University of California

Los Angeles

# Explaining Human Causal Learning using a Dynamic Probabilistic Model

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Statistics

by

**Randall Rojas Rojas**

2010

The dissertation of Randall Rojas Rojas is approved.

Ying Nian Wu

Keith J. Holyoak

Hongjing Lu

Allan L. Yuille, Committee Chair

University of California, Los Angeles

2010

*To John Miller and Vanessa Miller*

# Table of Contents

# List of Figures

# List of Tables

# Acknowledgments

I would like to first thank my advisor, Alan L. Yuille for his exceptional guidance, insightful discussions and support. His direction and valuable help were a central part of my training in Statistics and an inspiration to my future career path.

Hongjing Lu deserves a very special recognition for her endless assistance since the beginning of this endeavor. Her active involvement in every aspect of this dissertation, generous availability and admirable patience were crucial to my learning and highly influential in my decision to pursue this research work.

I would also like to express my gratitude to my committee members (current and former), Ying Nian Wu, Keith J. Holyoak, Bernard Balleine and Mark H. Hansen who have all been very instrumental in helping shape my dissertation work and improve its scientific quality.

Ivaylo Dinov also deserves recognition for encouraging me to pursue this Ph.D. and expressing his support of my application to the Statistics Department. Without his encouragement, I would not have been able to take this journey.

I am very grateful for all the help that Glenda Jones has provided during my graduate education in the Statistics Department. Without her help I would have been lost with all the administrative paperwork and questions which she patiently and regularly helped me with.

I would like to also acknowledge my employer, Raytheon, Space and Airborne Systems for their full financial support of this degree and incredible work schedule flexibility. It would have been impossible to accommodate my full-time student schedule with my full-time work responsibilities without their understanding and support.

I am indebted to Janelle Rodriguez for convincing me to pursue a degree in

Statistics instead of Mathematics, talking to Ivaylo Dinov on my behalf and her genuine support, friendship and confidence in me.

Maryna Taranova was a strong ally in this journey and helped me in every way she could. She saw to it that my efforts were always focused on my dissertation and future career in academia. Болшое спасибо! С Днем Рождения Обезьяна!

My father, Rodrigo Dondi Rojas, also played an important role in this endeavor by providing valuable support, company and advice. His incredible sense of humor helped me keep my sanity during these intense years. Grazie Papà per l'idea della scatola di ufficio postale!

I would like to thank my mother Arelys J. Rojas for raising me with incredible wisdom, love and care. She has always been a guiding spirit in my life, always there to give me advice, support and encouragement to pursue my dreams. ¡Gracias Mami nuevamente por todo lo que haces y has hecho por mi!

Words cannot express enough my gratitude towards John Miller and Vanessa Miller to whom this manuscript is dedicated. Their sincere love, support and involvement in my life have been very influential and greatly appreciated. Thank you both for everything you have helped me with all these years! In addition, my nieces, Kaelyn and Eliza provided many wonderful and cheerful moments filled with laughter and joy (and drawings) that contributed to my happiness and well being during this time.

Lastly, I would like to thank Paloma Franco for giving a new meaning to my life. Her unconditional love, support and devotion have brought out the best in me. Her presence in my life (and Emily's) has giving me the motivation to see this manuscript come to fruition, illuminated my career goals and established a deeper purpose for my life. Je vous aime de tout mon coeur!

# Vita

| | |
|---|---|
| 1972 | Born, San Jose, Costa Rica |
| 1991 | Diploma. (Mathematics), Universidad de Costa Rica. |
| 1992 | Diploma. (Physics), Universidad de Costa Rica. |
| 1997 | B.S. (Mathematics), University of California, Los Angeles. |
| 1997 | B.S. (Astrophysics), University of California, Los Angeles. |
| 1998–1999 | Teaching Assistant, Department of Physics, Drexel University. |
| 1999–2004 | Research Assistant, Department of Physics, Drexel University. |
| 2000 | M.S. (Physics), Drexel University. |
| 2001–2002 | Adjunct Faculty, Department of Physics, Drexel University. |
| 2004 | Ph.D. (Physics -concentration in Astrophysics/Cosmology), Drexel University. |
| 2004–present | Sr. Multi-Disciplined Engineer, Raytheon, Space and Airborne Systems. |
| 2005–present | Research Assistant, Department of Statistics, University of California, Los Angeles. |
| 2006 | M.S. (Statistics), University of California, Los Angeles. |

# Publications

**Rojas, R. R.,** Beckers, T., and Yuille, A. 2008, Sequential Causal Learning in Humans and Rats, Proceedings of the Twenty-ninth Annual Conference of the Cognitive Science Society.

## Abstract of the Dissertation

# Explaining Human Causal Learning using a Dynamic Probabilistic Model

by

**Randall Rojas Rojas**

Doctor of Philosophy in Statistics

University of California, Los Angeles, 2010

Professor Allan L. Yuille, Chair

Recent psychological experiments (Beckers, De Houwer, Pineño, & Miller, 2005; Beckers, Miller, De Houwer, & Urushihara, 2006) have revealed that pre- and/or post- training with unrelated cues can significantly modulate the performance of humans in causal learning tasks and rats in the standard Pavlovian conditioning paradigm. This modulation can be large enough that classical conditioning phenomena such as forward and backward blocking can vanish, contrary to expectations from traditional psychological theories of associative learning. In this work we present a novel Bayesian theory of sequential causal learning that explains these experimental results. In addition, we extend our theory to provide an account for the highlighting effect (Daw, Courville, & Dayan, (2007); Kruschke, 2006, 2001) and then generalize our formalism to model the case of multiple cues and outcomes in the learning framework. Our Bayesian theory assumes that humans and rats have available several alternative generative models (*linear-sum*, MAX, *noisy*-MAX, etc.) for causal learning. By exploring the model space, we narrow the plausible models to two possibilities (*linear-sum*, *noisy*-MAX) where the cues and outcomes are both continuous variables. We implement the models using two approaches: (1) discretize the cue and outcome variables (making sure the discretization is dense enough) and (2) use the particle filter algorithm as an approximation to statistical inference. Our results show that model selection and model averaging are able to capture the effects of pre- and post- training respectively. We conjecture that the choice between model selection and model averaging is determined by when the information for making this choice is available. For the experiments with pretraining, the information is available before the learning trails (by the pretraining), therefore, humans/rats know which model to use. For posttraining, the information is only made available after the learning trials, which requires humans/rats to make retrospective evaluations. Lastly, our generalization to multiple cues and outcomes is tested within the Highlighting paradigm and we show that this more robust approach, provides an excellent account of experimental findings.

# CHAPTER 1

# Introduction

Cognitive science, in the last couple of decades, has experienced a remarkable flourishing of ideas and advances. In particular, the topic of *Causal Learning* has evolved into an impressive interdisciplinary enterprize that combines principles from Psychology, Philosophy, Computer Science and Statistics in order to provide greater insight about the mechanisms by which people learn causal relationships and ultimately, understand the overall causal architecture of our thinking process.

A traditional approach to the study of causal learning and inference (see e.g., Arnheim, 1969), consists of a psychological experiment where subjects (people in the case of human causal learning) are given the learning task to infer the strengths (and existence) of causal relationships between cause-effect links, where the potential causes and effects have been a priori, established by the experimenter. Plausible causes and effects are commonly represented by binary variables to indicate the presence (value=1) or absence(value=0) of the respective event type. A rating system is typically employed to measure a person's judgement of a cause-effect link[1].

An explanation for how people solve this learning problem has led to the development of a large number of psychological accounts of human causal learning. However, there are two dominating schools of thought: (a) associative learning

[1] Depending on the nature of the experiment, continuous variables may also be used.

theories[2] (e.g., $\Delta P$ model (Allan, 1980) and Rescorla-Wagner model (Rescorla-Wagner, 1972)) which are based on the notion that humans track contingency and (b) Patricia Cheng's causal power theory (Cheng, 1997) which instead postulates that people assume that the influence of a cause on its effect is hidden, and therefore, the problem of causal learning is to ascertain the strength of the respective hidden influence.

A popular perspective that has been actively explored (which is also the one adopted in this manuscript) is to consider how a computer would learn a causal structure. The standard causal model framework (Spirtes, Glymour, & Scheines, 1993; Pearl, 2000; Glymour, 2001), commonly referred to as *Causal Bayes Net* framework, is a graphical model composed of links (which represent the causal relations) and nodes (which represent the variables of the system such as the cues and outcomes) that characterize the architecture of the causal system. The graphical model can be illustrated by a *directed acyclic graph* (DAG) as shown in Figure 1.1. Casual Bayes nets have served as a solid foundation for many psychological theories of causal learning and have also been applied with great success to many other fields (Danks, 2005; Bessler, 2003; Ramsey *et al.*, 2002).

Sloman (2005) identifies three main parts of a causal model (for a graphical representation see Fig. 4.1 in Sloman, 2005):

(1) The causal system being represented.

(2) A probability distribution over the variables.

(3) A DAG that represents the causes that generate the probabilities.

In the case of Bayes nets, these three parts are held together by two powerful

[2] Behaviorists divide associative learning into two categories: operant conditioning and classical conditioning.

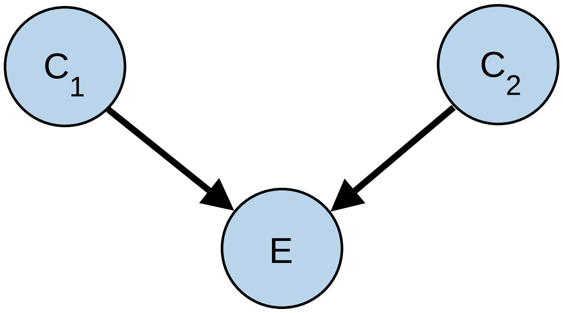

Figure 1.1: Directed Acyclic Graph (DAG). Illustration of a DAG consisting of two potential causes $C_1$ and $C_2$ depicted by the top nodes and respective effect $E$ indicated by the lower node. Edges represent conditional dependencies and arrows represent the causal relations (conditional probabilities).

assumptions: (a) the causal Markov assumption which states that every variable in a DAG is independent of all other variables in the graph conditional on its own direct causes and (b) the causal faithfulness assumption (also known as the stability assumption ) which asserts that the only conditional independencies are those that are consequences of the Markov assumption. Given the assumptions above and the ability of Bayes nets to model interventions (one of their key strengths, Danks, 2005), we can see that causal graphs are ideal for modeling complex systems of causal relations.

Studies have shown that intervention facilitates learning in many ways (Schulz & Gopnik, 2004). For example, interventions can help discriminate between competing causal structures that are otherwise indistinguishable based solely on observations. They also allow interveners to play an active role in systematic testing (Sloman, 2005). By allowing manipulations of the causal system variables, their effects can be more easily assessed. Therefore, a robust causal model framework has to have the flexibility to accommodate not only inferences based on observations but also, inferences about the underlying causal structure based on interventions that we may or may not engage in.

In the last two decades, many machine learning algorithms (Chickering, 2002; Spirtes *et al.*, 1993) have been developed to infer causal structure from patterns of correlations based on the Markov and faithfulness assumptions. There are two main categories into which these algorithms can be divided: (a) constraint-based which perform an exhaustive search for correlations and dependencies between variables to uncover the underlying causal structure and (b) Bayesian which takes on a holistic approach (Sloman, 2005; Danks, 2005) to exploring the space of causal models that could generate the data based on our prior beliefs.

Although there is no strong evidence to support one approach over another, for many situations, the Bayesian method does provide a better framework for performing causal learning. In favor of the Bayesian perspective, it has been suggested that human causal learning can be characterized as either a top-down search over the causal Bayes structures (Waldmann, 2000; Lagnado & Sloman, 2004) or as a rational analysis where a mixture of the two (constraint-based and Bayesian) methods is preferred (Gopnik *et al.*, 2001; Gopnik & Glymour, 2002; Tenenbaum & Griffiths, 2003).

Despite the many heroic attempts to understand how we think, the promise of a single theory capable of explaining our complex thinking process seems highly unlikely. However, the importance of causal learning models in every day life is eminent. Models of causal learning provide a window to our understanding of complex causal systems and have universal applicability that transcends disciplinary barriers.

The organization of the manuscript is as follows. In Chapter 2 we introduce the mathematical foundations (Bayesian Paradigm) of our simulations. Chapter 3 discusses details of the experiments simulated and our comparison of simulation results with experiments. An extension of our Bayesian sequential model to the modeling of the highlighting effect is introduced in Chapter 4 and a discussion of our conclusions and future work is given in Chapter 5.

# CHAPTER 2

# The Bayesian Paradigm

## 2.1 Introduction

Bayesian reasoning has its origins in the celebrated and influential work by Reverend Thomas Bayes. In particular, Proposition 9 in his famous essay, *An Essay Towards Solving a Problem in the Doctrine of Chances*, captures the main result of what nowadays is known as Bayes' Theorem (Bayes, 1763). Qualitatively, Bayes' theorem is a simple tool that allows us to update our prior belief in the light of new evidence (data) to yield a posterior belief. For a detailed discussion of Bayesian theory, see e.g., textbooks by Bernardo & Smith, (1998); Gelman *et al.*, (2003); O'Hagan, (1994); Press, (2003) and Robert, (2001).

An alternative to Bayesian statistics is the classical frequentist approach where only long-run frequencies of repeatable events have probabilities. From a frequentist perspective, probability statements about the respective parameters are not meaningful because the parameters are not random. This key difference between the Bayesian and frequentist view of parameters can be viewed as a dominant strength of the Bayesian paradigm. From a scientific perspective, a Bayesian approach has several advantages over traditional frequentist methods.

- One of the major benefits of the Bayesian approach is the ability to incorporate all available information. Bayesian statistics, by including the prior information ensures that no information is wasted, as opposed to frequentist statistics where prior information is disregarded in exchange for objectivity.

- A Bayesian analysis allows greater flexibility in the modeling of systematic errors (as well as hierarchical structures in the data).

- The Bayesian approach makes it possible to formulate direct statements about the parameters of interest, providing more intuitive and meaningful inferences and making it better suited for decision-making (O'Hagan, 2004; King *et al.*, 2000).

- Bayesian methods often yield better performance than frequentist methods and can answer complex questions cleanly and exactly. For example, the frequentist interpretation of confidence intervals is know to be a difficult task (except for the linear model). However, for a Bayesian analyst, the equivalent of classical confidence intervals, can be interpreted probabilistically in a straightforward way.

- The Bayesian framework relies on a single tool, Bayes theorem (Bolstad, 2004).

Despite its advantages, Bayesian theory did not gain much popularity until after its modern form was established by the French mathematician Pierre-Simon de Laplace (Laplace, 1812). Since then, there has been a plethora of applications of Bayesian reasoning (Chen, 2003; Press, 2003; Jaynes, 2003) including but not limited to topics in machine learning, pattern classification and recognition, decision theory, artificial intelligence, econometrics, epidemiology, non-parametric statistics, linear and non-linear regression, neuroscience, and cognitive science.

## 2.2 Bayesian Statistical Inference

In Bayesian inference, we treat the unknown parameters $\boldsymbol{\theta}$[1] (discrete-valued or continuous, time-varying or fixed) of the statistical model as random variables and the observations $\boldsymbol{y}$ as fixed, known quantities. We denote the respective prior distribution for the parameters $\boldsymbol{\theta}$ by the function $p(\boldsymbol{\theta})$. This function contains all the available information about the parameter values *prior* to observing the data. The prior distribution $p(\boldsymbol{\theta})$ is then updated with the new evidence $\boldsymbol{y}$ to yield the corresponding *posterior* distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$, which expresses our knowledge about the parameters after seeing the data. The update rule for $p(\boldsymbol{\theta}|\boldsymbol{y})$ is given by Bayes' Theorem (Bayes, 1763) according to:

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})}. \tag{2.1}$$

We can see that the posterior distribution is the conditional distribution of the parameters after observing the data. The conditional probability $p(\boldsymbol{y}|\boldsymbol{\theta})$ of the data $\boldsymbol{y}$ when $\boldsymbol{\theta}$ is assumed to be known, is commonly refereed to as the *likelihood* function. The normalization function $p(\boldsymbol{y})$ ensures that the posterior distribution will integrate (or sum in the case of discrete $\boldsymbol{\theta}$) to 1 and is obtained by integrating out $\boldsymbol{\theta}$ according to:

$$p(\mathbf{y}) = \int_{\boldsymbol{\theta}} p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{2.2}$$

From the posterior distribution, we can then perform model selection (for a detailed discussion see e.g., Zucchini, 2000; Linhart & Zucchini, 1986) or parameter estimation. For model selection, suppose we have to choose between two models (hypotheses) $M_1$ and $M_2$ based on a data set $\boldsymbol{D}$. We denote the marginal likelihood for model $i$ (where $i = 1, 2$) by $p(\boldsymbol{D}|M_i)$ and respective prior probability densities by $p(M_i)$. Note that in general, if $\boldsymbol{\theta}_i$ is the parameter vector

[1] Boldface symbols denote vectors or matrices.

of model $M_i$ and $p(\boldsymbol{\theta}_i|M_i)$ the prior distribution of $\boldsymbol{\theta}_i$, then the marginal likelihood of model $i$, $p(\boldsymbol{D}|M_i)$ is computed by integrating the likelihood (predictive probability) $p(\boldsymbol{D}|\theta_i, M_i)$ over the parameter space:

$$\begin{aligned} p(\mathbf{D}|M_i) &= \int p(\mathbf{D}, \boldsymbol{\theta}_i|M_i)d\boldsymbol{\theta}_i && (2.3) \\ &= \int p(\mathbf{D}|\boldsymbol{\theta}_i, M_i)p(\boldsymbol{\theta}_i|M_i)d\boldsymbol{\theta}_i. && (2.4) \end{aligned}$$

Next, we compute the corresponding Bayes Factor $B_{12}$ (Kass & Raferty, 1995; Jeffreys, 1935) and prior odds ($\Lambda_{12}$) for $M_1$ against $M_2$ defined respectively as:

$$\begin{aligned} B_{12} &= \frac{p(\mathbf{D}|M_1)}{p(\mathbf{D}|M_2)} && (2.5) \\ &= \frac{\int p(\boldsymbol{\theta}_1|M_1)p(\mathbf{D}|\boldsymbol{\theta}_1, M_1)d\boldsymbol{\theta}_1}{\int p(\boldsymbol{\theta}_2|M_2)p(\mathbf{D}|\boldsymbol{\theta}_2, M_2)d\boldsymbol{\theta}_2} && (2.6) \end{aligned}$$

and

$$\Lambda_{12} = \frac{p(M_1)}{p(M_2)}. \qquad (2.7)$$

Finally, model selection follows from computing the posterior odds for $M_1$ against $M_2$ given the data $\boldsymbol{D}$ according to:

$$\begin{aligned} \frac{p(M_1|\mathbf{D})}{p(M_2|\mathbf{D})} &= \frac{p(\mathbf{D}|M_1)}{p(\mathbf{D}|M_2)}\frac{p(M_1)}{p(M_2)} && (2.8) \\ &= B_{12}\Lambda_{12} && (2.9) \end{aligned}$$

where the Bayes factor $B_{12}$ provides a scale of evidence in favor of one model over another. Table (2.1) provides a list of various values for $B_{12}$ and their respective interpretation.

In the case of parameter estimation, given a statistic of interest, $f(\boldsymbol{\theta})$ and Eqs. 2.1 and 2.2 we can compute its expectation respective directly from

$$\mathbb{E}_{p(\boldsymbol{\theta}|\mathbf{y})}[f(\boldsymbol{\theta})] = \int_{\Theta} f(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}. \qquad (2.10)$$

| $B_{12}$ | Strength of Evidence |
|---|---|
| <1 | Supports $M_2$ |
| 1 to 3.2 | Barely worth mentioning |
| 3.2 to 10 | Substantial |
| 10 to 100 | Strong |
| >100 | Decisive |

Table 2.1: Table of Bayes factor values and their respective interpretation as given in Kass & Raferty, (1995).

## 2.3 Recursive Bayesian Estimation

Recursive Bayesian estimation deals with the problem of inferring knowledge about parameters (indirectly observable) recursively over time as new observations are collected using a mathematical process model. Both the sought parameters and observations are stochastic quantities where it is assumed that: (a) the true states $\mathbf{x}_t \in \mathbb{R}^n$ follow an unobserved Markov process and (b) the observations $\mathcal{Z}_t = \{\mathbf{z}_i\}_{i=1}^t$ are the observed states of a Hidden Markov Model (HMM) (Ho & Lee, 1964; Chen, 2003; Kramer & Sorenson, 1988; West, 1981; Jazwinski, 1970). From Bayes' Theorem (Eq. 2.1) it follows that

$$\begin{aligned}
p(\mathbf{x}_t|\mathcal{Z}_t) &= \frac{p(\mathcal{Z}_t|\mathbf{x}_t)p(\mathbf{x}_t)}{p(\mathcal{Z}_t)} \\
&= \frac{p(\mathbf{z}_t, \mathcal{Z}_{t-1}|\mathbf{x}_t)p(\mathbf{x}_t)}{p(\mathbf{z}_t, \mathcal{Z}_{t-1})} \\
&= \frac{p(\mathbf{z}_t|\mathcal{Z}_{t-1}, \mathbf{x}_t)p(\mathcal{Z}_{t-1}|\mathbf{x}_t)p(\mathbf{x}_t)}{p(\mathbf{z}_t|\mathcal{Z}_{t-1})p(\mathcal{Z}_{t-1})} \\
&= \frac{p(\mathbf{z}_t|\mathcal{Z}_{t-1}, \mathbf{x}_t)p(\mathbf{x}_t|\mathcal{Z}_{t-1})p(\mathcal{Z}_{t-1})p(\mathbf{x}_t)}{p(\mathbf{z}_t|\mathcal{Z}_{t-1})p(\mathcal{Z}_{t-1})p(\mathbf{x}_t)} \\
&= \frac{p(\mathbf{z}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathcal{Z}_{t-1})}{p(\mathbf{z}_t|\mathcal{Z}_{t-1})} \qquad (2.11)
\end{aligned}$$

and the respective prediction density $p(\mathbf{x}_{t+1}|\mathcal{Z}_t)$ is given by

$$p(\mathbf{x}_{t+1}|\mathcal{Z}_t) = \int_{\mathbb{R}^n} p(\mathbf{x}_{t+1}|\mathbf{x}_t)p(\mathbf{x}_t|\mathcal{Z}_t)d\mathbf{x}_t. \tag{2.12}$$

In general, Eqs. 2.11 and 2.12 can not be solved analytically and therefore, approximate methods must be employed such as Particle Filters (Gordon, 1993; Bergman, 1999; Doucet, 2001).

## 2.4 Monte Carlo Sampling

Monte Carlo[2] sampling (Metropolis & Ulam, 1949) is a technique used in the modeling of physical and mathematical processes based on repeated random sampling. Mathematically, Monte Carlo sampling can be formulated as follows: Consider the problem of computing the integral

$$\mathbb{E}[f(\mathbf{x})] = \int_{\mathcal{X}} f(\mathbf{x})p(\mathbf{x})d\mathbf{x} \tag{2.13}$$

where $\mathbf{x}$ is a continuous random variable, $p(\mathbf{x})$ its respective probability density function and $f(\mathbf{x})$ an integrable function of $\mathbf{x}$. We can approximate the integral given in Eq. 2.13 with its *Monte Carlo* estimate $\hat{f}_n(\mathbf{x})$ by taking $n$ independent and identically distributed (i.i.d) random samples $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\}$ from $\mathcal{X}$ and computing the mean of $f(\mathbf{x})$ over the samples according to:

$$\hat{f}_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}^{(i)}). \tag{2.14}$$

From the *Weak Law of Large Numbers*, if $\mathbb{E}[f(\mathbf{x})] < \infty$, then for any arbitrarily small $\epsilon$

$$\lim_{n\to\infty} P(|\hat{f}_n(\mathbf{x}) - \mathbb{E}[f(\mathbf{x})]| \geq \epsilon) = 0 \tag{2.15}$$

[2] The term Monte Carlo is believed to have been first used by S. Ulam and J. von Neumann while working in the Manhattan project in Los Alamos, New Mexico.

and provided $n$ is large enough and the variances Var$[f(\mathbf{x})]$ are finite, according to the *Strong Law of Large Numbers*, $\hat{f}_n(\mathbf{x}) \overset{a.s.}{\rightarrow} \mathbb{E}[f(\mathbf{x})]$ and its convergence rate, based on the *Central Limit Theorem*, ensures that

$$\sqrt{n}\left(\hat{f}_n(\mathbf{x}) - \mathbb{E}[f(\mathbf{x})]\right) \overset{\mathscr{L}}{\rightarrow} \mathcal{N}(0, \sigma^2) \tag{2.16}$$

where $\sigma^2$ =Var$[f(\mathbf{x})]$. It is also worth mentioning that $\hat{f}_n(\mathbf{x})$ is an unbiased estimator of $\mathbb{E}[f(\mathbf{x})]$ and the error rate is of order $\mathcal{O}(n^{-1/2})$, independent of the dimension of $\mathbf{x}$. A key issue in Monte Carlo sampling, is how to reduce the variance of the Monte Carlo estimators. Many variance-reduction techniques exist that are suitable for large variety of problems across a wide range of disciplines, however, three techniques that have been particularly useful for many applications are (a) Importance Sampling (IS, Marshall, 1956), (b) Sequential Importance Sampling (SIS) and Sampling Importance Resampling (SIR, Efron, 1982; Rubin, 1987). These methods are further discussed in the next section in the context of sequential Monte Carlo estimation. More formal discussions of Monte Carlo sampling methods can be found in Rippley, (1987); Hammersley & Handscomb, (1964); MacKAy, (1999), and Liu, (2001).

## 2.5 Sequential Monte Carlo Methods: Particle Filters

Sequential Monte Carlo (SMC) methods[3] are model estimation techniques that rely on simulations for sampling from a sequence of probability distributions. SMC methods for on-line learning within a Bayesian framework, can also be found in the literature under Particle Filters (Gordon, 1993), Bootstrap Filters (Green, 1995), Sampling Importance Resampling (SIR, Efron, 1982; Rubin, 1987), Condensation Trackers (Isard, 1998), Interacting Particle Approximations (Crisan,

[3] A valuable resource on this topic can be found at the University of Cambridge, Sequential Monte Carlo Methods homepage (www-sigproc.eng.cam.ac.uk/smc/).

1999) and Survival of the Fittest (Kanazawa, Koller, & Russel, 1995).

Traditionally, a large number of random samples (commonly referred to as *particles*) is used to approximate the desired sequence of probability distributions. The system then evolves in time (i.e., the particles are propagated) via sampling algorithms such as IS, SIS and SIR. Ideally, we would like to sample directly from the posterior distribution, unfortunately, this is usually not possible because the posterior distribution is unknown. Therefore, instead, we sample from a known, easy to sample *proposal* pdf distribution.

Mathematically, the canonical Bayesian Importance Sampling algorithm can be described as follows:

Let $p(\mathbf{x}_k|\mathcal{Z}_k)$ denote the posterior distribution from which it is difficult to sample from and $\pi(\mathbf{x}_k|\mathcal{Z}_k)$ the respective, easy to sample *proposal* pdf distribution. Samples $\{\mathbf{x}_k^i\}$ (where $i = 1, \ldots, n$) drawn from $p(\mathbf{x}_k|\mathcal{Z}_k)$ are assumed to be i.i.d. and therefore, the posterior distribution $p(\mathbf{x}_k|\mathcal{Z}_k)$ can be represented by

$$p(\mathbf{x}_k|\mathcal{Z}_k) \;\approx\; \frac{1}{n}\sum_{k=1}^{n} \delta(\mathbf{x}_k - \mathbf{x}_k^i) \tag{2.17}$$

and the expectation of any integrable function $f(\mathbf{x}_k)$ of $\mathbf{x}_k$, provided $n$ is large enough so that $\hat{p}(\mathbf{x}_k|\mathcal{Z}_k) \equiv p(\mathbf{x}_k|\mathcal{Z}_k)$, can be estimated by (Chen, 2003)

$$\begin{aligned}
\mathbb{E}[f(\mathbf{x}_k)] &\approx \int f(\mathbf{x}_k)\hat{p}(\mathbf{x}_k|\mathcal{Z}_k)d\mathbf{x}_k \\
&= \frac{1}{n}\sum_{i=1}^{n} \int f(\mathbf{x}_k)\hat{p}(\mathbf{x}_k|\mathcal{Z}_k)d\mathbf{x}_k \\
&= \frac{1}{n}\sum_{i=1}^{n} \int f(\mathbf{x}_k)\delta(\mathbf{x}_k - \mathbf{x}_k^i)d\mathbf{x}_k \\
&= \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_k^i) \equiv \hat{f}_n(\mathbf{x}) \qquad (2.18)
\end{aligned}$$

However, since it is easier to sample from the *proposal* distribution $\pi(\mathbf{x}_k|\mathcal{Z}_k)$, the mean of $f(\mathbf{x}_k)$ can be estimated as

$$\begin{aligned}\mathbb{E}[f(\mathbf{x}_k)] &= \int f(\mathbf{x}_k)\frac{p(\mathbf{x}_k|\mathcal{Z}_k)}{\pi(\mathbf{x}_k|\mathcal{Z}_k)}\pi(\mathbf{x}_k|\mathcal{Z}_k)d\mathbf{x}_k \\ &= \int f(\mathbf{x}_k)\frac{w(\mathbf{x}_k)}{p(\mathcal{Z}_k)}\pi(\mathbf{x}_k|\mathcal{Z}_k)d\mathbf{x}_k \\ &= \frac{\int f(\mathbf{x}_k)w(\mathbf{x}_k)\pi(\mathbf{x}_k|\mathcal{Z}_k)d\mathbf{x}_k}{\int p(\mathcal{Z}_k|\mathbf{x}_k)p(\mathbf{x}_k)d\mathbf{x}_k} \\ &= \frac{\int f(\mathbf{x}_k)w(\mathbf{x}_k)\pi(\mathbf{x}_k|\mathcal{Z}_k)d\mathbf{x}_k}{\int w(\mathbf{x}_k)\pi(\mathbf{x}_k|\mathcal{Z}_k)d\mathbf{x}_k} \\ &= \frac{\mathbb{E}_{\pi(\mathbf{x}_k|\mathcal{Z}_k)}[w_{k(\mathbf{x}_k)}f(\mathbf{x}_k)]}{\mathbb{E}_{\pi(\mathbf{x}_k|\mathcal{Z}_k)}[w_{k(\mathbf{x}_k)}]} \qquad (2.19)\end{aligned}$$

where

$$w_k(\mathbf{x}_k) = \frac{p(\mathcal{Z}_k|\mathbf{x}_k)p(\mathbf{x}_k)}{\pi(\mathbf{x}_k|\mathcal{Z}_k)} \qquad (2.20)$$

are know as the *importance* sampling weights. Therefore, by drawing the i.i.d. samples $\{\mathbf{x}_k^i\}$ from $\pi(\mathbf{x}_k|\mathcal{Z}_k)$

$$\begin{aligned}\mathbb{E}[f(\mathbf{x}_k)] &\approx \frac{\frac{1}{n}\sum_{i=1}^n w_k(\mathbf{x}_k^i)f(\mathbf{x}_k^i)}{\frac{1}{n}\sum_{i=1}^n w_k(\mathbf{x}_k^i)} \\ &= \sum_{i=1}^n f(\mathbf{x}_k^i)\tilde{w}_k(\mathbf{x}_k^i) \equiv \hat{f}(\mathbf{x}) \qquad (2.21)\end{aligned}$$

where

$$\tilde{w}_k(\mathbf{x}_k^i) = \frac{w_k(\mathbf{x}_k^i)}{\sum_{j=1}^n w_k(\mathbf{x}_k^j)}. \qquad (2.22)$$

When the *proposal* and posterior distributions, $\pi(\mathbf{x}_{0:k}|\mathcal{Z}_{0:k})$ and $p(\mathbf{x}_{0:k}|\mathcal{Z}_{0:k})$ respectively, can be factorized, a recursive estimate of the importance weights $w_k^i$ can be easily obtained. For example, assume that $\pi(\mathbf{x}_{0:k}|\mathcal{Z}_{0:k})$ can be factored as

$$\pi(\mathbf{x}_{0:k}|\mathcal{Z}_{0:k}) = \pi(\mathbf{x}_0)\prod_{j=1}^k \pi(\mathbf{x}_j|\mathbf{x}_{0:j-1}, \mathcal{Z}_{0:j}) \qquad (2.23)$$

and that the posterior $p(\mathbf{x}_{0:k}|\mathcal{Z}_{0:k})$ can be expressed as

$$p(\mathbf{x}_{0:k}|\mathcal{Z}_{0:k}) = \frac{p(\mathcal{Z}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{x}_{k-1})p(\mathbf{x}_{0:k-1}|\mathcal{Z}_{0:k-1})}{p(\mathcal{Z}_k|\mathcal{Z}_{0:k-1})} \qquad (2.24)$$

following a similar derivation as in Eq. 2.11. The importance weights are then obtained from

$$w_k^i = w_{k-1}^i \frac{p(\mathcal{Z}_k|\mathbf{x}_k^i)p(\mathbf{x}_k^i|\mathbf{x}_{k-1}^i)}{\pi(\mathbf{x}_k^i|\mathbf{x}_{k-1}^i, \mathcal{Z}_{0:k})}. \tag{2.25}$$

### 2.5.1 Sequential Importance Sampling (SIS) Algorithm

Following the general discussion of Monte Sequential methods from the preceding section, we now discuss a popular sampling scheme know as Sequential Importance Sampling (SIS). SIS is a widely-used method for randomly sampling from hard-to-sample distributions. This approach is based on the *importance* sampling idea (Marshall, 1956) that suggests sampling only the region(s) of importance in order to save computational resources (Liu, 2001).

Despite its successes, a common problem with the SIS filter is that as the number of iterations increases, only a few (or one) of the importance weights will be different than zero (Kong, Liu & Wong, 1994). This problem is known as the *weight degeneracy* problem (Rubin, 1987) and a proposed solution was introduced by Kong, Liu & Wong, (1994) where a measure of the difference between the trail and target distribution is quantified via the *effective sample size* ($\hat{N}_{eff}$) given by

$$\hat{N}_{eff} = \frac{1}{\sum_{k=1}^{n} (w_k^i)^2} \tag{2.26}$$

If $\hat{N}_{eff}$ is less than a previously specified threshold $N_{thresh}$, the respective sample is accepted, otherwise, resampling is done following the canonical SIS algorithm as described in the box below. This resampling step reduces the computation time and therefore, improves the efficiency of the algorithm.

**Sequential Importance Sampling Algorithm**

```
For steps $k = 1, 2, \ldots$
    1. For $i = 1, \ldots, n$ draw samples from the proposal distribution
           $\mathbf{x}_k^i \sim \pi(\mathbf{x}_k | \mathbf{x}_{k-1}^i, \mathcal{Z}_{0:k})$
    2. Set $\mathbf{x}_{0:k}^i = \{\mathbf{x}_{0:k-1}^i, \mathbf{x}_k^i\}$
    3. For $i = 1, \ldots, n$ update the importance weights
           $w_k^i = w_{k-1}^i \frac{p(\mathcal{Z}_k | \mathbf{x}_k^i) p(\mathbf{x}_k^i | \mathbf{x}_{k-1}^i)}{\pi(\mathbf{x}_k^i | \mathbf{x}_{0:k-1}^i, \mathcal{Z}_{0:k})}.$
       Note that when the transition prior $\pi(\mathbf{x}_k^i | \mathbf{x}_{k-1}^i, \mathcal{Z}_{0:k}) = p(\mathbf{x}_k^i | \mathbf{x}_{k-1}^i)$,
       the importance weights can be simplified to
           $w_k^i = w_{k-1}^i p(\mathcal{Z}_k | \mathbf{x}_k^i)$
    4. For $i = 1, \ldots, n$ compute the normalized importance weights
           $\tilde{w}_k^i = \frac{w_k^i}{\sum_{j=1}^n w_k^j}$
End
```

### 2.5.2 Sampling Importance Resampling (SIR) Algorithm

The Sampling Importance Resampling (SIR) filter (Gordon, 1993; Liu & Chen, 1998) is based on the idea of resampling from the observations (these methods are known as Bootstrap and Jacknife techniques (Efron, 1982)) and is used to circumvent the problem of degeneracy. By introducing the resampling step between two *importance* sampling steps (Rubin, 1987) particles with smaller weights are more readily eliminated and thereby, increasing those samples with larger weights. Another advantage of the SIR filter is that there is greater flexibility in the choice of resampling algorithm.

An important point to emphasize from the previous discussions is that particle filters, in general, can be regarded as the recursive version of the importance sampling scheme (Levy, Reali, & Griffiths, 2009) such as the SIR algorithm. Particle filters have been widely used for probabilistic inference across many scientific domains, in particular, their growing success in providing a means for exploration of the influence of memory limitations on probabilistic inference (Levy, Reali, & Griffiths, 2009; Daw & Courville, 2008; Sanborn, Griffiths, & Navarro, 2006), makes them highly attractive as a preferred choice of computational framework for the present work. Therefore, in this manuscript we implement our simulations using particle filters as an approximation (indicated by the number of particles) to human causal inference.

## Sequential Importance Resampling Algorithm

```
For steps $k = 1, 2, \ldots$

    1. For $i = 1, \ldots, n$ initialize samples and *importance* weights respectively:
           $\mathbf{x}_0^i \sim p(\mathbf{x}_0)$ and $w_0^i = 1/n$
    2. For $i = 1, \ldots, n$ draw samples from the proposal distribution
           $\mathbf{x}_k^i \sim \pi(\mathbf{x}_k | \mathbf{x}_{k-1}^i, \mathcal{Z}_{0:k})$
    3. Set $\mathbf{x}_{0:k}^i = \{\mathbf{x}_{0:k-1}^i, \mathbf{x}_k^i\}$
    4. For $i = 1, \ldots, n$ update the *importance* weights
           $w_k^i = w_{k-1}^i p(\mathcal{Z}_k | \mathbf{x}_k^i)$
    5. For $i = 1, \ldots, n$ compute the normalized *importance* weights
           $\tilde{w}_k^i = \frac{w_k^i}{\sum_{j=1}^n w_k^j}$
    5. Calculate the effective number of particles
           $\hat{N}_{eff} = \frac{1}{\sum_{k=1}^n \left(w_k^i\right)^2}$
    6. If $\hat{N}_{eff} < N_{thresh}$, perform resampling
        (i.) Draw $n$ particles $\{\mathbf{x}_k^j\}$ from the current particle set $\{\mathbf{x}_{0:k}^i\}$ by resampling with replacement with probabilities proportional to their weights
        (ii.) For $i = 1, \ldots, n$ set $\tilde{w}_k^i = 1/n$
    7. Repeat steps 2-6

End
```

# CHAPTER 3

# Sequential Causal Learning in Humans and Rats

## 3.1 Abstract

Recent experiments (Beckers, De Houwer, Pineño, & Miller, 2005; Beckers, Miller, De Houwer, & Urushihara, 2006) have shown that pretraining and posttraining 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. We also show that model averaging is able to capture the effects of posttraining. Detailed computer simulations are in good agreement with experimental findings.

## 3.2 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.*, (2005) showed that different pretraining conditions using unrelated cues could alter the learners 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.

## 3.3 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 learners 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.

### 3.3.1 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, $\omega_1$, $\omega_2$ analogous to causal strength (Cheng, 1997). The generative model of the data, as shown in Figure 3.1, is given by

$$P(O|\omega_1, \omega_2, x_1, x_2) = \int dR_1 \int dR_2 P(O|R_1, R_2) \prod_{i=1}^{2} P(R_i|\omega_i, x_i). \tag{3.1}$$

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 $\sigma_m^2$,

$$P(O|R_1, R_2) = \frac{1}{\sqrt{2\pi\sigma_m^2}} \exp\{-(O - R_1 - R_2)^2/(2\sigma_m^2)\}. \tag{3.2}$$

The complete first model is given by:

$$\begin{aligned} P(O|\omega_1, \omega_2, x_1, x_2) &= \int dR_1 \int dR_2 P(O|R_1, R_2)P(R_1|\omega_1, x_1)P(R_2|\omega_2, x_2), \\ &= \frac{1}{\sqrt{2\pi\sigma_f^2}} \exp\{-(O - \omega_1 x_1 - \omega_2 x_2)^2/(2\sigma_f^2)\}. \end{aligned} \tag{3.3}$$

In this case, we are able to exploit the fact that all the distributions are Gaussian in order to integrate out the hidden variables, with $\sigma_f = \sqrt{\sigma_m^2 + 2\sigma_h^2}$.

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 this model for continuous outcome variables, we express it as a MAX or noisy-MAX. This gives two related alternative expressions for $P(R|R_1, R_2)$:

$$P(O|R_1, R_2) = \frac{1}{\sqrt{2\pi\sigma_m^2}} \exp\{-(O - \max(R_1, R_2))^2/(2\sigma_m^2)\}, \tag{3.4}$$

$$P(O|R_1, R_2) = \frac{1}{\sqrt{2\pi\sigma_m^2}} \exp\{-(O - F(R_1, R_2; T))^2/(2\sigma_m^2)\}, \tag{3.5}$$

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

$$F(R_1, R_2; T) = R_1 \frac{e^{R_1/T}}{e^{R_1/T} + e^{R_2/T}} + R_2 \frac{e^{R_2/T}}{e^{R_1/T} + e^{R_2/T}}. \tag{3.6}$$

The parameter $T$ determines the sharpness of the noisy-MAX function. As $T \mapsto 0$, the noisy-MAX function becomes identical to the max function. By contrast, as $T \mapsto \infty$ the noisy-MAX function becomes the weighted average $(R_1 + R_2)/2$. Generally, the larger $T$ the softer the noisy-MAX.

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

$$P(R_i|\omega_i, x_i) = \frac{1}{\sqrt{2\pi\sigma_h^2}} \exp\{-(R_i - \omega_i x_i)^2/(2\sigma_h^2)\},\ i = 1, 2. \tag{3.7}$$

As before, we generate the full MAX and noisy-MAX models by using the formula:

$$P(O|\omega_1, \omega_2, x_1, x_2) = \int dR_1 \int dR_2 P(O|R_1, R_2) P(R_1|\omega_1, x_1) P(R_2|\omega_2, x_2), \quad (3.8)$$

using the appropriate formula for $P(R|R_1, R_2)$. In these cases, it is impossible to simplify the distribution $P(R|\omega_1, \omega_2, x_1, x_2)$.

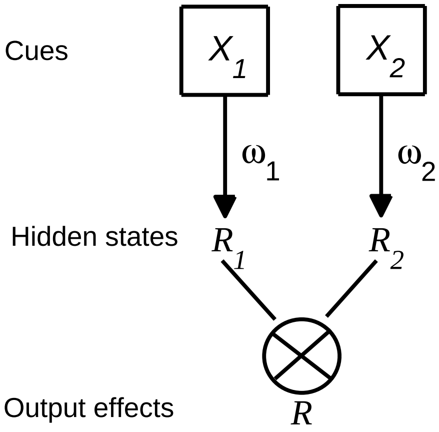

Figure 3.1: An illustration of the generative models. The different models combine $R_1$ and $R_2$ in different ways, a *linear-sum* or a *noisy*-MAX, to yield the output effect $O$.

### 3.3.2 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 $\sigma_p^2$. 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 $\omega_1$, $\omega_2$ over time (i.e.,trials), as in Dayan & Kakade (2000).

$$P(\omega_i^{t+1}|\omega_i^t) \quad = \quad \frac{1}{\sqrt{2\pi\sigma_T^2}} \exp\{-(\omega_i^{t+1} - \omega_i^t)^2/(2\sigma_T^2)\}, \quad i = 1, 2. \tag{3.9}$$

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

### 3.3.3 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:

$$P(\vec{\omega}^{t+1}|\{O_t\}, \{\vec{x}_t\}) = \int d\vec{\omega}_t P(\vec{\omega}^{t+1}|\vec{\omega}^t) P(\vec{\omega}^t|\{O_t\}, \{\vec{x}_t\}), \tag{3.10}$$

$$P(\vec{\omega}^{t+1}|\{O_t\}, \{\vec{x}_t\}) = \frac{P(O_{t+1}|\vec{\omega}^{t+1}, \vec{x}^{t+1}) P(\vec{\omega}^{t+1}|\{O_t\}, \{\vec{x}_t\})}{P(O_{t+1}|\{O_t\}, \{\vec{x}_{t+1}\})}. \tag{3.11}$$

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

The process is initialized by setting $P(\vec{\omega}^0)$ equal to the prior (i.e., product of

Gaussians with 0 means and variances $\sigma_p^2$).

We use Eq. 3.10 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. 3.11 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. 3.10-3.11 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, $\omega_{t+1}$. 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. 3.10-3.11 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. 3.10-3.11 are applied directly in the distribution updating.

### 3.3.4 Parameter Estimation and Model Selection

There are two types of inference that we can make from the posterior distributions $P(\vec{\omega}^t|\{O_t\}, \{\vec{x}_t\})$. 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:

$$\hat{\vec{\omega}}^t = \int d\vec{\omega}^t P(\vec{\omega}^t | \{O_t\}, \{\vec{x}_t\}) \vec{\omega}^t \tag{3.12}$$

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:

$$P(\{O_\tau\} | \{\vec{x}_\tau\}) = \prod_{t=0}^{\tau-1} P(O_{t+1} | \{O_t\}, \{\vec{x}_{t+1}\}), \tag{3.13}$$

with the convention that

$$P(O_1 | \{O_0\}, \{\vec{x}_1\}) = \int d\vec{\omega} P(O_1 | \vec{\omega}, \vec{x}_1) P(\vec{\omega}). \tag{3.14}$$

## 3.4 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 illustrate how our approach accounts for human and rat performance based on model selection and parameter estimation for sequential data.

We implement the models by discretizing the variable $\omega_1$, and $\omega_2$ and refer to this implementation as the 'bins' method. We perform the integrals in equations (3.10, 3.14) by discrete sums. This requires that the variables are restricted to lie within a fixed range depending on the experiment. The discretization must be sufficiently dense to ensure good quality results (an alternative is to use particle filtering techniques). Table (3.1) lists the respective parameters for the noisy-OR, human and rat experiments. These parameters include the observation variance $\sigma_h^2$, temporal prior variance $\sigma_T^2$, reward combination variance $\sigma_m^2$, range of values for the weights $\omega$, and the number of bins ($N$) used for discretizing $\omega$. To facilitate comparison with previous works, we list the corresponding standard deviations instead of the variances themselves.

| Data | $\sigma_h$ | $\sigma_T$ | $\sigma_m$ | T | $(\omega_{\min}, \omega_{\max})$ | N$^o$ of bins $(N)$ |
|---|---|---|---|---|---|---|
| **Noisy-OR** | 0.6 | 0.1 | 0.01 | 0.1 | [0, 1] | 100 |
| **Human** | 0.6 | 0.3 | 0.01 | 0.4 | [-1, 2] | 100 |
| **Rat** | 0.6 | 0.6 | 0.01 | 0.3 | [-2, 4] | 100 |

Table 3.1: Summary of model parameters. Listed are the parameters used for each of the three experiments, Noisy-OR, Human, and Rat.

We checked our implementation by comparing our results for the *linear-sum* model to those reported in Dayan & Kakade, (2000), and Daw, Courville, & Dayan, (2007) using the standard Kalman update equations. Our results were in very good agreement (provided the discretization was sufficiently dense).

The nature of the problem implies that the variables $\omega_1, \omega_2, R_1, R_2$ are essentially zero outside a small range. Hence we truncate the integrals outside these ranges (while checking to ensure that the distributions remain within them). We approximate the integrals by finite sums (we explored between 20 and 100 bins). We check that the distributions are normalized by computing the normalization factors explicitly. We also check that these normalization factors are close to 1 (if not, there is a risk of bugs). We are careful at computing terms such as $e^{R_1/T}/(e^{R_1/T} + e^{R_2/T})$ since such terms become unstable if $e^{R_1/T}$ and $e^{R_2/T}$ are small (like dividing zero by zero).

As discussed in Chapter 2, we also implement our simulations using particle filters. Therefore, in order to provide a unified account of the results from all the simulations, we report our results based on the simulations from the particle filter method when appropriate. We verified that both the bins and particle filter simulations yielded consistent results for every experiment. Simulations of experiments with the particle filter are performed with 4000 particles and the results shown are the averages from 25 Monte Carlo runs. In Chapter 4 we we provide further details of the particle filter implementation.

### 3.4.1 Forward/Backward Blocking

Conditioning paradigms provide a window to the investigation of natural inferences produced by causal learning. Two common paradigms, schematized in Table (3.2), 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.

Figure 3.2 shows simulations of learning of weight for cue A ($\omega_A$, solid) and cue X ($\omega_X$, dotted) as a function of trial number in forward blocking (black) and backward blocking (gray) designs. Figure 3.2A, 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

| **Blocking** | **Training** | **Training** | **Test** |
|---|---|---|---|
| **Paradigm** | phase 1 | Phase 2 | |
| Forward | 8A+ | 8AX+ | A, X |
| Backward | 8AX+ | 8A+ | A, X |

Table 3.2: Design summary for a typical blocking experiment. The numerical values indicate the number of trials, + indicates the presence of the outcome effect.

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.

### 3.4.2 Impact of Pretraining on Human Judgments

We simulated results of a pretraining study with humans by Beckers *et al.*, (2005). Table (3.3) 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 (3.3), 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

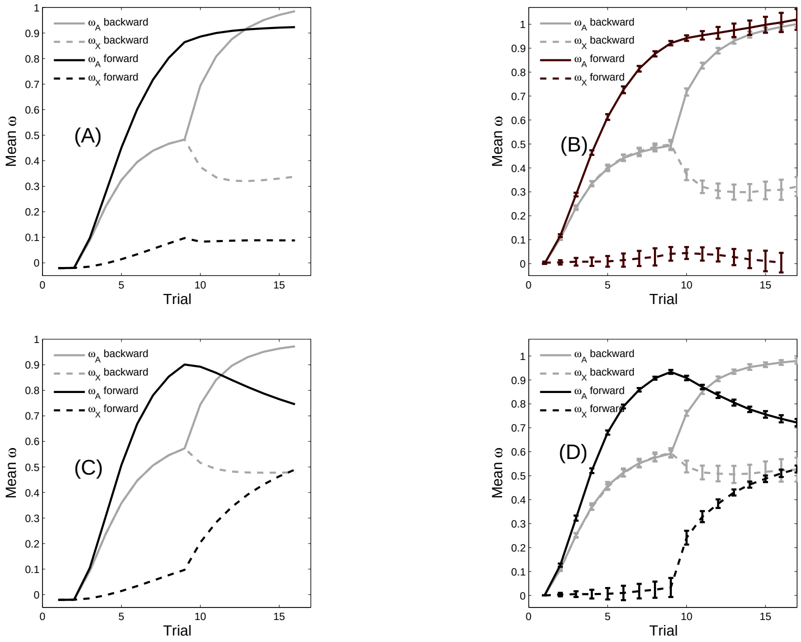

Figure 3.2: Predicted mean weights of each cue as a function of training trials in two different blocking paradigms. (A) *linear-sum* model; (C) *Noisy*-MAX model. The black lines indicate predictions for forward blocking paradigm (A+, AX+); the gray lines indicate predictions for backward blocking paradigm (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 $\omega_A$ and $\omega_X$ across the two blocking paradigms than does the *Noisy*-MAX. Plots (B) and (D) show the respective particle filter simulations of plots (A) and (C) using 4000 particles and averaging over 25 Monte Carlo runs. Also shown are the corresponding $1\sigma$ error bars.

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 3.4A, 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:** | **Phase 2:** | **Phase 3:** |
|---|---|---|---|
| | **Pretraining** | **Elemental Training** | **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 3.3: 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 $\omega$ for each cue. The mean of each $\omega$ 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. 3.10. The resulting plots are shown in Figure 3.3. In the simulation we used model parameters $\sigma_h = 0.6$, $\sigma_T = 0.3$, $\sigma_m = 0.01$, and $T = 0.4$. 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 3.8) 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.

We then computed the mean weights, using Eq. 3.12, for the models chosen by the model selection stage. These mean weights (see Figure 3.4B) constitute our simulations predictions for the causal ratings. The simulation results are in good agreement with the results for humans (Figure 3.4A). 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.

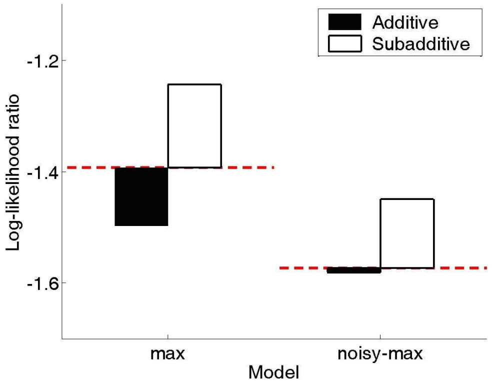

Figure 3.3: Log-likelihood ratios for the MAX and Noisy-MAX models relative to the *linear-sum* model for Experiment 2 of Beckers *et al.*, (2005). Two models were evaluated, MAX model and noisy-MAX model, as shown on the x axis. Black bars indicate model selection results for the additive group; white bars indicate the selection results for the subadditive group. The red lines indicate the thresholds for the log-likelihood ratio when only the data G+ and H+ had been shown. If the log-likelihood ratio is below the threshold line, the *linear-sum* model is preferred; if the ratio is above the threshold line, the MAX or noisy-MAX models are preferred. The results clearly show that the *linear-sum* model is selected in the additive group, whereas the MAX or noisy-MAX model is selected in the subadditive group.

### 3.4.3 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 (3.4) (Beckers *et al.*, 2006, Experiment 1). Animals in the

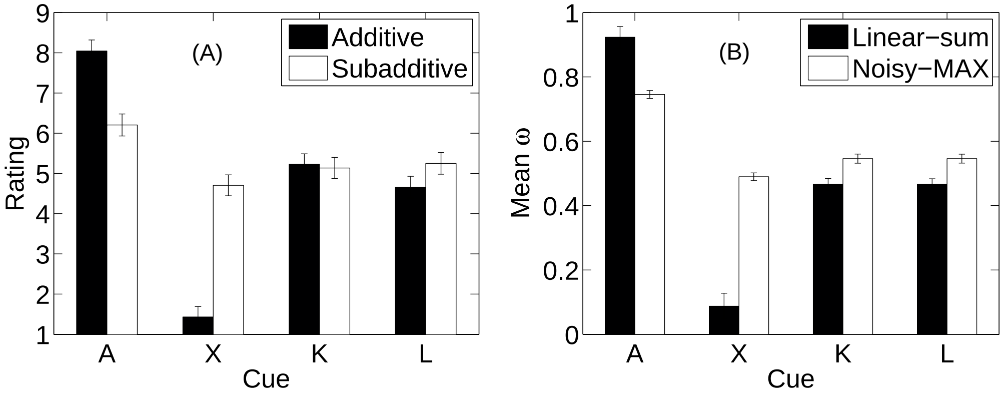

Figure 3.4: Mean causal rating for each cue. (A) Human ratings in Experiment 2 by Beckers *et al.*, (2005); see their Figure 3, p. 243. Black bars indicate the mean rating for additive pretraining group; white bars for sub-additive pretraining group. (B) Predicted ratings based on the selected model for each group. Black bars indicate the mean $\omega$ based on the *linear-sum* model, which gives a good fit for the human means in the additive group. White bars indicate the mean $\omega$ based on the *noisy*-MAX model, which give a good fit for the human means in the sub-additive group. Simulation results represent averages (with respective $1\sigma$ error bars) over 25 Monte Carlo runs using the particle filter implementation with 4000 particles.

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 subadditive 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.

We used the same translation to the model notation as before. We set the

| 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+ | 12A+ | 4AX+ |
| **Irrelevant Element** | | | |
| Experimental | 4C+/4D+/4E+ | 12A+ | 4AX+ |
| Control | 4C+/4D+/4E+ | 12A+ | 4AX+ |

Table 3.4: Design summary for the rat pretraining experiment by Beckers *et al.*, (Exp. 1, 2006).

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 loglikelihood ratios for the pre-testing data, using Eq. 3.13, to confirm that the *noisy*-MAX model was selected for the subadditive condition and the *linear-sum* model for the irrelevant condition. The results are shown in Figure 3.5. 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 3.7A 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. 3.12. 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:

$$\text{suppression ratio} = \frac{N - N\bar{\omega}_X}{N - N\bar{\omega}_X + N} = \frac{1 - \bar{\omega_X}}{2 - \bar{\omega_X}} \tag{3.15}$$

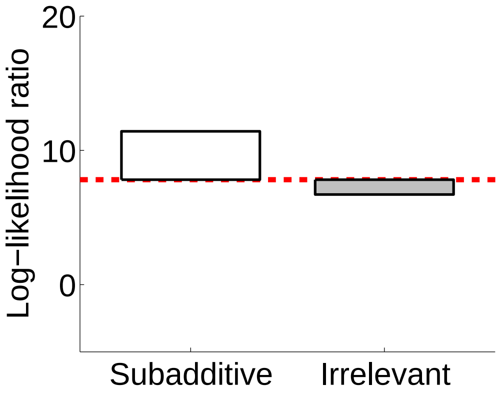

Figure 3.5: Log-likelihood ratios of *noisy*-MAX model relative to *linear-sum* model for the Subadditive condition (white), and the Irrelevant condition (gray) in the rat experiment (Beckers *et al.*, 2006, Exp. 1).

Figure 3.7B 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 3.7B, there was no significant difference

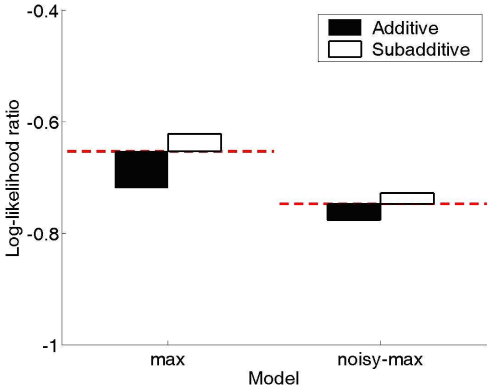

Figure 3.6: Log-likelihood ratios for the MAX and noisy-MAX models relative to the *linear-sum* model in the rat experiments performed by Beckers *et al.*, (2006). The same conventions are used as in Figure (3.3). The results clearly show that the *linear-sum* model is selected in the additive group, whereas MAX or noisy-MAX model is selected in the subadditive group.

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 Subadditive 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.

## 3.5 Comparison of Simulation Results with Experiments

Based on experimental results from Beckers *et al.*, (2005 -see Experiments 2, 3 and 4), we simulate their study of the effect of pre- and post- training on human judgements. Pretraining is simulated for (a) forward blocking (Experiment 2), (b) backward blocking (Experiment 3) and (c) release from overshadowing

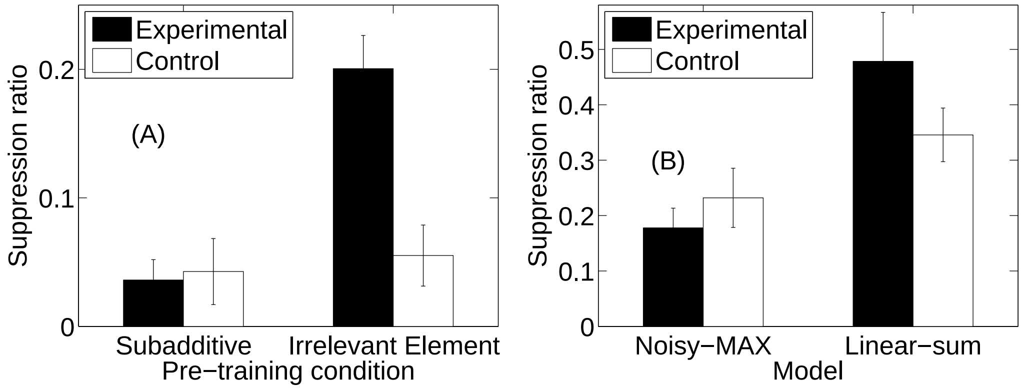

Figure 3.7: Mean suppression ratio for cue X in experimental and control groups by pretraining conditions in the subadditive condition and irrelevant condition (Beckers *et al.*, 2006, Exp. 1). Black/white bars indicate the experimental/control group, respectively. (A) Rat results; (B) Suppression ratio predicted by the *noisy*-MAX model (matched to Subadditive experimental condition), and predicted by the *linear-sum* model (matched to Irrelevant element condition). Simulation results represent averages (with respective $1\sigma$ error bars) over 25 Monte Carlo runs using the particle filter implementation with 4000 particles.

(Experiment 3). The effect of posttraining is modeled for forward blocking only (Experiment 4). Details of the experimental designs and experiments are given in Beckers *et al.*, (2005) and previous simulation results for the effect of pretraining in the forward blocking case are discussed in Lu *et al.*, 2008b.

### 3.5.1 Modeling of Pretraining Effect using Model Selection

The experiments conducted by Beckers *et al.*, 2005 consist of four different food cues: A, X, K, and L and allergic reactions to these cues are measured as moderate (+) or strong (++). In our notation, G+, H+, and GH+ correspond to $(x_1, x_2) =$

$(1, 0), (0, 1)$, and $(1, 1)$ respectively. The reaction strengths + and ++ correspond to $O = 1$ and $O = 2$ respectively. Human experiment consists of three phases: (1) pretraining, (2) elemental training and (3) compound training. Pretraining is performed with food cues, namely, G and H, and the subsequent training phases use different food cues. Pretraining conditions can be either additive (G+ → H+ → GH++) or subadditive (G+ → H+ → GH+). Cues K and L are only present in phase 3 and therefore serve as control cues.

Table 3.5 below shows the experimental design for the forward blocking experiment (Exp 2 in Beckers *et al.*, 2005) and the backward blocking and release from overshadowing experiments (Exp 3 in Beckers *et al.*, 2005). As discussed in greater detail in Lu *et al.*, (2008b), using the pretraining trials in Phase 1, we perform model selection, as shown in Figure 3.8.The simulation results 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. Next, we adopt the Bayesian sequential model to update posterior distributions of the weights $\omega$ for each cue presented in Phases 2 and 3. To compare our simulation results with the human ratings, we compute the mean of each $\omega$.

Figure 3.9 shows the mean causal rating for each cue. In the top panel, the left plot shows the human ratings in forward blocking experiment (Exp 2) by Beckers *et al.*, (2005), in which black bars indicate the mean rating for additive pretraining group; white bars for sub-additive pretraining group. The right plot shows the predicted ratings based on the selected model for each group. Black bars indicate the mean $\omega$ based on the *linear-sum* model, which gives a good fit for the human means in the additive group. White bars indicate the mean

Table 3.5: Design summary for human pretraining study in Beckers *et al.*, (2005) Experiment 2 & 3. The numerical values indicate the number of trials and + indicates the presence of the outcome effect.

| **Experiment** | **Blocking** | **Pretraining** | **Compound** | **Elemental** | **Test** |
|---|---|---|---|---|---|
| | **Paradigm** | **Phase 1** | **Phase 2** | **Phase 3** | |
| Exp 2 | Forward blocking additive | 8G+/8H+/8GH++ | 8A+ | 8AX+/8KL | A, X, K, L |
| | Forward blocking subadditive | 8G+/8H+/8GH+ | 8A+ | 8AX+/8KL | A, X, K, L |
| Exp 3 | Backward blocking additive | 8G+/8H+/8GH++ | 8AX+/KL+ | 8A+ | A, X, K, L |
| | Backward blocking subadditive | 8G+/8H+/8GH+ | 8AX+/KL+ | 8A+ | A, X, K, L |
| | Release from overshadowing additive | 8G+/8H+/8GH++ | 8AX+/KL+ | 8A- | A, X, K, L |
| | Release from overshadowing subadditive | 8G+/8H+/8GH+ | 8AX+/KL+ | 8A- | A, X, K, L |

$\omega$ based on the *noisy*-MAX model, which give a good fit for the human means in the sub-additive group. The simulation results are in good agreement with the results for humans. 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.

The bottom panel in Figure 3.9 show the results from human and model in backward blocking experiment (Exp 3) by Beckers *et al.*, (2005). Again, the model predictions agree with human performance well. A pretraining effect still preserves for both human and model, although the effect is much weaker than in forward blocking experiment (Exp 2 in Beckers *et al.*, 2005).

Another important human causal learning paradigm that has been experimentally observed (Exp. 3 in Beckers *et al.*, 2005) is the release from overshadowing effect. Similar to backward blocking, the release from overshadowing effect is another example of the more general retrospective revaluation effects. The key

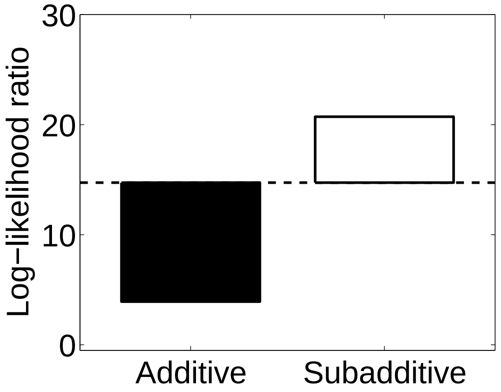

Figure 3.8: Log-likelihood ratios for the *noisy*-MAX model relative to the *linear-sum* model for experiment by Beckers *et al.*, (2005). Black bars indicate the ratio for the additive group; white bars for the sub-additive group. The dashed line indicates the threshold for model selection. These simulation results are in good agreement with experimental findings (see subsection 3.5.1).

difference between the two cue competition effects lies in the elemental training phase; in the case of release from overshadowing, trials of A- follow the compound training trials of AX+, as opposed to backward blocking where A+ trials follow the AX+ trials. Figure 3.10 shows our simulation results for the release from overshadowing experiment. Following the same convention as in Figure 3.9, the left plot shows the human ratings in the release from overshadowing experiment (Exp 3) by Beckers *et al.*, (2005) and the right plot the predicted ratings based on the selected model for each group. We can see that our model predictions, for cues A, X, and K accurately predict the observed human performance. In the case of cue L, there a small discrepancy, however, given the experimental $1\sigma$ error bars, are results are within a reasonable range.

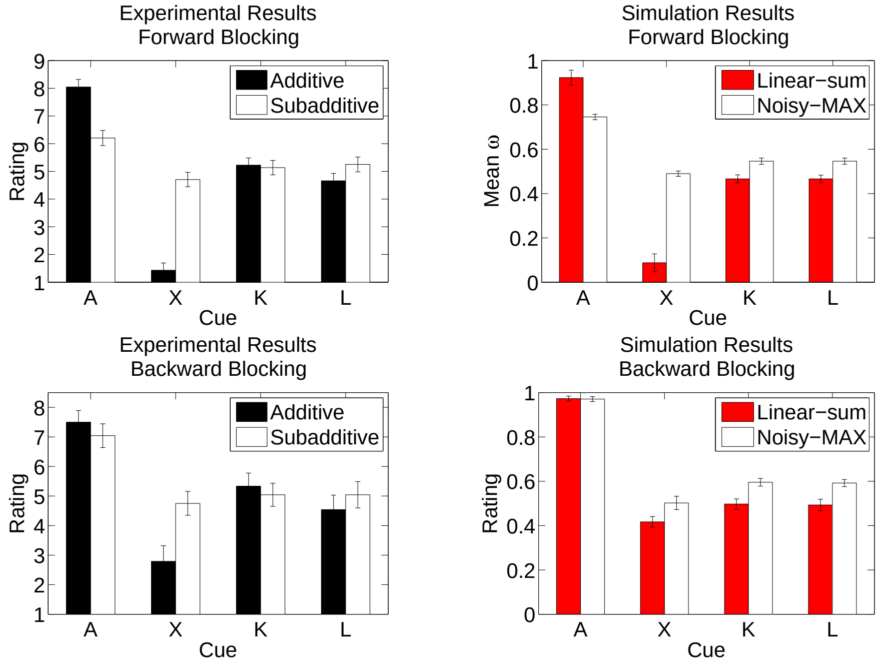

Figure 3.9: Left, Human ratings by Beckers *et al.*, (2005); Right, Predicted ratings based on the selected model for each group. Top panel, forward blocking experiment (Exp. 2 in Beckers *et al.*, 2005); Bottom panel, backward blocking experiment (Exp. 3 in Beckers *et al.*, (2005)). For further details see subsection 4.1.1. Simulation results represent averages (with respective $1\sigma$ error bars) over 25 Monte Carlo runs using the particle filter implementation with 4000 particles.

### 3.5.2 Modeling of Posttraining Effect using Model Averaging

Experiment 4 in the study conducted by Beckers *et al.*, (2005) reported that information about outcome additivity have an impact on blocking even if it is presented after the blocking training phases. As shown in Table 3.6, Phases 1 and 2 correspond to the elemental and compound training phases respectively with cue A and X, but Phase 3 is the posttraining phase with different cues (i.e. cue G and H). After the posttraining phase, human subjects were asked to evaluate

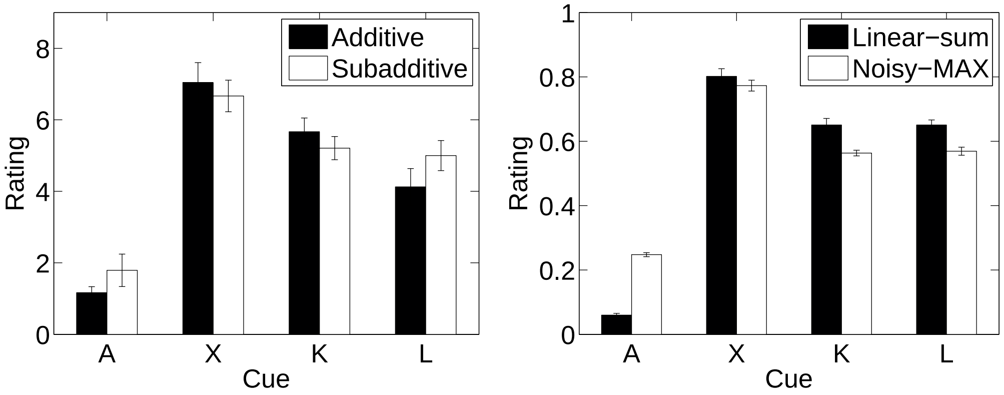

Figure 3.10: Release from Overshadowing results. Left, Human ratings from the release from overshadowing experiment (Exp. 3 in Beckers *et al.*, (2005); Right, Predicted ratings based on the selected model for each group. For further details see subsection 3.5.1. Simulation results represent averages (with respective $1\sigma$ error bars) over 25 Monte Carlo runs using the particle filter implementation with 4000 particles.

the causal power for cue A and X. In the other words, the design in Exp 4 is identical as it in Exp 1 described in Section 3.5.1, except reversing the order of the actual blocking training and the additivity training, effectively turning the additivity manipulation in a posttraining instead of a pretraining procedure.

In the first blocking training phases, we assume that humans update posterior distributions of causal strengths for models, *linear-sum* and *noisy*-MAX. When the posttraining phase is provided, model averaging is performed to combine the estimates of causal strengths from the two models as

$$\langle \omega \rangle = P(D|M_1)\bar{\omega}_{M_1} + P(D|M_2)\bar{\omega}_{M_2} \tag{3.16}$$

where $D$ is the data, $M_1$ and $M_2$ represent the *linear-sum* and *noisy*-MAX gate models respectively, and $P(D|M_i)$ is the 'evidence' for each model from observations in the posttraining phase. $\bar{\omega}_{M_i}$ is the estimated mean value of causal strength using each model from observations in the first two training phases.

Figure 3.11 shows our results for the posttraining experiment (Exp 4 in Beckers *et al.*, 2005). We can see that model averaging is able to capture the posttraining effects qualitatively, and correctly predict a weaker posttraining effect than the pretraining effect described in section 3.5.1.

Table 3.6: Design summary for human posttraining study in Beckers *et al.*, (2005) Experiment 4. The numerical values indicate the number of trials and + indicates the presence of the outcome effect.

| **Experiment** | **Group** | **Elemental Phase 1** | **Compound Phase 2** | **Posttraining Phase 3** | **Test** |
|---|---|---|---|---|---|
| Exp 4 | Forward blocking additive | A+ | 8AX+/8KL+ | 8G+/8H+/8GH++ | A, X, K, L |
| | Forward blocking subadditive | A+ | 8AX+/8KL+ | 8G+/8H+/8GH+ | A, X, K, L |

A summary of the values for all the model parameters used in our simulations of Beckers *et al.*, (2005, 2006) experiments using the particle filter method is given in Table 3.7 and in Table 3.8 we include the respective parameters for the bins implementation of the simulations. The particle filter parameters listed include the likelihood function variance $\sigma_m^2$, the reward combination variance $\sigma_m^2$, temporal prior variance $\sigma_T^2$, reward combination temperature $T$ for the Noisy-MAX likelihood function, the range for the causal weights $(\omega_{min}, \omega_{max})$, range of values for the reward $(R_{min}, R_{max})$, number of particles $N_p$ used and number of Monte Carlo runs $N_{MC}$. For the variance values, we list instead the respective standard deviation values in order to better match the algorithm implementation where standard deviation values are specified rather than variances.

A key observation from the parameter values in Tables 3.7 and 3.8 is that despite the inherent differences in the experiments (e.g., humans vs. rats, blocking

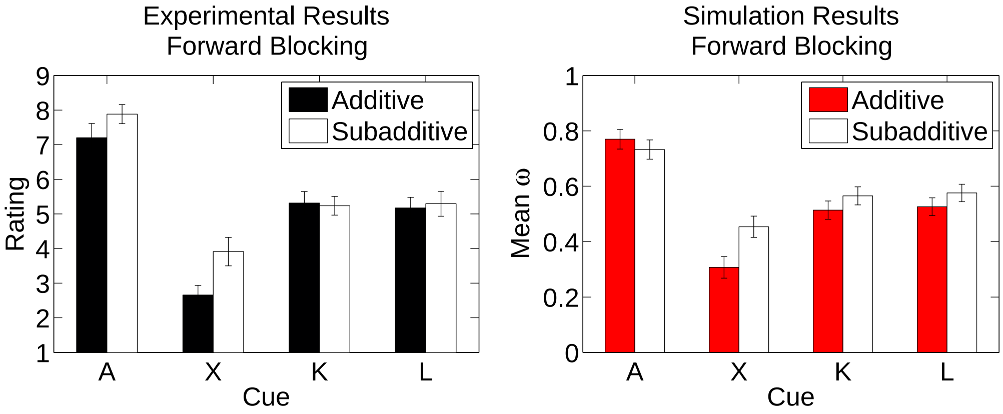

Figure 3.11: (A) Mean causal rating for each cue based on human subjects in Experiment 4 (Posttraining) of Beckers *et al.*, (2005) (see their Figure 5 p247) and (B) mean causal weights produced by the model as a function of different cues in the design of Experiment 4 (Beckers *et al.*, 2005). See section 3.2. Simulation results represent averages (with respective $1\sigma$ error bars) over 25 Monte Carlo runs using the particle filter implementation with 4000 particles.

experiment, etc.), the respective parameter values across the different experiments are not that different from one another. This property of our current implementation, highlights an important advantage of our models in that we are able to make more testable predictions than the experiments reveal (Lu, Weiden, & Yuille, 2009).

## 3.6 Analysis

### 3.6.1 Noisy-OR and MAX models

One interesting prediction is that the Noisy-OR model does not predict forward blocking. We compared the estimated causal weight of B in the binary output

Table 3.7: Summary of model parameters using the Particle Filter. Listed are the values of the parameters used for the simulations of Beckers *et al.* 2005, 2006 experiments using the method of Particle Filters for the simulations. The variables $N_p$ and $N_{MC}$ represent the number of particles used and the number of Monte Carlo runs respectively.

| **Experiment** | $\sigma_h$ | $\sigma_T$ | $\sigma_m$ | T | $(\omega_{min}, \omega_{max})$ | $(R_{min}, R_{max})$ | $N_p$ | $N_{MC}$ |
|---|---|---|---|---|---|---|---|---|
| **Exp 1, Rats (Beckers *et al.* (2006))** | 0.6 | 0.6 | 0.01 | 0.3 | [-2, 4] | [-2, 4] | 4000 | 25 |
| **Exp 2 Humans (Beckers *et al.* (2005))** | 0.6 | 0.3 | 0.01 | 0.4 | [-2, 4] | [-2, 4] | 4000 | 25 |
| **Exp 3 Backward (Beckers *et al.* (2005))** | 0.6 | 0.3 | 0.01 | 0.4 | [-2, 4] | [-2, 4] | 4000 | 25 |
| **Exp 3 Overshadowing (Beckers *et al.* (2005))** | 0.6 | 0.3 | 0.01 | 0.4 | [-2, 4] | [-2, 4] | 4000 | 25 |
| **Exp 4 Forward (Beckers *et al.* (2005))** | 0.6 | 0.3 | 0.01 | 0.4 | [-2, 4] | [-2, 4] | 4000 | 25 |

Table 3.8: Summary of model parameters using the bins method. Listed are the values of the parameters used for the simulations of Beckers *et al.* 2005 experiments using the bins implementation for the simulations.

| Experiment | $\sigma_h$ | $\sigma_T$ | $\sigma_m$ | T | $N$ | $(\omega_{min}, \omega_{max})$ | $(R_{min}, R_{max})$ |
|---|---|---|---|---|---|---|---|
| **Exp 1, Rats (Beckers *et al.* (2006))** | 0.6 | 0.6 | 0.01 | 0.3 | 100 | [-2, 4] | [-2, 4] |
| **Exp 2 Humans (Beckers *et al.* (2005))** | 0.6 | 0.3 | 0.01 | 0.4 | 100 | [-1, 2] | [-2, 4] |
| **Exp 3 Backward (Beckers *et al.* (2005))** | 0.6 | 0.3 | 0.01 | 0.4 | 100 | [0, 1] | [-2, 4] |
| **Exp 3 Overshadowing (Beckers *et al.* (2005))** | 0.6 | 0.3 | 0.01 | 0.4 | 100 | [-0.4, 2] | [-2, 4] |
| **Exp 4 Forward (Beckers *et al.* (2005))** | 0.6 | 0.3 | 0.01 | 0.4 | 100 | [-1, 2] | [-2, 4] |

case in two conditions, including standard forward blocking (A+, AB+) and its control condition (AB+). The noisy-OR model predicts the same causal weight of B in both conditions, suggesting a complete lack of blocking. This result is also replicated by the MAX and Noisy-MAX model. The simulation with small number of trials confirms the above theoretical prediction, as shown in Figure (3.12).

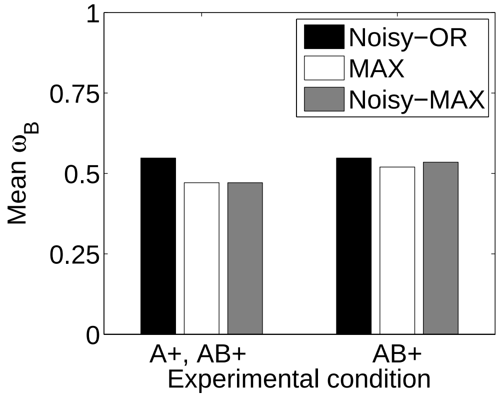

Figure 3.12: Model mean causal weights of cue B in two different conditions. The first one (A+, AB+) is the standard forward blocking paradigm; the second (AB+) is the control condition. The results show that MAX and noisy-MAX models are good approximation to the Noisy-OR model. As predicted, Noisy-OR model did not show any blocking effect. Max and noisy-MAX models predict weak blocking effect with 15 trials per condition.

### 3.6.2 Analysis without a Temporal Prior

The temporal prior plays a critical role in our models. As shown in Figure (3.2), the model with temporal prior correctly predicts the asymmetry between the forward and backward blocking. However, without the temporal prior, the model would give identical results for forward and backward blocking in disagreement with experimental findings Daw, Courville, & Dayan, (2007). To gain a deeper understanding of the model we analyze the case when the temporal prior is switched off.

We consider the limit case when the variance $\sigma_h$ is zero. The *linear-sum* model is expressed as:

$$P(r|x_1, x_2, \omega_1, \omega_2) \quad = \quad \frac{1}{\sqrt{2\pi}\sigma_m} \exp\{-(r - \omega_1 x_1 - \omega_2 x_2)^2/(2\sigma_m^2)\}. \quad (3.17)$$

Now suppose we have $N_1$ samples with $r = 1, x_1 = 1, x_2 = 1$ and $N_2$ samples with $r = 1, x_1 = 1, x_2 = 1$. Then the maximum likelihood (ML) estimates for the parameters $\omega_1, \omega_2$ are obtained by minimizing the energy:

$$E(\omega_1, \omega_2) \quad = \quad N_1(1 - \omega_1 - \omega_2)^2 + N_2(1 - \omega_1)^2. \quad (3.18)$$

For $N_2 > 1$, this is minimized by setting $\omega_1 = 1$ and $\omega_2 = 0$. Note that if $N_2 = 0$, then there is a family of solutions with $\omega_1 + \omega_2 = 1$.

There are two main points from this analysis. First, the results depend only on the numbers $N_1$ and $N_2$ of samples. In particular, the results are independent of the ordering of the samples and give the same results for both forward and backward blocking. Second, the results depend very weakly on $N_1, N_2$, and only on whether these numbers are zero or non-zero. A single example $N_2 = 1$ is sufficient to ensure that $\omega_1 = 1$ and $\omega_2 = 0$. This would imply that there would be very fast jumps in the estimates of $\omega_1, \omega_2$ as we change phases.

We can perform a similar analysis for the noisy-MAX. Again we set $\sigma_h = 0$. The noisy-MAX model is:

$$P(r|x_1, x_2, \omega_1, \omega_2) = \frac{1}{\sqrt{2\pi}\sigma_m} \exp\{-(r - \max(\omega_1 x_1, \omega_2 x_2))^2/(2\sigma_m^2)\}. \quad (3.19)$$

The values of $\omega_1, \omega_2$ are obtained by minimizing the energy:

$$E(\omega_1, \omega_2) \quad = \quad N_1(1 - \max(\omega_1, \omega_2))^2 + N_2(1 - \omega_1)^2. \quad (3.20)$$

For $N_1 > 0, N_2 > 0$, the solution is given by $\omega_1 = 1$ and $\omega_2 \leq 1$. For $N_2 > 0$ we have $\omega_1 = 1$. For $N_2 = 0$, we get $\max(\omega_1, \omega_2) = 1$.

This analysis yields similar results as above. The ordering of the samples does not matter. But the occurrence of a single example with $r = 1, x_1 = 1, x_2 = 0$ (i.e. $N_2 = 1$) is sufficient to ensure that $\omega_1 = 1$.

### 3.6.3 Need for Updating when No Data are Present

Finally, there is the issue of whether the temporal prior should be applied if the corresponding cause is not present (i.e., should we update $\omega_i$ if $x_i = 0$?). We have obtained results for either situation. The results do not differ very much at the level of analysis with which we are concerned. There are, however, various paradoxes associated with updating the temporal prior for $\omega_i$ in situations where $x_i = 0$. The reason is that the temporal prior acts as a type of blurring of the weight. If the distribution of $\omega_i$ is initially peaked sharply about zero, then it becomes increasingly broadly peaked as we run the temporal prior. This seems unintuitive. It is more plausible that the temporal prior is only activated when the cause is present. The idea is that the temporal prior is a luxury that the system can only afford if there are data coming in.

## 3.7 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 additive 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 learners 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.

In the next chapter we extend our applications of particle filters to the study of the limitations of human inference. In particular, we show that by using a limited number of particles (¡6000), we are able to capture important conditioning phenomena such as the highlighting effect and that when we increase this number substantially (e.g., $\geq 18,000$) to reflect exact inference, the effect vanishes as expected (see e.g., Daw *et al.*, 2005).

# CHAPTER 4

# Modeling Causal Generalization with Particle Filters

## 4.1 Abstract

We model the highlighting effect observed in Daw, Courville, & Dayan, (2007) using a particle filter algorithm as an approximation to exact statistical inference, in accord with the limited computational capacity of human cognition. We find that the inferential approximation based on particle filters predicts the highlighting effect. We also generalize our models to account for relations between multiple cues and multiple outcomes. We test our generalized framework with the highlighting effect and also show that our models are robust against variations in the likelihood function variance. Specifically, we demonstrate that the results obtained from assuming the likelihood function variance is fixed are consistent with those obtained from learning the likelihood variance by varying its value within in a reasonable range. Our results also show that when a significantly large number of particles ($\geq 18,000$) is used for simulation of exact statistical inference, the highlighting effect is no longer captured by our models.

## 4.2 Introduction

Human reasoning is adaptive, as exemplified by the reasoner's ability to generalize abstract cause-effect relations from one situation to another. The challenge in understanding causal generalization is to identify how humans acquire and propagate abstract causal knowledge across domains. Causal knowledge includes several key aspects, including causal structure as captured by graphical models of the existence of cause-effect links (Griffiths & Tenenbaum, 2005; Pearl, 2000), causal power as reflected in the strength of cause-effect links (Cheng, 1997), and causal integration rules that model how to combine multiple cause-effect relations (Cheng, 1997). In this paper, we will consider these three types of causal knowledge in the context of causal generalization.

A variety of "rational" models of causal learning have taken probabilistic approaches to explain how people acquire causal knowledge from covariational observations presented in the format of summarized contingency data (Griffiths & Tenenbaum, 2005; Cheng, 1997; Lu *et al.*, 2008a). Despite their success in capturing a variety of causal learning phenomena, these models leave open the question of how a learner can cope with non-summary data. In everyday life, people often receive observations incrementally. For such naturalistic learning situations, sequential models are required to account for the influence of the order of data presentation.

To address this issue, a Bayesian sequential model can be used as an inference engine to capture the propagation of causal knowledge over time. Kalman filtering (Kalman, 1960; Kamin, 1969) has been successfully applied in sequential causal learning, and has been used to explain various experimental phenomenon in animal conditioning (Dayan & Kakade, 2000; Daw, Courville, & Dayan, 2007). However, a limitation of previous work using Kalman filtering involves the assumption of Gaussian distributions with a *linear-sum* causal integration rule to combine multiple cause-effect relations. Many empirical studies have shown that the *linear-sum* rule is not able to account for human causal learning with binary variables (Buehner, 2003; Liljeholm, 2007). Accordingly, a more flexible inference model is required to account for a broader range of learning situations. In the present paper, we present a model of sequential causal learning based on particle filtering (Wood, 2007), a sequential method used for approximate probabilistic inference over time. This model is used to explore how different causal integration rules can be selected, and how causal knowledge can be propagated with increasing certainty as the number of sequential observations increases.

An additional critical issue concerns how to model the generalization of causal knowledge from one context to another. In the laboratory, researchers have designed behavioral experiments to measure causal generalizability in controlled environments. Beckers *et al.*, (2005) first trained human subjects with certain cue-outcome pairs, such as bacon (cue G) and eggs (cue GH) each paired with a moderate allergic reaction. The combination of the two cues, bacon and eggs (cue GH) was paired with either a moderate or a strong allergic reaction. The subjects were then transferred to a classic forward blocking paradigm with unrelated cues, such as cheese (cue A) paired with moderate allergy, and cheese and nuts (cue AX) also paired with moderate allergy. Finally, subjects were tested on how likely nuts alone (cue X) was to cause allergy. Human participants provided different ratings on the transfer test for cue X depending on whether cue combination GH has been paired with moderate or strong allergy during the pretraining. Beckers *et al.*,'s (2005) study provided empirical evidence that different pretraining conditions using unrelated causal cues can alter the reasoner's assumptions, and thereby change their subsequent causal inferences. From a computational perspective, the influence of pretraining conditions can be explained in terms of a Bayesian process of model selection, which operates to identify one of the most important aspects of causal knowledge, the causal integration rule, and transfer it to the subsequent inference task. An alternative Bayesian procedure would be model averaging (Erven, 2007), which accepts several causal integration rules with different calculated probabilities, and then averages the inference results across all the possible integration rules. In the final section, we will explore the possibility that particle filtering could serve as an approximation to rational inference while allowing for the limitations on human computational capacity (Wood, 2007).

In this chapter, we describe our computational theory in Section 4.3. Section 5 compares human results with model predictions in three experiments. Last, we show how the particle filter approach is able to explain highlighting phenomenon, which has been a challenge to Bayesian sequential learning model using Kalman filters with exact inference.

## 4.3 The Computational Theory

This section describes our computational theory. We specify two alternative models which compete to explain the data by model selection or cooperate to explain the data by model averaging, see subsection (4.3.1). We implement the theory using particle filters as described in subsection (4.3.2).

### 4.3.1 The Models

The experiments specify a sequence of input and output pairs $(\vec{x}_1, d_1), ..., (\vec{x}_t, d_t)$. The input $\vec{x} = (x_1, x_2)$ specifies which cause is present: (i) cause 1 if $x_1 = 1, x_2 = 0$ and, (ii) cause 2 if $x_1 = 0, x_2 = 1$. The output $d$ is a continuous variable. We use capital variables as shorthand for these sequences so that $\vec{X}_t = (\vec{x}_1, ..., \vec{x}_t)$ and $D_t = (d_1, ..., d_t)$.

Both models are parameterized by weight variables $\vec{\omega} = (\omega_1, \omega_2)$ which indicate the strength of the causes $x_1, x_2$ for causing the effect. We specify a prior $P(\vec{\omega})$ on the weights which is a Gaussian with zero mean and large covariance (making weak assumptions about the initial values of the weights). We specify a temporal prior $P(\vec{\omega}_{t+1}|\vec{\omega}_t)$ which allows the weights to change over time and means that the model is most influenced by the most recent data. The prior and temporal prior are specified by:

$$P(\vec{\omega}_1) = \frac{1}{2\pi\sqrt{|\boldsymbol{\Sigma_1}|}} \exp\{-(1/2)\vec{\omega}_1^T \boldsymbol{\Sigma}_1^{-1} \vec{\omega}_1\}, \tag{4.1}$$

$$P(\vec{\omega}_{t+1}|\vec{\omega}_t) = \frac{1}{2\pi\sqrt{|\boldsymbol{\Sigma}_2|}} \exp\{-(1/2)(\vec{\omega}_{t+1} - \vec{\omega}_t)^T \boldsymbol{\Sigma}_2^{-1} (\vec{\omega}_{t+1} - \vec{\omega}_t)\}, \tag{4.2}$$

where $\boldsymbol{\Sigma}_1 = \sigma_1^2 \mathbf{I}$ and $\boldsymbol{\Sigma}_2 = \sigma_2^2 \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. Hence the weights $\omega_1, \omega_2$ are decoupled in the priors. The respective $\sigma_1$ and $\sigma_2$ values are 0.05 and 0.1. It is important to point out that our sequential model is fairly robust to different initial priors. Although we only report results with one prior in this discussion, we have experimented with several different priors -including Gaussian, delta distributions centered at zero, and uniform distributions- and obtained similar results in all cases.

The likelihood functions are of form:

$$P(d|\vec{\omega}, \vec{x}, M) = \sum_{\vec{R}} P(d|\vec{R}, M) P(\vec{R}|\vec{\omega}, \vec{x}), \tag{4.3}$$

where $\vec{R}$ are the states of hidden units and $M$ indicates the model.

We define $P(\vec{R}|\vec{\omega}, \vec{x}) = \frac{1}{2\pi\sqrt{|\Sigma|}} \exp\{-(1/2)(\vec{R} - \vec{\omega} \bigotimes \vec{x})^T \Sigma^{-1} (\vec{R} - \vec{\omega} \bigotimes \vec{x})\}$, where $\vec{\omega} \bigotimes \vec{x} = (\omega_1 x_1, \omega_2 x_2)$.

The models $P(d|\vec{R})$ are of form:

$$P(d|\vec{R}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-(d - F(\vec{R}, M))^2/(2\sigma^2)\}, \quad (4.4)$$

where $F(\vec{R}, M = 1) = R_1 + R_2$ for the first model and $F(\vec{R}, M = 2) = R_1 \frac{e^{R_1/T}}{e^{R_1/T}+e^{R_2/T}} + R_2 \frac{e^{R_2/T}}{e^{R_1/T}+e^{R_2/T}}$ for the second model. The first model is the *linear-sum* model (Dayan & Kakade, 2000) and the second model is a *noisy*-MAX model (Lu *et al.*, 2008a) which is a generalization of the Noisy-OR model (Cheng, 1997). Figure 4.1 illustrates the two generative models, which employ different combination rules.

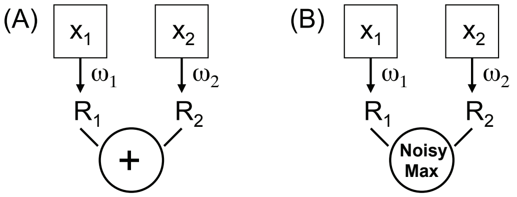

Figure 4.1: An illustration of the generative models. The different models combine R1 and R2 in different ways, a *linear-sum* (A) or a *noisy*-MAX (B), to yield the output effect R.

### 4.3.2 Inference by Particle Filtering

We use Bayes-Kalman to update the probabilities $P(\vec{\omega}_t|D_t, \vec{X}_t, M)$ when we receive new data. In the rest of the section we will drop the dependencies on $\vec{X}_t$ and $M$ to simplify the notation. Bayes-Kalman specifies that we update weights by a prediction stage followed by a measurement stage:

$$P(\vec{\omega}_{t+1}|D_t) = \int d\vec{\omega}_t P(\vec{\omega}_{t+1}|\vec{\omega}_t) P(\vec{\omega}_t|D_t) \quad \text{(prediction)} \quad (4.5)$$

$$P(\vec{\omega}_{t+1}|D_{t+1}) = \frac{P(d_{t+1}|\vec{\omega}_{t+1}) P(\vec{\omega}_{t+1}|D_t)}{P(d_{t+1}|D_t)} \quad \text{(measurement).} \quad (4.6)$$

We implement these equations using particle filters (Doucet, Freitas, & Gordon, 2001; Liu, 2001). Previous methods in the cognitive science literature are unsuited for this problem. Dayan & Kakade, 2000 used Kalman's algebraic update equations for the means and covariance of $P(\vec{\omega}_t)$, but this cannot be applied to our second model which is non-Gaussian. Lu *et al.*, (2008a) represented the distributions using a fixed lattice in $\vec{\omega}$ space, but this becomes problematic for the models described here (high accuracy requires a very dense lattice which leads to an extremely slow algorithm). By contrast, particle filters sample the space adaptively and are more efficient. (We validated particle filters by showing that they agree with these other methods when applicable).

Particle filters approximates distributions like $P(\vec{\omega}_t|D_t, M)$ by a set of discrete particles $\{\vec{\omega}_t^\mu : \mu \in \Gamma\}$. This enables us to approximates quantities such as $\int d\vec{\omega} g(\vec{\omega}_t) P(\vec{\omega}_t|D_t)$ by $(1/|\Gamma|) \sum_{\mu \in \Gamma} g(\vec{\omega}_t^\mu)$ for any function $g(.)$.

We initialize by drawing samples $\{\vec{\omega}_1^\mu : \mu \in \Gamma\}$ from the prior distribution $P(\vec{\omega})$. This is straightforward since the prior is a Gaussian.

Then we proceed recursively following the prediction and measurement stages of the Bayes-Kalman filter. Let $\{\vec{\omega}_t^\mu : \mu \in \Gamma\}$ be the set of particles representing $P(\vec{\omega}_t|D_t)$ at time $t$. Then we sample from the Gaussian distribution $P(\vec{\omega}_{t+1}|\vec{\omega}_t^\mu)$ for each $\mu$ to give a new set of particles $\{\vec{\bar{\omega}}_t^\mu : \mu \in \Gamma\}$ which represents $P(\vec{\omega}_{t+1}|D_t)$.

Next we compute the importance weights $\lambda^\mu = P(d_{t+1}|\vec{\bar{\omega}}_{t+1}^\mu)$ and normalize them to obtain $\bar{\lambda}^\mu = \lambda^\mu / (\sum_\mu \lambda^\mu)$. Then we re-sample with replacement from the set $\{\bar{\omega}_{t+1}^\mu : \mu \in \Gamma\}$ using probability $\bar{\lambda}^\mu$. This gives new set $\{\vec{\omega}_{t+1}^\nu : \nu \in \Gamma\}$ of particles which represent $P(\vec{\omega}_{t+1}|D_{t+1})$.

To compare to experiments, we need to measure the *model evidence* $P(D_t)$ for each model and to estimate the *mean values* of the weights $\vec{\bar{\omega}}_t = \int d\vec{\omega}_t \vec{\omega}_t P(\vec{\omega}_t|D_t)$. We compute these from the particles as follows.

The mean values are approximated by the average $(1/|\Gamma|)\sum_{\mu\in\Gamma}\vec{\omega}_t^{\mu}$.

The model evidence is expressed as $P(d_t|D_{t-1})P(d_{t-1}|D_{t-2})...P(d_1)$. We evaluate each term $P(d_{t+1}|D_t) = \int d\vec{\omega}_{t+1}P(d_{t+1}|\vec{\omega}_{t+1})P(\vec{\omega}_{t+1}|D_t)$ by $P(d_{t+1}|D_t) = \frac{1}{|\Gamma|}\sum_{\mu\in\Gamma}P(d_{t+1}|\vec{\omega}_{t+1}^{\mu})$.

The simulations are run using 6000 particles since beyond this value, even when performing numerous Monte Carlo runs (over 100 each with 6000 particles), the results do not show any significant variation in their outcomes. We also tested the model with a much smaller numbers of particles (100 and 500) with multiple simulation runs (i.e., the number of runs was matched to the number of subjects in the Beckers' experiments). The average performance with these numbers of particles was very close to the results obtained with 6000 particles. So a very large number of particles is not necessary to account for Beckers' experimental results. Our simulation of highlighting effects is instead performed with 1000 particles. Figure 4.2 illustrate the change of particle filters in a standard forward blocking paradigm over training trials.

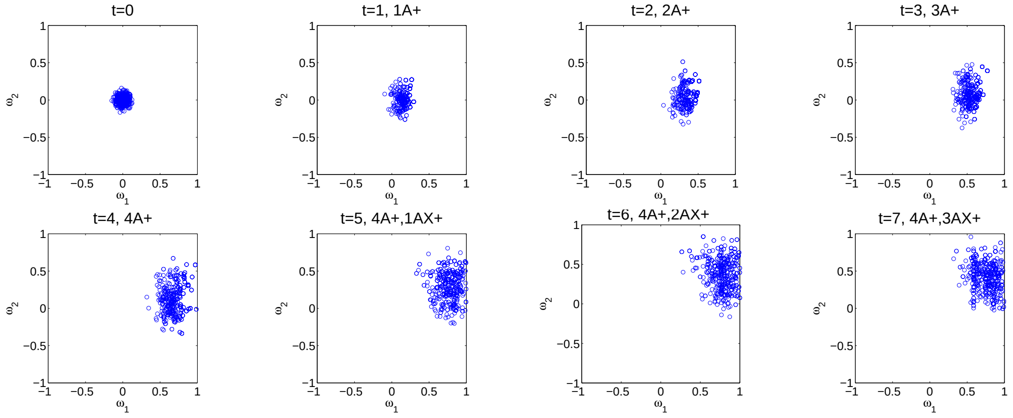

Figure 4.2: Particle filters in the simulation with forward blocking paradigm as a function of training trials.

## 4.4 Highlighting

In this section, we will demonstrate the use of particle filtering to explain another causal learning phenomenon, the highlighting effect reported by Kruschke, (2001), Kruschke, (2006) and Kruschke, (2008). Table 4.1 describes a canonical experimental design. The learner first sees 15 trials of cue A and B associated with outcome $O_1$, and 5 trials of cue A and C associated with outcome $O_2$; then in the second phase, the order of the training sessions is reversed. In the subsequent test phase, observers are asked to predict the outcome ($O_1$ or $O_2$) when showing cue A only and cue B and C together. The highlighting design equalizes the long-run base rates of the two outcomes, and the frequency of cue-outcome pairs (e.g., 20 trials in total for AC with $O_1$ and AB with $O_2$). Humans show a strong tendency to choose outcome $O_1$ for cue A, but a lower probability to choose $O_1$ for cue B and C.

Kruschke, (2006) developed a locally Bayesian learning model to account for the highlighting effect, by combining local Bayesian updating between layers and attention control through back-propagation. In contrast, Daw, Courville, & Dayan, (2007) argued that the highlighting effect could be explained by computation limitations of the human cognitive system. According to their account, human observers conduct inferential learning by an approximation to statistical models, such as Kalman filtering. Daw *et al.*, (2007) employed a rational model based on Kalman filtering with a *linear-sum* rule. With exact statistical inference, this rational model is unable to predict the highlight effect. However, including inferential approximations in their model using reduced-rank approximations was able to explain the highlighting effect.

In this section, we show that particle filtering, as an approximation to statistical inference, is able to predict the highlighting effect (Wood, 2007). The advantage of using particle filtering is that it makes it possible to control the degree of approximation to exact statistical inference, as simulations with a very large number of particles are closer to the rational inference model. To be consistent with the representations used by the two models described in the previous paragraph, we assume that outcome $O_1$ is indicated when a binary reward value is 1, and outcome $O_2$ is indicated when the reward value is 0. We thus focus the model on the case of multiple causes and a single effect, rather than extending it to the case of multiple causes and multiple effects. Given that the outcome variable is binary, we adopt two generative models, *linear-sum* (Daw, Courville, & Dayan, 2007) and noisy-logic (Cheng, 1997; Yuille, 2007), in the simulation of particle filtering.

The left plot in Figures 4.3 shows the learning curve of causal strengths for each cue and the combination of cue B and C as a function of training trials. Based upon the learned causal strengths for each cue, the model can predict how likely outcome $O_1$ will be chosen for each cue or cue combination. The highlighting effect is revealed by the difference between Cue A and BC predictions. We find that this difference is reduced with increases in the number of particles employed in the simulation, a result which agrees with the finding in Daw, Courville, & Dayan, (2007), summarized above.

The right plot in Figures 4.3 shows the probability of choosing $O_1$ for cue A and cue B&C as a function of different generative models, *linear-sum* and noisy-logic. Both models yield highlighting effects, namely higher $P(O_1)$ for cue A than for cue B&C, although the effect is weaker for the noisy-logic model.

In Figure 4.4 we show results from a simulation using 1000 (plot (A)) and 18,000 (plot (B)) particles. The experimental design is the same as in Daw, Courville, & Dayan, (2007). However, our implementation is performed with

Table 4.1: Highlighting Design by Daw, Courville, & Dayan, (2007)

| **Blocking Paradigm** | **Training Phase 1** | **Training Phase 2** | **Test** |
|---|---|---|---|
| Highlighting | 15 × ( AB → $O_1$) | 5 × ( AB → $O_1$) | A →?$O_1$ |
| | 5 × ( AC → $O_2$) | 15 × ( AC → $O_2$) | BC →?$O_2$ |

Table 4.2: Summary of model parameters. Listed are the values for the parameters used in each model.

| **Model** | $\sigma_h$ | $\sigma_T$ | $\sigma_m$ | $(\omega_{min}, \omega_{max})$ | $(R_{min}, R_{max})$ | N (particles) |
|---|---|---|---|---|---|---|
| **Noisy-OR** | 0.1 | 0.1 | 0.1 | [-1, 1] | [-1, 1] | 1500 |
| **Linear-Sum** | 0.1 | 0.1 | 0.1 | [-1, 1] | [-1, 1] | 1500 |

particle filters. We can see that the use of particle filters as an approximation to statistical inference is well motivated and justified as reflected in the causal strengths of cues A and BC in plots (A) and (B) where the presence of the highlighting effect is mediated via the number of particles.

Another important study we conducted was to investigate the effect of varying the value of the likelihood variance ($\sigma_h$) in our simulations of the highlighting effect. In our models we assumed $\sigma_h$ was fixed and equal to 0.1, however, this was motivated by the fact that when we varied its value in the interval [0.01, 1], which covers a large dynamic range, the results did not change by much. Figure 4.5 shows our results for the case when we fix $\sigma_h$ (plot (A)) and when we vary its value (plot (B)). In both case the causal strengths for the respective cues show almost indistinguishable behavior confirming that either approach is appropriate.

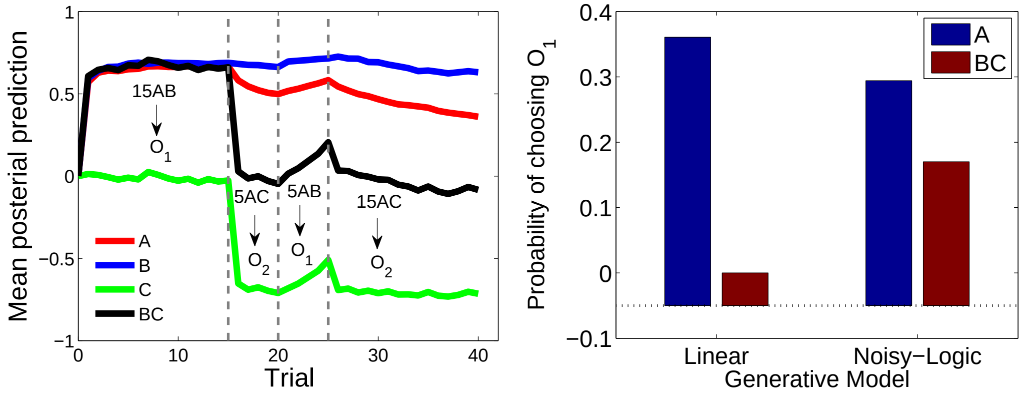

Figure 4.3: Highlighting results. Left, learning curve of causal strength when using *linear-sum* model. The highlighting effect is revealed by the difference between Cue A and BC predictions. With the increase of the number of particles, the highlighting effect will reduce. Right, comparison of highlighting effect between *linear-sum* model and noisy-logic model. See section 4.

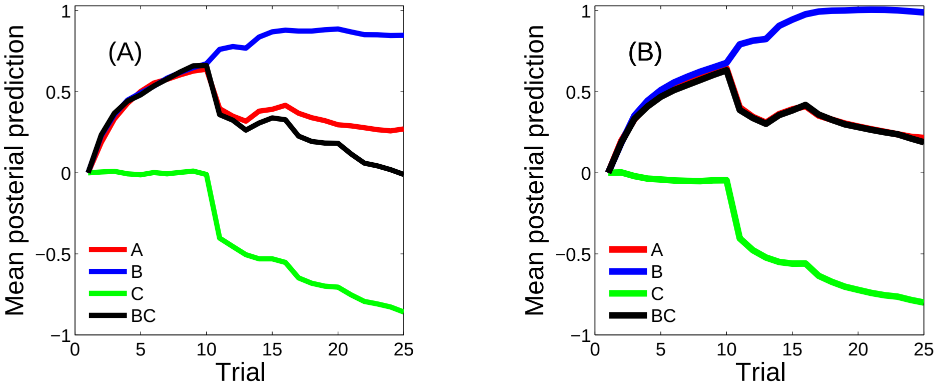

Figure 4.4: Highlighting results using 1000 (plot A) and 18,000 (plot B) particles. We can see that the highlighting effect vanishes when a very large number of particles is used as shown in Figure (B).

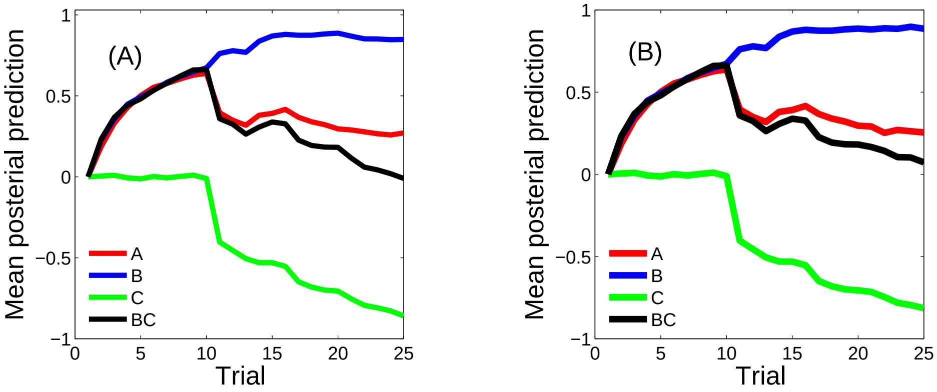

Figure 4.5: Highlighting results using a fixed likelihood variance value (plot A) and varying the variance in the range [0.01, 1] (plot B). We can see that the highlighting effect is clearly present in both cases (plots A and B) as reflected in the difference between the causal strengths of cues A and BC and the overall trends are about the same.

## 4.5 Conclusions

We propose that particle filter simulation could be a good candidate to mimic the limitations of computational capacity (in particular, working memory resources) in the human cognitive system. Accordingly, particle filtering may serve as a computationally realistic approximation to rational inference. This model of approximate inference is able to explain the highlighting effect and is a significant step forward in our understanding of human performance on causal reasoning.

# CHAPTER 5

## Conclusions and Future Work

The Bayesian theory of sequential causal learning described in the present work provides a unified explanation of important learning phenomena, using the implementation of a particle filter as an approximation to exact statistical inference. In particular, the theory accounts for influences of pretraining on subsequent learning as well as influences of posttraining on previous learning, with completely different stimuli. The key assumption is that learners have available multiple generative models, each reflecting a different integration rule for combining the influence of multiple causes. In particular, when the outcome is a continuous variable, humans have tacit knowledge that multiple effects may have an additive 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 continues to favor the most successful model during subsequent learning with different cues. In other situations, the learner uses both models to perform causal learning, but is able to retrospectively re-evaluate the estimations from different models when extra information about integration rules is provided by post-training with different cues. This post-training effect can be explained by model averaging.

Future work will entail extending our model to more complex causal networks that include multiple causes and multiple effects. In this case, rather than having one binary outcome $O$ (=0 or 1), we will have two binary outcomes, $E_1$(=0 or 1) and $E_2$(=0 or 1). For each cue, A, B and C, the corresponding causal strength is now associated with each respective binary outcome. To make the associations more transparent, we adopt the following notation: For a given cue, X, its causal strength with respect to an outcome $E_i$ will be denoted as $\omega_X^{E_i}$, where $i = 1, 2$. For example, the causal strength for cue B associated with the outcome $E_1$ would be denoted as $\omega_B^{E_1}$. To indicate the value that a particular outcome assumes, we use superscripts indicating a value of 1 or 0, e.g., if $E_1 = 0$, we denote this as $E_1^0$ and if $E_1 = 1$, then we write it as $E_1^1$.

In Figure 5.1 we show a plot of the causal network structure for three cues (A, B, and C), two binary outcomes ($E_1$ and $E_2$) and respective causal strengths $\omega_X^{E_i}$.

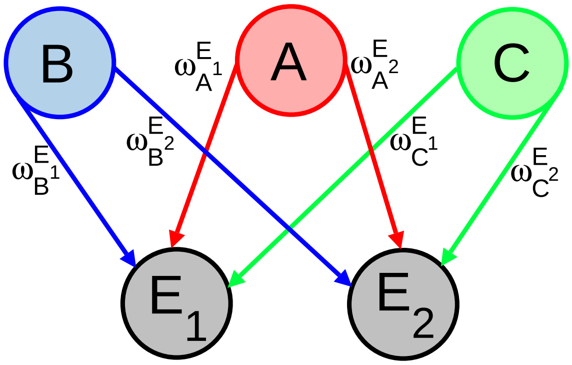

Figure 5.1: Causal network for multiple cues and multiple outcomes. The top nodes represent the three cues A, B and C, and the bottom nodes the two binary outcomes, $\mathrm{E}_1$ and $\mathrm{E}_2$. Arrows correspond to the respective associations between cues and outcomes and the respective causal strengths are indicated by the variables $\omega_X^{E_i}$, where $i = 1, 2$ and $X = \{A, B, C\}$.

The experimental design that we will explore for this framework is illustrated in Table 5.1 and can be regarded as a generalized version of that given in Table 4.1. To indicate the presence or absence of a particular cue, we use a value of 1 or 0 respectively as given by the numbers in parentheses in Table 5.1.

Table 5.1: Highlighting Design based on Daw, Courville, & Dayan, (2007)

| **Highlighting** | **No. of Trials** | **A** | **B** | **C** | $\mathbf{E}_1, \mathbf{E}_2$ |
|---|---|---|---|---|---|
| *Phase I.* | 15× | 1 | 1 | 0 | 1, 0 |
| | 5× | 1 | 0 | 1 | 0, 1 |
| *Phase II.* | 5× | 1 | 1 | 0 | 1, 0 |
| | 15× | 1 | 0 | 1 | 0, 1 |

$$\textbf{Test} \begin{cases} \mathbf{A} \rightarrow ?\mathbf{E}_1 \text{ or } \mathbf{A} \rightarrow ?\mathbf{E}_2 \\ \mathbf{BC} \rightarrow ?\mathbf{E}_1 \text{ or } \mathbf{BC} \rightarrow ?\mathbf{E}_2 \end{cases}$$

The implementation of this causal network is based on the mathematical principles discussed in Section 4.3. However, a key component of our future generalization is that we assume that the likelihood probability of the observed data $P(E_1, E_2|A, B, C, \omega_A^{E_2}, \omega_B^{E_1}, \omega_B^{E_2}, \omega_C^{E_1}, \omega_C^{E_2})$ can be factored according to:

$$\begin{aligned} P\left(E_1, E_2|A, B, C, \omega_A^{E_1}, \omega_A^{E_2}, \omega_B^{E_1}, \omega_B^{E_2}, \omega_C^{E_1}, \omega_C^{E_2}\right) &= P\left(E_1|A, B, C, \omega_A^{E_1}, \omega_B^{E_1}, \omega_C^{E_1}\right) \\ &\quad \times P\left(E_2|A, B, C, \omega_A^{E_2}, \omega_B^{E_2}, \omega_C^{E_2}\right) \end{aligned} \tag{5.1}$$

where $E_1, E_2, A, B, C \in \{0, 1\}$.

A natural extension of this work will be to test our approach on other conditioning paradigms such as upwards and downwards unblocking, second-order conditioning, and conditioning inhibition, and to integrate the collaborative interaction of our senses in the overall learning process. At the present, stimuli are modeled as inputs from one sense alone (e.g., visual or auditory). However, a more comprehensive approach will be to incorporate the participation of the other senses (even if their involvement is highly suppressed by a dominating sense) in the respective causal learning task.

# References

[1] Arnheim, R. (1969). Visual Thinking. Berkeley: University of California Press.

[2] Allan, L. G. (1980). Intuitive Theories of Events and Effects of Base-Rate Information on Prediction. *Journal of Personality and Social Psychology*, 35, 303-314.

[3] Bayes, T. R. (1763). Essay Towards Solving a Problem in the Doctrine of Chances. *Phil. Trans. Roy. Soc. Lond.*, Vol. 53, pp. 370418, 1763. Reprinted in Biometrika, vol. 45, 1958.

[4] 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*. Vol. 11. No. 2. pp 238-249.

[5] Beckers, T., Miller, R. R., De Houwer, J., & Urushihara, K. (2006). Reasoning Rats: Forward Blocking in Pavlovian Animal Conditioningis Sensitive to Constraints of Causal Inference. *Journal of Experimental Psychology*. General. Vol. 135. No. 1. pp 92-102. 2006

[6] Bergman, N. (1999). Recursive Bayesian Estimation: Navigation and Tracking Applications. Ph.D. Thesis, Linköping University, 1999. Dissertations No. 579

[7] Bernardo, J. M., & Smith, A. F. M. (1998). Bayesian Theory, 2nd ed., New York: John Wiley & Sons Inc.

[8] Blaisdell, A. P., Sawa, K., Leising, K. J., & Waldmann, M. R. (2006). Causal Reasoning in Rats. *Science*, 311(5763), 1020-1022.

[9] Bolstad, W. M. (2004), Introduction to Bayesian Statistics, Hoboken, NJ: John Wiley & Sons Inc.

[10] Buehner, M. J., Cheng, P. W., & Clifford, D. (2003). From Covariation to Causation: A Test of the Assumption of Causal Power. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29, 1119-1140.

[11] Chen, Z. (2003). Bayesian filtering: From Kalman Filters to Particle Filters, and Beyond. Technical Report, McMaster University.

[12] Cheng, P. W. (1997). From Covariation to Causation: A Causal Power Theory. *Psychological Review*, 104, 367405.

[13] 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.

[14] Chickering, D. M. (2002). Optimal Structure Identification with Greedy Search. *Journal of Machine Learning Research*, 3, 507-554.

[15] Crisan, D. (1999). Non-Linear Filtering using Branching and Interacting Particle Systems. *Markov Processes Related Fields*, Vol. 5, No. 3, pp. 293319.

[16] 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.

[17] Danks, D. (2005). Causal Learning from Observations and Manipulations, in Lovett, M., Shah, P. eds. Thinking with Data, Lawrence Erlbaum Associates.

[18] Daw, N. & Courville, A. (2008). The Pigeon as Particle Filter. In *Advances in Neural Information Processing Systems* 20, Cambridge, MA: MIT Press.

[19] Daw, N., Courville, A. C., & Dayan, P. (2007). Semi-rational Models of Conditioning: The Case of Trial Order. In M. Oaksford and N. Chater (Eds.). *The probabilistic mind: Prospects for rational models of cognition*. Oxford: Oxford University Press.

[20] 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.

[21] De Houwer, J., Beckers, T., & Glautier, S. (2002). Outcome and Cue Properties Modulate Blocking. *The Quarterly Journal of Experimental Psychologyl A*, 55A, 965-985.

[22] De Houwer, J., Beckers, T., & Vandorpe, S. (2005). Evidence for the Role of Higher Order Reasoning Processes in Cue Competition and Other Learning Phenomena. *Learning & Behavior*, 33(2), 239-249(11).

[23] Dickinson, A., Shanks, D., & Evenden, J. (1984). Judgement of Act-Outcome Contingency: The Role of Selective Attribution. *The Quarterly Journal of Experimental Psychology: Human Experimental Phycology*, 36A, 29-50.

[24] Doucet, A., de Freitas, N., & Gordon, N. J. (2001) *Sequential Monte Carlo Methods in Practice*, Springer-Verlag, New York, NY, 2001.

[25] Efron, B. (1982). The Jackknife, the Bootstrap, and Other Resampling Plans. Society for Industrial and Applied Mathematics, Philadelphia.

[26] Erven, T. V. , Grnwald, P. D. & Rooij, S. de. (2007). Catching up Faster in Bayesian Model Selection and Model Averaging. *Advances in Neural Information Processing Systems* (NIPS), 20.

[27] Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). Bayesian Data Analysis, (Texts in Statistical Science) (2 ed.). Chapman & Hall/CRC.

[28] Green, P. J. (1995). Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination, *Biometrika*, Vol. 82, pp. 711732.

[29] Glymour, C. (2001). The Mind's Arrows: Bayes Nets and Causal Graphical Models in Psychology. Cambridge, MA: MIT Press.

[30] 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.

[31] 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 Relations from Patterns of Variation and Covariation. *Developmental Psychology*, 37, 620-629.

[32] Gordon, N. J., Salmond, D. J., & Smith, A. F. M. (1993). A Novel Approach to Nonlinear/Non-Gaussian Bayesian State Estimation. In IEE Proceedings on Radar and Signal Processing, Vol. 140, pp. 107-113.

[33] Griffiths, T. L., & Tenenbaum, J. B., (2005). Structure and Strength in Causal Induction. *Cognitive Psychology*, 51, 334-384.

[34] Hammersley, J. M., & Handscomb, D. C. (1964). Monte Carlo Methods. London: Methuen & Co Ltd.

[35] 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.

[36] Isard, M., & Blake, A. (1998). CONDENSATION: Conditional Density Propagation for Visual Tracking. *Int. J. Comput. Vis.*, Vol. 29, No. 1, pp. 528.

[37] Jaynes, E. T. (2003). Probability Theory: The Logic of Science (Vol 1). Cambridge University Press

[38] Jazwinski, A. H. (1970). Stochastic Processes and Filltering Theory, Volume 64 of *Mathematics in Science and Engineering*. Academic Press, New York.

[39] Jeffreys, H. (1935). Some Tests of Significance, Treated by the Theory of Probability, *Proceedings of the Cambridge Philosophical Society*, Vol. 31, Issue 02, pp. 203-222

[40] Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME-Journal of Basic Engineering*, 82:35-45

[41] Kamin, L.J. (1969). Predictability, Surprise, Attention, and Conditioning. In B. A. Campbell & R. M. Church (Eds.), *Punishment and Aversive Behavior*. New York: Appleton-Century-Crofts, 279-296.

[42] Kanazawa, K., Koller, D., & Russel, S. (1995). Stochastic Simulation Algorithms for Dynamic Probabilistic Networks, in Proc. 11th Conf. UAI, pp. 346351.

[43] Kass, R. E. & Raftery, A. (1995). Bayesian Factors. *Journal of the American Statistical Association*, Vol. 90, No. 430, pp. 773-795

[44] King, G., Tomz, M., & Wittenberg, J. (2000). Making the most of statistical analyses: Improving Interpretation and Presentation. *American Journal of Political Science*, Vol. 44, No. 2, pp. 347-361.

[45] Kong, A., Liu, J. S., & Wong, W. H. (1994). Sequential Imputations and Bayesian Missing Data Problems, *J. Amer. Statist. Assoc.*, Vol. 89, pp. 278288.

[46] Kramer, S. C. & Sorenson, H. W. (1988). Recursive Bayesian Estimation using Piece-Wise Constant Approximations. *Automatica* (Journal of IFAC), Vol.24 No.6, pp.789-801

[47] Kruschke, J. K. (2001). Cue Competition in Function Learning: Blocking and Highlighting. Presented at the 3rd International Conference on Memory, July 2001, Valencia, Spain.

[48] Kruschke, J. K. (2006). Locally Bayesian Learning with Applications to Retrospective Revaluation and Highlighting. *Psychol Rev*, Vol. 113, No. 4, pp. 677-699

[49] Kruschke, J. K. (2008). Bayesian Approaches to Associative Learning: From Passive to Active Learning. *Learning & Behavior*, 36(3), 210-226.

[50] Lagnado, D. A., & Sloman, S. A. (2004). The Advantage of Timely Intervention. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 30, 856-876.

[51] Laplace, S-P de. (1812). Théorie Analytique des Probabilités.

[52] Levy, R., Reali, F., & Griffths, T. L. (2009). Modeling the Effects of Memory on Human Online Sentence Processing with Particle Filters. Advances in Neural Information Processing Systems 21.

[53] Liljeholm, M., & Cheng, P. W. (2007). When is a Cause the "Same"? Coherent Generalization Across Contexts. *Psychological Science*, 18, 1014-1021

[54] Liu, J. S. (2001). Monte Carlo Strategies in Scientific Computing, Springer-Verlag, New York, NY.

[55] Liu, J. S. & Chen, R. (1998). Sequential Monte Carlo Methods for Dynamical Systems. *J. Amer. Statist. Assoc.*, Vol. 93, pp. 10321044.

[56] Lovibond, P.F. (2003). Causal Beliefs and Conditioned Responses: Retrospective Revaluation Induced by Experience and by Instruction.*Journal of Experimental Psychology: Learning, Memory, & Cognition*, 29, 97-106.

[57] Linhart, H. & Zucchini, W. (1986). Model Selection, New York, NY: John Wiley & Sons, Inc.

[58] Lu, H., Weiden, M., & Yuille, A. L. (2009). Modeling the Spacing Effect in Sequential Category Learning. Advances in Neural Information Processing Systems (in press).

[59] Lu, H., et al. (2006). Modeling Causal Learning Using Bayesian Generic Priors on Generative and Preventive Powers. In R. Sun & N. Miyake (Eds.), *Proceedings of the Twenty-eighth Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum

[60] 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.

[61] Lu, H., Yuille, A., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian Generic Priors for Causal Learning. Psychological Review, 115(4), 955-984.

[62] Lu, H., Rojas, R. R., Beckers, T., & Yuille, A. (2008). Sequential Causal Learning in Humans and Rats. Proceedings of the Twenty-ninth Annual Conference of the Cognitive Science Society.

[63] 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.

[64] MacKay, D. J. C. (1999). Introduction to Monte Carlo Methods. *Learning in Graphical Models*, Jordan, M. I. Ed., MIT Press, Cambridge, MA.

[65] Malakoff, D. M. (1999). Bayes Offers a 'New' Way to Make Sense of Numbers, *Science*, Vol. 286, pp. 14601464.

[66] Marshall, A. (1956). The use of Multi-Stage Sampling Schemes in Monte Carlo Computations. In Symposium on Monte Carlo Methods, M. Meyer Ed. New York: Wiley, pp. 123140.

[67] Metropolis, N., & Ulam, S. (1949). The Monte Carlo Method. *J. Amer. Statist. Assoc.*, Vol. 44, No. 247, pp. 335-341

[68] O'Hagan, A. (1944). Kendall's Advanced Theory of Statistics, Vol. 2B, Bayesian Inference, Eward Arnold, London.

[69] O'Hagan, A. (2004). Bayesian Statistics and Quality Modelling in the Agro-Food Production Chain. In M. A. J. S. van Boekel, A. Stein &, A. H. C. van Bruggen (Eds.), *Bayesian Statistics: Principles and Benefits*, (pp. 31-45). Norwell, MA: Kluwer Academic Publishers.

[70] Pearl, J. (2000). Causality: Models, Reasoning, and Inference. Cambridge, England: Cambridge University Press.

[71] Press, S. J. (2003). Subjective and Objective Bayesian Statistics: Principles, Models, and Applications (2nd ed.), Hoboken, NJ: John Wiley & Sons Inc.

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

[73] 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.

[74] Ripley, B. D. (1987). Stochastic Simulation. New York: Wiley & Sons.

[75] Robert, C. P. (2001). The Bayesian Choice: A Decision-Theoretic Motivation (2nd ed.), New York: Springer.

[76] Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys, New York: John Wiley & Sons, Inc.

[77] Rubin, D. B. (1987). Comment on 'The Calculation of Posterior Distributions by Data Augmentation' by M. A. Tanner and W. H. Wong, *J. Amer. Statist. Assoc.*, Vol. 82, pp. 543546.

[78] Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2006) A More Rational Model of Categorization. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, Mahwah, NJ, Erlbaum.

[79] Schulz, L. E., & Gopnik, A. (2004). Causal Learning Across Domains. *Developmental Psychology*, 40, 162-176.

[80] Shanks, D. R. (1985). Forward and Backward Blocking in Human Contingency Judgement. *Quarterly Journal of Experimental Psychology*, 37, 1-21.

[81] 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.

[82] Shanks, D. R. (2004). Judging Covariation and Causation. In D. Koehler & N. Harvey (Eds.), *Blackwell Handbook of Judgment and Decision Making*. Oxford, UK: Blackwell.

[83] Sloman, S. (2005). Causal Models: How People Think about the World and Its Alternatives. Oxford University Press, Incorporated.

[84] Spirtes, P., Glymour, C., & Scheines, R. (1993). Causation, Prediction, and Search. Spinger Lecture Notes in Statistics. New York: Springer-Verlag.

[85] 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.

[86] Waldmann, M. R. (2007). Combining versus Analyzing Multiple Causes: How Domain Assumptions and Task Context Affect Integration Rules. *Cognitive Science*, 31, 233-256.

[87] Wheeler, D. S., Beckers, T., & Miller, R. R. (2008). The Effect of Subadditive Pretraining on Blocking: Limits on Generalization. *Learning & Behavior*, 36(4) 341-351

[88] Yuille, A. L., & Bulthoff, H. B. (1996) Bayesian Decision Theory and Psychophysics. In Perception as Bayesian Inference. Eds. D. Knill and W. Richards. Cambridge University Press.

[89] Yuille, A. L. (2004). The Rescorla-Wagner Algorithm and Maximum Likelihood Estimation of Causal Parameters. *Advances in Neural Information Processing Systems* (NIPS), Vol. 17. Cambridge, MA: MIT Press. A Bradford Book. Eds. L.K. Saul, Y. Weiss and L. Bottou. Pages 1585-1592. Proceedings of NIPS 2004.

[90] Yuille, A. L. (2005). Augmented Rescorla-Wagner and Maximum Likelihood Estimation. *Advances in Neural Information Processing Systems* (NIPS), Vol. 18. Cambridge, MA: MIT Press. A Bradford Book. Eds. Y. Weiss, B. Schoelkopf and J. Platt. Pages 1561-1568. Proceedings of NIPS 2005.

[91] Yuille, A. L., & Lu, H. (2007). The Noisy-Logical Distribution and its Application to Causal Inference. *Advances in Neural Information Processing Systems*(NIPS), Vol. 20. Cambridge, MA: MIT Press.

[92] 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.

[93] West, M. (1981). Robust Sequential Approximate Bayesian Estimation. *Journal of the Royal Statistical Society, Series B*, Vol. 43, pp. 157166

[94] Wood, F., & Griffiths, T. L. (2007). Particle Filtering for Nonparametric Bayesian Matrix Factorization. *Advances in Neural Information Processing Systems* (NIPS), Vol. 19.

[95] Zucchini, W. (2000). An Introduction to Model Selection. *Journal of Mathematical Psychology*, Vol. 44, No. 1, pp. 41-61