Sequential Causal Learning in Humans and Rats

Hongjing Lu (hongjing@ucla.edu)
Department of Psychology, UCLA

Randall R. Rojas (rrojas@stat.ucla.edu)
Department of Statistics, UCLA; Raytheon, Space and Airborne Systems

Tom Beckers (tom.beckers@psy.kuleuven.be)
Department of Psychology, K.U.Leuven

Alan Yuille (yuille@stat.ucla.edu)
Departments of Statistics, Psychology, and Computer Science, UCLA

Abstract

Recent experiments (Beckers, De Houwer, Pineño, & Miller, 2005;Beckers, Miller, De Houwer, & Urushihara, 2006) have shown that pretraining with unrelated cues can dramatically influence the performance of humans in a causal learning paradigm and rats in a standard Pavlovian conditioning paradigm. Such pretraining can make classic phenomena (e.g. forward and backward blocking) disappear entirely. We explain these phenomena by a new Bayesian theory of sequential causal learning. Our theory assumes that humans and rats have available two alternative generative models for causal learning with continuous outcome variables. Using model-selection methods, the theory predicts how the form of the pretraining determines which model is selected. Detailed computer simulations are in good agreement with experimental findings.

Keywords: Bayesian inference; model selection; sequential causal learning; animal conditioning

Introduction

For more than two decades, researchers in both animal conditioning and human causal learning have identified significant parallels between the phenomena observed in the two fields (see Shanks, 2004). It has even been suggested that rats in conditioning paradigms learn to relate cues to outcomes in a manner similar to the way a scientist learns cause-effect relations (Rescorla, 1988). At the same time, there have been strong disagreements about the theoretical basis for both human causal learning and animal conditioning. On the one hand, conditioning models (Rescorla & Wagner, 1972) have been applied to human causal learning (Shanks, 1985); on the other, models of human causal learning have been applied to animal conditioning (Blaisdell, Sawa, Leising, & Waldmann, 2006;Cheng, 1997).

A phenomenon that has received particular attention in both the human and animal literatures is the *blocking* effect (Kamin, 1969). Suppose that two cues, A and X, are repeatedly and consistently paired with a particular outcome O. X will be viewed as a weaker cause of O if A alone is repeatedly paired with O either before (forward blocking) or after (backward blocking) pairings of the AX compound with O. Some evidence has suggested that blocking is less pronounced in humans than in rats (De Houwer, Beckers, & Glautier, 2002). However, recent experiments by Beckers et al. (2005, 2006) indicate that apparent differences between humans and rats in the conditions that promote blocking may reflect different assumptions about the cue-reward relationship, rather than any basic difference in causal learning processes between species. For both species, Beckers et al. showed that different pretraining conditions using unrelated cues could alter the learner's assumptions and thereby prevent or promote the occurrence of classic phenomena such as forward and backward blocking (leading rats to behave more like humans, and vice versa).

The goal of this paper is to provide a computational explanation for these experimental findings based on Bayesian inference. Our theory proposes that experimental subjects, whether rats or humans, have available multiple models of cue integration appropriate for different situations (Waldmann, 2007; Lucas & Griffiths, 2007). From our computational perspective, pretraining influences the probability that causal learners will select a particular integration model during a subsequent learning session with different cues, and this choice in turn determines the magnitude of blocking effects.

Most previous statistical theories of human causal learning have focused on learning from summarized contingency data based on binary variables (Cheng, 1997;Griffiths & Tenenbaum, 2005). The computational theory described here instead provides a trial-by-trial model of learning from sequential data. For nonverbal animals, there is no obvious way to present summarized data; often, humans also must learn from sequential data. In particular, sequential models are required to account for influences of the order of data presentation (Danks, Griffiths, & Tenenbaum, 2003;Dayan & Kakade, 2000;Shanks, 1985). A computational theory should enable beliefs to be dynamically updated by integrating prior beliefs with new observations in a trial-by-trial manner. In addition, in conditioning experiments the outcomes (e.g., food reward) are generally continuous in nature (i.e., the magnitude of the reward may vary). A computational theory must therefore address continuous-valued as well as binary variables in order to integrate causal learning by humans with learning by other animals.

Bayesian Theory of Sequential Learning

Within our theory of causal learning, each *causal model* corresponds to a different probabilistic model for generating the data. For continuous-valued outcomes we use a *linear-sum* model (Dayan & Kakade, 2000), which has been used previously to explain many aspects of the blocking effect, and a *noisy-MAX* model, proposed here for the first time. The latter is a generalization of the noisy-OR model, which gives a good account of human causal learning about binary variables based on summarized contingency data (Cheng, 1997;Griffiths & Tenenbaum, 2005; Lu et al, 2007). The choice of model depends on the type of pretraining, and is determined by standard Bayesian model selection. These expectations based on pretraining carry over to influence the learner's judgments in the subsequent causal learning task, even though the specific cues differ from those used in the pretraining.

We first introduce likelihood functions for the two different causal models assumed by our theory. We then describe the priors, the resulting full models, and model selection. Finally, we report simulations of experimental data and discuss how the present theory relates to others.

Causal Generative Models as Likelihood Functions

We focus on the relationship between two binary-valued causes x_1, x_2 (i.e. $x_i = 1$ if cause i is present, and $x_i = 0$ otherwise) and a continuous-valued outcome variable O. We define two continuous-valued hidden variables R_1, R_2. The hidden variables correspond to internal states that reflect the magnitudes of the effect generated by each individual cause. Each such magnitude corresponds to the *weight* of the corresponding cause, ω_1, ω_2, analogous to causal strength (Cheng, 1997). The generative model of the data, as shown in Figure 1, is given by

$$P(O \mid \omega_1, \omega_2, x_1, x_2) = \int dR_1 \int dR_2 P(O \mid R_1, R_2) \prod_{i=1}^{2} P(R_i \mid \omega_i, x_i). \quad (1)$$

The first generative model is called the *linear-sum* model because the output O can be expressed as the sum of R_1 and R_2 plus Gaussian noise with mean 0 and variance σ_m^2,

$$P(O_i \mid R_1, R_2) \propto \exp\left\{-(O - R_1 - R_2)^2 / \left(2\sigma_m^2\right)\right\}. \quad (2)$$

The second generative model, termed the *noisy-MAX* model, is motivated by the successful noisy-OR model for causal reasoning with binary variables by humans (Cheng, 1997). To adapt the noisy-OR model for continuous outcome variables, we express it as a noisy-MAX,

$$P(O_i \mid R_1, R_2) \propto \exp\left\{-(O - F(R_1, R_2; T))^2 / \left(2\sigma_m^2\right)\right\}. \quad (3)$$

where the function $F(R_1, R_2; T)$ is a noisy-MAX function of R_1, R_2 specified by:

$$F(R_1, R_2; T) = R_1 \frac{e^{R_1/T}}{e^{R_1/T} + e^{R_2/T}} + R_2 \frac{e^{R_2/T}}{e^{R_1/T} + e^{R_2/T}} \quad (4)$$

The parameter *T* determines the sharpness of the noisy-MAX function. As $T \rightarrow 0$, the noisy-MAX function becomes identical to the MAX function, i.e., equal to the maximum value of R_1 and R_2. By contrast, as $T \rightarrow \infty$ the noisy-MAX function approaches the average $(R_1+R_2)/2$.

For both models, the hidden effects of the individual causes are assumed to follow a Gaussian distribution,

$$P(R_i \mid \omega_i, x_i) \propto \exp\left\{-(R_i - \omega_i x_i)^2 / \left(2\sigma_h^2\right)\right\}, \; i = 1, 2. \quad (5)$$

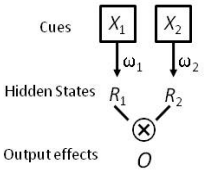

Figure 1. An illustration of the generative models. The different models combine R_1 and R_2 in different ways, a linear-sum or a noisy-MAX, to yield the output effect O.

Causal Priors

To perform Bayesian estimation we must specify prior distributions on the weights $P(\omega_1), P(\omega_2)$, which we define as Gaussians with 0 mean and small variance σ_p^2. This prior distribution expresses the default assumption that the weight of both causes is close to zero before observing any data.

For sequential presentation in a trial-by-trial dynamic manner, we also assume a *temporal prior* for the change of ω_1, ω_2 over time (i.e.,trials), as in Dayan and Kakade (2000).

$$P(\omega_i^{t+1} \mid \omega_i^t) \propto \exp\left\{-(\omega_i^{t+1} - \omega_i^t)^2 / 2\sigma_T^2\right\}, \; i = 1, 2 \quad (6)$$

These temporal priors imply that weights may be slowly varying from trial to trial. The amount of variation is controlled by the parameter σ_T^2. As $\sigma_T^2 \rightarrow 0$ the weight becomes fixed over trials, thus effectively switching off the temporal prior. For larger σ_T^2 the weights can change significantly over trials.

Combining the Likelihood and Priors

We use the standard technique for combining likelihoods with temporal priors for sequential data (Ho & Lee, 1964). The linear-sum model can be obtained from this formulation as the special case in which the likelihood, prior, and temporal priors are Gaussian.

To simplify the notation, we write $\vec{x} = (x_1, x_2)$, $\vec{\omega} = (\omega_1, \omega_2)$. We write $\{O_t\}$ and $\{\vec{x}_t\}$ to denote the set of rewards and causes on all trials up to and including trial t, i.e. $\{O_t\} = (O_t, O_{t-1}, \ldots, O_1)$.

The Bayesian formulation for updating the estimates of the weights is given in two stages:

$$P(\vec{\omega}^{t+1} \mid \{O_t\}, \{\vec{x}_t\}) = \int d\vec{\omega}_t P(\vec{\omega}^{t+1} \mid \vec{\omega}^t) P(\vec{\omega}^t \mid \{O_t\}, \{\vec{x}_t\}), \quad (7)$$

$$P(\vec{\omega}^{t+1} \mid \{O_{t+1}\}, \{\vec{x}_{t+1}\}) = \frac{P(O_{t+1} \mid \vec{\omega}^{t+1}, \vec{x}^{t+1}) P(\vec{\omega}^{t+1} \mid \{O_t\}, \{\vec{x}_t\})}{P(O_{t+1} \mid \{O_t\}, \{\vec{x}_{t+1}\})}. \quad (8)$$

Here we set $P(\vec{\omega}^{t+1} \mid \vec{\omega}^t) = P(\omega_1^{t+1} \mid \omega_1^t) P(\omega_2^{t+1} \mid \omega_2^t)$ assuming independence in the temporal prior.

The process is initialized by setting $P(\vec{\omega}^0)$ to equal the prior (i.e., product of Gaussians with 0 means and variances σ_p^2). We use Eq. 7 to predict a distribution on the weights $\vec{\omega}^1$ at time *t*=1 (with the convention that $\{O_0\}$ and $\{\vec{x}_0\}$ are empty

sets). Then we employ Eq. 8 to make use of the observed data on trial 1, O_1, x_1, to update the estimate of the weights, $\vec{\omega}^1$.

Eqs. 7-8 correspond to prediction and correction for each trial as a recursive estimator. That is, only the estimated weight distribution from the previous trial t and the current cue-outcome measurement, x_{t+1}, O_{t+1}, are needed to compute the weight estimate for the current trial, ω_{t+1}. Thus the model does not need to memorize cue-outcome pairs across all trials. If all the probabilities are Gaussian, then updating the probability distributions using Eqs. 7-8 simply corresponds to updating the means and covariance matrices using the standard Kalman filter equations (Dayan & Kakade, 2000). In the case of the noisy-MAX model, Eqs. 7-8 are applied directly in the distribution updating.

Parameter Estimation and Model Selection

There are two types of inference that we can make from the posterior distributions $P(\vec{\omega}^t \mid \{O_t\}, \{\vec{x}_t\})$. First, we can perform *parameter estimation* to estimate the weights $\vec{\omega}_t$, i.e., the weights of causes after *t* trials. Second, we can evaluate how well each model fits the data and perform *model selection* (i.e., choose between the linear-sum and noisy-MAX models). As discussed by Lu et al. (2007), different experimental paradigms can be modeled as parameter estimation or model selection.

Parameter estimation involves estimating the weight parameters $\vec{\omega}_t$. In our simulations, these estimates are the means of weights with respect to the distribution:

$$\vec{\omega}^t = \int d\vec{\omega}^t P(\vec{\omega}^t \mid \{O_t\}, \{\vec{x}_t\}) \vec{\omega}^t. \qquad (9)$$

Model selection involves determining which model is more likely to account for the observed sequence of data $\{O_t\}$ and $\{\vec{x}_t\}$. For each model (linear-sum or noisy-MAX), we compute:

$$P(\{O_\tau\} \mid \{\vec{x}_\tau\}) = \prod_{t=0}^{\tau-1} P(O_{t+1} \mid \{O_t\}, \{\vec{x}_{t+1}\}), \qquad (10)$$

with the convention that

$$P(O_1 \mid \{O_0\}, \{\vec{x}_1\}) = \int d\vec{\omega} P(O_1 \mid \vec{\omega}, \vec{x}_1) P(\vec{\omega}). \qquad (11)$$

Simulation of Blocking Experiments

We first report our simulations of traditional forward/backward blocking paradigms (Shanks, 1985) using linear-sum and noisy-MAX models. These two blocking effects provide a critical test for any sequential learning model. We then apply our Bayesian approach, e.g. using model selection, to a human experiment that employed pretraining (Beckers et al., 2005), and a similar conditioning experiment using rats (Beckers et al., 2006). The simulations ~~will~~ illustrate how our approach accounts for human and rat performance based on model selection and parameter estimation for sequential data.

Forward/Backward Blocking

Conditioning paradigms provide a window to the investigation of natural inferences produced by causal learning. Two common paradigms, schematized in Table 1, are *forward blocking* (A+, AX+) and *backward blocking* (AX+, A+). In both, the common finding is acquisition of a weaker weight between X and reward O than that between A and reward O (Kamin, 1969; Shanks, 1985). Note that backward blocking (typically weaker than forward blocking) implies that the weight of the absent cue X is updated as a result of a series of A+ trials. Any successful sequential learning model must explain the difference of weights associated with different cues in both blocking paradigms.

Blocking paradigm	Training phase 1	Training phase 2	test
forward	8A+	8AX+	A, X
backward	8AX+	8A+	A, X

Table 1: Design summary for a typical blocking experiment. The numerical values indicate the number of trials, + indicates the presence of the outcome effect.

Figure 2 shows simulations of learning of weight for cue A (ω_A, solid) and cue X (ω_X, dashed) as a function of trial number in forward blocking (black) and backward blocking (gray) designs. Figure 2A, B shows predictions based on the linear-sum and the noisy-MAX model, respectively. Both models predict the basic phenomena, as the weight associated with cue X is weaker than the weight for A in both forward and backward blocking paradigms, and more so in the former. However, the linear-sum model predicts a larger weight difference than does the noisy-MAX model in both paradigms. Furthermore, for the weight associated with cue X, the linear-sum model predicts a weaker weight in forward blocking (dashed black) than in backward blocking (dashed solid), which is an asymmetry between forward/backward blocking. The noisy-MAX model also predicts an asymmetry, although it diminishes as the number of trials increases. A novel prediction from the noisy-MAX in forward blocking is that the weight associated with cue A is expected to decrease to 0.5 after a large number of AX+ trials.

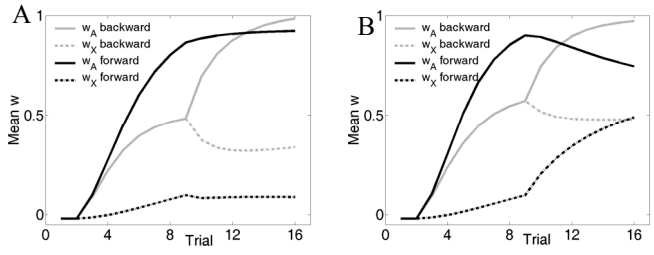

Figure 2. Predicted mean weights of each cue as a function of training trials in two different blocking paradigms. (A) linear-sum model; (B) noisy-MAX model. The black lines indicate predictions for forward blocking paradigm (A+, AX+); the gray lines indicate predictions for backward blocking paradigm (AX+, A+). The solid lines are estimates of weights for cue A; the dashed lines are estimates of weights for cue X. The linear-sum model predicts a larger difference between ω_A and ω_X across the two blocking paradigms than does the noisy-MAX.

Impact of Pretraining on Human Judgments

We simulated results of a pretraining study with humans by Beckers et al (2005). Table 2 schematizes the experimental design. G and H indicate different food cues: + and ++ indicate a moderate or a strong allergic reaction, respectively. As shown in Table 2, *additive* pretraining involved G+ trials followed by H+, and then followed by GH++. *Sub-additive* pretraining involved G+ trials followed by H+ trials, and then followed by GH+ trials.

The experiment included three phases: (a) pretraining, (b) elemental training, and (c) compound training. The elemental and compound training were always the same but the pretraining could be either additive or sub-additive for the two groups. In both groups, standard forward blocking trials with different food cues (A+ followed by AX+) were presented in phase 2 and 3. Note that the design used completely different cues in the pretraining phase 1 (cues G, H) and phases 2 and 3 (cues A, X, K, and L). If blocking occurs, we would expect the weight of cue X to be reduced by its pairing with cue A, due to the earlier elemental training on A in phase 2. K and L served as control cues, which were only presented in phase 3 as KL+ trials.

After completing these three phases, participants were asked to rate how likely each food cue separately would cause an allergic reaction. As indicated by the human results shown in Figure 4A, cue X was blocked after additive pretraining but not after sub-additive pretraining. More precisely, additive pretraining resulted in a lower rating for cue X than for the control cues, K and L, both of which in turn received significantly lower causal ratings than cue A. In contrast, after sub-additive pretraining there was little difference among the ratings for X, K, and L.

Group	Phase 1: Pretraining	Phase 2: Elemental Training	Phase 3: Compound Training
Additive	8G+/8H+/8GH++ /8I+/8Z-	8A+ /8Z-	8AX+/8KL+ /8Z-
Subadditive	8G+/8H+/8GH+ /8I++/8Z-	8A+ /8Z-	8AX+/8KL+ /8Z-

Table 2: Design summary for human pretraining experiment in Beckers et al. (Exp. 2, 2005).

The experimental design used by Beckers et al. (2005) can be translated into the notation of our model as follows. G+, H+, GH+ respectively correspond to $(x_1, x_2) = (1,0), (0,1),$ and $(1,1)$. The notation + and ++ correspond to $O = 1$ and $O = 2$, respectively. Using the pretraining trials in phase 1, we performed model selection to infer which model is more likely for the additive and sub-additive groups. With the models selected in the pretraining phase, we then used trials in phases 2 and 3 to estimate the distribution of the weights ω for each cue. The mean of each ω was computed to provide a comparison with human ratings.

We employed trials in the pretraining phase to compute the log-likelihood ratios for the noisy-MAX model relative to the linear-sum model using Eq. 10. The resulting plots are shown in Figure 3. In the simulation we used model parameters $\sigma_h = 0.6, \sigma_T = 0.3, \sigma_m = 0.01,\ T = 0.4$. To perform model selection, we need to impose a threshold on the log-likelihood ratios. We set the threshold to be the log-likelihood ratio obtained when only the data G+, H+ had been shown (as the experimental subject would have no basis for a preference between the two models at this stage). The simulation results (see Figure 3) show that the linear-sum model is selected if the pretraining is additive (i.e., G+, H+, GH++), because the corresponding ratio is below the threshold, whereas the noisy-MAX model is selected if the pretraining is sub-additive (i.e., G+, H+, GH+), because the corresponding ratio is above the threshold

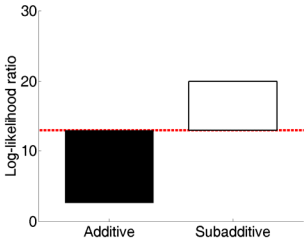

Figure 3. Log-likelihood ratios for the noisy-MAX model relative to the linear-sum model for the additive group (black) and the sub-additive group (white) in human experiment by Beckers et al. (2005). The dashed line indicates the threshold for model selection.

We then computed the mean weights, using Eq. 9, for the models chosen by the model selection stage. These mean weights (see Figure 4B) constitute our simulation's predictions for the causal ratings. The simulation results are in good agreement with the results for humans (Figure 4A). The linear-sum model generates accurate predictions for the additive group: the mean weight for X is much lower than weights for the control cues K and L, indicating blocking of causal learning for cue X. In contrast, the noisy-MAX model gives accurate predictions for the sub-additive group: the mean weight for X is about the same as the weights for the control cues K and L, consistent with absence of blocking for X.

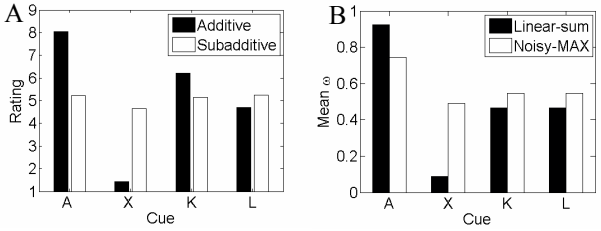

Figure 4. Mean causal rating for each cue. (A) Human ratings in Experiment 2 by Beckers et al (2005); see their Figure 3, p. 243. Black bars indicate the mean rating for additive pretraining group; white bars for sub-additive pretraining group. (B) Predicted ratings based on the selected model for each group. Black bars indicate the mean ω based on the linear-sum model, which gives a good fit for the human means in the additive group. White bars indicate the mean ω based on the noisy-MAX model, which give a good fit for the human means in the sub-additive group.

Impact of Pretraining on Rat Conditioning

Now we compare the predictions of the models to the experimental findings for a conditioning experiment with rats (Beckers et al., 2006). Animals were presented with cues that were associated with shocks while the animals pressed a lever

for water. We focus on two conditions: sub-additive and irrelevant element, as schematized in Table 3 (Beckers et al., 2006, Experiment 1). Animals in the experimental group received forward blocking training (A+ followed by AX+); control animals did not receive blocking training (B+ followed by AX+). Before the actual blocking training (phase 2 and phase 3), experimental and control animals in the sub-additive condition were exposed to a demonstration of two effective cues, C and D, that had sub-additive outcomes (i.e., C+, D+, CD+), or to an irrelevant pretraining (i.e. C+, D+, E+). The number of lever-press responses to X after phase 3 was measured for all animals.

Condition and group	Phase 1: Pretraining	Phase 2: Elemental Training	Phase 3: Compound Training
Subadditive			
Experimental	4C+/4D+/4CD+	12A+	4AX+
Control	4C+/4D+/4CD+	12B+	4AX+
Irrelevant element			
Experimental	4C+/4D+/4E+	12A+	4AX+
Control	4C+/4D+/4E+	12B+	4AX+

Table 3: Design summary for the rat pretraining experiment by Beckers et al. (Exp. 1, 2006).

We used the same translation to the model notation as before. We set the threshold such that without any training, the linear-sum model would be preferred over the noisy-MAX model, as evidence suggests that rats typically assume linear integration (Beckers et al., 2006, p. 98; see also Wheeler, Beckers, & Miller, 2008). We computed the log-likelihood ratios for the pre-testing data, using Eq. 10, to confirm that the noisy-MAX model was selected for the sub-additive condition and the linear-sum model for the irrelevant condition. The results are shown in Figure 5. We used model parameters $\sigma_h = 0.6, \sigma_T = 0.6,$ $\sigma_m = 0.01, T = 0.3$ in the simulations. Compared to the parameter set used for the human experiments, we increased the variance for the temporal prior to speed up causal learning of cues (perhaps reflecting the high salience of electric shock as an outcome).

Beckers et al. (2006) used the *suppression ratio* of cue X as a measure of rats' causal judgment about cue X. A value of 0 for the suppression ratio corresponds to complete suppression of bar pressing (i.e., high fear of cue X), and a value of 0.5 corresponds to a complete lack of suppression (i.e., no fear of X). Figure 6A shows the mean suppression ratios for experimental and control animals in Experiment 1 of Beckers et al. (2006).

We model the suppression ratio as a function of the predicted mean weight of cue X, $\bar{\omega}_X$ with Eq. 9. Assuming that the mean number of lever presses in the absence of cue X is N, the expected number of lever presses in the presence of cue X will be $N - N\omega_X$. Accordingly, the predicted suppression ratio can be computed as:

$$\text{suppression ratio} = \frac{N - N\bar{\omega}_X}{N - N\bar{\omega}_X + N} = \frac{1 - \bar{\omega}_X}{2 - \bar{\omega}_X} \quad (12)$$

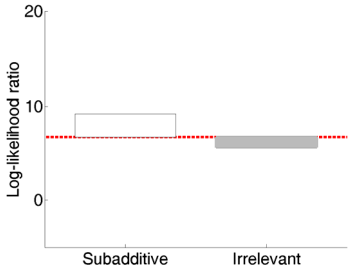

Figure 5. Log-likelihood ratios of noisy-MAX model relative to linear-sum model for the subadditive condition (white), and the irrelevant condition (gray) in the rat experiment (Beckers, et al., 2006, Exp. 1).

Figure 6B shows the predictions of selected models for the two conditions tested by Beckers et al. (2006). Similar to the results obtained when modeling the human data, the noisy-MAX model was selected for the sub-additive condition, and the linear-sum model for the irrelevant condition. Accordingly, the suppression ratio was estimated using the noisy-MAX model for the subadditive condition. The suppression ratio in the irrelevant condition was computed by the linear-sum, because the default model was assumed to favor the linear-sum given that irrelevant pretraining data did not provide clearly discriminative information for model selection. As shown in Figure 6B, there was no significant difference in the suppression ratio for the noisy-MAX model, in agreement with rat data showing no significant difference between the experimental and control groups with sub-additive pretraining. In contrast, suppression ratios differed between experimental and control groups using the linear-sum model in agreement with the rat data showing a significant difference between the experimental and control groups with irrelevant element pretraining.

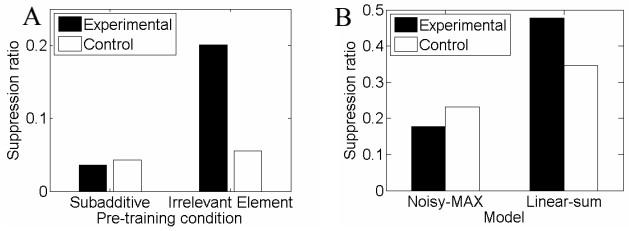

Figure 6. Mean suppression ratio for cue X in experimental and control groups by pretraining conditions in the subadditive condition and irrelevant condition (Beckers, et al, 2006, Exp. 1). Black/white bars indicate the experimental/control group, respectively. (A) Rat results; (B) Suppression ratio predicted by the noisy-MAX model (matched to sub-additive experimental condition), and predicted by the linear-sum model (matched to irrelevant element condition).

General Discussion

The Bayesian theory of sequential causal learning described in the present paper provides a unified explanation for important learning phenomena observed with both humans and rats. In particular, the theory accounts for influences of pretraining on subsequent learning with completely different stimuli (Beckers, et al., 2005, 2006). The key assumption is that learners have available multiple generative models, each reflecting a different integration rule for combining the influence of multiple causes (cf. Lucas & Griffiths, 2007; Waldmann, 2007). When the outcome is a continuous variable, both humans and rats have tacit knowledge that multiple causes may have a summative

impact on the outcome (linear-sum model). Alternatively, the outcome may be effectively "saturated" at a level approximated by the weight of the strongest individual cause (noisy-MAX). Using standard Bayesian model selection, the learner selects the model that best explains the pretraining data, and then employ the favored model in estimating causal weights with different cues during subsequent learning. Note that the information provided in Phases 2-3 is identical for both groups; hence only Phase 1 (pretraining) is relevant to model selection.

A key component of the sequential learning theory is the temporal prior, which controls dynamic updating of the estimated weight of each cue in a trial-by-trial manner. The temporal prior allows the theory to explain both forward and backward blocking effects, and more generally captures the influence of trial order on causal learning. Trial-order effects are outside the scope of models that only deal with summarized data (e.g., Cheng, 1997; Griffiths & Tenenbaum, 2005; Lu et al., 2007).

The present theory is also more powerful than previous accounts of sequential causal learning. The Rescorla-Wagner model (Rescorla & Wagner, 1972) and its many variants (see Shanks, 2004) only update point estimates of causal strength, and thus are unable to represent degrees of uncertainty about causal strength (Cheng & Holyoak, 1995). By adopting a Bayesian approach to learning probability distributions, the present theory provides a formal account of how a learner's confidence in the causal strength of a cue will be expected to change over the course of learning. The same limitation (updating point estimates of strength, rather than probability distributions) holds for a previous simulation of sequential learning based on the noisy-OR generative model (Danks, Griffiths & Tenenbaum, 2003). Most importantly, the present theory goes beyond all previous accounts of dynamical causal learning (e.g., Dayan & Kakade, 2000) in its core assumption that learners, both human and non-human, are able to flexibly select among multiple generative models that might "explain" observed data. The theory thus captures what appears to be a general adaptive mechanism by which biological systems learn about the causal structure of the world.

Acknowledgments

Preparation of this paper was supported by a Postdoctoral Fellowship and a Travel Grant from the Research Foundation - Flanders (FWO Vlaanderen) to TB, and a grant from the W. H. Keck Foundation to AY.

References

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

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

Blaisdell, A. P., Sawa, K., Leising, K. J., & Waldmann, M. R. (2006). Causal reasoning in rats. *Science, 311*(5763), 1020-1022.

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

Cheng, P. W., & Holyoak, K. J. (1995). Complex adaptive systems as intuitive statisticians: Causality, contingency, and prediction. In H. L. Roitblat & J.-A. Meyer (Eds.), *Comparative approaches to cognitive sciences* (pp. 271-302). Cambridge, MA: MIT Press.

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* (Vol. 15, pp. 67-74). Cambridge, MA: MIT Press.

Dayan, P., & Kakade, S. (2000). Explaining away in weight space. In T. K. Leen et al., (Eds.), *Advances in neural information processing systems* (Vol. 13, pp. 451-457). Cambridge, MA: MIT Press.

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

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

Ho, Y.-C., & Lee, R. C. K. (1964). A Bayesian approach to problems in stochastic estimation and control. *IEEE Transactions on Automatic Control, 9*, 333-339.

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

Lu, H., Yuille, A., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2007). Bayesian models of judgments of causal strength: A comparison. In D. S. McNamara & G. Trafton (Eds.), *Proceedings of the Twenty-ninth Annual Conference of the Cognitive Science Society* (pp. 1241-1246). Austin, TX: Cognitive Science Society.

Lucas C. G., & Griffiths, T. L. (2007). Learning the functional form of causal relationships. Poster presented in the Twenty-ninth Annual Conference of the Cognitive Science Society.

Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American Psychologist*, *43*, 151-160.

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

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

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, CA: Academic Press.

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

Waldmann, M. R. (2007). Combining versus analyzing multiple causes: How domain assumptions and task context affect integration rules. *Cognitive Science, 31*, 233-256.

Wheeler, D. S., Beckers, T., & Miller, R. R. (2008). The effect of subadditive pretraining on blocking: Limits on generalization. Manuscript accepted for publication, *Learning & Behavior*.