Learning Causal Structure through Local Prediction-error Learning

Sarah Wellen (swellen@andrew.cmu.edu)
Department of Philosophy, Baker Hall 135
Pittsburgh, PA 15213 USA

David Danks (ddanks@cmu.edu)
Department of Philosophy, Baker Hall 135
Pittsburgh, PA 15213 USA; and
Institute for Human & Machine Cognition, 40 S. Alcaniz St.
Pensacola, FL 32502 USA

Abstract

Research on human causal learning has largely focused on strength learning, or on computational-level theories; there are few formal algorithmic models of how people learn causal structure from covariations. We introduce a model that learns causal structure in a local manner via prediction-error learning. This local learning is then integrated dynamically into a unified representation of causal structure. The model uses computationally plausible approximations of (locally) rational learning, and so represents a hybrid between the associationist and rational paradigms in causal learning research. We conclude by showing that the model provides a good fit to data from a previous experiment.

Keywords: Causal learning; causal Bayes nets; prediction-error learning; algorithmic level

Introduction

From a young age, we spontaneously, and often effortlessly, come to understand the causal structure of the world, and then use that knowledge to both predict what might happen in the future and also design actions that will achieve our goals (e.g., Gopnik, *et al*., 2004; Sloman, 2005). Our focus here is causal learning from covariational data: how do people learn the causal structure of the world from a sequence of observations or interventions of that world?

Causal learning can usefully be separated into the related-but-distinct problems of representation and dynamics—*what* is learned and *how* is it learned. In this paper, we develop a novel account of causal learning that, at a high level, uses quasi-associationist processes to learn directed graph-like causal representations. It is thus a hybrid of the standard rationalist vs. associationist approaches to causal learning.

Representations of Causal Structure

The development of causal Bayesian networks prompted a major advance in our understanding of causal knowledge. A causal Bayes net has two components: (i) a directed acyclic graph (DAG) whose nodes represent variables and directed edges represent direct causal relations (see Figure 1); and (ii) a probability distribution that encodes how causes influence their effects. These two elements represent qualitative and quantitative causal structure, respectively, and are connected by a pair of assumptions (Markov and Faithfulness) that capture the ways in which causal structure manifests in observed data. Sloman (2005) and Spirtes, Glymour, & Scheines (1993) provide useful expositions of the causal Bayes net framework.

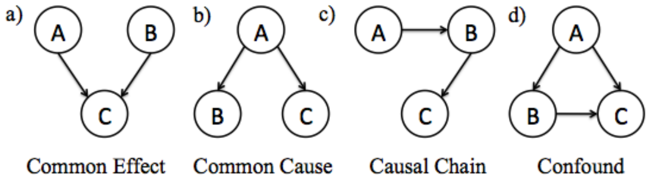

Figure 1: Prototypical 3-variable causal Bayes nets

There is substantial evidence that the type of structural knowledge captured by a causal Bayes net—or at least, the directed graphical model part—is necessary to account for many causal reasoning abilities. One hallmark of *causal* reasoning, rather than correlational, is that cases involving observations vs. interventions are treated differently (Sloman & Lagnado, 2005; Waldmann & Hagmayer, 2005). For example, one can infer, from an observation of a professor's gray hair, that she likely has many publications. No such inference follows if she instead intervened to dye her hair gray. Causal Bayes nets can straightforwardly account for this difference, as interventions are represented by 'graph surgery,' where a variable that is intervened upon is separated from its typical causes (Spirtes, *et al*., 1993). This surgery changes the informational relations, and so one's inferences can be different in the two situations.

Some aspects of causal knowledge are not easily represented by this formalism (e.g. the spatiotemporal relations between causes and effects), but it seems to provide a good account of people's representations of causal structure. Thus, we aim to develop a theory of causal learning in which people learn a directed graph (perhaps acyclic, though we will allow for cyclic structures).

Dynamics of Causal Structure Learning

Theories about how people use covariation to learn directed graph representations can be divided roughly into *rational* and *heuristic* accounts of causal learning. Rational accounts model causal learning as rational inference. These include constraint-based algorithms (e.g., Glymour, 2003; Gopnik, *et al*., 2004), and those based on Bayesian inference (e.g., Steyvers, *et al*., 2003; Griffiths & Tenenbaum, 2005). They are usually intended at the computational level of analysis, as they show how the cognitive system's performance solves the problem faced by that system, but do not attempt to characterize the underlying cognitive processes. There have been some recent attempts to develop algorithmic (i.e. process) models of causal learning based on approximations of Bayesian inference (e.g., Bonawitz, *et al*., 2011). These models have so far only addressed causal strength learning, and it is not clear how to extend them (in a computationally tractable manner) to structure learning.

Heuristic accounts of causal learning propose that people use various cues to suggest and modify causal hypotheses in a not-necessarily-rational (though presumably sensible) manner. Causal model theory (Waldmann, 1996) proposes that learners use cues such as covariation, temporal order, and spatial proximity to select an initial causal structure and adjust it in the face of inconsistent data (Lagnado, Waldmann, Hagmayer, & Sloman, 2007). Causal model theory has never been entirely formally specified, though some parts have received formal treatment.

The local computations model (Fernbach & Sloman, 2009) attempts to explain how learners use data from interventions to learn a causal structure. The key idea is that, when a variable is intervened upon and other variables change, the learner infers that the intervened-upon variable caused those other variables. Critically, all learning in this model is local, as people evaluate individual causal relations rather than entire graphs. The model we present here adopts this important insight and extends it to all covariation-based structure learning, including learning from observations.

The single-effect learning model (Waldmann, *et al*., 2008) also assumes that people focus on evaluating single causal relations. It is a model of learning from observations, and proposes that learners estimate the causal power (Cheng, 1997) of each potential cause of an effect. If a variable has sufficient (estimated) causal power, then the learner accepts the causal relation and integrates it with her previous causal knowledge. This model has found some empirical support in both humans and rats (Waldmann, *et al*., 2008).

Our model adopts the single-effect learning model's focus on causal power, and the integration of these individually learned relations into a unified causal structure. However, the standard causal power theory is a computational theory that makes no commitment to underlying processes. Danks, Griffiths, & Tenenbaum (2003) provided a prediction-error-based model of causal strength learning whose equilibrium states are causal powers, and so their model can be viewed as an algorithmic implementation of the causal power theory. Moreover, its basis in prediction-errors is consistent with neuroscientific evidence that the right lateral prefrontal cortex encodes prediction-error signals during causal learning (Corlett, *et al*., 2004; Turner, *et al*., 2004).

Another lacuna in the single-effect learning model is that it does not explain how the learner uses a causal power estimate to determine whether a link actually exists. We thus provide a decision procedure for causal relation acceptance based on both the learner's point estimate and her confidence in that estimate. This addition allows us to model the dynamics of learning for directed graphs that are more complex than the single-effect structure.

The LPL Model

The Local Prediction-error Learning (LPL) model aims to explain how observations and interventions are used to learn causal structure when one has relatively little prior knowledge. We do not model many other relevant sources of information, including verbal communication, reasoning, or spatiotemporal information. The model does assume that the learner knows the functional form of the causal relations and (when relevant) the expected temporal delay between causes and effects.

The LPL model begins with an initial causal structure hypothesis: a directed graph representing the individual's prior beliefs, where an edge indicates an *a priori* belief that there is a causal connection, and absence indicates only agnosticism.[1] For typical experiments in which participants have little prior knowledge, this will be an empty graph. The model alters this causal structure hypothesis by adding or removing single edges, thereby reducing the structure learning problem to the simpler task of evaluating individual causal relations. Multiple experimental results suggest that learners focus primarily on single causal relations (e.g., Gopnik *et al*., 2004; Waldmann, *et al*., 2008), presumably because of the computational complexity of evaluating larger structures.

Figure 2 shows a high-level overview of the LPL algorithm. The key pieces to be explained are the Causal Strength Estimates, and how the Decision Procedure changes the Causal Structure Hypothesis.

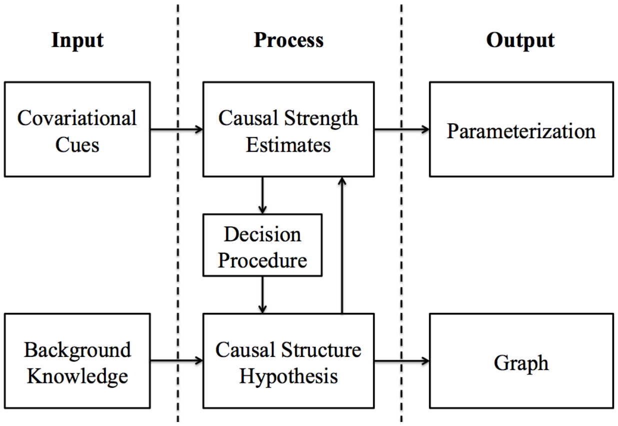

Figure 2: A high-level description of the LPL model

[1] The model can also encode *a priori* belief of definite edge absence, though we omit this complication for reasons of space.

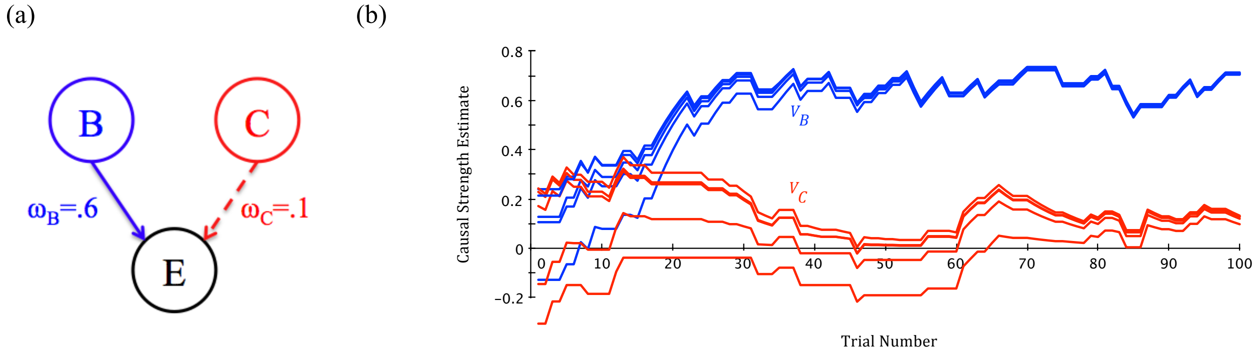

Figure 3: (a) Example local learning context (solid / dashed arrows indicate known / potential causal relations); (b) Causal strength estimates of *B* (blue) and *C* (red) using five particles per edge

Causal Strength Estimates

The LPL model generates causal strength estimates for each possible cause-effect pair that is not ruled out *a priori*. That is, for each pair of variables (*A*, *B*), the learner estimates the causal strength of *A*→*B* and *B*→*A* unless she has prior knowledge about potential edge direction. We assume here that the correct functional form for causal relations is noisy-OR, and so causal strengths are causal powers (Cheng, 1997; Griffiths & Tenenbaum, 2005), though this can change based on background information.

Like the single-effect learning model, we assume that the appropriate scope for learning causal strength includes the potential cause, the effect, and any other definite causes of that effect. For instance, consider Figure 3(a), where the learner believes that *B* causes *E* and is trying to determine whether *C* also causes *E*. Unlike the single-effect learning model, however, causal strength estimates for *C* are generated by a mechanism similar to particle filters in approximate Bayesian inference, though our "particles" move by associationist learning.

The LPL model initially draws *n* particles for each possible causal relation from a prior strength distribution determined by the learner's background knowledge. The learner's current beliefs about whether *C* causes *E* are represented by these particles $\{V_C^1,\ldots,V_C^n\}$. There is a corresponding set $\{V_B^1,\ldots,V_B^n\}$ of particles for *B*. The use of multiple particles enables the model to capture both strength estimates and confidence in those estimates. The mean particle value, $\overline{V}_C = \frac{1}{n}\sum_{i=1}^{n} V_C^i$, is the point estimate of *C*'s causal strength. The average squared deviation of the particles, $D_C = \frac{1}{n}\sum_{i=1}^{n}\left(V_C^i - \overline{V}_C\right)^2$, is the learner's (lack of) confidence: low values of D_C indicate high confidence.

We define a *layer i* of particles for an effect *E* as the *i*-th particle from each known and potential cause of *E*. In Figure 3(a), for example, layer *i* would be $\{V_B^i, V_C^i\}$. A layer of particles is a specific hypothesis about the strengths of all known and potential causes of *E*. Each layer is updated independently after each data point by prediction-error learning. Such learning can be represented schematically as:

$$V_C^{i,t+1} = V_C^{i,t} + \alpha\,(observed - expected)$$

The learning rate (α) is a free parameter, and *observed* has the value 1 if the effect occurs and 0 if it does not. The value of *expected* is typically the expected value of the effect variable, calculated using the functional form for the cause-effect relation. Many associationist learning models fit this schema, including the classic Rescorla-Wagner model and the causal power estimator of Danks, *et al*. (2003).

The *expected* value is computed separately for each potential cause in a layer. The current structure hypothesis has an influence because *expected* is based on only definite, known causes and the particular target potential cause for that update; other variables are ignored. This restriction reduces the computational demands on the learner, and fits real-world contexts where the learner cannot simultaneously attend to all the potential causes in her environment. If the causes combine as causal powers, then the expected value of *E* (for layer *i* and potential cause *C*) is:

$$expected = \prod_{K=present}\left(1-V_K^{t-1}\right)\left(1-\prod_{J=present}\left(1-V_J^{t-1}\right)\right)$$

where *J* (*K*) is the set of *E*'s generative (preventive) causes.

Figure 3(b) shows how initial causal strength estimates can change over time. Data were generated by Figure 3(a) with a noisy-OR functional form. At first, the particles are spread widely around zero, representing the learner's uncertainty in her estimate. As the learner observes more data points, prediction-error learning brings the particles closer to the true parameter values. The layers of particles that are further from the true values will generally have greater errors and thus will shift more towards the true values during learning. As a result, the estimates in different layers converge,[2] representing the learner's increasing confidence. This process gives no account of structure learning, however, so we turn to that now.

[2] Though they only stabilize around equilibrium values. If the learning rate is based on the learner's current (lack of) confidence, then true convergence is possible.

Causal Structure Judgments

The LPL model has a single, definite structure hypothesis at each point in time, which can then be modified by either adding or removing an edge. These modifications are based on a decision procedure applied to the causal strength estimates after each update.

Since an edge with a causal strength of zero is equivalent to no edge, the decision procedure uses a t-test on each set of particles with the null hypothesis that the particles are drawn from a distribution with mean $\mu = 0$. The outcome of this test depends on both the particles' mean and deviation. A free parameter $p_{critical}$ guides the decision procedure. If there is no edge in the graph and the t-test rejects the null hypothesis (i.e., the p-value p of the test statistic is less than $p_{critical}$), then an edge is added. If there is an edge present and the t-test does not reach significance (i.e., $p > p_{critical}$), then the edge is removed from the graph.

If a $C \rightarrow E$ edge is added or removed, future calculations of *expected* change for *other* potential causes of E, as those involve only the known causes of E. Crucially, this form of causal structure learning satisfices: the learner accepts the most plausible structure as a working hypothesis rather than representing and evaluating all possible structure hypotheses (as in standard Bayesian models).

Other Factors

Temporal information and the data source can influence the interpretation of covariational data, and so are also incorporated into the LPL model.

Interventions Given an observation about C and E, the LPL model updates the causal strength estimates for both $C \rightarrow E$ and $E \rightarrow C$ whenever the model does not yet know which direction the causal influence flows (if any). If C's value is instead set by intervention, then one knows that C is severed from its normal causes. Thus, one should not update causal strength estimates for potential causes of C. Operationally, if given data about an intervention on C, the LPL model updates only the $C \rightarrow E$ particles, and not the $E \rightarrow C$ ones.

Temporal Information Temporal delays between the cause and effect influence contingency learning, though mediated by the learner's expectations (Buehner & May, 2003; Buehner & McGregor, 2006). The LPL model compares the observed temporal difference $d_{E\text{-}C}$ between a potential cause and the effect to the expected temporal difference d_{typ}. If the learner expects the delay to always be d_{typ}, then the causal strength estimates update only when that delay occurs. If the learner expects the timeframe of the causal mechanism to be noisy, then the model reduces the salience of C—captured in the learning rate α—as a potential cause of E in proportion to $d_{err} = d_{E\text{-}C} - d_{typ}$. We define a learning rate α' that decreases exponentially as d_{err} increases: $\alpha' = \alpha e^{-\frac{d_{err}}{s}}$, where s is a scaling parameter that determines how sharply α' drops off as d_{err} increases.

Evaluating the LPL Model

Data

We evaluate the LPL model using data from Lagnado & Sloman (2006). In this experiment, participants had to discover the causal connections between four computers by sending 100 text messages to computer A and observing whether those messages were sent on to other computers. The true causal system is shown in Figure 4, where the arrows represent noisy causal relations. Messages always reached computer A, and the probability of a message being transmitted from one computer to the next was 0.8. Messages never spontaneously occurred.[3] Trial order was randomized both for participants and for modeling.

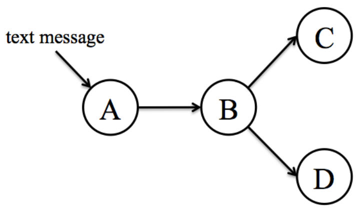

Figure 4: Causal structure from Lagnado & Sloman (2006)

The original experiment contrasted temporal and covariational information, so there were four conditions that varied the temporal order in which messages appeared. Condition 1 involved no timing information, but conditions 2-4 did (with different delays[4]).

LPL Model

Participants had no prior knowledge of causal structure, so the initial model was the empty graph (i.e., agnosticism). Connections between the computers were clearly generative, so the model only considered causal strength estimates between 0 and 1. For each possible edge, five particles were drawn from a truncated Gaussian ($\mu = 0$, $\sigma^2 = .2$).

The LPL Model has four other free parameters. The expected temporal delay d_{typ}, and the temporal scaling parameter s are not used with simultaneous occurrences (as in condition 1). We thus first determined the values for the learning rate α and the critical significance level $p_{critical}$ by maximizing model fit (via a grid search) for condition 1 only. Model fit was based on R^2 values[5] for the proportions, over all possible causal relations CR, of (a) 1000 model runs that yielded CR, and (b) experimental participants that

[3] The resulting case distribution (N = 100) was: 51 cases with ABCD; 13 AB¬CD; 13 ABC¬D; 3 AB¬C¬D; and 20 A¬B¬C¬D.

[4] The messages always appeared in the same order within conditions: A-B-D-C in Condition 2, A-D-C-B in Condition 3, and A-B-CD (C and D simultaneous) in Condition 4.

[5] $R^2 = 1 - (SS_{err} / SS_{tot})$, where SS_{err} and SS_{tot} are the sum of squared differences between the participant endorsement frequencies and the model proportion or mean endorsement, respectively.

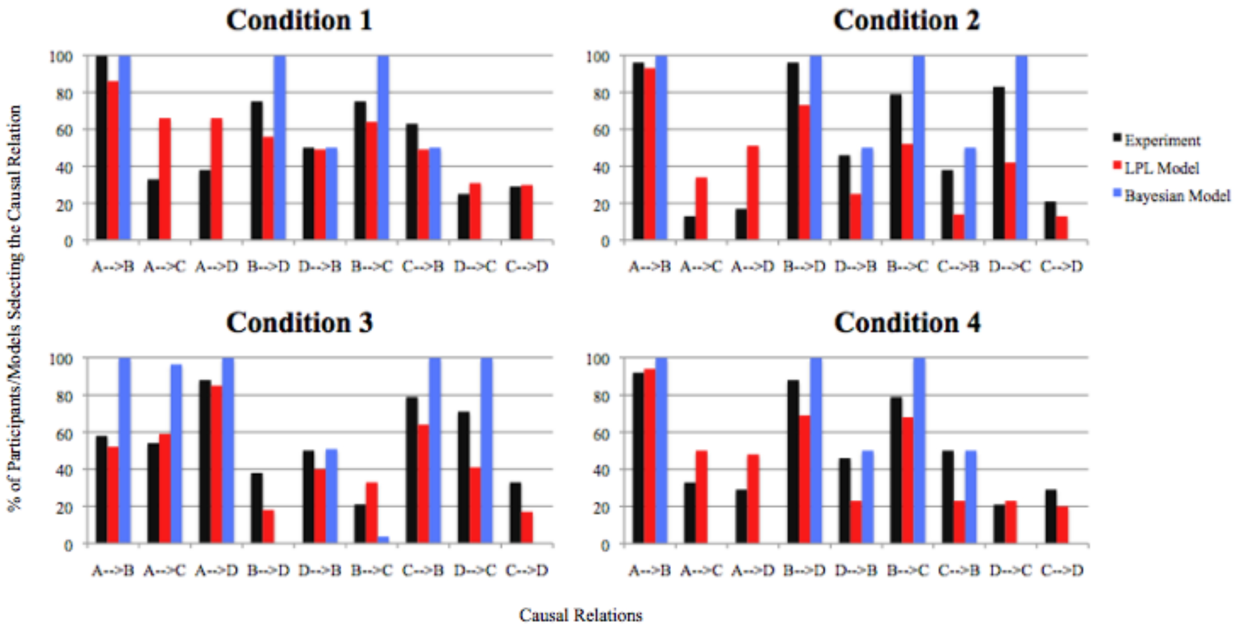

Figure 5: Proportion of causal relation endorsements by the LPL model, Bayesian model, and experimental participants

endorsed *CR*. The optimal model fit (R^2 = .47) was with $\alpha = 0.1$ and $p_{critical} = 7 \times 10^{-5}$.

These parameter values were used for all subsequent simulations. Model results for conditions 2-4 thus provide cross-validation for those parameter values. We set $d_{typ} = 1$, as the natural temporal delay between a computer sending a text message and one receiving it would be one time-step. We then searched and found that $s = 7$ optimized model fit across conditions 2-4 (R^2 = .47).

Bayesian Model

We compare the LPL model to a standard Bayesian model of causal structure learning. The model used a uniform prior over all possible graphs (cyclic and acyclic) over the four variables. The posterior probability of a graph *H* given the *j*-th datapoint is:

$$P(H_i|d_j) = \frac{P(d_j|H_i, do(A))P(H_i)}{P(d_j)}$$

If t_V denotes the time of *V*, then the likelihood is given by:

$$P(d_j|H_i, do(A)) = P(b,c,d,t_B,t_C,t_D|H_i, do(A))$$
$$= P(b,c,d|H_i, do(A))P(t_B,t_C,t_D|b,c,d,H_i, do(A))$$

Participants were told the true parameterization, so we use that distribution to calculate $P(b,c,d|H_i, do(A))$. For temporal sequences, the Bayesian model also assumed that delay probabilities followed an exponential decay function:

$$P(d_{E-C}) = \frac{1}{2s} e^{-\frac{|d_{err}|}{s}}$$
.[6]

This adjustment introduces a new free parameter, *s*, that was estimated by maximizing model fit across conditions 2-4 (s = 2, R^2 = .23). To determine Bayesian model predictions, we assumed that people probability match: the proportion of "Bayesian endorsements" for each causal relation *CR* was simply the posterior probability of *CR*.

Results and Discussion

Figure 5 shows the LPL and Bayesian model predictions, as well as the actual participant data. R^2 values for the models for each condition are shown in Table 1.

Table 1: R^2 values for the models

	LPL Model	Bayesian Model
Condition 1	.47	-.03[7]
Condition 2	.40	.81
Condition 3	.46	-1.01
Condition 4	.59	.36
Overall	.47	.23

The LPL model explains roughly half the variance in participant responses across all conditions, whereas the Bayesian model fit varies widely. Moreover, the Bayesian model does much worse than the LPL model in Condition 1

[6] The probability of a temporal sequence is complicated for cyclic graphs, as one must consider multiple ways to generate a temporal sequence. Technical details are available upon request.

[7] If $R^2 < 0$ then the mean predicts more variance than the model.

(i.e., with no temporal information), suggesting that the modification of the Bayesian model to allow for temporal delays does not explain the poor fit.

At the same time, both models provide good qualitative fits to the data: the model-participant correlations are $r = .74$ for the LPL model and $r = .97$ for the Bayesian model. However, only the LPL model predicts the appropriate variability in the participants' responses. For instance, the data are sufficient in Condition 1 for a Bayesian learner to determine the true causal structure (except for $D \rightarrow B$ and $C \rightarrow B$, about which it is indifferent), and so even probability matchers should exhibit relatively little variation. However, many experimental participants select causal relations that are not part of the true structure, and some omit relations that are. Participants do not seem to be fully rational learners, and the LPL model is able to explain the types of errors that occur.

Conclusion

The LPL model aims to provide a formal algorithmic model of the mechanisms underlying covariation-based causal structure learning. It provides a computationally well-specified dynamical model that learns directed graphs, and so potentially captures the cognitive mechanisms underlying causal learning. Moreover, this model predicts some of the *sub*-optimal learning behaviour exhibited by participants. Open questions remain about, for example, the suitability of the t-test-based decision procedure. But the LPL model provides a model that bridges the gap between associationist and rational models of causal learning.

Acknowledgments

This research was partially supported by a Scholar Award from the James S. McDonnell Foundation.

References

Bonawitz, E., Denison, S., Chen, A., Gopnik, G., & Griffiths, T. L. (2011). A simple sequential algorithm for approximating Bayesian inference. *Proceedings of the Thirty-third Cognitive Science Society*.

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

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

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

Corlett, P. R., Aitken, M. R. F., Dickinson, A., Shanks, D. R., Honey, G. D., Honey, R. A. E. (2004). Prediction error during retrospective revaluation of causal associations in humans: fMRI evidence in favor of an associative model of learning. *Neuron*, *44*(5), 877-888.

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

Fernbach, P. M., & Sloman, S. A. (2009). Causal learning with local computations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 678.

Glymour, C. (2003). Learning, prediction and causal Bayes nets. *Trends in Cognitive Sciences*, *7*(1), 43-48.

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), 1-31.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334-384.

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.

Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., & Sloman, S. A. (2007). Beyond covariation. In A. Gopnik, & L. Schultz (Eds.), *Causal learning: Psychology, philosophy, and computation,* 154-172.

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

Sloman, S., & Lagnado, D. A. (2005). Do we 'do'. *Cognitive Science*, *29*, 5-39.

Spirtes, P., Glymour, C. N., & Scheines, R. (1993). *Causation, prediction, and search*. Cambridge, MA: The MIT Press.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E. J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*(3), 453-489.

Turner, D. C., Aitken, M. R. F., Shanks, D. R., Sahakian, B. J., Robbins, T. W., & Schwarzbauer, C. (2004). The role of the lateral frontal cortex in causal associative learning: Exploring preventative and super-learning. *Cerebral Cortex*, *14*(8), 872-880.

Waldmann, M. R. (1996). Knowledge-based causal induction. *Psychology of Learning and Motivation*, *34*, 47-88.

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, *The Probabilistic Mind: Prospects for Bayesian Cognitive Science*. Oxford: 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.