12] That is, this work focuses on understanding the high-level relationship between the observed cases and people's causal

[12] These papers use a mix of constraint-based and Bayesian learning algorithms. However, the differences are not as great as they might appear. In particular, since constraint-based algorithms do not care about the source of the association/independence judgments, one could easily use Bayesian statistics to calculate the associations and independencies. See Danks (2004) for more details.

judgments, without necessarily arguing for any particular algorithm by which those judgments are reached. Although specific learning algorithms are, of course, used in deriving predictions, no particular descriptive claim is intended by their use. Rather, the claim here is that people act rationally (i.e., according to the normative prescriptions of causal Bayes net learning algorithms), even though they might not actually be using causal Bayes net learning algorithms. For example, Tenenbaum & Griffiths (2001) argue that people make judgments as though they are Bayesian learners of causal Bayes nets, but note that $\chi^2$ estimates are a simple, close approximation of the "rational" prediction, and would be indistinguishable for all of their experiments.

A quite different use of the causal Bayes net framework is to argue that people are essentially doing top-down search over causal Bayes net structures (Lagnado & Sloman, 2002, 2004; Waldmann, 2000, 2001; Waldmann & Martignon, 1998). Sometimes referred to as "causal model theory," the representation of causal beliefs was inspired in part by causal Bayes nets (particularly Pearl, 1988), but uses an independently developed learning mechanism (Waldmann, 1996; Waldmann & Holyoak, 1992; Waldmann, *et al.*, 1995). Specifically, learning is top-down: people use prior beliefs (e.g., beliefs about temporal order, intervention status, or other prior domain knowledge) to determine an initial causal structure, estimate the strengths of influences based on that structure, and then revise their beliefs about the underlying causal structure only as necessary. This last step is obviously of central importance, but has not been significantly explored. This approach can also be represented as a set of restrictions on the learning process in the rational analysis approach (e.g., as a highly skewed prior distribution for Bayesian learning of causal Bayes nets).

The causal Bayes net approaches are not necessarily direct competitors of the other psychological theories. In particular, both of the causal Bayes net approaches typically assume

some situation-dependent prior knowledge constraints on the particular form(s) that causal influence can take, where these restrictions often correspond to one of the other extant psychological theories. That is, the causal Bayes net approaches typically agree with one or another of the other psychological theories about causal influence estimation *once the causal structure has been learned*. The causal Bayes net approaches differ in that they posit a structure learning step that is conceptually—and sometimes algorithmically—prior to the parameter learning step. They argue that people do not infer the quantitative causal influence (the parameters) until they determine that there is a qualitative causal influence (the graphical structure). We will return to this connection in Section 4.

A Metatheoretic Structure Based On Causal Bayes Nets

This connection between the causal Bayes net approaches and the other extant psychological theories points to an idea: perhaps *all* of the other psychological theories can be completely explained as doing parameter estimation on some fixed causal structure. This intuition turns out to be exactly right, as we can represent the range of theories in a single, metatheoretical structure using the causal Bayes net framework. Consider the DAG in Figure 1, where $B$ is some always-occurring background variable, and $w_C$ and $w_B$ are parameters associated with the edges (and used below).

---

INSERT FIGURE 1 ABOUT HERE

---

To turn this DAG into a causal Bayes net, we must also provide a probability distribution for $C$ and $E$. Since $C$ is an exogenous variable (i.e., one with no cause within the system), we need provide only its base rate. For the distribution for $E$, we can specify a function whose "free

parameters" are the $w_B$ and $w_C$ parameters in Figure 1, and whose input variables are whether each parent variable occurs.[13] For example, the probability of $E$ might be given by $w_B$, plus $w_C$ when $C$ occurs: $P(E) = w_B + w_C \times \delta(C)$. (Remember that $\delta(X) = 1$ if $X$ is present, 0 otherwise, and that $B$ is always present.) For this function for $P(E)$, $\Delta P_C$ is the maximum likelihood estimate of $w_C$ (Tenenbaum & Griffiths, 2001).[14] That is, the one-potential cause $\Delta P$ theory can be interpreted as a maximum likelihood estimate of a parameter in a fixed-structure causal Bayes net with a particular functional form.

In fact, we can provide an even stronger result. Consider the DAG shown in Figure 2.

---

INSERT FIGURE 2 ABOUT HERE

---

The conditional $\Delta P$ for variable $X$ is a maximum likelihood estimate of $w_X$ when the probability of $E$ is the sum of the $w$-parameters for the occurring variables (Tenenbaum & Griffiths, 2001):

$$P(E) = \sum w_X \times \delta(X) \qquad (3.2.1)$$

And given the equivalence between conditional $\Delta P$ and the equilibrium states of the R-W model (and VHW and Tassoni variations), we can reinterpret these dynamical theories as algorithms for learning the maximum likelihood values of the parameters. That is, all of the theories in section 2.1 are slightly different ways to do parameter estimation in a fixed-structure, fixed-functional form causal Bayes net.

Alternately, suppose the functional form for $P(E)$ in Figure 1 is given by:

$$P(E) = w_C \times \delta(C) + w_B - w_C \times w_B \times \delta(C) \qquad (3.2.2)$$

[13] This is not the only way to specify $P(E)$, but it is convenient for our purposes.
[14] All proofs are omitted due to space considerations, but are available upon request from the author.

The causal power in Cheng's power PC theory is the maximum likelihood estimate of $w_C$ in this function (Glymour, 1998; Tenenbaum & Griffiths, 2001). For multiple causes, causal power is the maximum likelihood estimate of the *w*-parameters in Figure 2, where the functional form is the natural multivariate extension of equation (3.2.2): a multivariate noisy-or gate for generative causes, or a noisy-and gate for preventive causes (Danks, *et al.*, 2003). No results are known about the "parameter estimation properties" of the recent extension of the power PC theory to interactive causes (Novick & Cheng, 2004). The power PC theory and corresponding dynamical theory are thus maximum likelihood parameter estimators for the exact same causal structure as the conditional $\Delta P$ theory; they make different predictions because they assume different functional forms for $P(E)$ in that causal Bayes net.

The story is a bit more complicated for Pearce's theory and its equilibrium states because of the generalization parameters. For one potential cause (Figure 1), $V_C$ in Pearce's theory is the maximum likelihood estimate of the $w_C$ term (and $V_X$ for the $w_B$ term) for the $P(E)$ function:

$$P(E) = S(CB \rightarrow B)^{\delta(C)} \times w_B + \frac{S\left(B \rightarrow CB\right)^{\left(1-\delta(C)\right)}}{S\left(CB \rightarrow B\right)} \times w_C. \qquad (3.2.3)$$

That is, in Pearce's theory, the probability of the effect is a weighted sum of both *w*-parameters, where the weights depend on whether *C* is present or absent. It is currently unknown whether the above equation for $P(E)$ can be extended to the multiple-cause situation depicted in Figure 2.

Not all of the previously discussed theories can be represented in this way, however. In particular, the pCI theory cannot be represented as a parameter estimator for the causal Bayes net in Figure 1. Since it is unclear how to extend pCI to multiple potential causes, we focus here on the one-cause situation. In the causal Bayes net framework, if *C* and *E* are unconditionally independent, then the causal Faithfulness assumption implies that there cannot be a graphical edge between them (i.e., *C* does not cause *E*), and so there must be a zero *w*-parameter in Figure

1. Therefore, for any theory that can be represented as a causal Bayes net parameter estimator, it must be the case that unconditional independence between $C$ and $E$ (in the one-cause situation) implies a zero $w$-parameter. pCI fails this requirement: as noted earlier, if $P(C) \neq 0.5$ and $C$ and $E$ are independent, then pCI $\neq 0$. Therefore, there cannot be a (Faithful) functional form for the causal Bayes net in Figure such that pCI is the maximum likelihood estimate of $w_C$.[15]

Given all of these results, we can place these various theories into three distinct columns in the single, metatheoretic structure shown in Table 1, where names are provided for the more common theories. (Literature references for each cell can be found in the sections above. Numbers in parentheses indicate equations.)

---

INSERT TABLE 1 ABOUT HERE

---

Each row of the table represents a class of theories: the first four rows contain various parameter estimators, and the fifth row describes the "native" causal Bayes net structure learning algorithms. For the parameter estimators (the first four rows), there are shared relationships between the columns: (i) the long-run behavior (typically, the equilibrium state) of the dynamical models is the asymptotic model; (ii) the asymptotic model is a maximum likelihood estimate of the $w$-parameters in the Bayes net function (for the causal Bayes net of Figure 2); and (iii) the Bayes net function is—for the first three rows—the prediction function for the error-correction term of the dynamical models. For the structure learning algorithms, the relationships are a bit different, since they are also learning the graphical structure of the causal Bayes net. That row is included primarily to highlight the contrasts with the parameter estimation theories.

[15] Griffiths & Tenenbaum (2005) have provided a rational "approximate justification" of pCI as causal Bayes net structure learning, but their reconstruction requires one to make highly implausible assumptions.

## Not All Data Are Created Equal: Learning from Manipulations

The data used by the theories in Section 2 are purely observational: cases where we only see the naturally occurring values of each variable. But we often get data from our manipulations of the causal systems around us; a simple example is flipping a light switch to figure out what it causes. Moreover, there can be a substantial difference between observing and manipulating a variable. The observation that someone has nicotine stains on her fingers licenses the inference that she (probably) smokes; intervening to force someone to have nicotine stains on her fingers eliminates the support for the inference to her smoking. The variable values in an observation are due to the causal structure that is "in the world"; in contrast, manipulating a variable changes the causal structure so that a variable's value depends on us, rather than its normal causes. In this example, the nicotine stains in the second case are due solely to our manipulation, which is why we cannot infer anything about the person's smoking behavior.

In this example, the manipulation yielded less information than the observation, but sometimes manipulations are more informative. Consider the simple case in which $X$ and $Y$ are observed to be associated. All we can conclude is that there is some causal connection between $X$ and $Y$, but we don't know what it is: (i) $X$ causes $Y$; (ii) $Y$ causes $X$; (iii) an unobserved common cause of $X$ and $Y$; or some combination of these possibilities. Now suppose that we can manipulate $X$ and $Y$ independently, and then check whether they are associated. The outcomes of the manipulations will depend on the underlying causal structure, and so we summarize the inferences we can make in each possible outcome pair in Table 2.

---

INSERT TABLE 2 ABOUT HERE

---

For example, suppose that *X* = nicotine stains (present or absent) and *Y* = smoking (present or absent). These variables will be associated if we manipulate *Y*, but not if we manipulate *X* (for the reasons discussed earlier). We are thus in the upper-right hand cell of the table, and so we can (correctly) conclude that smoking causes nicotine stains, and that there might also be an unobserved common cause of the two variables. Observations alone would only tell us that there is some causal connection between them, but not its form. Being able to manipulate the variables thus led to more learning than given observations. In general, manipulations give us more information, particularly about direction, for individual causal connections, but at the cost of changing the causal structure. Observations show us the full causal structure, but at the cost of reduced information about the specific causal connections. Sometimes manipulations are the best way to learn; sometimes observations are superior; often, a combination is best.

We might wonder whether people can exploit this informational difference in learning. In fact, recent research suggests that we learn causal structure substantially *better* when we can manipulate the causal system (Gopnik, *et al.*, 2004; Kushnir, Gopnik, Schulz, & Danks, 2003; Lagnado & Sloman, 2004; Schulz & Gopnik, 2004; Sloman & Lagnado, 2004; Sobel & Kushnir, 2003, In press; Steyvers, *et al.*, 2003). Furthermore, we can understand this advantage by considering the representation of manipulations within the causal Bayes net framework (Pearl, 2000; Spirtes, *et al.*, 1993). A manipulation on target *X* is represented by the introduction of a new direct cause of *X* that represents the manipulation occurring or not. When the manipulation does not occur, the causal system functions as normal; the causal influence of the manipulation is simply inactive. When the manipulation occurs, the other causes of the manipulated variable no

longer matter, and so we can remove (or "break") those edges in the causal Bayes net.[16] This transition is shown in Figure 3.

---

INSERT FIGURE 3 ABOUT HERE

---

And notice that smoking and nicotine stains will be independent in the right-hand causal system, since there is no causal connection between the two. The causal Bayes net representation of manipulations thus correctly captures our intuitions.

In addition to providing an excellent representation of the impact of manipulations, the causal Bayes net framework also gives a natural account of learning from manipulations. That is, the learning algorithms discussed in Section 3.1 can be straightforwardly adjusted to incorporate exclusively post-manipulation data, or even mixtures of observational and manipulation data. Moreover, there are also causal Bayes net accounts of "active learning": choosing the manipulation or experiment (or series of manipulations) that maximally reduces one's uncertainty about the underlying structure (Eberhardt, Glymour, & Scheines, 2006; Tong & Koller, 2001). Because of this natural integration of manipulations into the causal Bayes net framework, no adjustment is needed for any of the psychological accounts of human causal learning that are based directly on that framework.

The story is more complicated for the "traditional" psychological theories (i.e., those in Sections 2.1-2.4), since none of them explicitly discuss the observation/manipulation distinction. In fact, the lack of this distinction in the traditional theories, and the importance of the distinction

[16] As a technical aside, a variable being "edge-breaking" is actually sufficient but not necessary for it to count as a "manipulation" on this scheme. A more precise characterization of 'manipulation' can be given in terms of sources of variation in the target variable that are independent of the other variables in the system. See the Manipulation Theorem of Spirtes, *et al*. (1993) for a precise statement.

in the causal Bayes net framework, has been a crucial motivation of much recent experimental research on human causal learning. However, the re-description of those theories in terms of parameter estimation in a specific causal Bayes net provides one explanation for the lack of focus on this distinction: namely, that there is no observation/manipulation distinction for the potential causes in the fixed causal Bayes net in Figure 2. Given that we know (or assume) that the causal system has the structure in Figure 2, we can make exactly the same inferences—either about the parameters or about the likelihood of the effect occurring—given either (a) observations that the potential causes have some values; or (b) manipulations to force the potential causes to have exactly the same values. In addition, if we want to know which variable to manipulate to bring about the effect, we can simply use the observational probabilities to determine which variable would be most efficacious. This result might seem a bit surprising, but notice that the manipulations all take place on variables that have no edges directed into them, and so the manipulation does not break any causal connections (within the system). The manipulation/observation distinction only matters when the manipulation is on some variable that has causes in this causal structure. None of the potential causes in Figure 2 meets this requirement, and so the distinction is not relevant for learning or inference. Of course, the parameter estimation theories could resist this reinterpretation, but then they must provide some explanation of the observation/manipulation distinction, which seems to be quite important in human causal learning.

If we apply this reinterpretation to the traditional psychological (parameter estimation) theories, then they can explain the manipulation data and experiments, though at a cost. First, there is a potential rhetorical cost. Several of the theories were originally developed within the animal learning community (e.g., Pearce, 1987; Rescorla & Wagner, 1972), and so are

sometimes accompanied by rhetoric about there being no distinction between learning covariations and causation, or about causal learning being just a type of covariation learning (De Houwer & Beckers, 2002). That rhetoric is no longer legitimate in this reinterpretation of the parameter estimation theories, since the observation/manipulation distinction is now being explicitly drawn in the framework. We just happen to know (or assume) a causal structure in which we can learn, predict, and make decisions equally well given the two kinds of information. In this framework, there really is a difference between beliefs about correlations and beliefs about causation, but they happen to coincide in these particular learning situations.

More importantly, the traditional psychological theories must make a choice about theoretical scope in this reinterpretation. One option is to argue that the parameter estimation theory explains and predicts all parts of data-driven causal learning. That is, to argue that people assume the causal structure in Figure 2, and so cannot learn different causal structures. This strategy is unlikely to succeed, since there is substantial experimental evidence that people can learn other causal structures, such as a chain, or a common cause (Lagnado & Sloman, 2002; Steyvers, *et al.*, 2003; Waldmann, *et al.*, 1995), and even that rats can learn such structures (Blaisdell, Sawa, Leising, & Waldmann, 2006). Alternately, one could narrow the scope of the parameter estimation theory to apply only *after* the causal structure has been determined (where some other mechanism handles the structure learning). This option results in a theory such as Waldmann & Martignon (1998), in which people estimate each "edge parameter" according to the power PC theory, but use another algorithm to determine the structure in which estimation occurs (see also Danks, *et al.*, 2003; Tenenbaum & Griffiths, 2003). In this option, the parameter estimation theories explain less than has previously been thought.

## Conclusion

Many of our cognitive activities presuppose beliefs about causal relationships in the world, and a range of theories have been proposed to explain how we make causal inferences from our observations and manipulations of the world around us. The primary concern in the psychological literature to this point has been on the successes and failures of these theories at predicting various experimental data. This focus has led to less exploration of the relationships among the theories. Despite the fact that many of the theories were independently proposed, there are numerous interesting relationships among them. In particular, many of the theories are dynamical or long-run versions of each other; there are connections across explanatory levels that provide a better understanding of the "theory space," as well as support the design of better crucial experiments among the theories. For example, we can see that there is little point to performing an experiment to distinguish between the R-W model and (the long-run version of) the conditional $\Delta P$ theory, since the former is a dynamical version of the latter.

More importantly, we can use the framework of causal Bayes nets to demonstrate that most of the extant psychological theories have essentially the same structure: they are parameter estimators for a fixed-structure, fixed-functional form causal Bayes net, where the precise functional form differs between the theories. These theories (almost) all focus on the estimation of quantitative strengths of causal influence, and thereby infer causal structure only indirectly (through inference of zero causal strength). Moreover, various theoretical considerations – particularly the observation/manipulation distinction – point towards the vital importance of correct inference of causal structure. But rather than concluding (as one might) that the lack of structure learning in the parameter estimation theories implies that they are deeply flawed, we can again use the causal Bayes net framework to show the vital role played by these theories: they provide accounts of the types of causal "functions" that people will consider when inferring

causal structure. The “theory space” of functional forms has been extensively explored in the past fifteen years of research on human causal learning; distinguishing the various possible structure learning algorithms and determining their empirical accuracy remains a substantial open research problem.

Acknowledgements

Maralee Harrell, Clark Glymour, Alison Gopnik, and Michael Waldmann provided useful comments on earlier drafts. Thanks also to the audience of the 33rd Carnegie Symposium on Cognition, particularly Luis Barrios, Christian Schunn, and Priti Shah, for valuable questions and comments.

References

Ahn, W.-K., & Bailenson, J. (1996). Causal attribution as a search for underlying mechanisms: An explanation of the conjunction fallacy and the discounting principle. *Cognitive Psychology, 31,* 82-123.

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

Baker, A. G., Mercier, P., Vallee-Tourangeau, F., Frank, R., & Pan, M. (1993). Selective associations and causality judgments: Presence of a strong causal factor may reduce judgments of a weaker one. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 19,* 414-432.

Bessler, D. A. (2003). *On world poverty: Its causes and effects*. Rome: Food and Agricultural Organization of the United Nations.

Blaisdell, A. P., Denniston, J. C., & Miller, R. R. (2001). Recovery from the overexpectation effect: Contrasting performance-focused and acquisition-focused models of retrospective revaluation. *Animal Learning & Behavior, 29,* 367-380.

Blaisdell, A. P., & Miller, R. R. (2001). Conditioned inhibition produced by extinction-mediated recovery from the relative stimulus validity effect: A test of acquisition and performance models of empirical retrospective revaluation. *Journal of Experimental Psychology: Animal Behavior Processes, 27,* 48-58.

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

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 19th*

*annual conference of the cognitive science society* (pp. 55-60). Mahwah, NJ: LEA Publishers.

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, & Cognition, 29,* 1119-1140.

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

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

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

Cartwright, N. (1989). *Nature's capacities and their measurement*. Oxford: Oxford University Press.

Catena, A., Maldonado, A., & Candido, A. (1998). The effect of the frequency of judgment and the type of trials on covariation learning. *Journal of Experimental Psychology: Human Perception and Performance, 24*, 481-495.

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

Chapman, G. B., & Robbins, S. J. (1990). Cue interaction in human contingency judgments. *Memory & Cognition, 18*, 537-545.

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

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

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

Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of Machine Learning Research, 3*, 507-554.

Collins, D. J., & Shanks, D. R. (2006). Conformity to the Power PC theory of causal induction depends on the type of probe question. *The Quarterly Journal of Experimental Psychology, 59*, 225-232.

Conati, C., Gertner, A., Van Lehn, K., & Druzdzel, M. J. (1997). On-line student modeling for coached problem solving using Bayesian networks. In *Proceedings of the 6th international conference on user modeling* (pp. 231-242). Vienna: Springer-Verlag.

Cooper, G. F., Fine, M. J., Gadd, C. S., Obrosky, D. S., & Yealy, D. M. (2000). Analyzing causal relationships between treating clinicians and patient admission and mortality in low-risk pneumonia patients. *Academic Emergency Medicine, 7*, 470-471.

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

Danks, D. (2004). Constraint-based human causal learning. In M. Lovett, C. Schunn, C. Leviere & P. Munro (Eds.), *Proceedings of the 6th international conference on cognitive modeling* (pp. 342-343). Mahwah, NJ: Lawrence Erlbaum Associates.

Danks, D. (2005). The supposed competition between theories of human causal inference. *Philosophical Psychology, 18*, 259-272.

Danks, D., Griffiths, T. L., & Tenenbaum, J. B. (2003). Dynamical causal learning. In S. Becker, S. Thrun & K. Obermayer (Eds.), *Advances in neural information processing systems 15* (pp. 67-74). Cambridge, MA: The MIT Press.

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

Eberhardt, F., Glymour, C., & Scheines, R. (2006). N-1 experiments suffice to determine the causal relations among N variables. In D. Holmes & L. Jain (Eds.), *Innovations in machine learning*. Berlin: Springer-Verlag.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117,* 227-247.

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

Glymour, C. (1999). Rabbit hunting. *Synthese, 121,* 55-78.

Glymour, C., & Cooper, G. F. (1999). *Computation, causation, & discovery*. Cambridge, MA: AAAI Press & The MIT Press.

Gopnik, A., & Glymour, C. (2002). Causal maps and Bayes nets: A cognition and computational account of theory-formation. In P. Carruthers, S. Stich & M. Siegal (Eds.), *The cognitive basis of science* (pp. 117-132). Cambridge: Cambridge University Press.

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

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

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

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

Heckerman, D. (1998). A tutorial on learning with Bayesian networks. In M. I. Jordan (Ed.), *Learning in graphical models* (pp. 301-354). Boston: Kluwer.

Hempel, C. (1965). *Aspects of scientific explanation*. New York: Free Press.

Hume, D. (1748). *An enquiry concerning human understanding*. Oxford: Clarendon.

Kushnir, T., Gopnik, A., Schulz, L. E., & Danks, D. (2003). Inferring hidden causes. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the 25th annual meeting of the cognitive science society* (pp. 699-703). Boston: Cognitive Science Society.

Lagnado, D. A., & Sloman, S. A. (2002). Learning causal structure. In *Proceedings of the 24th annual conference of the cognitive science society*.

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

Lerner, U., Moses, B., Scott, M., McIlraith, S., & Koller, D. (2002). Monitoring a complex physical system using a hybrid dynamic Bayes net. In A. Darwiche & N. Friedman (Eds.), *Uncertainty in artificial intelligence: Proceedings of the 18th conference* (pp. 301-310). San Francisco: Morgan Kaufmann.

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

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

Lopez, F. J., Shanks, D. R., Almaraz, J., & Fernandez, P. (1998). Effects of trial order on contingency judgments: A comparison of associative and probabilistic contrast accounts. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 24,* 672-694.

McCarthy, J., & Hayes, P. J. (1969). Some philosophical problems from the standpoint of artificial intelligence. In D. Michie & B. Meltzer (Eds.), *Machine intelligence 4* (pp. 463-502). Edinburgh: Edinburgh University Press.

McKenzie, C. R. M., & Mikkelsen, L. A. (2000). The psychological side of hempel's paradox of confirmation. *Psychonomic Bulletin & Review, 7,* 360-366.

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

Nodelman, U., Shelton, C. R., & Koller, D. (2002). Continuous time Bayesian networks. In *Proceedings of the 18th international conference on uncertainty in artificial intelligence* (pp. 378-387).

Nodelman, U., Shelton, C. R., & Koller, D. (2003). Learning continuous time Bayesian networks. In *Proceedings of the 19th international conference on uncertainty in artificial intelligence* (pp. 451-458).

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

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

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

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

Perales, J. C., & Shanks, D. R. (2003). Normative and descriptive accounts of the influence of power and contingency on causal judgement. *The Quarterly Journal of Experimental Psychology, 56A*, 977-1007.

Ramsey, J., Gazis, P., Roush, T., Spirtes, P., & Glymour, C. (2002). Automated remote sensing with near infrared reflectance spectra: Carbonate recognition. *Data Mining and Knowledge Discovery, 6*, 277-293.

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 research and theory* (pp. 64-99). New York: Appleton-Century-Crofts.

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

Shanks, D. R. (1993). Associative versus contingency accounts of causal learning: Reply to Melz, Cheng, Holyoak, and Waldmann (1993). *Journal of Experimental Psychology: Learning, Memory, & Cognition, 19*, 1411-1423.

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

Shanks, D. R. (2004). Judging covariation and causation. In N. Harvey & D. Koehler (Eds.), *Blackwell handbook of judgment and decision making*. Oxford: Blackwell.

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

Shipley, B. (2000). *Cause and correlation in biology: A user's guide to path analysis, structural equations and causal inference*. Cambridge: Cambridge University Press.

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

Sloman, S. A., & Lagnado, D. A. (2004). Causal invariance in reasoning and learning. In B. H. Ross (Ed.), *The psychology of learning and motivation, vol. 44* (pp. 287-325): Elsevier.

Smith, V. A., Jarvis, E. D., & Hartemink, A. J. (2002). Evaluating functional network inference using simulations of complex biological systems. *Bioinformatics, 18*, S216-224.

Sobel, D. M., & Kushnir, T. (2003). Interventions do not solely benefit causal learning: Being told what to do results in worse learning than doing it yourself In R. Alterman & D. Kirsh (Eds.), *Proceedings of the 25th annual meeting of the cognitive science society*. Boston: Cognitive Science Society.

Sobel, D. M., & Kushnir, T. (In press). The importance of decision-making in causal learning from interventions *Memory & Cognition.*

Spellman, B. A. (1996). Conditionalizing causality. In D. R. Shanks, K. J. Holyoak & D. L. Medin (Eds.), *Causal learning: The psychology of learning and motivation, vol. 34* (pp. 167-206). San Diego, CA: Academic Press.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search*. Berlin: Springer-Verlag.

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

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review, 88,* 135-170.

Tassoni, C. J. (1995). The least mean squares network with information coding: A model of cue learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 21,* 193-204.

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

Tenenbaum, J. B., & Griffiths, T. L. (2003). Theory-based causal inference. In S. Becker, S. Thrun & K. Obermayer (Eds.), *Advances in neural information processing systems 15* (pp. 35-42). Cambridge, MA: The MIT Press.

Tong, S., & Koller, D. (2001). Active learning for structure in Bayesian networks. In *Proceedings of the 17th international joint conference on artificial intelligence* (pp. 863-869). Seattle: AAAI 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,* 127-151.

Waldemark, J., & Norqvist, P. (1999). In-flight calibration of satellite ion composition data using artificial intelligence methods. In C. Glymour & G. F. Cooper (Eds.), *Computation, causation, & discovery*. Cambridge, MA: AAAI Press & The MIT Press.

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

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

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

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

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

Waldmann, M. R., & Martignon, L. (1998). A Bayesian network model of causal learning. In M. A. Gernsbacher & S. J. Derry (Eds.), *Proceedings of the 20th annual conference of the cognitive science society*. Mahwah, NJ: Lawrence Erlbaum.

Wasserman, E. A., Kao, S.-F., Van Hamme, L. J., Katagiri, M., & Young, M. E. (1996). Causation and association. In D. R. Shanks, K. J. Holyoak & D. L. Medin (Eds.), *Causal learning: The psychology of learning and motivation, vol. 34* (pp. 207-264). San Diego: Academic Press.

White, P. A. (2000). Causal judgment from contingency information: Relation between subjective reports and individual tendencies in judgment. *Memory & Cognition, 28,* 415-426.

White, P. A. (2003a). Causal judgement as evaluation of evidence: The use of confirmatory and disconfirmatory information. *The Quarterly Journal of Experimental Psychology, 56A,* 491-513.

White, P. A. (2003b). Effects of wording and stimulus format on the use of contingency information in causal judgment. *Memory & Cognition, 31,* 231-242.

White, P. A. (2003c). Making causal judgments from the proportion of confirming instances: The pci rule. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 29,* 710-727.

Woodward, J. (2003). *Making things happen: A theory of causal explanation.* Oxford: Oxford University Press.

Table 1: Metatheoretic Structure

| **Dynamical Model** | **Long-run Model** | **Causal Bayes net function** |
|---|---|---|
| R-W and variants (2.1.1) | Conditional Δ*P* (2.1.2) | Sum of present cue strengths (3.2.1) |
| (Generalized) Pearce (2.2.1) | One-cue: (2.2.2) & (2.2.3); Multiple-cue: General procedure, but no equations | One-cue: Equation (3.2.3); Multiple-cue: Unknown |
| Equation (2.3.2) | Power PC (2.3.1) | Noisy-OR/AND (3.2.2) |
| Equation (2.4.2) | pCI/belief adjustment (2.4.1) | None exists |
| Bayesian updating; dynamic estimation of independencies and associations; testing the current causal model | Arbitrary causal Bayes net structure learning | Various possible functions, depending on prior knowledge or biases |

Table 2: Possible Causal Models Given Manipulations

| | *Independent after Y manip.* | *Associated after Y manip.* |
|---|---|---|
| *Independent after X manip.* | Unobserved common cause | *Y* causes *X*; and perhaps an unobserved common cause |
| *Associated after X manip.* | *X* causes *Y*; and perhaps an unobserved common cause | Each one causes the other (feedback loop); and perhaps an unobserved common cause |

Figure 1: Causal Bayes Net for Parameter Estimation

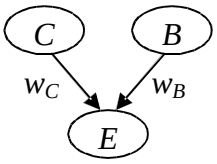

Figure 1: Full One-Layer Causal Bayes Net for Parameter Estimation

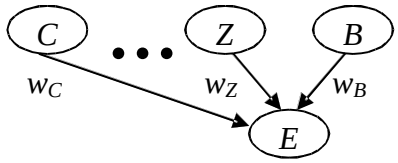

Figure 3: Example of Manipulation Representation