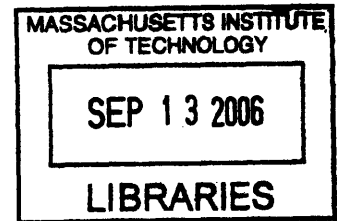


# The Role of Temporal Factors and Prior Knowledge in Causal Learning and Judgment

by

Tevye Rachelson Krynski

B.S. Computer Science  
Cornell University, 1996



SUBMITTED TO THE DEPARTMENT OF BRAIN AND COGNITIVE SCIENCES IN  
PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY  
AT THE  
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

SEPTEMBER, 2006

© 2006 Tevye Rachelson Krynski. All rights reserved.

The author hereby grants to MIT permission to reproduce and to distribute publicly paper and electronic copies of this thesis document in whole or in part in any medium now known or hereafter created.

Signature of Author: \_\_\_\_\_  
Department of Brain and Cognitive Sciences  
August 29<sup>th</sup>, 2006

Certified by: \_\_\_\_\_  
Joshua B. Tenenbaum  
Paul E. Newton Assistant Professor of Computational Cognitive Science  
Thesis Supervisor

Accepted by: \_\_\_\_\_  
Matt Wilson  
Professor of Neurobiology  
Chairman, Committee for Graduate Students

# THE ROLE OF TEMPORAL FACTORS AND PRIOR KNOWLEDGE IN CAUSAL LEARNING AND JUDGMENT

by

TEVYE RACHELSON KRYNSKI

Submitted to the Department of Brain and Cognitive Sciences  
on August 22, 2006, in partial fulfillment of the  
requirements for the degree of Doctor of Philosophy in  
Cognitive Science

## ABSTRACT

Causal relationships are all around us: wine causes stains; matches cause flames; foods cause allergic reactions. Next to language, it is hard to imagine a cognitive process more indicative of human intelligence than causal reasoning. To understand how people accomplish these feats, two major questions must be addressed: how do people acquire knowledge of causal relationships (causal learning), and how do people use that knowledge to make predictions and draw inferences (causal judgment)?

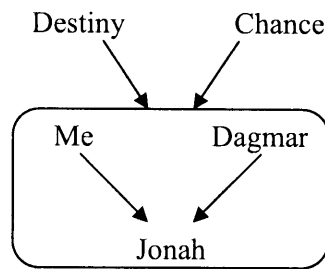
The first part of this thesis is concerned with causal learning, and draws on the foundation of Bayesian inferential frameworks (e.g., Tenenbaum, Griffiths, & Kemp, 2006) to explain how observable data can be used to infer causal relationships between events. I will argue that rapid causal learning from small samples can be understood as rational inference over a representation of causality that includes a temporal delay between cause and effect. Experimentally, I show that people learn causal relationships faster when the temporal delay between cause and effect is less variable, just as is predicted by a rational statistical model of event causation. I argue that people's tendency to learn better from short delays is an artifact of the fact that short delays are inherently less variable.

The second part of this thesis is concerned with causal judgment, and draws on the foundation of knowledge-based Bayesian networks to show that it is often more rational to make judgments using causal frameworks than purely statistical frameworks. Deviations from traditional norms of judgment, such as "base-rate neglect" (Tversky & Kahneman, 1974), can be explained in terms of a mismatch between the statistics given to people and the causal models they intuitively construct to support probabilistic reasoning. Experimentally, I provide evidence that base-rate neglect may be an artifact of applying causal reasoning to purely statistical problems. Six experiments show that when a clear mapping can be established from given statistics to the parameters of an intuitive causal model, people are more likely to use the statistics appropriately, and that when the classical and causal Bayesian norms differ in their prescriptions, people's judgments are more consistent with causal Bayesian norms.

Thesis Supervisor: Joshua B. Tenenbaum

Title: Paul E. Newton Assistant Professor of Computational Cognitive Science

*For my family*



# Acknowledgements

It was towards the end of my undergraduate years that I first became aware of the field called cognitive science. At that time, I was struggling to make sense of philosophy of mind and the hard problem of consciousness, quantum physics, and artificial intelligence. On a thanksgiving trip to New York in 1997, I happened to pick up a book at the airport that changed my life: *The Language Instinct*, by MIT professor Steven Pinker. Within its covers was a treasure trove of scientific studies and theoretical claims that examined in exquisite detail how the mind could be understood as an information processing device. I knew then that I wanted to explore that world. I thank Steve for writing so lucidly and entertainingly about our field.

My California comrade, Aaron Ross, was a sponge for my musings about philosophy of mind and consciousness, at a time when I didn't even know what cognitive science was. I thank him for his awe at my journey.

The person who encouraged me most to go to graduate school was my grandfather, Boris. He vigorously championed the idea that my education was more important than anything else, and without his uncompromising guidance I might not have had the courage to enter MIT. He deserves a piece of the credit for wherever I end up.

Whitman Richards found in my MIT application a paper I'd written in college on agents, and saw the seeds of something interesting. I owe him a deep debt for opening the door to MIT and advising me to work with Josh Tenenbaum.

I am in awe of my advisor, Josh Tenenbaum, a man who somehow juggles his obligations to family, graduate students, journals, conferences, grants, and the department, and still has a smile on his face. He has been an absolutely wonderful mentor, and my compass for good

cognitive science in an often bewildering sea of literature. I am immensely grateful for his patience, endurance, wisdom, integrity, and invaluable advice.

I'd like to thank the members of my thesis committee. Tom Griffiths' work inspired much of the first part of this dissertation. He has been a model for superior computational theory and scientific achievement. Molly Potter and Laura Schulz have graciously donated their time to serving on my committee, and have been extremely valuable resources in helping me focus my research ideas.

The cocosci lab has a higher percentage of smart people than any place I've worked, making it very difficult to pretend I know what I'm talking about. I'd like to specifically thank Tom Griffiths, Charles Kemp, Liz Baraff, Noah Goodman, Konrad Koerding, and Vikash Mansinghka for their help and support of my research. My undergraduate research assistants, Brigid Dwyer, Suzanne Luther, Laila Shabir, Sadik Antwi-Boampong, and Stephanie Brenman, provided essential legwork where I couldn't, and helped me talk through ideas that were too hard to think through on my own.

My MIT friends, Neville Sanjana, Emily Hueske, Nathan Wilson & Dana Hunt have been silly, sarcastic, sympathetic and supportive. They have been a constant in a fast-changing life, and I will make sure we get together and reminisce about the good ol' days on a regular basis.

My mother, Rachel, raised me to be totally secure in the idea that I was as smart as they came (but not smarter). Her encouragement of my intellectual growth as a young child cannot be underestimated as a distant cause of my current activity.

My father, Peter, has provided unconditional support of my goals, ideas, and activities. He has shown me that it is possible to keep getting better with age.

My enchanting wife, Dagmar, has taught me as much about human emotion as my studies have about human intelligence. Despite accompanying me on an intensely grueling journey, she has never swayed in her support for the pursuit of my PhD and our uncertain future, because it was important to her that I love what I do. I am committed to giving her the freedom to explore her own muses as we proceed to the next phase of life.

# Contents

Acknowledgements.....	4
Contents .....	7
Part I. The role of temporal factors in causal learning.....	12
1 Introduction.....	13
1.1 Causal learning.....	16
1.2 The importance of temporal delay .....	18
1.2.1 Previous studies of the role of temporal delay in causal learning.....	24
1.2.2 Previous studies of the role of temporal delay in associative learning.....	27
1.3 Previous models of causal learning.....	29
1.3.1 Causal power theory (White, 1995).....	30
1.3.2 The probabilistic dependency view .....	30
1.3.3 The mechanism view (Ahn & Kalish, 2000).....	36
1.3.4 Causal grammars (Tenenbaum, Griffiths, & Niyogi, in press).....	37
1.3.5 Theory-based causal induction (Tenenbaum & Griffiths, 2003).....	38
1.3.6 Intervention vs. temporal order (Langado & Sloman, 2004).....	39
1.4 Previous models incorporating temporal delay.....	40
1.4.1 Rate models.....	40
1.4.2 Delay models .....	43
1.4.3 The co-occurrence contingency model .....	45
2 A new framework for event-based causal learning.....	47
2.1 Bayesian inference for causal learning .....	49

2.1.1	The likelihood ratio.....	52
2.2	Causal attribution during learning.....	53
2.2.1	The delay distribution.....	55
2.3	Causal discovery: learning without priors.....	56
2.4	The short delay advantage: Why short delays result in faster learning.....	58
3	Experiments in causal learning.....	63
3.1	Experiment 1: causal discovery.....	67
3.1.1	Method.....	67
3.1.2	Results.....	69
3.1.3	Discussion.....	71
3.2	Experiment 2: causal classification after training.....	73
3.2.1	Method.....	76
3.2.2	Results.....	79
3.2.3	Discussion.....	86
3.3	Experiment 3: causal attribution.....	87
3.3.1	Method.....	89
3.3.2	Results.....	92
3.3.3	Discussion.....	97
3.4	Experiment 4: rates and strengths.....	97
3.4.1	Method.....	98
3.4.2	Results.....	99
3.4.3	Discussion.....	100
4	General discussion.....	101



4.1	The event-based framework: beyond the chain model .....	103
4.2	Object vs. Classes .....	103
4.3	Solving to the mechanism paradox .....	104
4.4	Anecdotal evidence as rational statistical inference .....	105
	Conclusion .....	106
	References.....	107
	Part II. The role of prior knowledge in causal judgment .....	112
5	Introduction to judgment under uncertainty.....	113
5.1	Statistical frameworks for judgment under uncertainty.....	114
6	A causal Bayesian framework for judgment under uncertainty.....	124
6.1	Causal Bayesian inference as a rational method of judgment under uncertainty .....	125
6.2	Causal Bayesian inference as a new normative standard.....	129
7	Experiments .....	133
7.1	Experiment 1 .....	136
7.1.1	Method.....	139
7.1.2	Results.....	140
7.1.3	Discussion .....	142
7.2	Experiment 2.....	143
7.2.1	Method .....	144
7.2.2	Results.....	146
7.2.3	Discussion .....	147
7.3	Experiment 3.....	148
7.3.1	Method.....	149

7.3.2	Results.....	151
7.3.3	Discussion.....	152
7.4	Experiment 4.....	152
7.4.1	Method.....	156
7.4.2	Results.....	158
7.4.3	Discussion.....	159
7.5	Experiment 5.....	160
7.5.1	Method.....	164
7.5.2	Results.....	167
7.5.3	Discussion.....	170
7.6	Experiment 6.....	172
7.7	Method.....	174
7.8	Results.....	175
7.9	Discussion.....	177
8	General Discussion.....	177
8.1	Relation to the heuristics and biases view.....	179
8.2	Relation to the natural frequency hypothesis.....	180
8.3	Explaining other apparent errors of judgment under uncertainty.....	181
8.4	Learning Structure from Statistical Data.....	183
8.5	Deterministic Mechanisms and Randomly Occurring Causes.....	184
8.6	Making Statistics Easy.....	185
	Conclusion.....	188
	References.....	189

Appendix A: A hierarchical generative model for event-based causal learning .....	194
8.6.1 Specification of the generative model.....	195
8.6.2 Inference of causal powers.....	196

## Part I. The role of temporal factors in causal learning

In which the causal learning can be explained as Bayesian inference over causal relations  
between objects from event-based data.

# 1 Introduction

Causal relations are all around us. Spills cause stains; guitar strings cause sounds; buttons cause changes in machines; medications cause side-effects. Knowing about causal relations is crucial for successfully predicting the consequences of observed events, and of one's own potential actions. People routinely make predictions about complex systems they've never encountered before (e.g., this restaurant is well-reviewed; it will probably fill up early), and draw inferences from scant evidence (e.g., this restaurant is well-reviewed; the chef is probably famous). Causal reasoning drives technical innovation, enabling engineers to invent better mousetraps and clinicians to diagnose disease. It enables us to make predictions about the future, and inferences about the past. Part I of this dissertation addresses the question of how people acquire knowledge of which causal relations exist, and how they leverage that knowledge in learning new causal relations.

Causal learning can be analyzed most generally as the inference of the existence and/or strength of a causal relationship between two things. Many factors influence the degree to which people will believe a causal relationship exists. These include the contingency of one state on another, the temporal contiguity of two events, the spatial contiguity of two objects, the rates of occurrence of events, the plausibility of a relationship between two classes, and the overall quantity of available data. Three branches of psychological inquiry have produced evidence for various aspects of causal learning. First, the associative learning literature (Gallistel, 1990; Gibbon et al., 1977; Pavlov, 1927; Rescorla & Wagner, 1972; Shanks, 1995; Skinner, 1938) has extensively studied the effect of temporal contiguity, repetition, domain, and contingency on the speed of acquisition and subsequent strength of associations between cues and outcomes.

Second, the causal learning literature (Ahn, Kalish, Medin, & Gelman, 1995; Cheng, 1997; Griffiths & Tenenbaum 2005; Shanks, Pearson, & Dickinson, 1989; Waldmann, 1996) has focused on learning causal relationships, and their subsequent strengths, in many cases distinguishing these from the predictions of associative accounts (Cheng, 1997; Waldmann & Holyoak, 1992). Third, the literature on intuitive theories (Carrie, 1987; Griffiths, 2005; Gopnik & Glymour, 2002; Gopnik, Sobel, Schulz, & Glymour, 2001; Tenenbaum & Griffiths, 2003; Tenenbaum, Griffiths, & Niyogi, in press) has recently made use of the tools of causal learning as a computational grounding of theory acquisition.

In cognitive science there has been a recent resurgence of research and interest in how people learn about causal relationships (e.g., Cheng, 1997; Gopnik et al., 2004; Griffiths, 2005; Griffiths & Tenenbaum, 2005; Novick & Cheng, 2004). Typically, these approaches have assumed that a causal relationship is tantamount to a contingency of one variable on another (which is not conditional on any third variable). However, people tend to know much more about specific causal relations than mere contingency. With this additional knowledge, there are many ways more ways to detect causation than just via contingency, just as there are many ways to detect a house (by seeing a front door, windows, street number, etc.). For instance, when a glass of wine spills and creates a stain, it is not just the contingency of the stain on the spill that compels us to infer a causal connection between the two. It is also that the color of the stain matches the wine, the shape of the stain is consistent with liquid spilling, the fact that the wine is missing from the glass, that the table is wet, and that the tablecloth now smells like wine. We are far from developing a formal account of how people know which of these cues are relevant, but it is clear that additional knowledge about causal processes influences causal learning.

I argue for an approach that is in principle capable of naturally accommodating such additional knowledge about causal processes, and I take one step in the direction of modeling that additional knowledge. The approach relies on a generative model of causal processes, with causal learning resulting from statistical inference over hypothesized causal relations. Provided one knows what kinds of evidence causal relations could produce, and with what probability (the probabilistic generative model), one can work backwards from the evidence to infer which causal relations exist (using Bayesian inference). This inference forms the essence of what we will mean by “causal learning”. In formalizing this inference, I develop a rational Bayesian framework for causal learning from event-based evidence. I will argue that rapid causal learning from small samples can be understood as rational inference over a representation of causality that includes a temporal delay between cause and effect.

Four experiments will be presented to demonstrate that people’s judgments from dynamic event data are better accounted for by this framework than by previous proposals. The experiments were all conducted using computer animation depicting explosions occurring at a popular fishing pond. In Experiment 1 the explosions are said to be mysterious, and the participants must discover that one of the fishermen is using explosive lures. In Experiment 2, the participants are trained that some of the fishermen are using explosive lures and they are tasked with determining which of the fishermen in the display are using explosive lures. In Experiment 3, participants must determine which of two people caused a given explosion. In Experiment 4, participants must determine whether a particular fisherman is using explosive lures when a scuba diver is known to be causing explosions in the background.

Part I of the dissertation is organized into four sections. Section 1 reviews the existing literature on causal and associative learning, focusing on studies and models that incorporate

temporal delay. Section 2 presents a new computational framework for event-based causal learning, which provides a rational statistical basis for the role of temporal delay in causal learning. Section 3 presents four experiments that test the predictions of the model against human judgments, distinguishing between the current model and previous proposals. Section 4 discusses possible extensions of the framework and relations to other aspects of the causal learning literature.

## 1.1 Causal learning

Ever since Hume (1739/1978), scientists and philosophers alike have been guided by the notion that when one stimulus quickly and reliably follows another, a causal process is often at work (e.g., eating causes satiation). But they have been equally troubled by the lack of logical principles justifying this intuitive inferential rule. First, speed is not necessary. A causal process need not be quick; indeed, sometimes a long time must pass before one should infer a causal relationship (e.g., does the manipulation of interest rates by the Federal Reserve Bank cause economic growth? Does a high-fat diet cause heart disease?). Second, reliability is not sufficient: two stimuli can reliably follow each other without being causally connected (e.g., eating reliably follows cooking, but cooking does not cause eating; satiation reliably follows hunger, but hunger does not cause satiation).

Given that these cues are either unnecessary or insufficient to infer a causal relationship, two questions arise: (1) is there a rational process by which true causal relations can be inferred from the evidence available, and (2) how well do people's inferences conform to this process? Much of the philosophical and psychological literature has been concerned with the second cue: reliability. In particular, two questions have been addressed. First, given a reliable relationship, how can one distinguish true causal relations from non-causal (but real) contingencies? For



example, how can one distinguish cooking followed by eating (mere contingency resulting from a common cause) from eating followed by satiation (true causation)? The work in this area has shown that people can distinguish contingencies due to common causes from those due to true causation by controlling for third variables (Gopnik & Glymour, 2001; Waldmann & Hagmayer, 2001) and making interventions (Sloman & Lagnado, 2005, Schulz & Gopnik, in submission). The second question about reliability that has been addressed is how one can determine the existence or strength of a causal relationship from evidence of contingency (e.g., does a particular medication cause a particular side effect? how reliably?). Formal models have been developed for estimating both the strength (or reliability) of a causal relationship (Cheng, 1997) and the evidential support for the existence of a causal relationship (Griffiths & Tenenbaum, 2005).

In contrast to much of the causal learning literature, I will focus not on the cue of reliability, but on the other major intuitive cue to causation: time. I propose a rational approach to inferring the existence of a causal relationship based on the temporal delay between the occurrence time of a candidate cause and the occurrence time of an effect. Although the role of temporal delay has been nearly absent from the causal learning literature, the associative learning literature has amassed a body of evidence suggesting that the delay between cue and outcome is a primary determinant of the rate of acquisition and ultimate strength of a learned association. Associative learning can be fruitfully viewed as causal learning when the cue is a possible cause (or an observable indication of an unobservable possible cause) of the outcome.

The associative learning literature has typically implicated biological constraints to explain why certain temporal delays produce faster learning than others (e.g., the optimal delay may be dependent on the neural plasticity mechanisms of LTP). But at least one finding turns

this idea on its head. In a demonstration dubbed the Garcia effect (Garcia, Ervin, & Koelling, 1966), rats associated nausea not with something they ate seconds earlier, but with something they ate several hours earlier. The Garcia effect obviously has adaptive value, and many have speculated an evolutionarily developed process for making such associations. But if biological processes are capable of associating events separated by long time intervals, then we cannot explain the appeal of short delays in associative learning as an inherent property of neurons. The Garcia effect suggests that if biology can learn from any kind of delay, then attending to short delays in for general kinds of associations may also result from the adaptive value of short delays. We must then ask: what is the rational statistical basis for making inferences from temporal delays, why are shorter temporal delays generally so much more compelling than longer ones, and when (and why) can longer delays be more compelling than shorter ones?

## 1.2 The importance of temporal delay

Most previous accounts of causal learning have implicitly assumed that the primary evidence produced by causal relations is the co-variation of cause and effect. This assumption is valid, however, only when the data are observed at fixed moments in time. In this case, co-variational evidence can be summarized in a contingency table consisting of the number of individuals in which (a) the cause and effect are both present, (b) the cause is present and the effect is absent, (c) the cause is absent and the effect is present, and (d) the cause and effect are both absent (see Figure 1).

	E	$\neg$ E
C	a	b
$\neg$ C	c	d

**Figure 1: Contingency table representing the number of trials in which both C and E are present (a), C is present without E (b), E is present without C (c) and neither C nor E are present (d).**

However, when cause and effect are events that can occur at different points in time, the co-variation of cause and effect is no longer well-defined, because it is not clear what it means for a cause and an effect to occur together. To overcome this problem, early models of associative learning (i.e., classical conditioning) (Pavlov, 1927) assumed that close “temporal contiguity” between cue and outcome was a requirement for an association to be acquired. In fact, until Rescorla's (1967) landmark study, the paradigm for associative learning experiments consisted of an experimental condition in which the cue was closely followed by the outcome, and a control condition in which the cue and outcome were widely separated in time (Gallistel, 2002). The assumption that associations are learned only between closely occurring events is reflected in the format of the data that is presumed to be available to the organism during associative learning: the co-occurrence table. The actual occurrence times of the events are not included in this data, only the number of occasions in which the cue and outcome occur together within a short period of time.

There is nothing inherently wrong with choosing a window of time within which the effect must occur after the cause in order to count as a co-occurrence. However, different causal processes have different timescales, and it is far from clear how one should go about choosing this window (e.g., illness can occur hours, days or even years after exposure to the cause). Although the Garcia et al. (1966) study famously demonstrated that associations between taste and illness can be learned with a delay of several hours, this has generally been viewed as a special case that evolved to handle food-borne illness. But the Garcia effect represents a deeper problem with co-occurrence tables as a rational basis for learning: if there are cases of causation in the world in which the cause and effect are widely separated in time, then a rational learner

ought to be capable of inferring causal relationships that unfold over such intervals. The co-occurrence table format ignores any relationship that does not fall within the prescribed temporal window, and depending on which window one chooses one could end up with vastly different co-occurrence counts. As Shanks (1995) points out, the normative theory of causal learning based on contingency cannot handle variable delays between cause and effect because there is no method for determining the appropriate temporal window within which to count co-occurrences.

Perhaps because of this problem, most causal learning experiments from co-occurrence evidence have focused on static causes and effects that do not occur in time (e.g., inferring whether students with healthier diets have higher test scores). The data are typically single values of the cause and effect variables for each individual, and are provided for multiple individuals, rather than multiple time points. For example, in determining whether a particular drug causes a particular side effect, typically one compares the number of people taking the drug who exhibit the side effect within a fixed amount of time to that of a control group taking a placebo. But information about how soon the side effect occurred after taking the drug, which seems so intuitively important to inferring whether a particular individual's side effect was caused by the drug, is ignored by these methods. In fact, the standard contingency table does not even distinguish between side effects that were present before taking the drug and those that occurred afterwards: it merely records the presence or absence of the variable.

While these static cause and effect variables fit nicely into contingency tables, they represent a rather limited subset of cause-effect relationships. Causation is a dynamic process; causes generate their effects in time, and the time interval between cause and effect is an immensely useful cue to causation. For instance, if one feels lightheaded moments after taking a new medication, this evidence is intuitively much stronger than if one feels lightheaded some

time within the next 3 months after taking a new medication, yet a contingency table examining the side effects of a medication occurring over a 3 month period would not distinguish between these cases. What is needed to account for this intuitive difference is a data format that includes the occurrence times of cause and effect, and a model of causal relations that generates causes and effects in time. Armed with such a model and data format, we can evaluate whether two events are causally connected by evaluating the extent to which the temporal delays between them in the data is representative of the temporal delays predicted by the model.

Most existing approaches to causal learning model causal relations as an increase in the probability of an effect, contingent on a cause. Here, I take a step beyond existing approaches by including an additional dimension in the representation of causal relations: the temporal delay between cause and the effect. Formally, for each causal relation, the model specifies a probability density over the temporal delay between the occurrence times of the cause and the effect. An extended example will illustrate this concept. First, suppose that we somehow know what the exact probability density is. In this case, we can use it to compute the likelihood of observing the cause and effect occurring with various delays between them. For instance, if we know that snake venom causes paralysis in 30-40 seconds, and we model this as a normal distribution with mean 35 and standard deviation 5, we can compute that a 30-second delay between snake venom and paralysis is about 10,000 times more likely than a 60-second delay.

The primary question of interest for causal learning is whether there is in fact a causal relationship between two events. For instance, given the evidence that I was bitten by a snake and that I later experienced paralysis, is the snake capable of causing paralysis? For this question, we can use Bayesian inference to determine the posterior odds of a causal relation existing between two events,  $A$  and  $B$ , written as  $A \rightarrow B$  (see Eq 1):

$$\frac{P(A \rightarrow B | D)}{P(A \not\rightarrow B | D)} = \frac{P(D | A \rightarrow B) P(A \rightarrow B)}{P(D | A \not\rightarrow B) P(A \not\rightarrow B)} \quad \text{Eq 1}$$

This equation states that the posterior odds of a causal relation  $A \rightarrow B$  existing, given some data, is equal to the likelihood ratio of the data given the hypotheses (where the two hypotheses under consideration are that the causal relation exists and that it does not) times the prior odds of the causal relation existing. The prior odds should depend on one's beliefs about the classes of objects in question (e.g., before getting bitten by the snake, what is your belief that it is poisonous?). We will discuss how prior odds can be computed using class models in the general discussion (Section 4). For now, we will assume for simplicity that one has a 50% prior belief that the causal relation exists, allowing us to drop the prior odds from the equation.

To learn using this kind of model, we must explicitly include in the data information about temporal contiguity. Rather than traditional co-occurrence frequencies (e.g., "*I once got a snake bite and paralysis*"), we must include the times that various events occur, (e.g., "*I got a snake bite at 12:15:02, and paralysis at 12:15:35*"). The likelihoods of the data under each hypothesis can be computed using the probability density we specified earlier, plus some additional background probability. The background is necessary because we need to specify a probability of the effect occurring in the absence of the causal relation. Returning to our example, we'll need the probability of getting paralysis in the absence of a snake bite. This background probability is appropriately specified mathematically using a Poisson process. Supposing that paralysis spontaneously happens to one in 100 people once in their lifetime, the rate of paralysis is about once every 10,000 years.

Learning whether a causal relation exists can be accomplished by computing the evidential support for a causal relation, or "causal support" (Griffiths & Tenenbaum, 2005), which is the likelihood ratio of the observed data. The higher the likelihood ratio, the greater the

likelihood is of a causal relation existing. To learn whether the snake is poisonous from a single example, we can compute the likelihood ratio for the observed delay between the snake bite and paralysis. Under the assumed density (a normal distribution with mean of 30 seconds and standard deviation of 5 seconds) and the assumed background probability (once every 10,000 years), the likelihood ratio for a delay of 30 seconds between snake bite and venom is approximately  $10^{10}$ , while for 60 seconds it is approximately  $10^{-2}$ , and for 90 seconds it is 0.

The result that a delay of 90 seconds provides solid evidence against the snake being venomous might strike you as being unintuitive; after all, maybe the venom took longer than usual. It still seems like an extremely suspicious coincidence that one happened to become paralyzed, purely by chance, a mere 90 seconds after being bitten by a snake. We can address this problem by relaxing our assumption that we know the probability density over the delay between cause and effect. Without knowledge of the density, it is impossible to compute the above likelihood ratio, but we can hypothesize several different densities ( $f$ ) and sum their likelihoods, weighting each by our belief that it is the true density (see Eq 2).

$$\frac{P(D | A \rightarrow B)}{P(D | A \not\rightarrow B)} = \sum_f \frac{P(D | A \rightarrow B, f)}{P(D | A \not\rightarrow B)} P(f) \tag{Eq 2}$$

By integrating these likelihoods across a space of hypothesized probability densities, we can arrive at an inference about whether a causal relation really exists, as well as an inference about the probability density that best characterizes the temporal delay. Presumably if we integrate over the right set of hypothesized probability densities, we would get the intuitive result that for a 90 second delay the snake was likely poisonous, but for a delay of 3 years, the snake was likely not poisonous. But which probability densities are right?

One particular family of probability densities is particularly appropriate from a rational perspective for characterizing the temporal delay between cause and effect: the family of gamma

distributions. A gamma distribution with rate parameter  $\lambda$  and shape parameter  $n$  specifies the probability distribution over the temporal delay until the  $n^{\text{th}}$  occurrence in a Poisson process with rate  $\lambda$ . If a causal relation is physically realized as a chain of mechanisms (e.g., snake bite causes insertion of venom which causes venom to reach critical cells which causes paralysis), and if each mechanism operates via a Poisson process with the same rate, then the temporal delay between cause and effect is governed by the gamma distribution.

Although a multi-step Poisson process is clearly an idealized representation of a physical causal system (e.g., intervening mechanisms need not operate at the same rate), this model can naturally account for the primary influence of temporal delay on learning: the finding that short delays result in faster acquisition of associations and causal relations. In our example, by integrating over gamma distributions ranging from 1 second to 90 years, we obtain the result that paralysis after 90 seconds provides a likelihood ratio that the snake was poisonous of  $10^8$ , while getting paralysis after 3 years provides a likelihood ratio of only  $10^2$ . A computer program in Matlab<sup>®</sup> has been developed that implements a more general version of this model for an arbitrary number of causes, an arbitrary number of occurrences for each cause, and an arbitrary number of effect occurrences. The computational model will be presented in Section 2.

### **1.2.1 Previous studies of the role of temporal delay in causal learning**

This thesis is motivated by previous findings that people learn causal relations more rapidly than contingency-based models can account for, and that this rapid learning is facilitated by short delays between cause and effect. A telling example comes from Schulz and Gopnik (2004), who reported that after seeing one example of the experimenter activating a machine by talking to it, most children who were asked to make the machine stop chose to talk to it rather than push a button. Since the children's declared prior beliefs were that talking would not



activate the machine, we cannot appeal to prior knowledge to explain how quickly children learned the causal relation.

A number of causal learning studies provide participants with temporal evidence, but the authors often do not highlight or even seem to recognize this aspect of the experiment. Instead, they analyze the learner's performance using standard contingency models. For instance, in a series of experiments dubbed the "blicket detector" studies (Gopnik et al., 2001), participants were trained that blickets activate a blicket detector, and then were asked to judge whether subsequent objects (blocks) were blickets or not. One task was to infer which of two blocks was a blicket, given evidence that block *A* activated the detector once, block *B* did not activate the detector once, and blocks *A* and *B* together activated the detector twice. Children were able to infer that block *A* was a blicket and block *B* was not, despite having observed only 4 events involving two blocks. Contingency-based learning methods cannot explain this without giving the sample size a "large fictitious multiplier" (Gopnik et al., 2004, p. 21). But arbitrarily multiplying sample sizes would result in learning far too many causal relations from insignificant samples. For instance, if block *A* were cubic and block *B* were cylindrical, using sample size multipliers would lead to the inference that cubic blocks are blickets and cylindrical blocks are not, but this inference is clearly unjustified with such a small sample.

What needs to be explained is not why people are willing to generalize from small samples, but why it seems rational in some cases and not others. For instance, why does it seem justifiable to infer that object *A* has the causal power to activate the machine, but not to infer that *cubes* have the causal power to activate the machine? Our framework accounts for this intuition by inferring causal powers of objects, but not of object classes, from temporal evidence. When an effect occurrence coincides precisely with a cause occurrence, this produces overwhelming

evidence in just one trial for a causal relationship between the cause and effect objects. This is because the probability of the effect occurring due to a background Poisson process at the exact moment the cause occurs is extremely low. The framework only attempts to determine whether the specific object was a cause, and does not generalize to other objects.

Few studies have specifically explored how variations in the delay between cause and effect influence causal learning. An exception, Shanks et al. (1989), investigated how causal strength judgments are influenced by the delay between cause and effect. They showed participants a figure on a computer screen which could be caused to appear by pressing the spacebar. By parametrically varying the delay between the key-press and the appearance of the figure, Shanks et al. found that people's judgments of causal strength were lower with longer delays. Shanks (1995) explains that contingency-based models are ill-equipped to handle this kind of learning, as they required discrete trials. Even if one conceives of trials as time-slices, there is no principle by which one could decide *a priori* how long a time-slice should be, and the evidence for a causal relation can change dramatically depending on the length of the time slice. Shanks (1995) leaves this as an unaddressed challenge for causal learning models, and I view my proposal as a response to this challenge. After I present my framework in Section 2, I offer a more sophisticated version of contingency learning, which integrates over multiple possible time slice lengths to infer a causal dependency. However, I will show that even this model does not account for some of the experimental data as well as my proposal.

A recent study by Griffiths, Baraff & Tenenbaum (2004) also investigated causal learning from temporal data, although the length of the delay was not varied. In their experiment, participants were trained that a set of "nitro-x" cans were explosive. Sometimes cans would explode randomly, but the experimenter could force a can to explode by clicking on it. Cans also

transmitted blast waves, so that nearby cans would also explode after a specific delay, presumably governed by the rate of transmission of the blast wave. By learning the rate of transmission, participants were able to judge whether two exploding cans were caused by a blast transmission or by coincidental random explosions, based on the temporal delay between their explosions. The most interesting condition of this experiment focused on discovering a hidden cause. When multiple cans exploded simultaneously, people inferred that an unobserved cause was responsible for making both cans explode, and their inferences were stronger when more cans exploded simultaneously. Since the cans exploded simultaneously, they could not have caused each other to explode, and therefore the remaining hypotheses are that the cans coincidentally exploded simultaneously, or that there was an unobserved common cause. This kind of causal discovery, however, is different from the one we will be investigating. In our experiments, there are specific observable potentially causal events and a specific effect that occurs at some temporal delay after its cause. The task involves discovering which of the causal events may have caused the effect, even when it is known that the effect may have unobservable causes as well.

### **1.2.2 Previous studies of the role of temporal delay in associative learning**

Associative learning is the process by which organisms learn to associate a cue with an outcome. The role of delay has been extensively studied in associative learning. Although a number of researchers have argued that models of causal learning account for data that associative learning models cannot account for (e.g., Waldmann, 2000), the experiments conducted in the associative learning literature can be productively viewed as relevant to causal learning when looking at single cause-effect relationships. This is because associative learning is identical to causal learning when the cue is a potential cause of the outcome and it is known that

there is no common cause of both the cue and the outcome. The primary findings of associative learning are:

- (1) A cue can be associated with an outcome through repeated exposure to the occurrence of the outcome in the presence of the cue. Behavioral measurements of association can be made by presenting the cue and measuring if and when the animal responds to the anticipated outcome.
- (2) Numerous studies have found that longer delays between the cue onset and outcome result in slower acquisition (e.g., Gibbon et al., 1977). In the case of Pavlovian (classical) conditioning, when the outcome occurs regardless of whether the animal responds, associations are learned fastest when the outcome is presented a short time after the cue. Holding the inter-trial interval (I) fixed, the number of reinforcements required to learn an association increases linearly with the delay of reinforcement (T) (see figure 11, solid line, in Gallistel & Gibbon, 2000, p. 299.). Holding T fixed, the number of reinforcements required for acquisition decreases as the inter-trial interval becomes longer (the trial-spacing effect, Barela, 1999). Acquisition rate remains constant if the I/T ratio remains fixed (Gallistel and Gibbon, 2000, call this “timescale invariance”, see figure 11, dashed line, p. 299).
- (3) In operant conditioning, the outcome only occurs if the animal responds to the cue. The response behavior is best reinforced if the outcome occurs shortly after the animal responds. In a fixed interval schedule of reinforcement, animals can learn that the response will not produce an outcome for certain time after the cue appears, and will adjust their responding rate to be maximal at the time the outcome is expected to occur. After an animal is conditioned to a given delay, the animal displays a characteristic conditioned response (CR) in close proximity to the delay, which indicates the animal has learned the specific delay. The

probability of exhibiting the CR increases as the expected time approaches and then trails off as time passes without an outcome occurrence (e.g., Gibbon, Fairhurst & Goldberg, 1997).

The variance in the response time increases proportionally to the delay (Gibbon et al., 1977).

In my proposed model the delay can be variable; the variability of the delay is a primary influence on the rate of acquisition. It is not clear whether researchers have ever systematically studied how the variance of the delay influences associative learning. One study conditioned rabbits to a bi-modal distribution, showing two different delays to the same stimulus. The rabbits were conditioned to expect the US at either 400ms or 900ms after the CS onset, and the results after training show that the rabbits blinked much more at both of those delays than in between (Kehoe, Graham-Clarke, and Schreurs, 1989). Aside from this study, however, I know of no reports indicating how the variance of the delay influences associative learning.

### 1.3 Previous models of causal learning

Many models of causal learning exist, and it is useful to situate our proposed framework within the space of other causal learning theories. Causal learning theories can generally be classified as either (1) process-level accounts, which provide algorithmic descriptions, or (2) computational level accounts, which provide rational frameworks, often based on normative statistical principles. The framework I propose here is a computational level account, in that we attempt to characterize people's judgments as rational approximations to an optimal statistical standard. In our case, the standard we adopt is Bayesian statistical inference over a space of possible causal representations. Here, we will provide an overview of the major proposals for causal representations and learning.

### 1.3.1 Causal power theory (White, 1995)

According to White’s (1995) causal power theory, objects have properties that enable them to participate in causal relations. An object with a certain causal power can produce an effect in an object with a corresponding liability, given that certain releasing conditions hold. This conceptual framework is the inspiration for our computational framework, in which we explain causal processes by appealing to “powers” of objects that take part in them. The full causal power theory is too rich for the kinds of experiments we will be modeling. Specifically, we limit our investigation to inferences about causal powers, but not liabilities or releasing conditions. However, our framework could naturally be extended to handle these inferences in a similar manner.

### 1.3.2 The probabilistic dependency view

The probabilistic view of causation states that causes are things that raise the probability of their effects:  $P(E | C) > P(E)$ . Causation is not the only thing that can make this probabilistic inequality true; while it is consistent with  $A$  causing  $B$ , it is also consistent with  $B$  causing  $A$ , or something else ( $C$ ) causing both  $A$  and  $B$ . A number of philosophical theories have been advanced to distinguish between these cases (e.g., Reichenbach, 1956). These theories have primarily been concerned with determining whether the increase in the probability of  $B$  is due to a true causal relation between  $A$  and  $B$  or whether it is the result of a common cause. Since our focus is not on distinguishing between true cause and common cause networks, we will not delve further into these theories. We will instead consider the representation itself as a basis for several relevant causal learning methods:

### 1.3.2.1 Hypothesis testing (Frequentist and Bayesian)

In science, causal inference is typically done by creating two randomized groups, making an intervention, and testing for statistical significance. In the simplest case with a binary outcome, a chi-square test can tell you whether the probability of the outcome in the experimental group is different from the probability of the outcome in the control group, and with what level of significance. This kind of test is a statistically normative method of verifying that a probabilistic dependency exists with a certain level of confidence. It is therefore a crucial computational component to the probabilistic dependency view. With the Bayesian version of hypothesis testing, one can also include prior beliefs about whether the dependency exists. More extensive Bayesian frameworks, like the one I propose here, also perform a kind of hypothesis testing, although the hypotheses they consider are much more specific and plentiful than the standard statistical test which examines just the likelihood of the data under the null hypothesis that there is no dependence relationship.

### 1.3.2.2 Power-PC theory (Cheng, 1997; Glymour & Cheng, 1998; Novick & Cheng, 2004)

Combining causal power theory with probabilistic dependency, Cheng (1997) introduced a method to infer causal powers from probabilistic contrasts, dubbed Power-PC theory. The idea is that causal power can be represented by a number between 0 and 1, corresponding to the probability that of the effect in the presence of the cause and in the absence of any other possible causes (including hidden background causes). This probability is labeled  $q$ . When other generative causes are present, the probability of the effect is necessarily increased, and can be computed precisely using a probabilistic or:  $P(E | C_1 \dots C_N) = 1 - (1 - q_1)^{C_1} \times \dots \times (1 - q_N)^{C_N}$ , where  $C_i$  is 1 if the  $i^{\text{th}}$  cause is present and 0 if the  $i^{\text{th}}$  cause is absent, and  $q_i$  is the causal power of the  $i^{\text{th}}$  cause.

### 1.3.2.3 Causal Support (Griffiths & Tenenbaum, 2005)

Griffiths & Tenenbaum (2005) point out that the Cheng’s Power-PC theory is designed to infer the *strength* of a cause’s causal power, but not *whether* the cause has the power. They advance an approach that distinguishes between these two notions. For the purposes of this dissertation, we will use the term “causal strength” to refer to the probability that a cause will produce an effect, and “causal power” to refer to the property of an object that enables it to produce a cause. For instance, almost all matches have the causal power to produce flames, but some do not (e.g., wet matches). For those matches with this causal power, the causal strength is about 75% (e.g., on any given attempt to light a match, there is a 75% chance that a flame will be produced).

Griffiths & Tenenbaum (2005) have advanced a fully Bayesian approach to the learning of causal structure and strength. Inspiration for my framework derives from their notion of causal support, which is the likelihood ratio of the data under the hypotheses that the causal relations

does or does not exist:  $\frac{P(D|H)}{P(D|\neg H)}$ . Intuitively, the likelihood ratio provides a measure of how

strongly one should believe in a causal relation, given the data and uniform prior beliefs about whether or not a causal relation exists. If one starts out believing that causal relations are rare, and therefore unlikely to exist in any particular case, then the likelihood ratio can be multiplied by the ratio of prior beliefs to obtain the posterior odds, which is the fully Bayesian measure of

the odds that the causal relation exists:  $\frac{P(H|D)}{P(\neg H|D)} = \frac{P(D|H)}{P(D|\neg H)} \frac{P(H)}{P(\neg H)}$ .

Causal support has primarily been applied to model human causal induction from contingency data, but it has recently been used in continuous time environments by applying the framework to rate data (Griffiths & Tenenbaum, 2005). Rate data consist of measuring the rate



of occurrence of an effect in the presence and absence of a candidate cause. To compute causal support, one must assume an underlying process by which the data are generated. An obvious approach is to model this process using a Poisson distribution, which is a natural extension of the noisy-or generative model for continuous time (Griffiths, 2005). To compute the rate of an effect in the presence of two causes, one can simply add the Poisson rates of the two causes separately, which is equivalent to OR-ing the causal strengths in a noisy-or model. The model will be discussed in detail below in the context of temporal models (Section 1.4.1).

*1.3.2.4 Causal maps and Bayes nets (Pearl, 2000; Glymour, Spirtes & Scheines, 2000; Gopnik et al., 2004; Gopnik & Glymour, 2002)*

One major problem with all approaches to causal learning is the possibility of confounding. There could be a common cause of both the candidate cause and the effect, which cannot be ruled out unless it, too, is tracked. For example, if insomnia is contingent on wine-drinking, then it could be the case that wine-drinking causes insomnia, but it could also be the case that partying causes both wine-drinking and insomnia. It is therefore unclear from the contingency of insomnia on wine-drinking whether wine-drinking in fact causes insomnia or not.

In the psychological literature, this problem has recently been tackled for large samples by using Bayes nets, a representation from computer science that account for certain important properties of causal relations (Pearl, 2000). Bayes nets provide an elegant formalism of the probabilistic view of causation, accounting for many of the important findings in causal reasoning. In Bayes nets, causes are represented as directed graphs depicting relations between causes and effects. The formalism predicts that a common cause  $C$  of  $A$  and  $B$  renders  $A$  and  $B$  conditionally independent given knowledge of  $C$ , and explains our intuition that  $B$  is irrelevant to judging  $A$  once  $C$  is known.

To infer causation from contingency, one must ensure that there is no common cause of both. The approach taken by constraint-based learning is to learn a causal network by considering all possible variables and finding conditional independencies among them. When variable  $A$  is found to be conditionally independent of variable  $B$  given variable  $C$ , there are several possible causal networks that are consistent with this evidence:  $A \leftarrow C \rightarrow B$ ,  $A \rightarrow C \rightarrow B$ , and  $A \leftarrow C \leftarrow B$ . To distinguish between these networks, one can use temporal order or other cues to the directionality of the causal relationship.

Recently, an attempt has been made to use constraint-based approaches as a model of theory-learning. Gopnik et al. (2004) proposes that “theories” can be modeled as “causal maps”, which are conceived of as representing the true causal structure of the world and are modeled as causal Bayes nets. This idea is based on the “theory” theory (Gopnik, 2003) that data-driven learning can lead to theory-like knowledge in children the same way that experimental data leads to theories in the scientific community. However, it is not clear that scientific theories derive from experimental data. Often scientists develop theories from a small number of observations or thought experiments, and then test them experimentally using controlled studies. How the scientist comes up with the theory prior to running the experiment is unclear. Furthermore, people often do not seem to have the ability to make the kinds of statistical inferences that are required for valid scientific conclusions, as evidenced by the vast numbers of medical treatments people believe to work that scientific methods have shown to be no better than placebo (e.g., recent studies show no benefit from Echinacea, the most popular natural product in America, used by 14 million people; Kolata, 2005).

The Bayes net representation provides intuitive rules for causal inference (i.e., prediction & judgment), including screening off (due to common causes) and handling interventions

differently from observations. However, methods for learning Bayes nets have been somewhat less successful at capturing human intuition, especially in the number of samples required. Inevitably, one must introduce statistically unsound tricks to get the Bayes net algorithm to learn as quickly as people do. For instance, in explaining people's rapid inferences aboutblickets causing detectors to activate, Gopnik et al. (2004) invoke a "large fictitious multiplier" to obtain statistical significance (p. 21). The problem with this approach is not just that it is statistically unsound – it also cannot explain why people learn rapidly in some cases but not in others. For instance, we would not want to conclude that having blue eyes causes a person to enjoy hamburgers on the basis of one or two examples of a blue-eyed hamburger-lover.

Learning algorithms for Bayes nets also have problems handling confounding with small samples. These methods infer causal relations from data by exploiting the fact that a common-cause structure creates conditional independence between the effects given knowledge of the cause. But this solution does not help us in the case of small samples, when confounding can occur even without a common cause structure. For example, one could get sick after eating cotton candy and going on a rollercoaster, but one could not conclusively determine which caused the sickness, even though neither is a cause of the other. If the two variables really are unrelated then with enough samples one will be able to determine that sickness is conditionally dependent on its true cause, given the other cause, but this requires many more samples than people often have. Thus, to determine from just one observation which of these variables is a true cause of sickness, one must use additional information and knowledge. Furthermore, one must actually *attribute* the effect to one or the other of the candidate causes, rather than merely noticing a contingency. This is consistent with a point made in Ahn et al. (1995), which found that in causal attribution, people cared more about factors present in the occasion of interest than

covariations of those factors with the effect on other occasions. We investigate the role of temporal delay in causal attribution in Experiment 3.

### **1.3.3 The mechanism view (Ahn & Kalish, 2000)**

Under the mechanism view, causation is represented by a mechanism enabling a cause to transmit force to an effect. For example, if a person is sneezed on and then later gets sick, the mechanism is the transmission of a germ. On this view, if the germ (or some equivalent mechanism) is not part of one's representation, then one cannot represent a causal relation between getting sneezed on and getting sick. The proposal has primarily been advanced as an alternative to causal learning from covariation. Ahn et al. (1995) proposed that causal learning can often be better characterized as abduction, or inference to the best explanation, rather than induction, or generalizing from examples. They found that when people are faced with determining what caused a given effect (causal attribution), people tend to seek out information about which mechanisms were at work in that specific case rather than which causes tend to covary with the effect in other cases.

In a more theoretical treatment, Ahn and Kalish (2000) argue that learning from covariations cannot explain much of human causal learning. In order to confirm a causal relation exists between  $A$  and  $B$  using covariation information alone one must rule out the possibility that there is a third variable,  $C$ , given which  $A$  and  $B$  are conditionally independent. But to do so would require exhausting all possible variables in the world, of which there are essentially an infinite number. Mechanism knowledge, it is argued, is crucial for deciding which variables, if any, should be considered as possible common or mediating causes. Of course, the mechanism knowledge must itself be learned, and therefore this view can appear circular, in that it does not

provide a proposal for how mechanism knowledge can be acquired. However, a recent proposal has been advanced that addresses this need: causal grammars.

#### **1.3.4 Causal grammars (Tenenbaum, Griffiths, & Niyogi, in press)**

Causal grammars are a proposal for representing theory-like knowledge that can constrain which types of causal relations are plausible, and thereby enable learning from smaller samples. With a causal grammar, one can specify a probability that a member of one class can causally influence a member of another class. For instance, one could specify that there is a 1% chance that a member of the food class has the power to cause allergic reactions, while there is a 0% chance that a member of the silverware class has that power. Therefore, when eating food using silverware, one should attribute any observed allergic reaction to one of the foods, rather than the silverware.

Research into these causal grammars has looked at both learning the grammars themselves, and how a particular grammar, once learned, influences the learning of causal relations. The grammars specify a prior probability of a causal relation existing between two entities. This prior can then be used with any other form of causal learning from evidence. Since our framework is Bayesian, it requires a prior, and the causal grammar formalism would be an excellent way to set the priors. Also, because our framework provides a method for rapidly learning individual causal relations, it can provide the causal relation data required for learning the causal grammars themselves. In this way, our framework and the causal grammar framework are complementary, and could (and should) be fruitfully combined. We discuss this idea in more detail in the general discussion (Section 4.2).

### 1.3.5 Theory-based causal induction (Tenenbaum & Griffiths, 2003)

Griffiths and Tenenbaum (under review) argue that the kind of learning exemplified by the blicket detector experiments (Gopnik et al., 2001) requires pre-existing theory-like knowledge (which may come from the cover story) in order to learn rapidly in just a few trials. In the case of the blicket detector experiments, this theory-like knowledge asserts that some of the blocks are blickets, the machine is a blicket detector, and blickets activate blicket detectors with a noisy-or functional form and no background cause. Their prime example comes from the backward blocking experiment (Sobel, Tenenbaum & Gopnik, 2004), in which people are given two demonstrations of *A* and *B* activating the detector followed by one demonstration of *A* activating the detector alone. People then usually infer that *B* is not a blicket (the rational basis of this inference is that people should provide their prior belief in any block being a blicket, which in this case happened to be low).

While this theory-like knowledge may be required to explain rapid learning when the data are described in a trial-based format, it is not necessary if the data are represented temporally. When block *A* is observed to activate the detector, the dynamic evidence is overwhelming for a causal relationship between block *A* and the detector; it is not necessary to assume that there is no background cause, because it is so unlikely that the background cause would occur at the exact same moment that the block is placed on the detector. One could construe my proposed framework as a possible method for learning the theory-like knowledge proposed in Tenenbaum & Griffiths, however I do not pursue a formalization of such a construal in this dissertation.

### 1.3.6 Intervention vs. temporal order (Langado & Sloman, 2004)

Interventional accounts of causal learning place emphasis on the utility of intervention to discover causal relations. The primary reason interventions are useful is that they distinguish between common cause and direct causation models, even for unanticipated potential common causes. In contrast, constraint-based approaches can only rule out common causes that were previously anticipated and controlled for. Furthermore, one need not control any potential common causes; merely intervening to make the first variable occur is enough. And, even if one infers that there is no common cause of two contingent variables, it is still a problem to determine which of the variables caused the other. Intervention also helps here. If  $A \rightarrow B$  then intervening on  $A$  will change the probability distribution over  $B$ , but not vice versa, because the link from  $A$  to  $B$  is severed once  $B$  is intervened on.

In the case of the experiments presented here, we assume that participants understand that the candidate causes are interventions by invisible people. The lures, divers, and airplanes in our experiments are all controlled by people who are making interventions, and therefore there is no common cause of explosions and lures. This enables us to study temporal delay independently of the issue of confounding due to common cause structures. Recently, researchers have addressed the relationship between temporal order and intervention. Lagnado and Sloman (2004) investigated the extent to which the advantages of interventional learning can be explained by the direct knowledge of temporal order that it provides. However, they did not specifically study the role of temporal delay between cause and effect, and therefore their studies investigate an orthogonal issue to ours. All of the experiments I present here have identical temporal orderings, therefore temporal order can be ruled out as an explanation of any difference between conditions.

## 1.4 Previous models incorporating temporal delay

Although most causal learning models ignore temporal factors, several models in the associative and causal learning literatures have been proposed to account for the role of temporal delay in learning. These models can be understood as resulting from different assumptions about whether the causes and effects are states that persist through time or events that occur at specific time points (see Table 1). We have already seen that when both cause and effect are states (each is either present or absent), a standard contingency model can be used. But when both cause and effect are events that occur at specific points in time, the co-occurrence of cause and effect is not well-defined. In this case, a model that incorporates the delay between cause and effect is appropriate. There is a third type of situation: when the cause is a state and the effect is an event, the presence of the cause can increase the rate of the effect. This is the causal process behind a Geiger counter: the presence of a radioactive substance increases the rate of clicking in the counter. Next, we will present these alternative models, while exploring some of the limitations that led to the need for our newly proposed model. These models will also be used to analyze the experimental data alongside our new event-based framework.

		Effect	
		State	Event
Cause	State	Standard contingency model	Rate models
	Event	N/A	Delay models Co-occurrence contingency model Event-based framework

**Table 1: Space of model types for causal learning.**

### 1.4.1 Rate models

Both Griffiths & Tenenbaum (2005) and Gallistel (1990) have proposed models of causal (or associative) relations in which the rate of an effect (or outcome) is increased by the presence



of a cause (or cue). Gallistel's Rate Estimation Theory (RET) conceives of an association as an increase in the rate of an outcome in the presence of the cue. According to RET, an association is acquired when a learner's estimate of the ratio of the outcome rate in the presence of the cue to outcome rate in the absence of the cue exceeds some threshold. Gallistel & Gibbon (2000) refer to this as the "whether" criteria: whether to infer that there is a relationship between the cue and the outcome. Although traditional contingency models use a specific time slice length to define a co-occurrence, Gallistel (1990; 2002) argues that the data from the associative learning literature actually show learning to be timescale-invariant. Gallistel's RET theory is timescale-invariant, meaning the same number of outcomes is required to learn a fast rate as a slow rate. This is because taking the ratio of rates eliminates the time interval used in the denominator of each ratio; a ratio of twice per second to once per second is learned in just as many trials as a ratio of twice per day to once per day. Because it is timescale-invariant, RET gracefully overcomes the problem of choosing a time interval for co-occurrence tables.

Griffiths & Tenenbaum (2005) have also proposed a rate model, but unlike Gallistel they have proposed the model within a rational statistical framework. They apply Bayesian statistical inference to the problem of inferring whether the rate of an effect is greater in the presence of a cause than in its absence, and use this model to account for people's judgments about a Geiger counter. Griffiths (2005) analyzes the rate model as a continuous version of a contingency model, because the Poisson rate is the continuous analogue of the discrete Bernoulli trial.

Like our model, these rate models predict a natural advantage for shorter temporal delays between cause and effect. In the case of Gallistel's (1990) model, a learner acquires an association when the estimate of the ratio of the rate of the outcome in the presence of the cue to the rate in the absence of the cue exceeds a pre-specified threshold. Because a shorter observed

temporal delay between cue and outcome represents a higher rate of outcome occurrence in the presence of the cue, the ratio will exceed the threshold faster than for a longer temporal delay. In the case of Griffiths' (2005) model, one integrates over all hypothesized rates, including both low and high rates. Shorter delays again have an advantage because the likelihood of a short delay under a high rate is much higher than the likelihood of a long delay under a low rate.

The rate models break down, however, in two cases. First, when the cause does not always produce the effect, then sometimes the cause will occur without an effect occurrence. If the effect then occurs on its own, after some extremely long delay, this would greatly reduce one's estimate of the rate. If the cause only rarely produces the effect (such as in the case of a medication producing a side effect), the overall increase in rate may not be significant because there are so many occasions on which the effect never occurs. We investigate this issue in Experiment 4. Second, when the delay is long, but fixed, then the rate model will learn just as quickly as if the delay is variable, provided the mean delay is the same. However, a rational model should be capable of learning that the delay is fixed and therefore should be quicker to infer a causal connection if the delay is fixed than if it is variable. This is because the likelihood of the fixed delay under a model that generates a fixed delay is much higher than the likelihood of the mean delay under a model that generates a variable delay centered around that mean. It is perhaps for this reason that both Gallistel & Gibbon (2000) and Griffiths (2005) have embraced models that specifically learn fixed delays. Neither, however, appear to have integrated these approaches into a single rational model. We investigate the issue of long but fixed delays in Experiment 2.

### 1.4.2 Delay models

The rate models nicely extend existing accounts based on covariation to handle continuous time and are applicable to situations in which the cause produces the effect at a specific rate (Poisson process). However, they do not provide a means to learn the delay between two events that occur at fixed points in time, which is the subject of our inquiry here. The rate model is designed to handle data that associate cause presence or absence with the rate of effect occurrences. In contrast, our model is designed to handle cases in which the cause occurs at a particular moment in time, is only present for that brief moment, and produces a single effect occurrence at a later time. For instance, it would be odd to speak of the rate of paralysis in the presence of a snake bite.

Despite the distinction between rates and delays, one can coerce the rate model into handling event data, with limited success. This is done by assuming that the cause is present from the time it occurs until the time the effect, if any, occurs, at which point the cause becomes absent. In our experiments, this presence is explicitly shown because the explosive lures remain in the water until an explosion occurs. In the case of the nitro-X experiments (Griffiths, Baraff, & Tenenbaum, 2004; Griffiths, 2005), one can assume that since the cause was an explosion, a blast wave radiated out and although it was invisible, it persisted through time. In this situation, however, one does not necessarily know when the cause stops being present. Imagine a situation in which a lure fails to explode and remains in the water until the next lure causes an explosion. It would be erroneous to treat the unexploded lure as a present cause up until the next explosion because this could result in dramatically different rate estimates depending on when the subsequent lure is cast. Similarly, if a Nitro-X blast fails to cause its neighbor to explode, one would not want to treat the blast wave as continuing to be present until the next explosion occurs.

For learning a fixed delay between two events, both Gallistel & Gibbon (2000) (borrowed from Gibbon et al., 1977) and Griffiths (2005) have proposed additional models. Gibbon et al.'s (1977) Scalar Expectancy Theory (SET) is based on Weber's law that the error in one's estimate of a time interval is proportional to the length of the interval. This model only applies to mature responding, after the association has already been learned, and accounts for the "when" criteria: when to respond to a cue, or more generally, the expected occurrence time of the outcome once the cue has occurred. The model remembers previous delays between cue and outcome (the reinforcement intervals) and predicts responses to the cue in anticipation of an expected outcome. It incorporates Weber's law to account for the fact that the variability in the distribution of responses is scalable – it is proportional to the mean delay, or in the case of a fixed interval, it is proportional to that interval. Griffiths (2005) proposes a rational model of learning a fixed delay, without estimation error. In his model the estimate of the delay is characterized by a dirac delta function. This can be reduced to Gibbon's proposed Gaussian distribution with a variance of zero.

If the effect reliably follows the cause after a fixed delay, then a model in which causes generate their effects after a delay is more appropriate than a rate model. Such a model can be capable of learning causal relations very rapidly; even in just one trial. For example, suppose you find a remote car door unlocker and you want to figure out which car it controls, if any. If you press the button and moments later a nearby car beeps, you may be happy to conclude that you've figured out which car it controls. However, in some cases one trial is not enough. For example, if your mobile phone rings just after you press the button on the car door unlocker, you would likely consider this to be a coincidence. However, if it happens on a second occasion exactly the same way, you may discover that in fact the button does cause your phone to ring.

With short delays, just two trials can often produce significant evidence of a causal relationship, provided the delay is identical on each trial.

Although it is an excellent characterization of cases where the delay is truly fixed, the fixed delay model suffers from over-fitting. First, if the delay truly is fixed, the model may infer a causal relationship faster than people would, especially for long delays. One can remedy this by introducing a prior against long delays (as in Griffiths, 2005) but the rational basis for this prior is unclear. Alternatively, this could be remedied by using a Gaussian density (as in Gibbon et al., 1977), rather than a delta function, with variance proportional to the mean, but this is motivated by presumed biological constraints (i.e., Weber's law applied to remembered temporal intervals), rather than having a rational basis. We will see in Experiments 1 and 2 that neither of these methods accounts for the experimental data as well as the event-based causal learning framework. Second, if the effect does not always occur with the exact same delay after the cause, the model will only be capable of treating one of the observed delays as evidence of a causal relationship, and because of this, depending on the assumed causal strength, may fail to infer any relationship whatsoever. In contrast, people should be capable of learning a causal relationship in which the temporal delay varies. We will see in Experiment 2 that people's ability to learn causal relationships does not degrade significantly when the delay is variable.

### **1.4.3 The co-occurrence contingency model**

Contingency models learn from co-occurrence data defined in terms of trials. One way to use a contingency model with event-based data is to divide the data into time slices, with each time-slice representing a trial. If cause and effect both occur within the same time slice, this is counted as a co-occurrence. The full dataset includes all co-occurrences, single occurrences and non-occurrences, and can be represented in a co-occurrence table, as depicted in Figure 1. The

standard method of learning a causal relation from this data is to compute a significance value from the table using a chi-square ( $\chi^2$ ) test. This significance value represents the probability of obtaining a dataset as extreme or more extreme than the one observed under the null hypothesis ( $H_0$ ) that there is no causal relationship. In standard hypothesis testing, one typically adopts a pre-specified significance criteria after which one would accept the alternative hypothesis that there is a causal relationship (e.g., accept if  $P(x>data|H_0) < .05$ ). The test can be made consistent with Bayesian principles by introducing a prior probability representing our prior belief that there is in fact a causal relationship, and comparing the significance value with the corresponding prior.

As Shanks (1995) has articulated, the use of contingency models with temporal data suffers from a chicken-and-egg problem: we must specify the size of the time slice ahead of time, yet the data will generally determine the timescale of the process, and consequently the size of time slice that would be best. We can, however, remedy this problem in one of two ways. First, we can examine the data and pick the one time slice that makes  $p(\text{data}|H_0)$  maximal. Second, we can integrate over all hypothetical time slice sizes, just as the event-based causal learning framework integrates over all possible densities. Both of these methods will result in models that learn faster from shorter delays and can also handle variable delays. However, like the rate model, they suffer in cases of a long but fixed delay, learning no faster than in cases of a long but variable delay. We will see in Experiment 2 that even when integrating over multiple window sizes, contingency models do not account for the data as well as the event-based causal learning framework.

## 2 A new framework for event-based causal learning

We have just reviewed several previous models of causal learning. Most have ignored the role of temporal delay in causal learning. Those that incorporate temporal delay have focused on one or another specific probability density over temporal delay that happens to be suited to the task at hand: the Poisson distribution for the rate model, the dirac delta function for a fixed delay, or the normal distribution for a fixed delay with estimation error. For these previous models, learning consists of estimating a specific parameter. In the case of the Poisson distribution for the rate model, the parameter is the increase in rate of the effect due to the presence of the cause. For the fixed delay models, the parameter to be estimated is the delay itself. In the case of SET, the uncertainty in the parameter estimate is modeled as being proportional to the parameter value representing the fixed delay.

Here I propose a new framework that offers a rational account of causal learning from temporal data when the delay is not assumed to be fixed. In contrast to previous models of temporal delay, I propose that people are sensitive not just to one parameter, but to the entire probability distribution of possible delays between cause and effect. In practice, this amounts to learning which probability density, or mixture of densities, best characterizes the distribution of delays, in addition to the most likely parameters of those densities. In principle, the density could be arbitrarily complex, such as a mixture of densities from different families. To keep our hypothesis space from becoming unwieldy, however, in my computational model I specify a set of possible densities and a range of parameters, and integrate over both the densities and the

parameters. I have not included mixtures of densities in the computational model, but this is a straightforward extension of the general framework.

In the spirit of rational analysis (Anderson, 1990), I develop a model that is designed to make judgments that are optimal with respect to the environment. In the case of causal learning from temporal data, the environment determines the kinds of delay distributions that one is likely to encounter. A judgment that is optimal with respect to this environment should make some assumptions about what distributions exist in the environment, and then interpret data as being generated from these distributions. In our case, we assume that the causal processes in the environment can be characterized as chains of Poisson processes, and therefore the delay distributions can be characterized as gamma distributions. We will refer to the use of gamma distributions in the event-based framework as the “event-based chain model”.

The event-based *chain model* makes a number of interesting predictions. Each of these points will be investigated in our experiments (Section 3).

1. Shorter delays provide more evidential support for a causal relationship. This is because it is impossible to know the actual distribution of temporal delays, and by integrating over all possible gamma distributions, the shorter means have higher peak values, making shorter delays more likely overall (see Figure 6).
2. A smaller variance in inferred distribution over delays will make subsequent observations more informative. This is because distributions with smaller variances are more peaked, thus subsequent delays that fall near the mean have a higher likelihood compared to distributions with larger variance, while those that fall far from the mean have a lower likelihood compared to distributions with larger variance. Since shorter delays naturally have smaller variance, this explains the advantage of shorter delays in causal learning. However, if a



distribution of long delays with a small variance is known, it should provide an equally compelling advantage for inferring causal relationships.

3. The presence of alternative causes should reduce the evidential support for a causal relationship. This is because causes compete as explanations for the observed effect. This is the idea behind the well-known phenomenon of blocking in associative learning, in which the presence of a known cause of an effect blocks subjects from learning about an additional cause. However, to my knowledge there have been no investigations of how the temporal delay between the causes and the effect influence blocking. In the event-based framework, a short delay between a new candidate cause and an effect can overcome the effect of blocking by a previously known cause provided that the delay between the previously known cause and the effect is not typical of the delay for that cause.

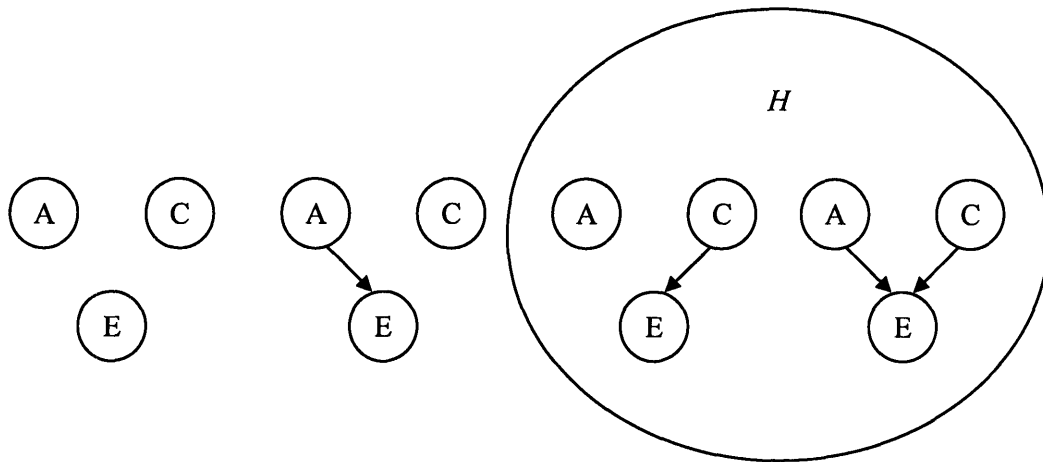
## 2.1 Bayesian inference for causal learning

The basic representation of causation in this framework derives from White's (1995) causal power theory. According to this theory, objects can have properties called causal powers. If an object has such a causal power, it can cause an effect within an object having a corresponding liability once certain releasing conditions are met. In the framework developed here, I assume for the sake of simplicity that the liabilities and releasing conditions are unambiguously observable, whereas the causal powers are uncertain and must be inferred. For example, if two objects enter a pond, and then an explosion occurs in the pond, it is unclear which of the two objects caused the explosion, but it is perfectly clear that the pond has the liability of being "explodable", and that the releasing condition is entry of the causing object into the pond. We extend this conceptual framework by introducing specific parameters associated with causal power. Each causal power in the framework has a characteristic distribution over the

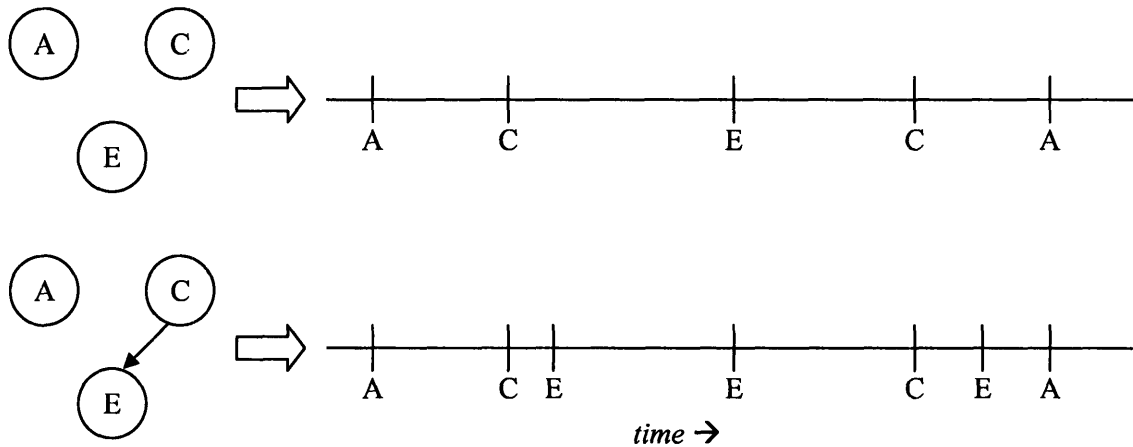
temporal delay between the occurrence of the cause (or, more specifically, the occurrence of the releasing condition) and the occurrence of the effect.

In this section, we introduce the important ideas underlying inferences in the framework. For the technically advanced reader, a formal specification of the model is provided in Appendix A. Inference in this framework is based on the tradition of Bayesian inference over spaces of hypotheses generated by hierarchical probabilistic generative models (e.g., Griffiths, 2005; Kemp et al., 2004). At the first level of the hierarchy, the model generates a set of causal powers within the objects of the world. Objects with a particular causal power are capable of producing a particular effect, while those that do not have the causal power cannot produce the effect. Each assignment of causal powers to objects represents a hypothesis about the truth of the world (see Figure 2).

At the second level of the hierarchy, the model generates events. First, it generates cause event occurrences. A cause event is similar to the releasing conditions of White's (1995) causal power theory. In our stimuli, the cause event is the entry of an object into the pond. If a cause object has the causal power for a particular effect, then some time after the cause event occurs, the effect will be generated, with probability determined by the causal strength  $q$ , and after a delay determined by the distribution over temporal delay (see Figure 3). With this model, Bayesian inference can be used to arrive at a posterior probability of each hypothesis, determining which assignment of causal powers to objects is the best explanation for the observed data.



**Figure 2: Level 1 of the hierarchical generative model: a set of causal powers is generated. An assignment of causal power to an object is depicted as an arrow from the object (A or C) to the effect (E). The hypothesis space depicted here consists of all combinations of causal power assignments to a candidate cause C, alternate cause A, and an effect E.  $H$  corresponds to the set of hypotheses in which C causes E.**



**Figure 3: Level 2 of the hierarchical generative model: a set of event occurrences is probabilistically generated for the hypothesis of causal relations generated in Level 1. If a causal relation exists, the effect is probabilistically generated following its cause.**

Under our framework, causal learning is formalized as inference about the probability that a particular set of hypotheses  $H$  is true: those in which a particular object has the causal power to produce an effect (see Figure 2). Using Bayesian inference, the posterior odds of  $H$  can

tell us the likelihood of this set of hypotheses being true; if the posterior odds is greater than one,  $H$  is more likely to be true, and if less than one,  $H$  is more likely to be false. The Bayesian formula for the posterior odds is shown in Eq 3. It states that the posterior odds of  $H$  being true (after having observed some event data  $D$ ) is equal to the product of the prior odds of  $H$  being true (before  $D$  was observed) times the likelihood ratio of observing the data under the  $H$  vs.  $\neg H$ .

$$\underbrace{\frac{P(H | D)}{P(\neg H | D)}}_{\text{posterior odds}} = \underbrace{\frac{P(H)}{P(\neg H)}}_{\text{prior odds}} \times \underbrace{\frac{P(D | H)}{P(D | \neg H)}}_{\text{likelihood ratio}} \quad \text{Eq 3}$$

One way to formalize the notion of learning a new causal relationship is computing a posterior odds to be greater than one.

### 2.1.1 The likelihood ratio

The likelihood ratio is computed using the second level of our hierarchical generative framework, in which events are generated from causal relations. The data in this case is the observed occurrence times of the all the cause and effect events. When  $H$  is false (i.e., the object of interest does not have the power to cause the effect), the likelihood of the data is equal to the combined probability of all events occurring spontaneously. Here we describe the case for just one cause of interest and one effect; in Appendix A, the model is presented for an arbitrary number of causes. If we assume that a Poisson process can characterize the spontaneous occurrence of events, each event occurrence is independent of each other, and the likelihood of the data is equal to the product of the probability of each event occurrence (see Eq 4).

$$P(D | \neg H) = P(C, E | \neg H) = \left( \prod_{c_j \in C} P(c_j) \right) \left( \prod_{e_k \in E} P(e_k) \right) \quad \text{Eq 4}$$

When  $H$  is true, and the object of interest does have the causal power to produce the effect, the likelihood of the data is equal to the probability of the cause occurrences multiplied by the conditional probability of the effect occurrences given the cause occurrences (see Eq 5).

$$P(D | H) = P(C)P(E | C, H) = \left( \prod_{c_j \in C} P(c_j) \right) \left( \prod_{e_k \in E} P(e_k | C, H) \right) \quad \text{Eq 5}$$

The likelihood ratio now simplifies to Eq 6.

$$\frac{P(D | H)}{P(D | \neg H)} = \frac{\left( \prod_{c_j \in C} P(c_j) \right) \left( \prod_{e_k \in E} P(e_k | C, H) \right)}{\left( \prod_{c_j \in C} P(c_j) \right) \left( \prod_{e_k \in E} P(e_k | \neg H) \right)} = \frac{\prod_{e_k \in E} P(e_k | C, H)}{\prod_{e_k \in E} P(e_k | \neg H)} \quad \text{Eq 6}$$

## 2.2 Causal attribution during learning

The likelihood of the effect occurring some time after the cause occurs will necessarily depend on whether the effect was actually produced by that cause occurrence or whether it occurred spontaneously. We must therefore consider both possible explanations of the effect; this is the process of causal attribution, and it plays a fundamental role in our computation of likelihoods. For spontaneous occurrence, the likelihood of the effect is defined by the Poisson process, assuming the effect randomly occurs at some rate  $\lambda$ . For effect occurrences that are produced by the cause, the likelihood can be computed if we specify a distribution  $f_\theta$  over the temporal delay between cause and effect. For simplicity, we assume that if a particular cause occurs more than once before the effect occurs, only the most recent cause occurrence can produce the effect (in our stimuli, a lure cannot enter the water twice prior to an explosion). Furthermore, we assume that if once an effect occurs, the cause occurrences prior to that effect occurrence are no longer active, regardless of whether the effect was produced by the cause or

not (in our stimuli, once an explosion happens, all lures in the water are destroyed, regardless of which lure caused the explosion). We introduce a new variable,  $\psi$ , to specify whether a particular effect occurrence was produced by the cause, or occurred spontaneously:  $\psi_k = 1$  if the  $k^{\text{th}}$  occurrence of the effect was produced by the cause, and  $\psi_k = 0$  if it occurred spontaneously.

If  $H$  is true, and the effect is produced by the cause, as opposed to occurring spontaneously, then the time of effect occurrence is defined by the function  $f_\theta$  as in Eq 7 (for the purposes of this section, we assume that causal strength,  $q$ , is 1; in Appendix A we relax this assumption):

$$P(e_k | C, H, \psi_k = 1) = f_\theta(e_k - c_*) \quad \text{Eq 7}$$

[ $c_*$  is the cause occurrence nearest to the effect occurrence]

$$P(e_k | C, \neg H, \psi_k = 1) = 0$$

If the effect occurs spontaneously, the probability density function characterizing the likelihood of the effect occurrence is defined by the Poisson distribution:

$$p(e_k | C, H, \psi_k = 0) = p(e_k | \psi_k = 0, \neg H) = \lambda e^{-\lambda(e_k - e_{k-1})} \quad \text{Eq 8}$$

Summing over these possibilities, the probability of an effect occurrence is given in Eq 9:

$$\begin{aligned} p(e_k | C, H) &= p(e_k | C, H, \psi = 0)P(\psi = 0 | C, H) + p(e_k | C, H, \psi = 1)P(\psi = 1 | C, H) \\ &= \lambda e^{-\lambda(e_k - e_{k-1})} \left( 1 - \int_0^{e_k - c_*} f_\theta(t) dt \right) + e^{-\lambda(e_k - e_{k-1})} f_\theta(e_k - c_*) \end{aligned} \quad \text{Eq 9}$$

If  $H$  is false, then the effect must have occurred spontaneously because the cause object does not have the power to produce the effect. In this case, the likelihood is given by the poisson process:

$$P(e_k | \neg H) = \lambda e^{-\lambda(e_k - e_{k-1})}$$

The likelihood ratio now becomes Eq 10.

$$\begin{aligned}
\frac{P(D | H)}{P(D | \neg H)} &= \frac{\prod_{e_k \in E} P(e_k | C, H)}{\prod_{e_k \in E} P(e_k | \neg H)} = \frac{\prod_{e_k \in E} \lambda e^{-\lambda(e_k - e_{k-1})} \left( 1 - \int_0^{e_k - c_*} f_\theta(t) dt \right) + e^{-\lambda(e_k - e_{k-1})} f_\theta(e_k - c_*)}{\prod_{e_k \in E} \lambda e^{-\lambda(e_k - e_{k-1})}} \quad \text{Eq 10} \\
&= \prod_{e_k \in E} \left( 1 - \int_0^{e_k - c_*} f_\theta(t) dt + \frac{1}{\lambda} f_\theta(e_k - c_*) \right)
\end{aligned}$$

We have just described how causal attributions play a role in learning. But the framework also can account for how causal attributions are computed. In this framework, determining whether the effect should be attributed to the cause corresponds to computing the odds that  $\psi_k = 1$  (see Eq 11).

$$\frac{P(\psi = 1 | D)}{P(\psi = 0 | D)} = \frac{P(\psi = 1 | H, D)P(H | D) + P(\psi = 1 | \neg H, D)P(\neg H | D)}{P(\psi = 0 | H, D)P(H | D) + P(\psi = 0 | \neg H, D)P(\neg H | D)} \quad \text{Eq 11}$$

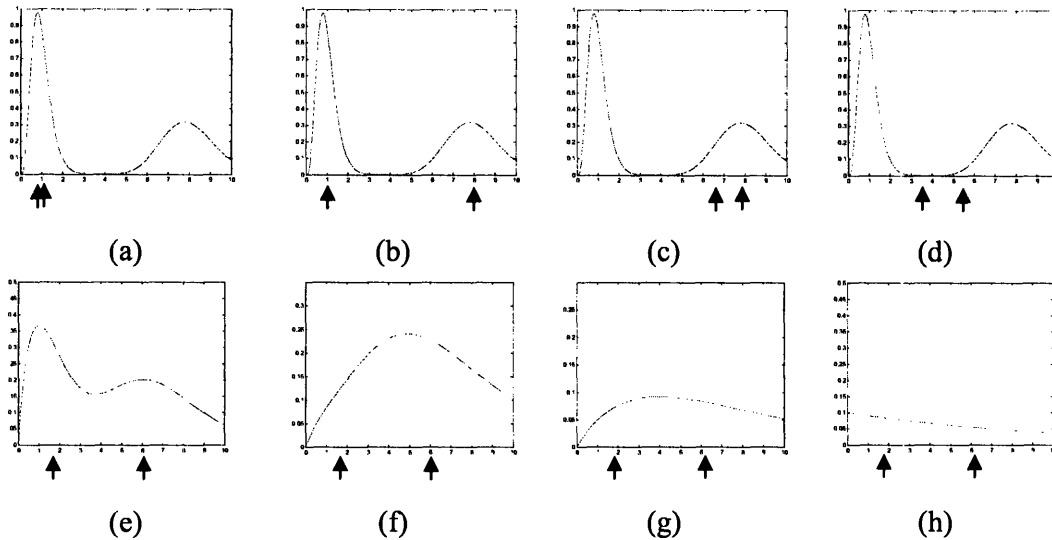
Each of the above probabilities can be computed using Bayes rule, using the formulas we derived earlier (e.g., see Eq 12).

$$P(\psi = 1 | H, D) = \frac{P(D | \psi = 1, H)P(\psi = 1 | H)}{P(D | H)} \quad \text{Eq 12}$$

### 2.2.1 The delay distribution

When  $H$  is false (i.e., the object of interest does not cause the effect) the effects can only arise via a Poisson process. This means the occurrence times are independent of each other, and no occurrence time is more likely than any other, therefore for any fixed number of effect events, the denominator of the likelihood ratio will always be the same regardless of the occurrence times of candidate cause or effect. Thus, for a given number of effect events, it is the numerator that primarily determines the value of the likelihood ratio. For a given hypothesized distribution, it is the actual delays between cause and effect in the event data that determine the value of the likelihood ratio (see Figure 4(a-d)). Conversely, for a given set of candidate cause and effect

events, it is the hypothesized distribution over temporal delay that primarily determines the value of the likelihood ratio (see Figure 4(e-h)).



**Figure 4: the delay distribution and the actual observed delays determines the likelihood ratio. The distribution appears in blue, and two example delays between cause and effect are depicted by the black arrows. (a)-(d) show different delays for a single distribution, in order of decreasing likelihood ratio. (e)-(h) show the same delay with different distributions, in order of decreasing likelihood ratio.**

### 2.3 Causal discovery: learning without priors

Causal discovery is a particularly important kind of causal learning because it can account for true bottom-up acquisition of causal knowledge. It is the earliest form of causal learning, when there is no prior knowledge, not even a prior belief that a causal relation exists to be discovered. If one has no prior belief in a causal relation existing, immense skepticism is appropriate, and new causal relationships should be inferred only with overwhelming statistical significance. Otherwise, one would discover causal relations among all sorts of random events, merely because of chance co-occurrences. Formally this means the prior odds should be very low. To discover the causal relation, the posterior odds should be greater than one, which means



the likelihood ratio must be very high. In practical terms it is difficult to obtain such significance without temporal evidence. Contingency-based accounts from real-world data typically fall far short of providing the required evidential support, unless they have very large samples, or specifically take into account temporal evidence by using time slices as trials.

Under the event-based chain model, the likelihood ratio will be high if the distribution assigns high likelihood to short delays and the actual delays in the data are in fact short. Similarly, if the distribution assigns high likelihood to long delays of a specific duration and the actual delays in the data match that duration, the likelihood ratio will also be high. Qualitatively speaking, for the likelihood ratio to be high enough to discover causes, the delay distribution should have a relatively high peak at whichever delays occur frequently in the data (as in Figure 4(a, b, and e)).

In real-world causal discovery, however, one usually does not actually know the distribution over temporal delay between cause and effect. To solve this problem in a Bayesian framework, we can consider a potentially large set of possible distributions, compute the likelihood for each, and then sum them, weighted by a prior probability that each is the true distribution (Eq 13).

$$\frac{P(D|H)}{P(D|\neg H)} = \int_{f_\theta} \frac{P(D|H, f_\theta)}{P(D|H)} P(f_\theta) \tag{Eq 13}$$

If we have absolutely no information about the correct distribution we could weight all distributions equally. But often we know something about the kind of causal power we are considering, and this provides us the possibility of weighting some distributions differently than others. For instance, we might expect a distribution for a match lighting to have a peak at roughly half a second after striking, a distribution for food poisoning symptoms to have a wide

peak at a delay of two to six hours after eating the poisoned food, and a distribution for a kitchen timer to have an extremely narrow peak at the delay specified when the timer was set.

A natural choice for characterizing the delay of physical causal processes is the gamma distribution. If we integrate the likelihood ratio over all gamma distributions with a uniform prior probability, we can obtain the probability of the data under  $H$  even without knowing ahead of time anything about what the temporal distribution is. The use of the gamma distribution leads to a natural advantage for shorter delays: data with only short delays will have a higher likelihood after integrating over gamma distributions than data with only long delays. This is due to a property of gamma distributions that reflects the natural world: the variability of a delay is proportional to its mean. This leads to a situation in which the peak of a distribution favoring short delays is higher and tighter than the peak of a distribution favoring long delays. Of course, there are causal processes with long but predictable delays, and a nice feature of using the gamma family is that with sufficient evidence one can learn such a distribution. The fact that it requires more evidence to learn a long but tightly peaked distribution is consistent with intuition, and will be tested empirically in Experiment 2.

## 2.4 The short delay advantage: Why short delays result in faster learning

In the study of associative learning, it has been clear from the earliest studies that the delay between cue and outcome was a primary determinant of acquisition rate, and that shorter delays tended to work better than longer delays. For the purposes of this dissertation, we will call this phenomenon the “short delay advantage”. From a purely rational perspective, it is not clear why one should be less inclined to believe in causal relationships characterized by long delays

between cause and effect. For example, for a model with a fixed delay and uniform prior over all delays, a reliable 10 second delay between a cause and effect should carry just as much statistical evidence for a causal relationship as a reliable 1 second delay. Thus, any model of causal learning from temporal data must offer an explanation for why short delays should provide such an advantage.

Gallistel's (1990) RET model predicts the short delay advantage because the estimated rate of the outcome in the presence of the cue is higher with shorter delays. Although RET is a process-level account, it is rooted in a rational analysis of learning in which the cue increases the rate of an effect. However it is not entirely rational to apply RET to a case where there is a fixed interval. If one already knows that the interval is fixed, then one could rationally learn what the interval is in just a few trials. In such a case, acquisition would occur in the same number of trials, regardless of the delay. Thus a rational version of the RET approach should only predict a short delay advantage in cases where it is rational to assume that the cue increases the rate of the outcome, rather than produces the outcome after a fixed interval. Furthermore, if one looks at the data to decide whether a rate model or a fixed model is more appropriate, the fixed delay model will provide a much better fit to fixed delay data and will therefore be selected as the most appropriate model. Thus, although it may predict the behavioral pattern of responses in the fixed interval case, the rationality of using of rate estimation in this case is questionable.

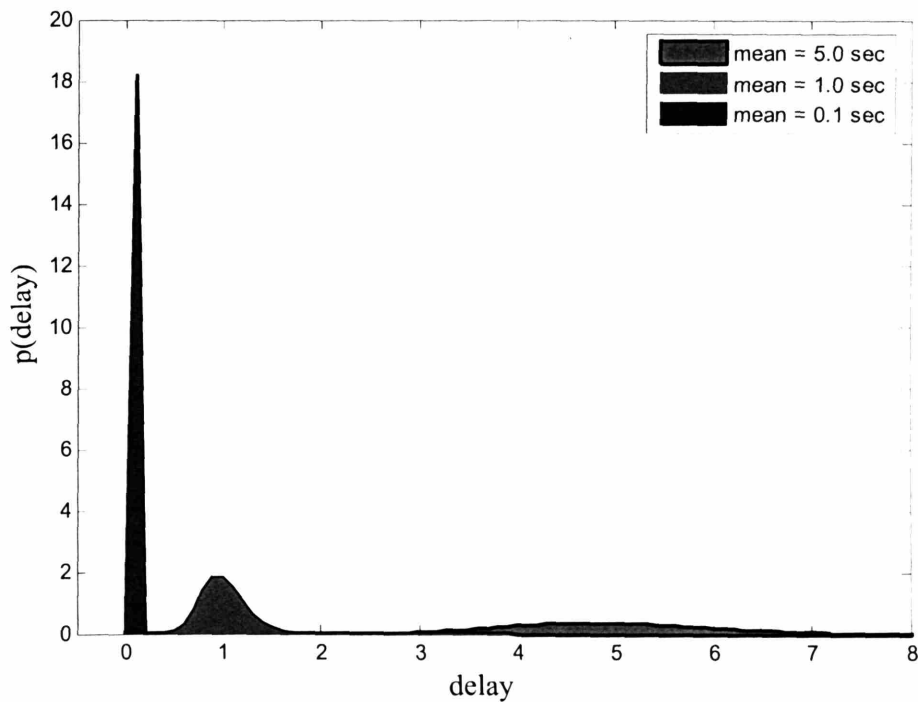
Gibbon et al.'s (1977) SET theory applies to fixed intervals. According to SET, Weber's law predicts that the standard deviation in one's estimate of an interval is proportional to the length of the interval. Thus, SET accounts for the fact that shorter intervals result in greater certainty of the interval duration. However, SET is not a model of acquisition, only of mature

responding. Thus, it does not predict that shorter delays will result in faster acquisition, which is what the short delay advantage represents.

In contrast to SET, which is a process-level model, the framework I develop here is a computational-level account, which seeks to explain causal learning as a rational inference. Griffiths (2005) has also proposed a rational approach to causal learning in the fixed interval case. Unlike RET, it can model acquisition of a fixed interval, and unlike SET, it models the acquisition of a causal relation, rather than the estimation of temporal delay after acquisition has occurred. However, it does not naturally predict the short delay advantage. It can be modified to account for the short delay advantage by adding into the model a prior favoring short delays, but the motivation for this prior is unclear.

Unlike these earlier models, I seek to explain the short delay advantage as resulting from rational statistical inference of causal relations from temporal evidence. In the proposed event-based chain model, the short delay advantage results from assuming that the probability density characterizing the temporal delays between cause and effect is a gamma distribution. Gamma distributions characterize the time to the first occurrence in a multi-step Poisson processes. Under a gamma distribution, short delays are inherently less variable than long delays, therefore the peak of the delay is narrower and higher, resulting in a higher likelihood of the data. To see this, consider the gamma distributions depicted in Figure 5. These particular distributions depict the probability density over the second occurrence time in a Poisson process (the shape parameter is 2 which corresponds to the second occurrence). Three distributions are depicted, with rates of 0.1sec, 1sec, and 5sec. The likelihood of observing a particular delay corresponds to the height of the distribution at that delay. The likelihood of observing a delay equal to the mean of the distribution is proportional to the mean of the distribution. Thus, the likelihood of

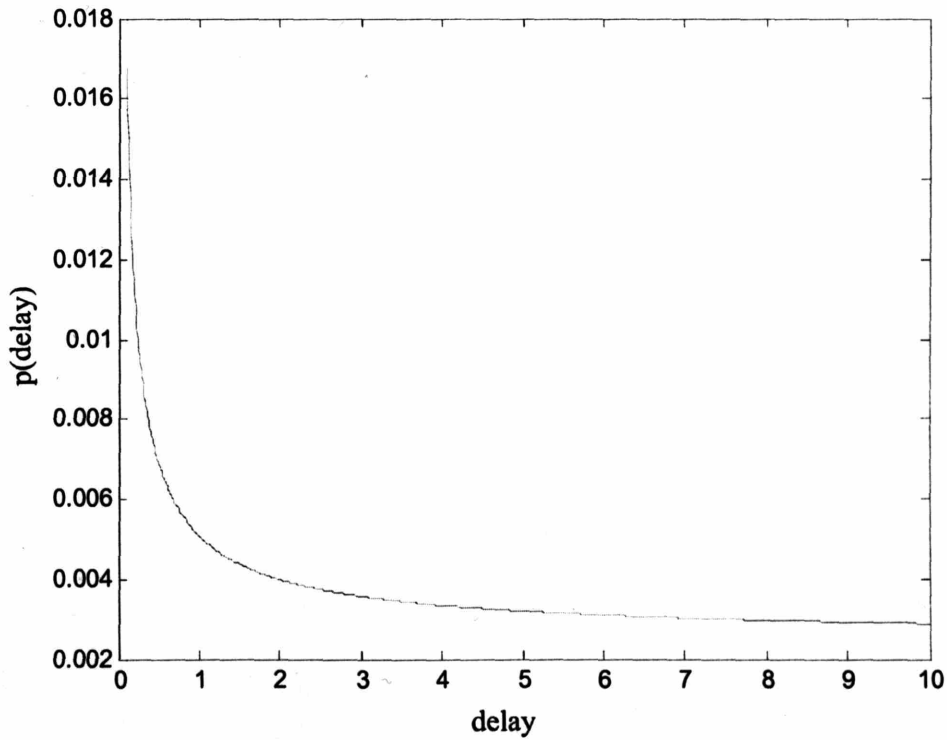
observing a 0.1sec delay when the mean delay is 0.1sec is 10 times higher than the likelihood of observing a 1sec delay when the mean delay is 1sec, and 50 times higher than the likelihood of observing a 5sec delay when the mean delay is 5sec.



**Figure 5: three gamma distributions with means of 0.1sec, 1sec, and 5sec, and shape parameter integrated uniformly from 1 to 21.**

If one has no prior information about the number of steps in the process, or about the Poisson rate of each step, then a rational approach would be to integrate over all possible gamma distributions, summing the likelihood of the data under each to compute the overall likelihood of the data. In this case, a shorter observed delay will naturally have a higher likelihood, and should therefore appear more like a causal process, even having no prior information or expectations about the actual distribution over delays. We can see in Figure 6 that by summing over many

different gamma distributions, we get a combined probability density favoring shorter delays, even with no prior bias towards shorter delays.



**Figure 6: the likelihood of various temporal delays, created by summing over all gamma distributions with location parameter from 0.1 to 10 and shape parameter from 1 to 21.**

## 3 Experiments in causal learning

Four experiments in causal learning were conducted to test the predictions of the event-based computational model. In Experiment 1, we investigate the process of causal discovery, showing that people are influenced by both temporal delay and alternative candidate causes when making causal discoveries. In Experiment 2, we use several training examples to provide participants evidence for a particular distribution over temporal delay, and then investigate how that learned distribution influences judgments of whether future objects have causal power. In Experiment 3, we show that the occurrence times of alternative causes influence people's judgments about the causal power of a particular cause of interest. In Experiment 4, we contrast the predictions of the event-based chain model to those of alternative models when the overall rate of an effect in the presence of the cause is held constant, but the individual delays between cause and effect occurrences vary. The results will be compared to the predictions of the event-based chain model<sup>1</sup>, as well as to those of three other models capable of handling temporal data: (1) Griffiths and Tenenbaum's (2005) rate model using causal support, (2) the fixed delay model with a prior favoring short delays, and (3) the contingency model using a variable timeslice.

Using these four experiments I seek to investigate several important phenomena in causal learning, focusing on how well people's judgments can be predicted by the event-based chain model. The primary phenomenon I seek to investigate is the finding that people are able to discover new causal relations from small samples, often from just a single trial (e.g., Schulz &

---

<sup>1</sup> In all predictions that follow, the location parameter of the gamma distribution is integrated uniformly from 0.1 to 10.0, the scale parameter is integrated uniformly from 1 to 21, and the causal strength is integrated uniformly from 0.1 to 0.9.

Gopnik, 2004). This is poorly explained by contingency models of causal learning, which generally require many trials to distinguish the probability of the effect in the presence of the cause from the probability of the effect in its absence. In contrast, the event-based chain model predicts that causes can be discovered in a very small number of trials, provided that there is a short delay between cause and effect. Experiment 1 investigates the role of temporal delay in causal discovery, examining whether the event-based chain model accounts for the influence of shorter delays on people's causal discoveries from very small samples.

A second phenomenon is the finding that in both causal and associative learning, acquisition is faster with short delays. This is not predicted by models that assume a fixed delay between cause and effect (unless they incorporate a prior favoring short delays), nor by models that count co-occurrences when the effect occurs within a fixed window of time after the cause. It is predicted by rate models (Gallistel, 2002; Griffiths & Tenenbaum, 2005) as well as contingency models in which one integrates over the temporal window that defines a co-occurrence. The event-based chain model predicts this effect by using gamma distributions to model the delay between cause and effect; when the actual distribution is unknown, integrating over all gamma distributions results in short delays being more likely than long delays. In Experiment 1 we test the models' predictions for the influence of delay length on causal discovery when the delays are fixed. In Experiment 2, we examine what happens when people are trained on a particular delay distribution, investigating the extent to which the advantage provided by short delays depends on training.

A third phenomenon is that short delays are not always more indicative of real-world causal relationships than long delays, as exemplified by the Garcia effect. It should be possible to learn that a particular causal relationship has a characteristically long delay. The event-based



chain model's predictions can change depending on initial evidence supporting certain delay distributions over others. In Experiment 2, we test whether the advantage for short delays disappears when people are given training examples that suggest the delay is unlikely to be short, again comparing people's judgments to the model's predictions.

A fourth phenomenon is that people, but not most models, tend to make causal attributions during causal learning. Contingency models generally ignore the presence of alternative causes (except when they are not probabilistically independent of the cause of interest). Rather than attempt to attribute effects to specific causes, these models rather track only whether the cause of interest increases the probability of the effect. In fact, it is generally considered philosophically untenable to attribute the occurrence of the effect on a particular trial to a particular cause (this is called the problem of actual causation). With the large samples required for contingency-based methods to detect causal relations, the influence of alternative causes can safely be ignored, provided they occur independently of the cause of interest. But when one is learning from just a few trials, it becomes extremely important to consider alternative causal explanations for the occurrence of the effect, otherwise one would infer many more causal relations than actually exist in the world. In Experiment 1, we investigate the extent to which the presence of alternative candidate causes impedes the discovery of a primary candidate cause.

Causal attributions are also influenced by temporal factors. Thus far we have suggested that it can be rational to learn a causal relationship from just a single example, provided the candidate cause of interest occurs shortly before the effect. However, it is clearly irrational if there is another event known to be a cause of the effect that also occurs shortly before the effect. Although it is well-known phenomenon from associative learning that new associations are

blocked by a previously known cue, the role of temporal delay in blocking has not been examined. If the delay between the known cause and the effect is uncharacteristic, then it may be rational to infer a relationship between the candidate cause of interest and the effect, provided the effect follows it with a characteristic delay. The length of the characteristic delay should also influence these judgments. In Experiment 3, we investigate how the learned distribution of temporal delays, and a subsequent observed delay between a known cause, a new candidate cause, and the effect, influence judgments about whether the new candidate is in fact a cause.

Even when no alternative causes are observable, causal attributions can be important during learning, because they will impact one's inferred distribution over temporal delay. Consider a situation in which a cause produces an effect within one second, but only 50% of the time, and the effect also occurs spontaneously every few hours. On the occasions when the cause does not produce the effect, one might attribute a spontaneous occurrence of the effect to the cause, making it seem as though the cause sometimes produces the effect after one second, and sometimes after a delay of several hours. The event-based chain model can resolve such problems by considering all possible assignments of each effect occurrence to a specific cause occurrence or to the background (spontaneous occurrence). When this is done, assignments in which a cause produces an effect after a very long delay will naturally have very low likelihood, much less than assignments in which the effect spontaneously occurs. By considering all possible assignments, the model gives less weight to unlikely assignments, avoiding the pitfalls that can result from rate models. In Experiment 4, we investigate this issue by examining people's judgments of causal power when the overall rate of the effect is identical across conditions but the individual delays between cause and effect vary.

## 3.1 Experiment 1: causal discovery

Hume observed that when one event reliably follows another, people often feel compelled to infer a causal relationship. For instance, suppose you press the button on a remote car unlocker and moments later a nearby car beeps; in such a case, you would likely infer that the button controls the car. This is partly because the beep closely followed the button-press, but also because you already know that buttons on remote car unlockers often cause cars to beep. Now suppose you press the same button and moments later your mobile phone rings; in this case, you would likely consider the events to be unrelated, unless you try it several more times and each time the mobile phone rings just after the button is pressed. If you begin to believe that the button controls your phone, you have made a “causal discovery”. We reserve the term “causal discovery” for cases when there is no reason to suspect a causal relationship prior observing the events. In Experiment 1, we investigate the process of human causal discovery. Two factors are examined: (1) the length of the temporal delay between cause and effect, and (2) the number of alternative causes present that could alternatively explain the occurrence of the effect.

### 3.1.1 Method

A computer animation was presented to participants on a web site. In the animation, participants saw a fishing pond with two fishermen. The top half of the pond was covered by a cloud (see Figure 7). In some conditions they also saw a scuba diver in the pond and planes above the pond (see Figure 7(b)). The two fishermen’s lures were shown pseudo-randomly going in and out of the water. Participants were told that mysterious explosions were happening in the pond, and their job was to determine predict when the explosions would occur. One of the fisherman was, unbeknownst to the participants, using “explosive” lures. Half of the displayed explosions were timed to coincide with the entry of an “explosive” lure into the water.

Explosions occurred throughout the test phase of the experiment (see Figure 7(c)). The other half of the displayed explosions happened pseudo-randomly, but never while the explosive lure was in the water. Participants received points for every prediction made, according to how close their prediction came to the actual explosion time. After they made 10 predictions, they were asked to give a free response answer to the question: “what do you think is causing the explosions.”

#### *3.1.1.1 Participants*

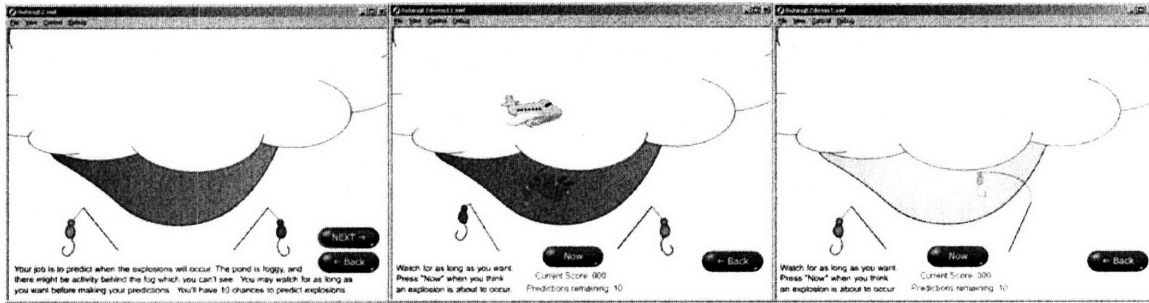
The 222 participants were MIT undergraduates, recruited by bulk email sent to their dormitory. They were compensated with a \$3 payment via the MIT debit account system.

#### *3.1.1.2 Materials*

The computer animation was developed using Macromedia® Flash® 8.0. Participants accessed the animation by clicking on a URL contained in an email solicitation. Prior to the animation, participants were asked to agree to a consent form, and following the animation they were debriefed.

#### *3.1.1.3 Design*

The experiment had a between-subjects design consisting of 12 conditions. Two factors were crossed: (1) the temporal delay between the entry of an “explosive” lure into the water and the explosions (0.1sec, 1sec, 5sec, 10sec), and (2) the number of observable alternative causes (0, 1, or 2). In the condition with 1 alternative cause, participants saw a scuba diver circling within the pond, while in the condition with 2 alternative causes participants saw planes flying above the pond in addition to the scuba diver.



(a)

(b)

(c)

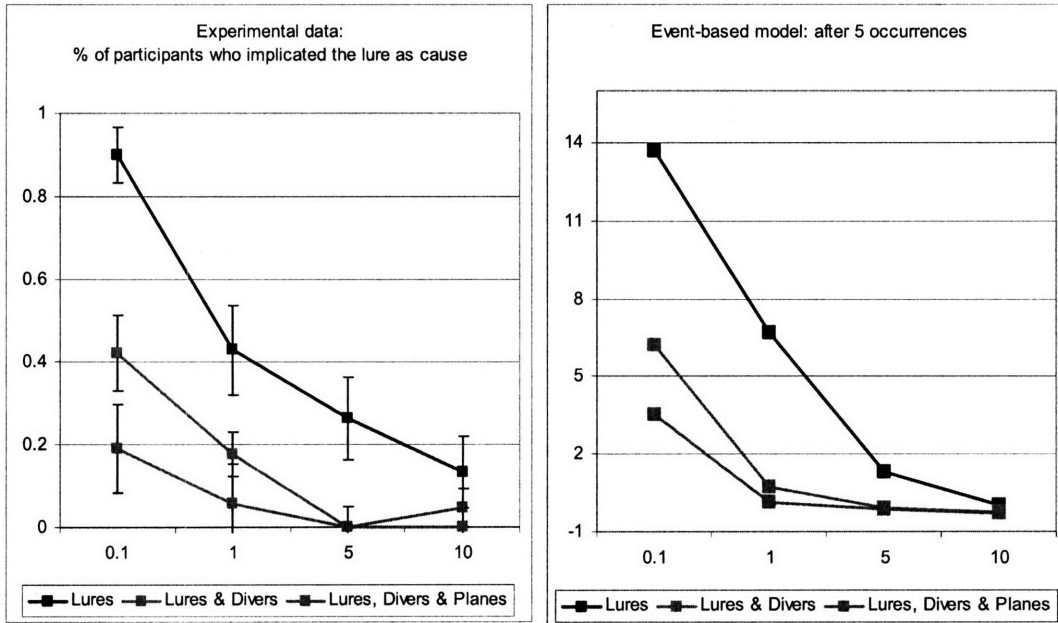
**Figure 7: Experiment 1 stimuli. (a) Two fisherman are shown at a pond. No alternative causes are present. Participants are instructed to watch the activity, and press the “now” button when they think an explosion is about to occur. (b) The left lure is being cast into the water, while the right lure is stationary. Two alternative causes are shown. (c) The explosive lure enters the water, and an explosion occurs shortly thereafter.**

### 3.1.2 Results

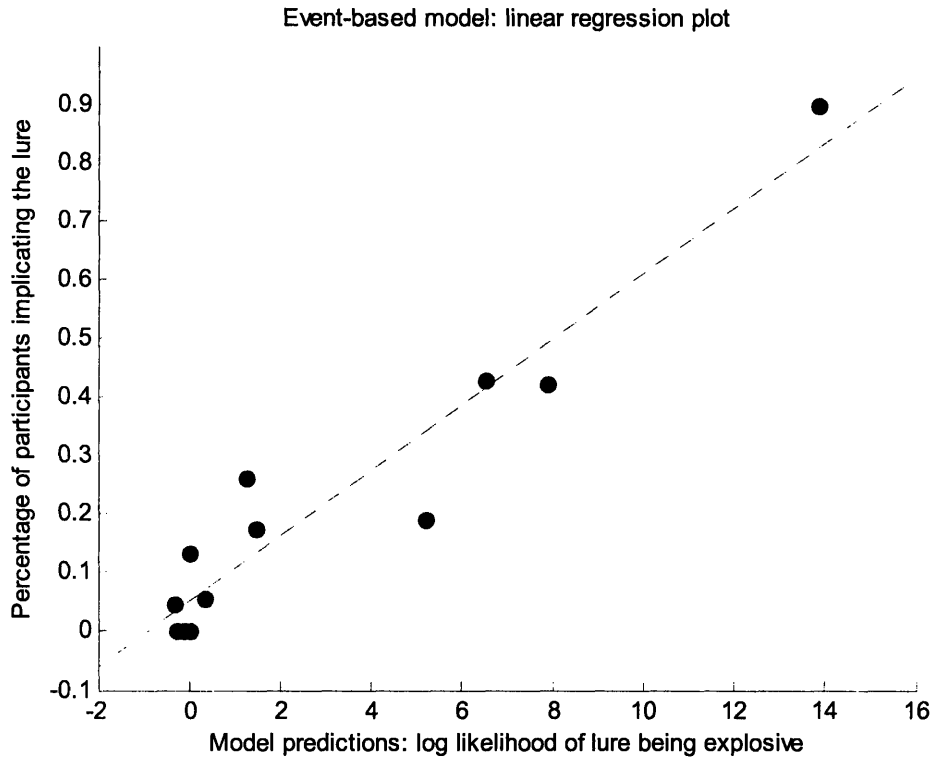
The free response answers were analyzed by two independent raters, who coded them according to whether they implicated the lure as a cause of the explosions. Responses that implicated causes in addition to the lure were treated as implicating the lure (e.g., “An explosion is caused every time the green fishhook is put in the water and also sometimes when the diver swims by and the red fishhook is in the water.”). However, responses that implicated a combination of events, including the lure, as a cause for an explosion were not treated as implicating the lure (e.g., “An explosion occurs if the green hook is in the water and the diver leaves to the right.”).

For the model predictions, the location parameter of the gamma distribution was varied from 0.1 to 10.0 (the range of explosion delays), the shape parameter from 1 to 21, and the background rate of the alternative causes from once every 20 seconds to once every 5 seconds. The results show that the percentage of participants implicating the lure as a cause is highly

correlated with the model's computation of evidential support for the lure being a cause after 5 occurrences of the lure entry followed by the effect ( $r^2=0.91$ ,  $p<.001$ ; see Figure 8 and Figure 9).



**Figure 8: human judgments and model predictions for discovering the lure as a cause of explosions. Three conditions are shown: the lures alone, the lures with divers, and the lures with divers and planes. The model predictions show the evidential support (as a log likelihood ratio) for the lure having explosive power. The experimental data show the percentage of participants who implicated the lure as a cause of the explosions.**

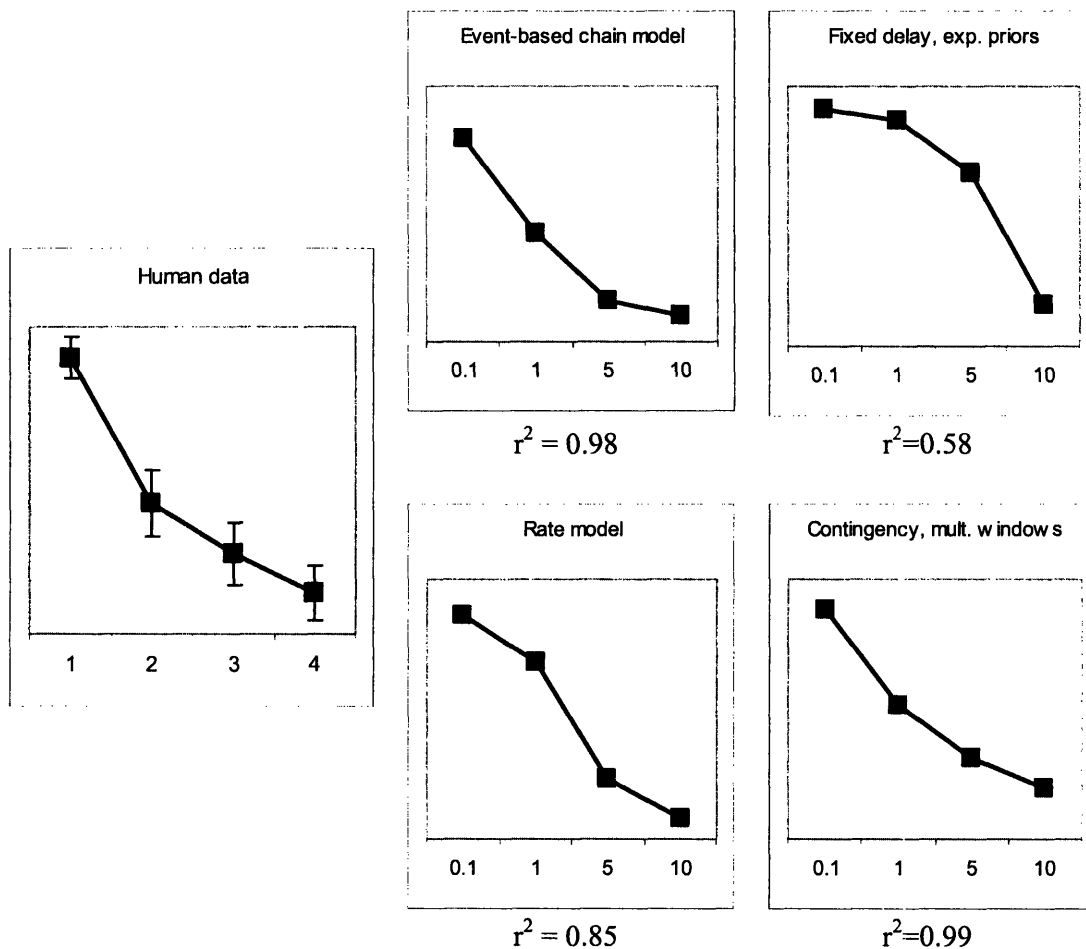


**Figure 9: Scatter plot and linear regression of event-based chain model predictions and human judgments. The correlation is highly significant ( $r^2=0.91$ ,  $p<.001$ ).**

### 3.1.3 Discussion

The results demonstrate that people’s ability to discover the lure as a cause of the explosions was influenced both by the length of the delay and the number of alternative causes. The event-based chain model’s predictions show a similar pattern to the human judgments, and there is a very strong correlation between them ( $r^2=0.91$ ). Of the other temporal models presented earlier, none include the presence of alternative causes as a factor. They cannot account for the influence of alternative candidate causes on causal discovery because they do not make causal attributions during the learning process.

To compare previous models' predictions to the event-based chain model's, we analyze just the condition with no alternative causes. In these conditions, Griffiths' (2005) fixed delay models can predict the influence of shorter delays on learning only by applying an exponential prior on the delay. However, even with this prior, the correlation between this model's predictions and the experimental data is not as strong as with the event-based chain model ( $r^2=0.65$  vs  $r^2=0.98$ , see Figure 10). The rate model does better ( $r^2=0.87$ ), and the contingency model integrating across multiple windows does very well ( $r^2=0.99$ ).



**Figure 10: Correlations between model predictions and human judgments.**



## 3.2 Experiment 2: causal classification after training

In Experiment 1, people's prior expectations of the delay between an explosion and its cause may have influenced their ability to discover the cause of the explosion. One possible explanation for the preference for short delays is that people expected a short delay between an explosion and its cause. In Experiment 2, we control for people's expectations by training them on the distribution of delays to expect. Participants were told that some fishermen were using explosive lures, and were shown either one or three demonstrations (depending on the condition) of an explosive lure entering the water followed by an explosion. These demonstrations serve as evidence for the distribution over delays that participants should expect from explosive lures. In some conditions the delay was variable, while in others it was fixed. Following the training phase, we tested people's classification of whether a new lure was explosive based on one demonstration of the new lure entering the water followed by an explosion. We made it clear that the explosion could have been caused by a second explosive lure hidden behind a cloud, thus participants had to judge whether the explosion was the result of the new lure or a hidden cause based on the delay between the lure's entry and the subsequent explosion.

Under any temporally sensitive model of causal learning, if the delay between a new lure's entry into the water and a subsequent explosion matches the delay one expects, the lure should be judged more likely to be explosive. But under the event-based chain model, even when the delay matches the expected delay, shorter delays often still provide more support for a causal relationship, because the peak of the inferred distribution is higher for shorter delays. This advantage for short delays is driven entirely by the height of the delay distribution at the observed delay. With sufficient training, the model can learn a distribution with a peak at a longer delay, in which case longer delays should be more indicative of a causal process than

shorter delays. In this experiment, we attempted to influence which delay was most indicative of a causal process by training people on different delay distributions. We then tested whether subsequent delays appeared causal in light of the training.

Four types of delay variability were investigated, each with three different mean delays, for a total of 12 conditions. In the “fixed” type of variability, participants were informed that the lures were precision-manufactured to explode after a specific delay (0.1s, 1.0s, or 5.0s). They were then shown three demonstrations, along with a stopwatch display, of the lure exploding after the specified number of seconds. During the test phase, the stopwatch was no longer displayed, and participants were asked to judge whether a new lure was explosive based on observing the lure enter the water followed by an explosion (which could have been caused by a different lure behind a cloud). The conditions with “fixed” variability enable us to test whether longer delays are inherently more difficult for people to judge. Gibbon et al.’s (1977) SET model predicts that because of Weber’s law, one’s ability to estimate a particular delay can be modeled using a normal distribution in which the variance is proportional to the mean, thus shorter delays are much easier to estimate than longer delays. The extent to which people’s judgments are more confident for a short delay than for a long delay can tell us the extent to which Weber’s law accounts for the short delay advantage.

In the “proportional” type of variability, the standard deviation of the delay during training was proportional to the mean of the delay. Participants who were assigned a higher mean delay were shown greater variability in the training delays. No stopwatches were used during the training or test phase. The proportional variability conditions were intended to be representative of physical causal processes because for a gamma distribution characterizing a Poisson process, the standard deviation of is proportional to the mean.

In the “single” type of variability, participants only observe one training example, and therefore they have no evidence of the variability of the delay. No stopwatches were used during the training or test phase. These conditions enable us to test people’s expectations about the variability of a delay based on only a single example. According to the fixed delay models, a single training example is sufficient to learn the exact delay, thus participants’ judgments should be identical to those of the “fixed” variability conditions. Furthermore, there should be no effect of training delay on participants’ ratings for a new lure whose delay matches the training delay; if the delay matches, the lure should be judged explosive, and if it does not match, it should be judged normal. In contrast, under the event-based chain model shorter delays are naturally more indicative of a causal relation because higher delays tend to come from distributions with larger variability. Hence, judgments should look more like the “proportional” variability conditions than the “fixed” variability conditions. The event-based chain model predicts that shorter training delays will result in a higher likelihood that a new lure whose delay matches the training delay is explosive. It also predicts that a new lure with a short delay in a long training delay condition should be judged more likely explosive than a new lure with a long delay in a short training delay condition.

In the “inverse” type of variability, participants were shown a smaller variability for a longer delay, which is the opposite of the gamma distribution’s natural pattern of variance being proportional to the mean. However, as the shape parameter of the gamma distribution increases, the variance decreases, and with a large enough shape parameter the gamma distribution can be very peaked, even for a long delay. Thus, the event-based chain model is capable of inferring a peaked distribution at a long delay, although it will naturally be less peaked than for a short delay. The rate model and the contingency model will naturally assign higher likelihood to

shorter delays, regardless of the training variability, thus they will be poor at learning a long fixed delay.

### **3.2.1 Method**

A computer animation was presented to participants on a web site(see Figure 11). In the animation, participants saw a fishing pond with fishermen. During the training phase, they were told that some fishermen were using explosive lures, and then were shown either one or three examples (depending on condition) of an explosive lure entering the water followed by an explosion after some delay. After the training phase, participants were shown a new lure at the bottom of the pond, and an animated cloud came over the top of the pond to cover two existing lures. They were then shown one explosion occurring spontaneously, after which the new lure entered the water and a second explosion occurred after some specific delay. They were then asked to judge the likelihood that the new lure was explosive.

#### *3.2.1.1 Participants*

The 279 participants were MIT undergraduates, recruited by bulk email sent to their dormitory. They were compensated with a \$3 payment via the MIT debit account system.

#### *3.2.1.2 Materials*

The computer animation was developed using Macromedia® Flash® 8.0. Participants accessed the animation by clicking on a URL contained in an email solicitation. Prior to the animation, participants were asked to agree to a consent form, and following the animation they were debriefed.

### 3.2.1.3 Design

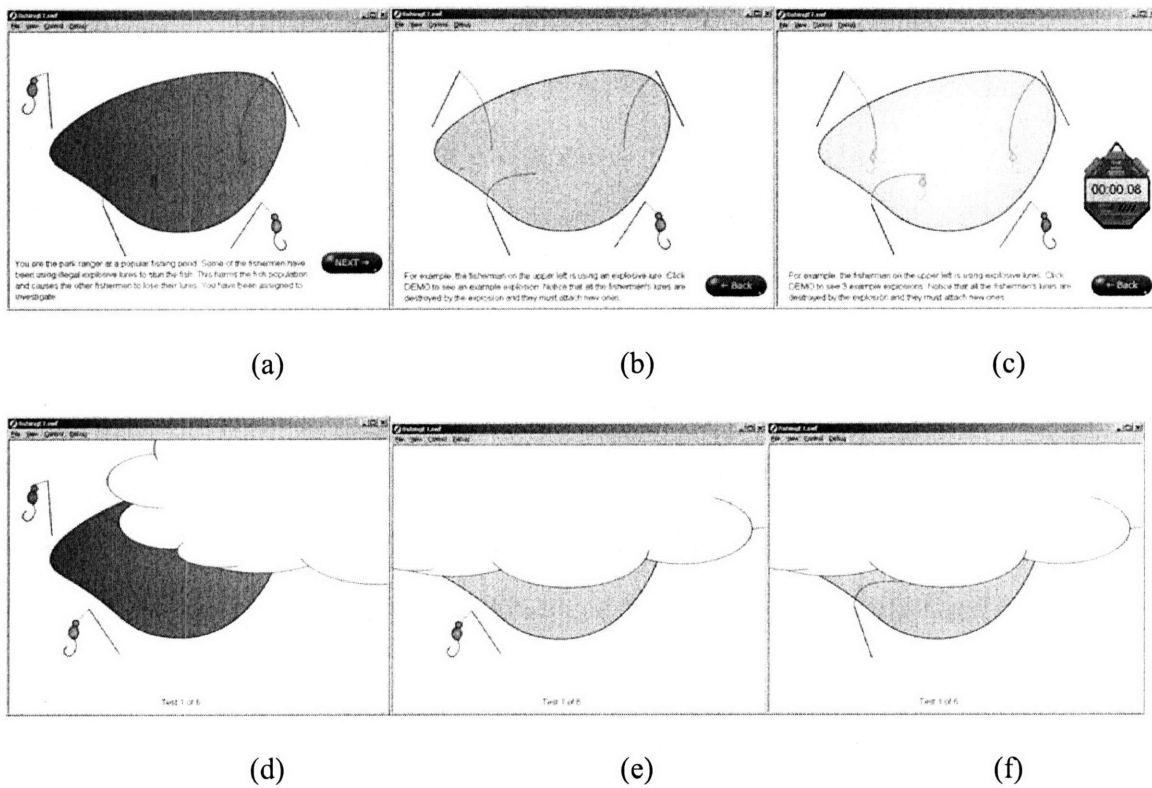
There were twelve total conditions, randomly assigned using a between-subjects design, as two factors were crossed: (1) the mean/minimum delay (0.1sec, 1sec, or 5sec), and (2) the variance of the delay (fixed, variable, inverse, or single). Each condition corresponded to a different distribution of training delays. A summary of the training is shown in Table 2. For the “fixed” distribution, we trained participants that the lures are precision-manufactured to have a fixed delay, and gave them three training examples of that delay. During this training, a stopwatch was displayed on the screen to demonstrate that the delay was identical on each example (see Figure 11(c)). For the “variable” distribution, we provided participants with three training examples, and the standard deviation was proportional to the mean. For the “inverse” distribution, we trained participants on a distribution that has a lower standard deviation for higher means. For the “single” distribution, we trained participants with only a single example, thus they were not provided with any information about the variability.

Variability of delay	Condition	Training examples		
1. Fixed: precision-manufactured (training with stopwatch)	0.1s mean, 0.0s dev	0.1s	0.1s	0.1s
	1.0s mean, 0.0s dev	1.0s	1.0s	1.0s
	5.0s mean, 0.0s dev	5.0s	5.0s	5.0s
2. Variable: standard deviation proportional to mean	0.1s mean, 0.1s dev	0.1s	0.0s	0.2s
	1.0s mean, 0.3s dev	1.0s	0.7s	1.4s
	5.0s mean, 1.7s dev	5.0s	3.5s	7.0s
3. Inverse: low standard deviation for high mean	5.0s mean, 0.1s dev	5.0s	4.9s	5.1s
	3.0s mean, 2.0s dev	1.0s	3.0s	5.1s
	2.6s mean, 2.5s dev	0.1s	2.6s	5.1s
4. Single: single training example	0.1s	0.1s		
	1.0s	1.0s		
	5.0s	5.0s		

**Table 2. Training examples for the 12 conditions of Experiment 2**

After the training phase, participants were shown six test trials. In each test, a new fisherman arrived at the pond, and a cloud came over the pond covering all but the new

fisherman. Then, participants observed a single explosion, presumably caused by a fisherman behind the cloud, after which the new fisherman's lure was cast and a subsequent explosion occurred (see Figure 11(d-f)). The delay between the entry of the new lure and the subsequent explosion was varied on each test trial. The delays tested were 0.1s, 1.0s, 5.0s (in counter-balanced order), followed by 6.0s, 0.5s, and 2.0s (in that order). In each test trial, participants were asked to rate the extent to which they believed the new fisherman was using explosive lures.



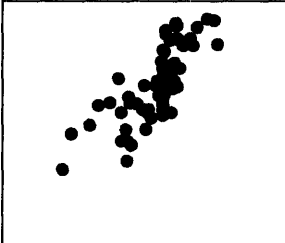
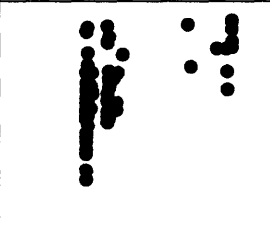
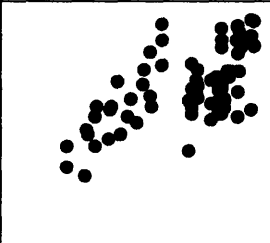
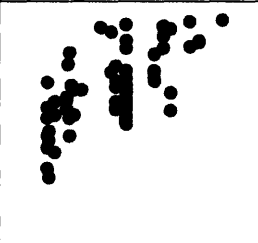
**Figure 11: Experiment 2 stimuli.** The top row depicts the training phase, and the bottom row depicts the test phase. (a) In the training phase, participants see a fishing pond with several fishermen. (b) They then observe either one or three training examples of one lure entering the water and causing an explosion after some delay. (c) In the conditions with a fixed delay, participants see a stopwatch during training so they can verify that the delay is fixed. (d) In each test trial, a new lure is shown at the bottom of the pond, and a cloud comes

- over the existing lures. (e) An explosion then occurs, presumably caused by a hidden lure behind the clouds.  
 (f) The new lure then enters the pond, followed by an explosion after some delay.

### 3.2.2 Results

An ANOVA analysis of the results to all conditions shows a highly significant main effect of variability ( $F(3,1338) = 13.45, p < .0001$ ), a highly significant main effect of the test delay ( $F(5,1338) = 13.02, p < .0001$ ), and a highly significant interaction between the training delay and the test delay ( $F(10,1338) = 13.00, p < .0001$ ). These results suggest that the distribution of delays during the training phase was a strong influence on subsequent judgments during the test phase.

For the model predictions, the location parameter of the gamma distribution was varied from 0.1 to 6.0, the shape parameter from 1 to 21, and the background rate of the alternative cause from once every twenty seconds to once every ten seconds. An analysis of the correlation between participants' judgments and the event-based chain model revealed significant correlations in all conditions. Overall, the correlation between the event-based chain model predictions and human judgments across all conditions was  $r^2 = 0.55$ . The other three models considered all had lower overall correlations with the human judgments, although the contingency model with multiple windows also had a good overall correlation of  $r^2 = 0.53$  (see Figure 12).

Event-based	Fixed delay, exponential prior	Griffiths' rate model	Contingency, multiple windows
			
$r^2 = 0.55$	$r^2 = 0.33$	$r^2 = 0.38$	$r^2 = 0.53$

**Figure 12: Correlations between model predictions and human judgments in experiment 2.**

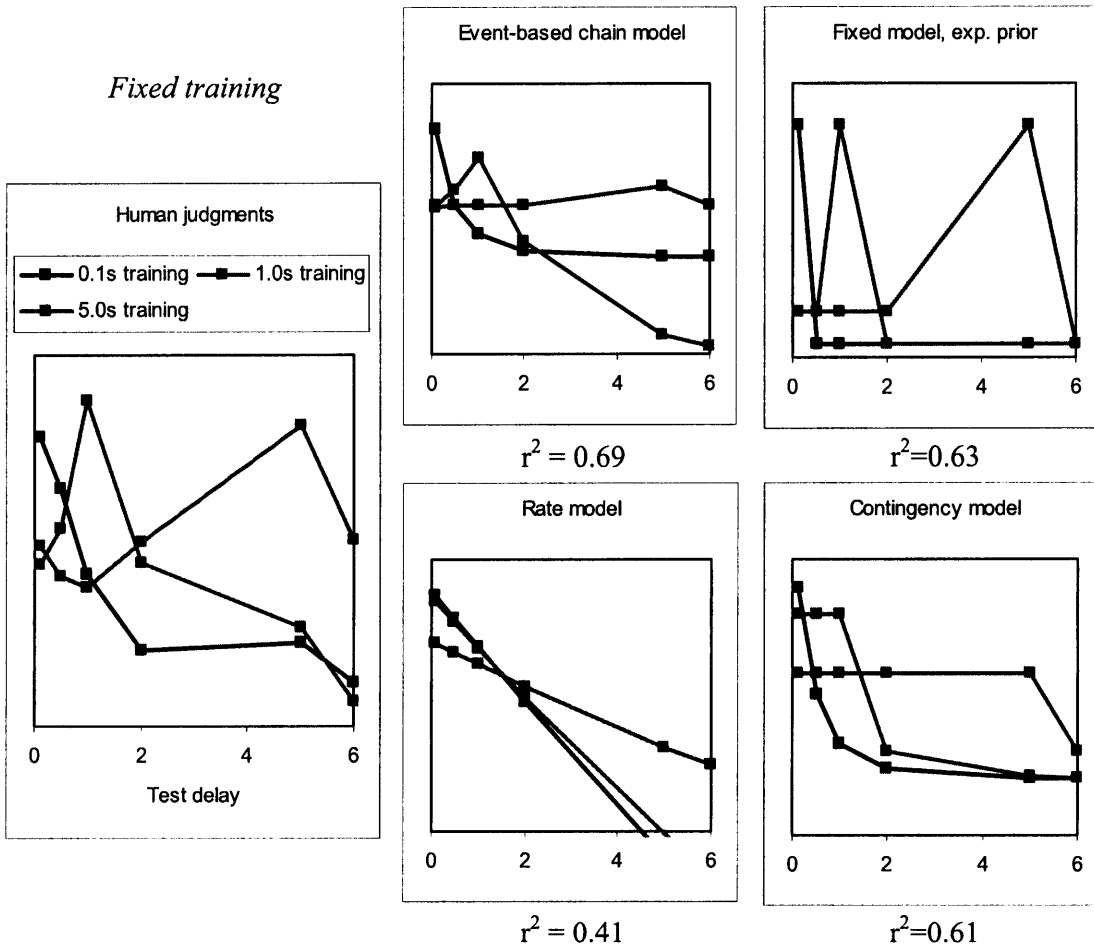
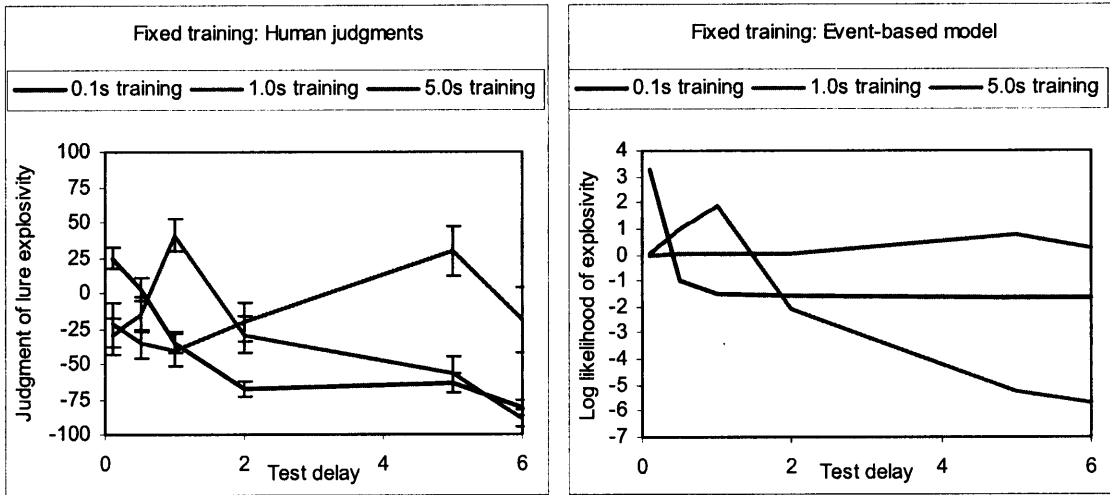
There were 126 participants in the fixed training conditions. An ANOVA found no effect of training delay on peak judgments (when test delay matched training delay). This was expected because the training made it clear that only one specific delay was possible for an explosive lure. The event-based chain model is actually ill-suited to model this task because it assumes that delays are characterized by gamma distributions and the variability increases with the mean delay. Nevertheless, the model fit ( $r^2=0.69$ ) was better than the fixed model ( $r^2=0.63$ ), which is the most appropriate model for this kind of training (see Figure 13).

There were 54 participants in the variable training conditions. An ANOVA revealed a significant difference in peak judgments (when test delay matched training delay) ( $F(2,48)=4.06$ ,  $p<.05$ ), although there was a clear trend in which peak judgments were higher for shorter training delays (see Figure 14). The correlation with the event-based chain model was good ( $r^2=0.57$ ), and was much higher than the fixed delay model with exponential priors ( $r^2=0.31$ ). The correlation coefficient of the rate model ( $r^2=0.63$ ) and the contingency model with multiple windows ( $r^2=0.71$ ) were a bit higher than the event-base chain model.

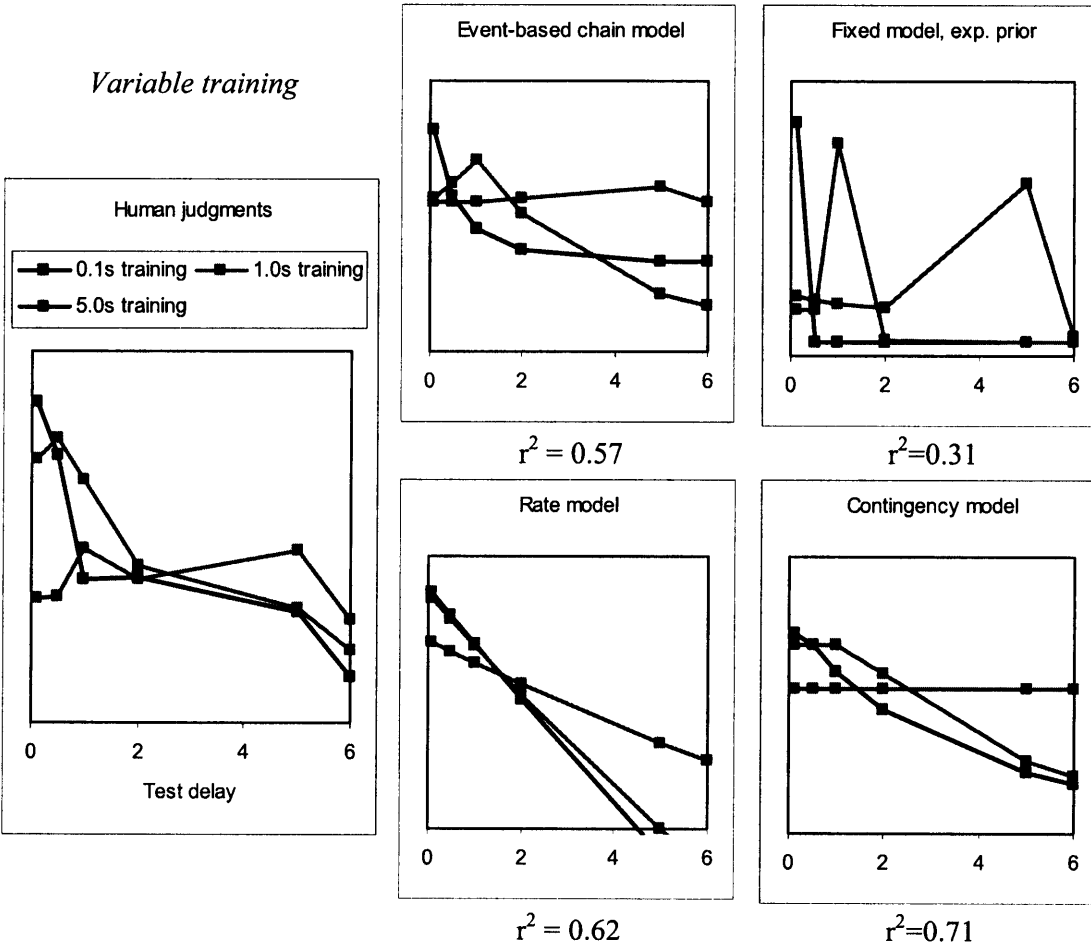
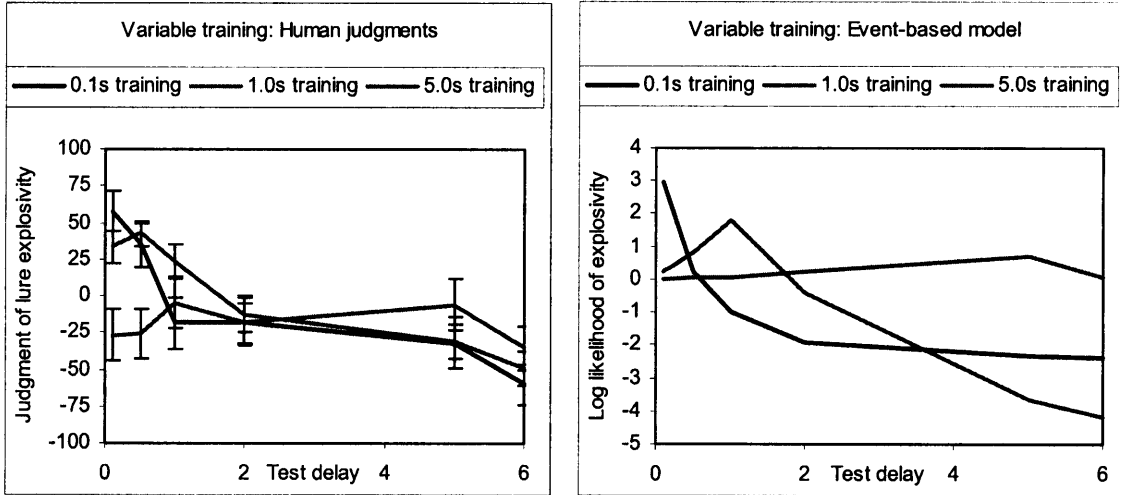
There were 49 participants in the single training example conditions. An ANOVA showed a marginally significant effect of training delay on subsequent peak judgments (when test delay matched training delay) ( $F(2,46)=2.76$ ,  $p=0.07$ ). There was a clear trend in the data for higher peak judgments on shorter training delays, which is predicted by the event-based chain model. The event-based chain model's correlation coefficient was  $r^2=0.65$ , which was much higher than that of the fixed delay model with exponential prior ( $r^2=0.27$ ), but not much higher than that of the Griffith's rate model ( $r^2=0.63$ ) or the contingency model with multiple windows ( $r^2=0.71$ ) (see Figure 15).



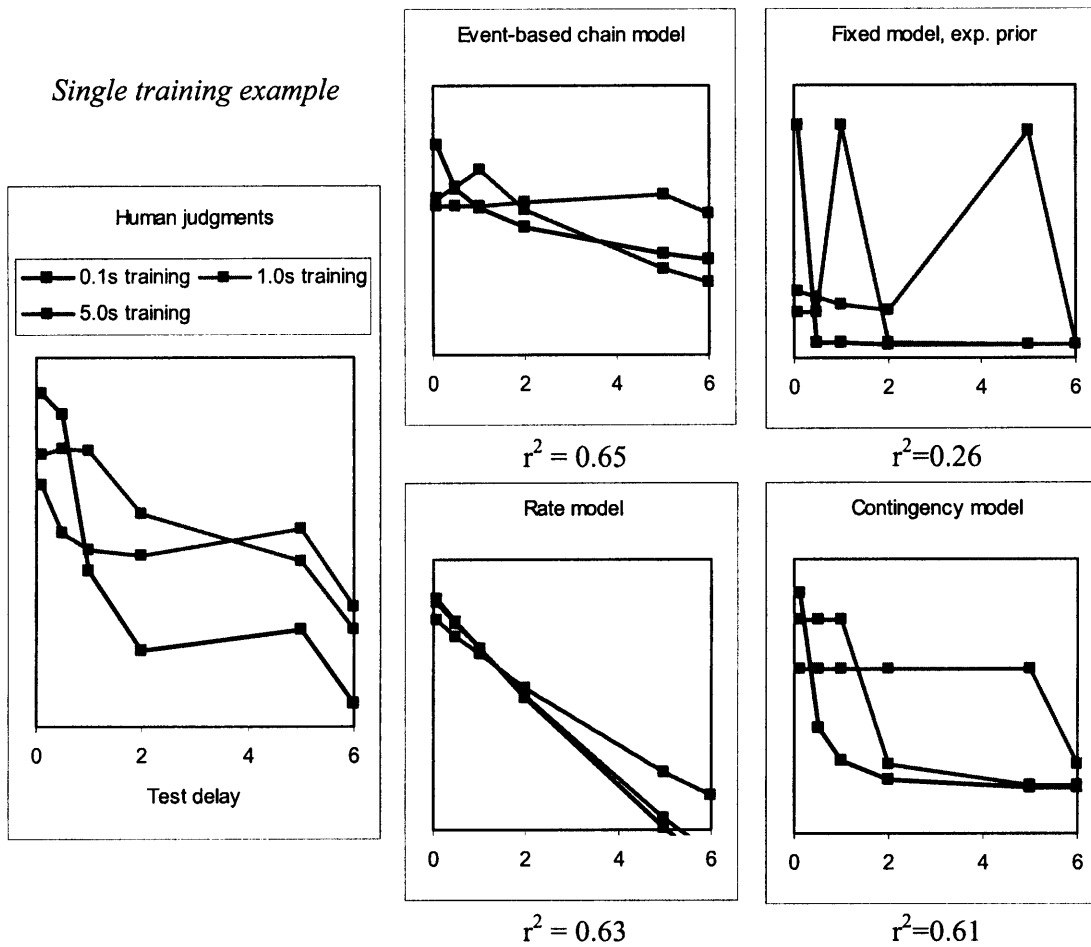
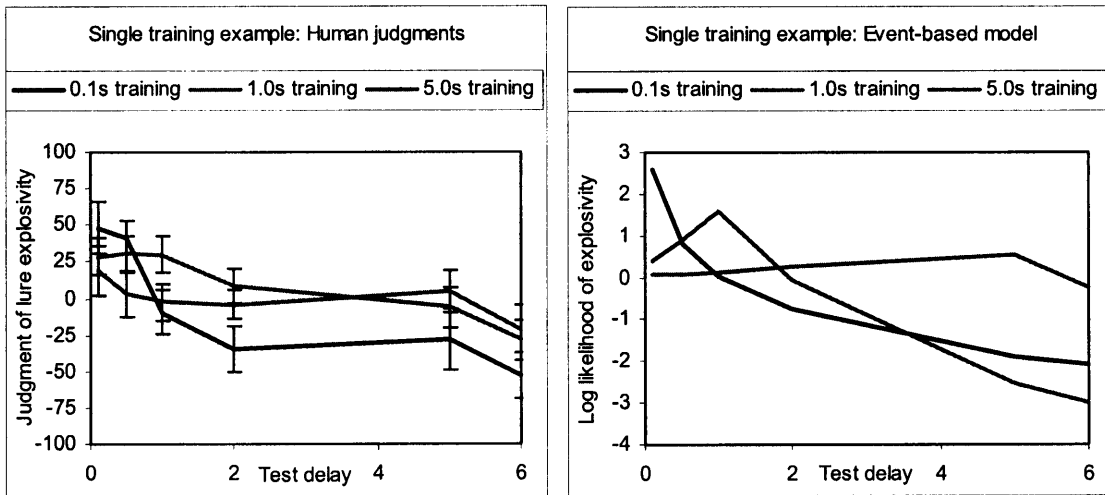
There were 50 participants in the inverse training conditions. An ANOVA revealed a significant effect of training delay on judgments in which the test delay matched the training delay minimum ( $F(2,47)=3.36, p<.05$ ). The only strong judgments in favor of the lure being explosive were made on the 0.1s delay in the condition in which a 0.1s delay appeared, and on the 5.0s delay in the condition in which all training delays were approximately 5.0s. On this condition in particular, participants seem to have learned that the delay was 5.0s with little variability, as their mean judgment of the lure being explosive on the 5.0s test trial was nearly as high as any judgment in the experiment. The correlation coefficient of the event-based chain model was good ( $r^2=0.45$ ), and much higher than that of any other model (see Figure 16).



**Figure 13: Experiment 2, fixed training conditions. A significant correlation was found between human judgments and the event-based chain model predictions ( $r^2=0.69$ ,  $p<.001$ ).**



**Figure 14: Experiment 2, variable training conditions. A significant correlation was found between human judgments and the event-based chain model predictions ( $r^2=0.61$ ,  $p<.001$ ).**



**Figure 15: Experiment 2, single training example conditions. A significant correlation was found between human judgments and the event-based chain model predictions ( $r^2=0.59$ ,  $p<.005$ ).**

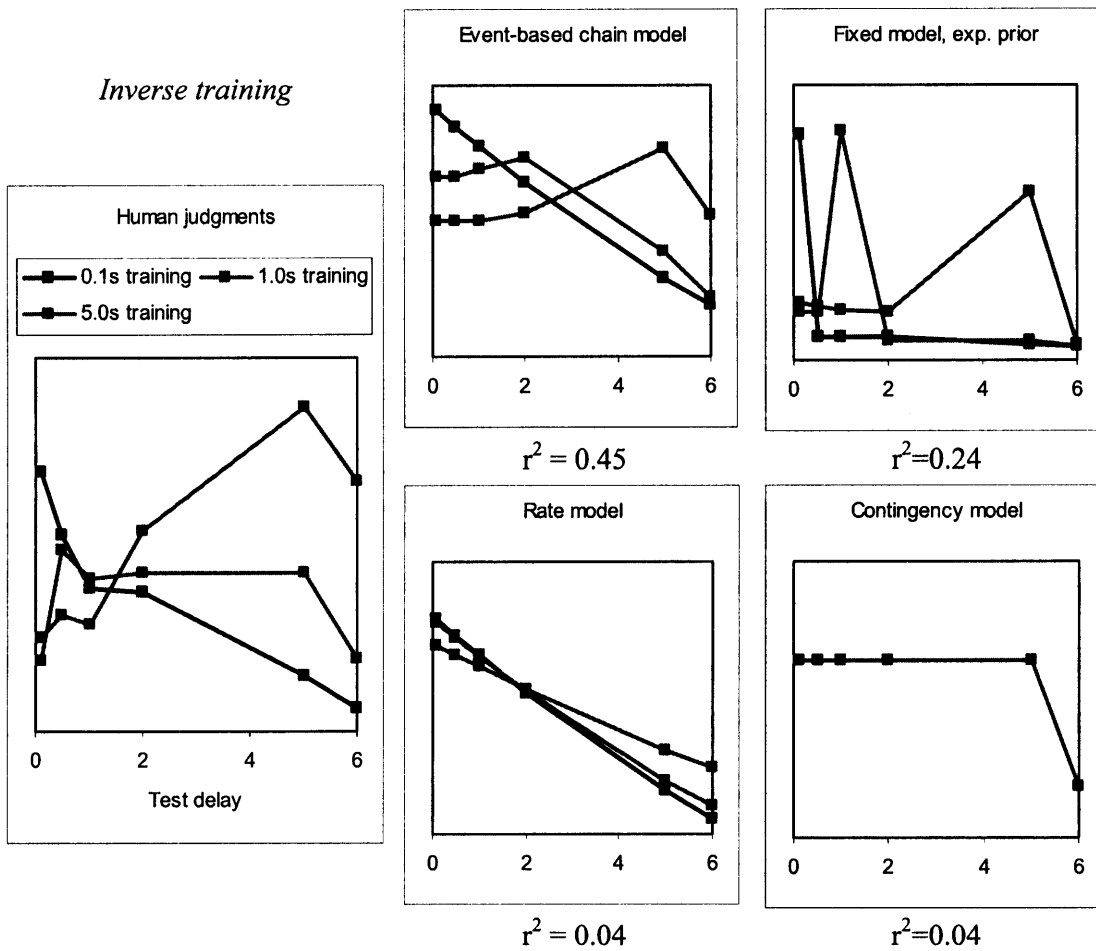
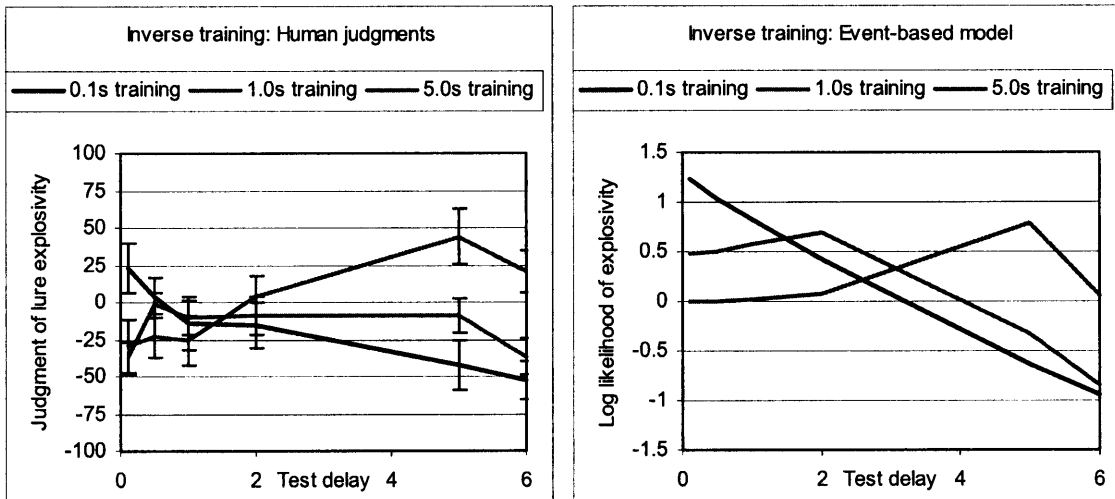


Figure 16: Experiment 2, inverse training conditions. A significant correlation was found between human judgments and the event-based chain model predictions ( $r^2=0.45$ ,  $p<.001$ ).

### 3.2.3 Discussion

In 12 conditions we trained people on different distributions of training delays, and found that their judgments were strongly influenced by these distributions. In the event-based chain model, the evidential support for a lure being explosive is primarily determined by how likely the observed delay is under the delay distribution inferred from prior experience (in this case, training). Overall, the model predictions were correlated with human judgments to a greater degree than competing models.

The single training example conditions provide a clear test of whether the event-based chain model is more appropriate than the fixed model for modeling human causal learning. If people's judgments correspond to the fixed delay model's predictions, then after a single training example they should only treat subsequent delays matching that training example as causal, and the judgment should be the same for all delays. In contrast, if people's judgments correspond to the event-based chain model's predictions, then the peak judgments (where the test delay matches the training) will be higher for shorter delays. The results clearly show that peak judgments were higher for shorter training delays. Furthermore, by examining the fixed training conditions we can see that this pattern is not due to difficulty judging longer temporal delays, as training delay did not influence peak judgments in those conditions.

In the variable delay and single training example conditions, the correlation coefficient of the event-based chain model is not very different from that of the rate model and the contingency model with multiple windows. However, inspection of the graphs (Figure 14 and Figure 15) reveals that the shape of the rate model clearly does not capture the overall pattern of the judgments as well as the event-based chain model. The rate model predicts linearly decreasing judgments, whereas both participant's ratings and the event-based chain model show peaks at or

near the training delay mean. The contingency model with multiple windows also misses the peaks in the data because it predicts monotonically decreasing judgments. Thus, only the event-based chain model has both a strong correlation with the data and a pattern of predictions with peaks at the same locations as the data.

The strongest support for the event-based chain model comes from the “inverse” variability conditions, in which the variance of the distribution of delays was inversely proportional to the mean delay. In those conditions, participants who were trained on 3 examples of a nearly 5-second delay judged a subsequent (5-second) delay to be much more highly indicative of a causal relationship than other subsequent delays. The rate model and the contingency model failed to predict this pattern, because for them, longer delays are always less indicative of causal relationships than shorter delays. Only the event-based chain model and the fixed model predicted that repeated exposure to a fixed long delay would support an inferred distribution that assigns high likelihood to that long delay. Between the two, the event-based chain model had a much better fit to the data ( $r^2=0.45$  vs.  $r^2=0.24$ ), and also had a pattern of peaks that was similar to the data (see Figure 16).

### 3.3 Experiment 3: causal attribution

We have argued that our event-based causal learning model provides a rational statistical basis for inferring the existence of causal relations in just a handful of trials, particularly when the effect follows shortly after the cause. However, this can create problems when a candidate cause coincidentally occurs just prior to an effect that was caused by something else. For instance, suppose someone eats strawberries, then washes their hands with soap, then gets a rash moments later. If the person has never developed such a rash before, they might think the soap caused it. But if the person knows they are allergic to strawberries, they should infer that the rash

was caused by the strawberries, even though the use of soap occurred closer to the onset of the rash.

This issue of potentially confounding causes is important when learning from small samples because there can be many candidate causal events prior to the occurrence of an effect. Previous researchers have proposed models that only address situations in which there is no confounding (e.g., Cheng, 1997; Novick & Cheng, 2004) or that specifically test for potential confounds by examining situations in which they are held constant (e.g., Gopnik & Glymour, 2002). However, both of these solutions require large samples, large enough to either determine that there is no confounding, or to test for contingency when holding potential confounds constant. To illustrate, consider an effect which occurs just once after two potential causes,  $C1$  and  $C2$ , occur. Determining whether  $C2$  is a cause can only be assessed if one can determine whether  $P(E|C1, C2)$  equals  $P(E|C1, \neg C2)$ . However, this probability is not computable, because we have only observed one occurrence of the effect.

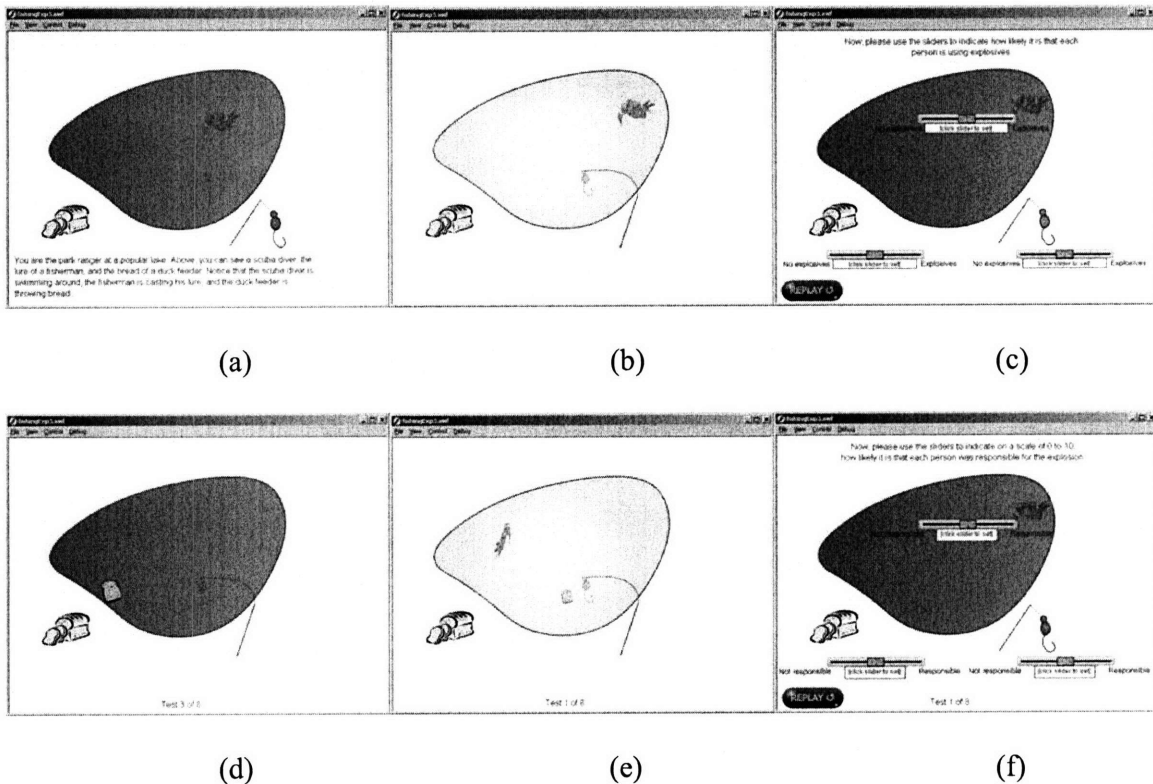
In our event-based chain model, this issue of confounding is handled by the process of causal attribution. During the process of causal learning, for each effect occurrence one must go through all previous cause occurrences, hypothesizing each as a possible cause of that effect. By entertaining all possible attributions of the effect occurrence to various cause occurrences, one can compute the sum of the likelihood of the data over each possible attribution. Contingency models of causal learning are able to avoid making causal attributions during the learning process because they count occasions in which cause and effect co-occurred rather than cases in which the cause actually produced the effect. In our case we have no co-occurrences, only occurrences. Therefore, to determine whether a causal relation exists, we must attribute effect occurrences to cause occurrences, naturally weighting each by its likelihood under the model.



In this experiment, we test people's causal attributions when two candidate causes both precede an effect. During an initial training phase, participants receive evidence that one of the candidate causes is a true cause, while the other may not be. They also learn something about the delay distribution for the true cause. During the test phase, we examine how the delay between the two candidate causes and the effect influence participants causal attributions in eight test trials.

### **3.3.1 Method**

Participants were shown a display with a fishing pond containing a scuba diver, the lure of a fisherman, and the bread of a duck feeder (see Figure 17). They were instructed that people have reported mysterious explosions occurring in the pond, and they were shown a demonstration explosion. In the training phase of the experiment, their task was to watch the activity for 60 seconds, and then decide whether the scuba diver, the fisherman, the duck feeder, or some combination was using explosives (see Figure 17(c)). The activity consisted of the duck feeder's bread entering the water at random times, the fisherman's lure entering the water at random times, and explosions occurring. The timing was such that either the bread or the lure was explosive, and the explosion always happened after some fixed delay from its entry into the water (either 0.1s, 1.0s, or 3.0s). In the test phase of the experiment, they were shown eight additional explosions, each of which was preceded by both the duck feeder's bread and the fisherman's lure entering the water. Early in the testing, the trained explosive object has the training delay and the alternate object has various other delays. Later in the test, the alternate object has the training delay and the trained object has various other delays. In each test case, participants were instructed to determine which of the people (including the scuba diver) caused the explosion (see Figure 17(f)).



**Figure 17: Experiment 3 stimuli.** The top row is the training phase and the bottom row is the test phase. (a) In the training phase, a fisherman, a scuba diver, and a duck feeder are at a popular pond. (b) Participants observe 60 seconds of activity in which either the bread or the lure causes four explosions with a characteristic delay. (c) They are then asked to judge which person is causing the explosions. The test phase consists of eight trials. (d) In each trial both the lure and the bread enter the pond, often at different times. (e) Then an explosion occurs. (f) Participants are then asked to judge which person caused the explosion.

### 3.3.1.1 Participants

The 49 participants were MIT undergraduates, recruited by bulk email sent to their dormitory. They were compensated with a \$3 payment via the MIT debit account system.

### 3.3.1.2 Materials

The computer animation was developed using Macromedia® Flash® 8.0. Participants accessed the animation by clicking on a URL contained in an email solicitation. Prior to the

animation, participants were asked to agree to a consent form, and following the animation they were debriefed.

### 3.3.1.3 *Design*

There were three conditions, given in a between-subjects design. Participants were randomly assigned to one of the three conditions. One factor was varied: the delay between the entry of the explosive item into the water and the subsequent explosions during training (values of 0.1s, 1.0s, and 3.0s). The explosive item (randomized to be either the bread or the lure) entered four times during training, and was always followed by an explosion after the appropriate delay. The non-explosive item entered twice during training, once with a very long delay, and once at the end with no explosion following it. After the training phase, participants were asked to judge which of the three people (including the scuba diver) was using explosives, on a scale of -100 (definitely not using explosives) to +100 (definitely using explosives) (see Figure 17(c)). Following the training phase, there was a test phase with eight tests, each of which depicted both objects entering the water, often at different times, followed by an explosion. After each test the participants were asked to rate how strongly they believed each of the people to be causing the explosion, on a scale of 0 to 100 (see Figure 17(f)). The test phase was organized as depicted in Table 3:

Displayed delay between entry of objects and explosions, for trained object (T) and alternate object (A)		Training delay						
		0.1s		1.0s		3.0s		
			<i>T</i>	<i>A</i>	<i>T</i>	<i>A</i>	<i>T</i>	<i>A</i>
Test type	Both items have training delay.	<i>TA</i>	<i>0.1s</i>	<i>0.1s</i>	<i>1.0s</i>	<i>1.0s</i>	<i>3.0s</i>	<i>3.0s</i>
	Trained object has training delay. (order randomized)	<i>TA<sup>-2</sup></i>					<i>3.0s</i>	<i>0.1s</i>
		<i>TA<sup>-1</sup></i>	<i>0.1s</i>	<i>0.0s</i>	<i>1.0s</i>	<i>0.1s</i>	<i>3.0s</i>	<i>1.0s</i>
		<i>TA<sup>+1</sup></i>	<i>0.1s</i>	<i>1.0s</i>	<i>1.0s</i>	<i>3.0s</i>	<i>3.0s</i>	<i>6.0s</i>
		<i>TA<sup>+2</sup></i>	<i>0.1s</i>	<i>3.0s</i>	<i>1.0s</i>	<i>6.0s</i>		
	Alternate object has training delay (order randomized)	<i>T<sup>-2</sup>A</i>					<i>0.1s</i>	<i>3.0s</i>
		<i>T<sup>-1</sup>A</i>	<i>0.0s</i>	<i>0.1s</i>	<i>0.1s</i>	<i>1.0s</i>	<i>1.0s</i>	<i>3.0s</i>
		<i>T<sup>+1</sup>A</i>	<i>1.0s</i>	<i>0.1s</i>	<i>3.0s</i>	<i>1.0s</i>	<i>6.0s</i>	<i>3.0s</i>
		<i>T<sup>+2</sup>A</i>	<i>3.0s</i>	<i>0.1s</i>	<i>6.0s</i>	<i>1.0s</i>		
	Max vs. min	<i>T<sup>+1</sup>A</i>	<i>6.0s</i>	<i>0.1s</i>	<i>6.0s</i>	<i>0.1s</i>	<i>6.0s</i>	<i>0.1s</i>

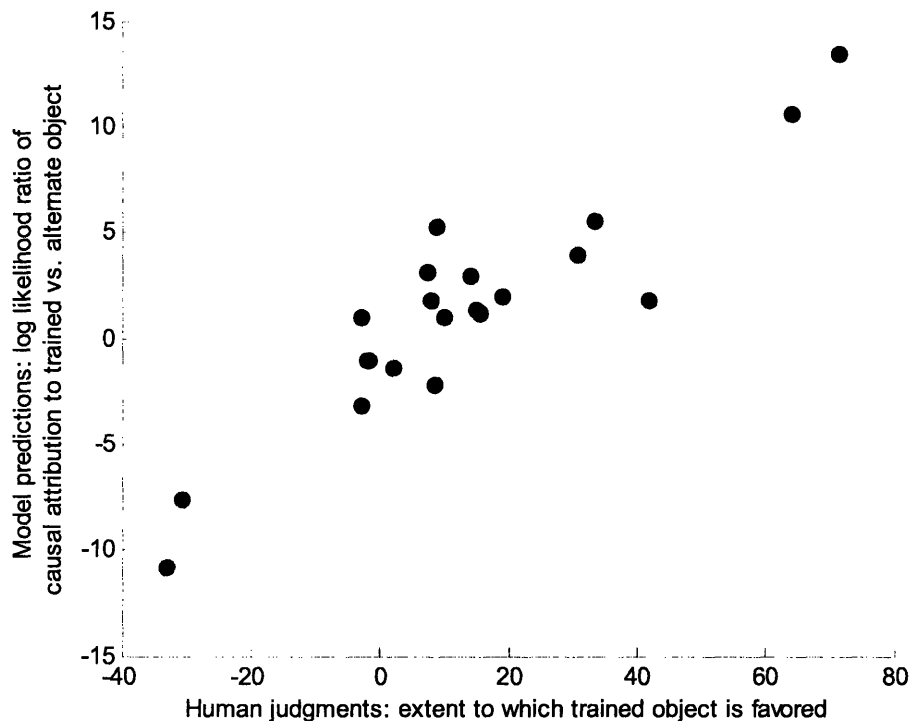
Table 3: the eight tests for experiment 3.

### 3.3.2 Results

The results indicate that overall, participants' causal attributions were influenced by the how closely the temporal delay in the test cases matched the training delays. Judgments were coded by taking the response to the trained object (on a scale of 0 to 100) and subtracting the response to the alternate object (on a scale of 0 to 100) to obtain the extent to which the trained object was favored on each judgment, on a scale from -100 to 100. An ANOVA revealed a significant effect of delay on judgments of which object was explosive after training ( $F(2,46) = 4.42, p < .05$ ), with judgments highest for the explosive objects when the delay was shortest. There was no effect of training delay on judgments of which item caused the explosion on the first test, when both items entered the water at the same time and the explosion occurred after the training delay. Regardless of training delay, judgments were 15 to 19 points higher for the trained object than for the alternate object on this test. The ANOVA also revealed a significant effect of training delay on the final test (trained object having 6.0s delay, alternate object having

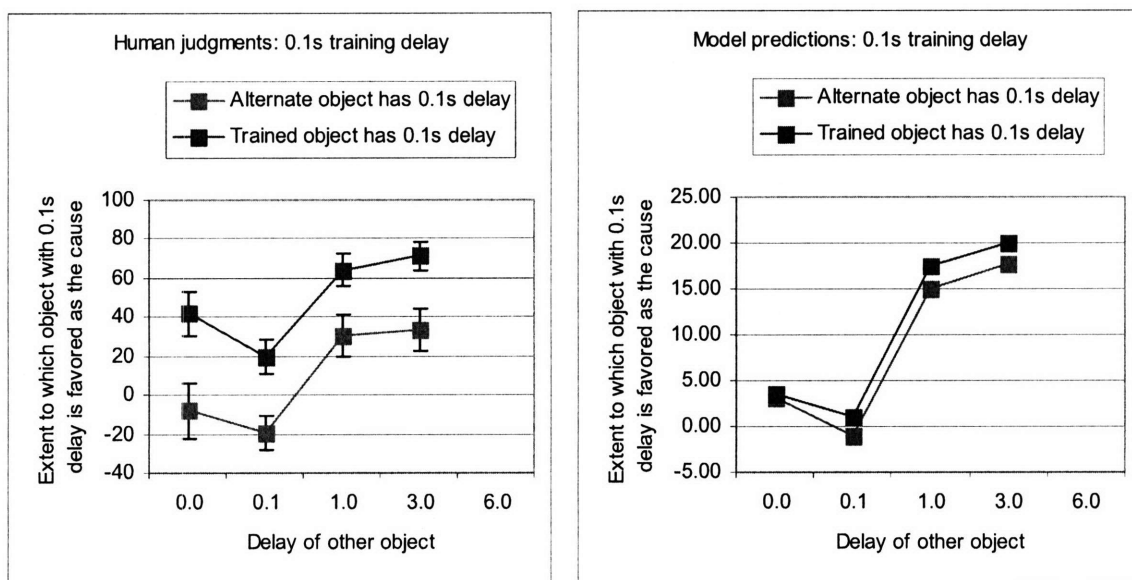
0.1s delay) with judgments favoring the alternate object significantly more for the shorter training delay ( $F(2,41)=3.61, p<0.05$ ). No other tests were directly comparable via ANOVA because the test delays depended on the training delay. However, they are included in our analysis of model predictions.

The event-based chain model predictions were computed by taking the log likelihood ratio of the probability that the effect was caused by the trained object to the probability that the effect was caused by the alternate object. The location parameter of the gamma distribution characterizing the temporal delay was varied from 0.1 to 6.0, while the shape parameter was varied from 1 to 21. A comparison of people's judgments to the event-based chain model predictions revealed a strong overall correlation ( $r^2=0.88, p<.001$ , see Figure 18).

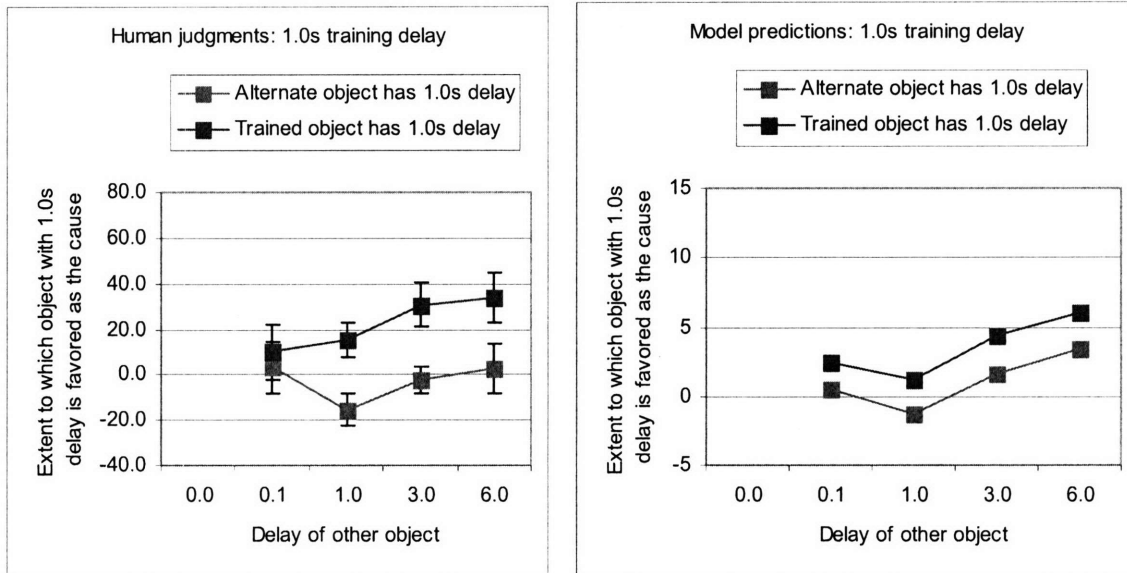


**Figure 18: Experiment 3: Correlation between human judgments and event-based chain model predictions. The correlation is significant ( $r^2=0.88, p<.001$ ).**

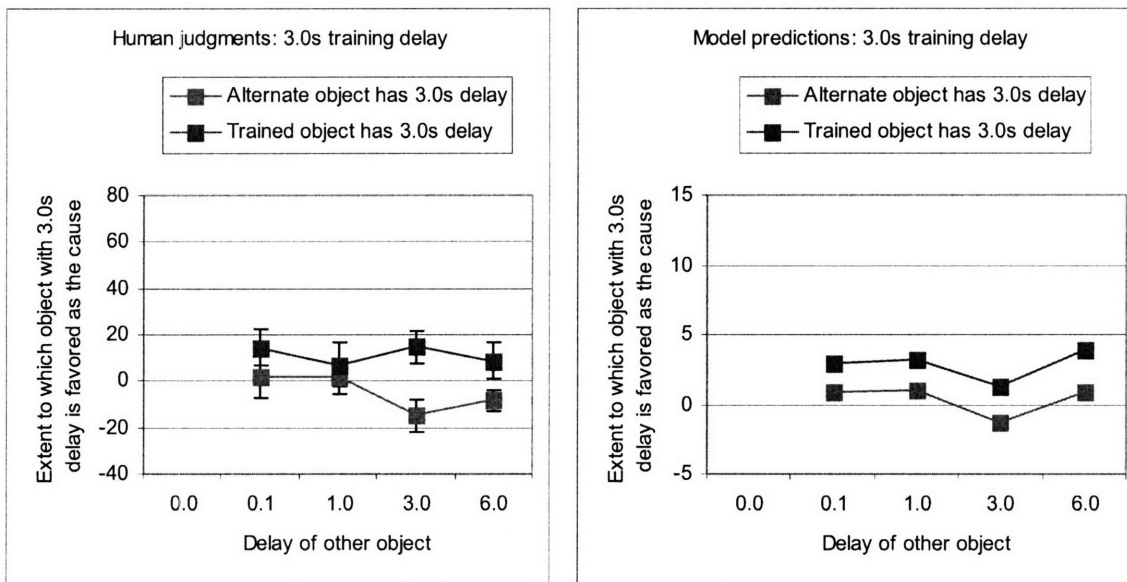
The plots below show the human causal attributions alongside the event-based chain model predictions for each of the training delays. The vertical axis represents the extent to which the object whose delay matched the training delay was favored as the cause of the observed explosion. The blue line represents trials on which the delay between the trained object's entry into the pond and the explosion matched the training delay, while the delay between the alternate object's entry and the explosion was varied. The red line represents trials on which the delay between the alternate object's entry and the explosion matched the training delay, while the delay between the alternate object's entry and the explosion was varied.



**Figure 19: Experiment 3 results for training delay of 0.1s. Results show attributions favoring the object having the training delay as the delay of the other object varies.**

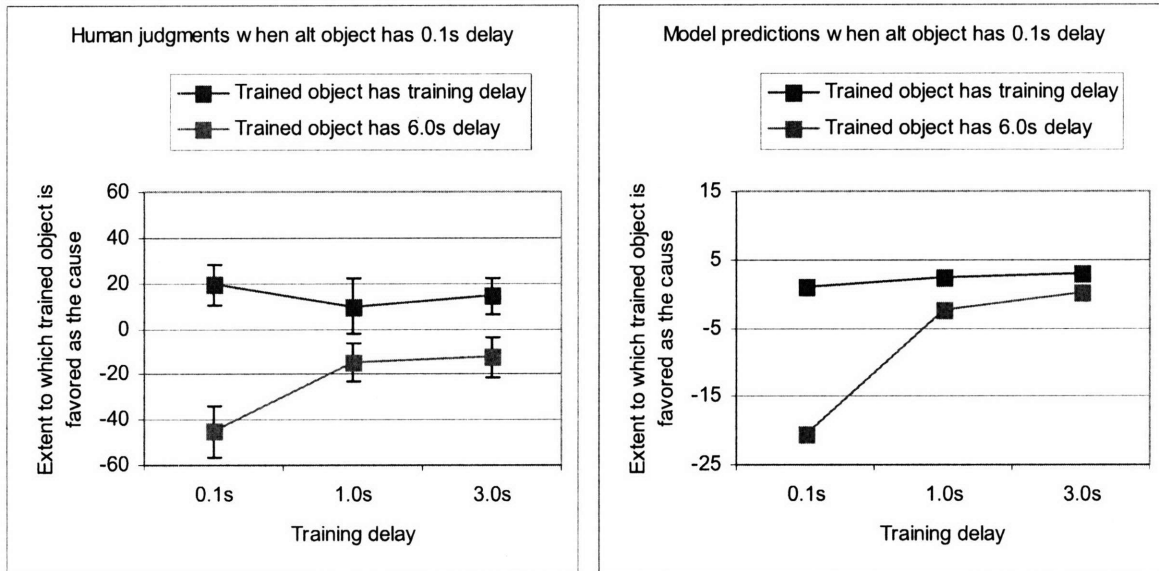


**Figure 20: Experiment 3 results for training delay of 1.0s. Results show attributions favoring the object having the training delay as the delay of the other object varies.**



**Figure 21: Experiment 3 results for training delay of 3.0s. Results show attributions favoring the object having the training delay as the delay of the other object varies.**

We also obtained results for cases on which the alternate object entered the pond with a very short 0.1s delay before the explosion. On these test cases, participants were significantly less likely to say the object was the cause of the explosion when the trained object entered with the trained delay than when the trained object entered with a delay of 6 seconds (see Figure 22, ANOVA:  $F(1,89) = 23.14, p < .001$ ). Overall, participants rated the trained object as more likely than the alternate object when the trained object entered the water with the trained delay and the alternate object entered the water with 0.1s delay (mean advantage for trained object was 15.2 points). However, participants rated the alternate object as more likely than the trained object when the trained object had a 6.0s delay and the alternate object had a 0.1s delay (mean advantage for alternate object was 21.7 points). In fact, there was no effect of training delay on these judgments when the trained object occurred with the trained delay. For all conditions, the trained object had an advantage of 10-20 points. The event-based chain model predictions were highly correlated with these judgments ( $r^2=0.90, p < .005$ ).



**Figure 22: Experiment 3 results for cases in which the alternate object has a 0.1s delay. Results are shown for cases in which the trained object has the training delay (blue line), and for cases in which the trained object**



has a 6.0s delay (red line). There is a strong correlation between model predictions and human judgments ( $r^2=0.80$ ,  $p<.005$ ).

### 3.3.3 Discussion

Overall, the event-based chain model was a strong predictor of human causal attributions. In the event-based chain model, causal attributions are a fundamental process that occurs during learning, as hypothesized distributions are evaluated according to whether they predict certain hypothetical assignments of effect occurrences to specific cause occurrences. In contrast, no other computational model considered thus far makes predictions for the influence of delay on causal attributions.

The trials in which the alternate object entered the pond 0.1s before the explosion are particularly interesting. Participants generally did not attribute the explosion to the alternate object when the trained object entered with its typical delay. This suggests that people's use of short temporal delays as a cue to causation depends on whether there is a previously known cause present that occurs with its expected delay. While short temporal delays are clearly an important cue to causation, the event-based chain model accurately predicts that this cue is modulated by the presence of other causes that can explain the effect.

## 3.4 Experiment 4: rates and strengths

Although our model accounted for a number of important trends in the data in Experiments 1 through 3, the rate model also provides a good fit to much of the data. This is in part because, like the event-based chain model, it provides a natural advantage for short delays. However, it cannot handle a situation in which the causal strength of a cause is small (i.e., the probability of any given cause occurrence producing an effect occurrence is small). Consider the case of a medication which produces migraines in one percent of patients. Even if the rate of

migraines in the patients who get them is high, the other ninety-nine percent of patients who do not develop migraines will effectively reduce the estimated rate to the point where the rate may not be significantly different statistically from chance.

In Experiment 4, we investigate whether people's judgments deviate from the rate model when the causal strength is less than one. Specifically, we test a situation in which the total rate of the effect in the presence of the cause is the same across all conditions. This is achieved by making the cause present for the same amount of time in each condition, and making the number of effect occurrences the same. The major difference between the conditions is how soon the effect occurs after the cause on some proportion of the trials.

### **3.4.1 Method**

Participants were shown a fishing pond, a fisherman's lure, and a scuba diver swimming in the pond. They were instructed that the scuba diver is using explosives to destroy poisonous coral that has infested the pond. However, the park trustees suspect that some of the fisherman's lures might also be explosive. Participants were then shown activity for 60 seconds, after which they were asked to judge whether the fisherman was using explosive lures, and if so, what percentage were explosive.

#### *3.4.1.1 Participants*

The 50 participants were MIT undergraduates, recruited by bulk email sent to their dormitory. They were compensated with a \$3 payment via the MIT debit account system.

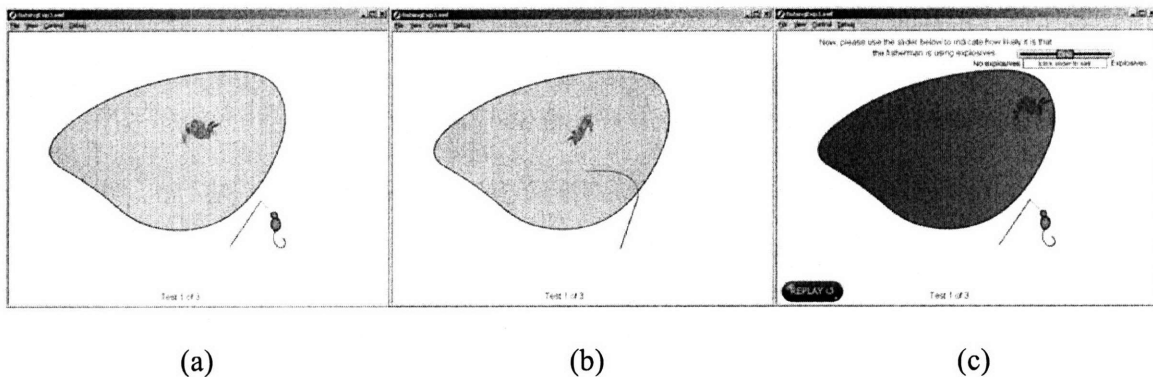
#### *3.4.1.2 Materials*

The computer animation was developed using Macromedia® Flash® 8.0. Participants accessed the animation by clicking on a URL contained in an email solicitation. Prior to the

animation, participants were asked to agree to a consent form, and following the animation they were debriefed.

### 3.4.1.3 Design

There were 9 total conditions as two factors were crossed: the causal strength of the lure (40%, 60%, or 80%) and the temporal delay between entry of the lure and the explosion on those trials when the lure produced an explosion (0.1s, 1.0s, or 3.0s). Each participant was shown 3 blocks, each having a different causal strength value and a different delay value. The order was counter-balanced across participants. Five examples of the lure entering the water were shown, each followed by either the temporal delay (in those cases where the lure produced the explosion) or by a very long delay (in those cases where the lure failed to produce an explosion and the scuba diver later caused the explosion). The total time that the lure was in the water was the same across all conditions, as was the total number of explosions.

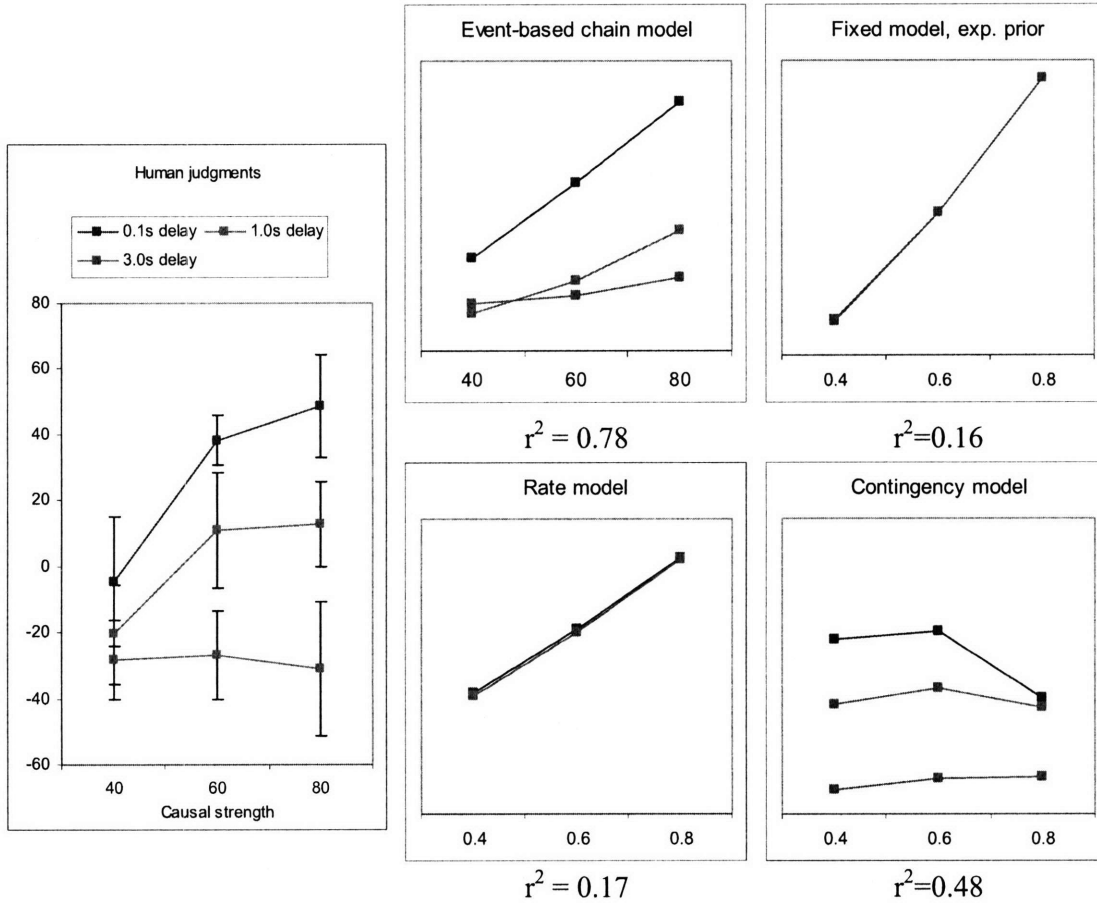


**Figure 23: Experiment 4 stimuli. (a) explosions sometimes occur with just the scuba diver in the pond. (b) explosions sometimes occur after the lure enters the pond. (c) participants judge whether the lure is explosive.**

### 3.4.2 Results

An ANOVA revealed significant main effects of both delay ( $F(2,140) = 10.82, p < .001$ ) and causal strength ( $F(2,140) = 3.37, p < .05$ ) on judgments of whether the fisherman was using

explosive lures. The results were strongly correlated with the event-based chain model's predictions ( $r^2=0.78$ ,  $p<.001$ ). Other models did not correlate as strongly (see Figure 24).



**Figure 24. Experiment 4 results and model predictions. An ANOVA showed significant main effects of delay and causal strength. There was a strong correlation of human judgments to model predictions ( $r^2=0.78$ ).**

### 3.4.3 Discussion

Participants were much more likely to believe the fisherman was using explosive lures when the delays were short, even though the total amount of time in the water and the total number of explosions was identical in all conditions. Because the overall rate of the effect in the

presence of the cause was the same in all conditions, the rate model predicts no effect of delay on causal learning in this experiment. However, the results clearly show a large effect of delay, which strongly suggest that the rate model is insufficient to capture people's judgments.

## 4 General discussion

It has been known since the earliest days of classical conditioning that shorter temporal delays between cue and outcome resulted in faster learning. However, the typical explanation from the associative learning literature has implicated biological constraints. We have proposed a new computational model based on a rational analysis of causal learning from temporal data. This framework explains the short delay advantage (that shorter delays result in faster acquisition), by appealing to the fact that under a generative model of event causation as a series of Poisson processes, shorter delays provide more evidential support for a causal relationship.

We tested the event-based chain model's predictions in four experiments designed to explore different aspects of causal learning from temporal data. In Experiment 1, we showed that people's ability to discover new causes could be predicted by the length of the delay between cause and effect and by the number of alternative candidate causes present. No other model predicts an effect of the presence of alternative causes on the discovery of a new cause, and no other model correlates as well with human judgments as the event-based chain model ( $r^2=0.91$ ). In Experiment 2, we showed that by training people on the distribution of delay to expect we can influence their judgments about the degree to which a particular delay is indicative of a causal relationship. Most interestingly, when trained on a long delay with small variance, participants judged long delays that matched training to be more indicative of a causal relationship than short delays. Most other models predict only that short delays are better indicators of causal

relationships, but the event-based chain model accurately predicts that with sufficient training, long delays can be better indicators of causal relations than short delays. In Experiment 3, we showed that training about temporal delay influences people's subsequent judgments during causal attribution. Of particular interest is the finding that people only perceive a short delay to indicate a causal relationship when known alternative causes do not occur with their typical delay. In Experiment 4, we showed that in the case of probabilistic causation (causal strength less than 1), people's judgments are not unduly influenced by the long delays that will inevitably occur when a cause fails to produce its effect. The rate model can be affected by long delays, but by using causal attribution during the learning process the event-based chain model accurately predicts that the length of a delay does not matter when the cause fails to produce the effect.

Until now, no theory has provided a rational account for the short delay advantage in the case of event causation. Gallistel's (1990) Rate Estimation Theory and Griffiths & Tenenbaum's (2005) rate model both predict that shorter delays between cue onset and outcome occurrence should result in faster learning in the case where cues are either present or absent for extended periods, but neither predicts an effect of shorter delays on learning rate when both cause and effect are both events. Griffiths' (2005) event model only predicts a short delay advantage if a prior favoring short delays is included, while Gibbon & Gallistel's (2000) Scalar Expectancy Theory, which implicates Weber's law as an explanation for increased error in estimating longer time intervals, only applies to mature responding and therefore does not predict anything about acquisition rate. Furthermore, no previous model accounts as well as the event-based chain model for the combined experimental findings presented in Experiments 1-4.

## 4.1 The event-based framework: beyond the chain model

The event-based chain model provided good fits to the data in many of the experiments, but the event-based framework can be used with other kinds of temporal distribution models as well. This would be desirable for Experiment 2, in which the fixed training conditions informed participants that the lures were precisely timed to explode after a fixed delay. With such training, the model should be able to infer that a fixed delay distribution is most appropriate, rather than the gamma distributions of the chain model. It would be desirable to populate the event-based framework with many such alternative distributions. Ideally, these distributions should be weighted by their representation in the environment, which can be done by assigning different prior probabilities to the different distributions. One could even start with a completely uniform prior, and use data to infer which families of distributions are most common in the environment. This would require extending the hierarchical framework to include a level in which the distribution family is generated.

## 4.2 Object vs. Classes

Much of real-world causal learning relies on using the class of an object to infer its causal powers. For instance, one expects wine to cause drunkenness, but not grape juice. Thus, if one feels dizzy after drinking grape juice, one might want to check into the emergency room. The model presented here only enables learning causal relations among objects, but it can be extended to learn something about classes. Specifically, I propose extending it along the lines of the relational block model (Tenenbaum & Niyogi, 2003; Kemp et al., 2004) and causal grammars (Tenenbaum, Griffiths, & Niyogi, in press). According to these models, an object's class determines the probability that it will have a particular causal power. Thus, if 25% of

snakes are poisonous, then a newly generated snake will have a 25% chance of being poisonous. This simple assumption can enable transferring knowledge learned about one object to another object by noting that they are members of the same class.

To learn which classes of objects are causally related to which other classes, the block model relies on data indicating which objects are causally related to each other. However, it has traditionally been unclear how to rationally obtain knowledge of object-level causal relations. In fact, the proponents of the block model assume that data are straightforwardly observable when one observes an object activating another object on contact. The event-based causal learning framework provides a rational basis for making inferences about object-level causal relations which can serve as relational data for the higher-level inferences made by the block model. It also provides the necessary statistical basis for understanding how temporal delay between cause and effect influences learning with the block model.

### 4.3 Solving to the mechanism paradox

Proponents of the mechanism view (Ahn & Kalish, 2000) have argued that people only believe covariations to be evidence for causal relations when they believe that an intervening mechanism exists by which the effect could have produced the cause. Proponents of the covariational view (e.g., Glymour & Cheng, 1998) have in turn argued that knowledge about mechanisms could come from covariational learning, whereas the mechanism view has not provided a coherent proposal for how people could learn about mechanisms independently of covariations. I believe that knowledge of mechanisms can be learned via class-level knowledge derived from object-level causal relations (as described above). For instance, if it is discovered via event-based causal learning that a certain food causes an allergic reaction, one can store this as knowledge that foods in general are capable of causing allergic reactions, or in mechanistic



language, that a mechanism exists by which food can produce allergic reactions. In the future, when one encounters a non-temporal covariation between a certain food and an allergic reaction, one can be more likely to infer causation because one already believes that a mechanism exists.

#### 4.4 Anecdotal evidence as rational statistical inference

Scientific inquiry is often cast as an antidote to the ill-founded conclusions that people often draw from anecdotal evidence. But often the kinds of anecdotal evidence that people find compelling are often incompatible with traditional contingency-based causal learning. Our event-based framework offers insight into certain kinds of non-normative leaps that people tend to make. A particularly important domain is medical reasoning, in which people are notoriously prone to believing in treatments that are in fact no better than placebo. If recovery follows within a short period of time after treatment, the event-based framework provides a rational statistical basis for inferring a causal connection, sometimes even with just one example. Similarly, if a severe side effect occurs within a short period of time after treatment, it can be rational to infer a causal connection. This is particularly important for rare side effects, which may not be statistically significant with standard sample sizes. In fact, it is often anecdotal evidence from just a handful of rare but intuitively compelling occurrences that prompts researchers to conduct a clinical investigation using traditional scientific inquiry. By formalizing the statistical basis of anecdotal evidence, the event-based framework provides a rationale for the normative use of individual cases as evidence for causal relationships.

# Conclusion

The ability to reason causally is one of the hallmarks of human intelligence. I have proposed a new rational model to explain how people can be capable of learning new causal relations much more quickly than traditional models, and with much less evidence required. Like many current trends in both cognitive science and artificial intelligence, I demonstrate here that causal learning can be understood as rational statistical inference. These findings have wide implications, not just for the causal learning and the associative learning literatures, but for philosophy of science, and the role of temporal factors in scientific causal discovery. The kinds of stories that end up getting relegated to the status of “anecdotal evidence” are of often great value to clinicians, because they represent statistically compelling evidence despite being wholly incompatible with traditional contingency analysis. By incorporating temporal delay into scientific data, and using a statistical model that represents distributions over delays, we could discover causes faster, more reliably, and in better alignment with human intuition.

# References

- Ahn, W., & Kalish, C. (2000). The role of mechanism beliefs in causal reasoning. In R. Wilson, & F. Keil (Eds.) *Explanation and Cognition*, Cambridge, MA: MIT Press.
- Ahn, W., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation vs. mechanism information in causal attribution. *Cognition*, *54*, 299-352.
- Carey, S. (1987). Theory changes in childhood. In B. Inhelder, D. Caprona & A. Cornce-Wells (eds.), *Piaget Today*. Hillsdale, NJ: Erlbaum, 141-163.
- Cheng, P.W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*, 367-405.
- Gallistel, C. R. (2002). Frequency, contingency and the information processing theory of conditioning. In *Etc. Frequency Processing and Cognition*, Sedlmeier & Betsch (Eds.), pp. 153-171.
- Gallistel, C. R. (1990). *The organization of learning* (Cambridge, MA, Bradford Books/MIT Press)
- Gallistel, C. R., & Gibbon, J. (2000). Time, rate and conditioning. *Psychological Review*, *107*, 289-344.
- Garcia, J., Ervin, F. R., & Koelling, R. A. (1966). Learning with prolonged delay of reinforcement. *Psychonomic Science*, *5*, 121-122
- Gibbon, J., Fairhurst, S., Goldberg, B., CM Bradshaw (1997). Cooperation, conflict and compromise between circadian and interval clocks in pigeons. In C.M. Bradshaw & E. Szabadi (Eds). *Time and Behaviour: Psychological and Neurobehavioural Analyses*. Elsevier: London.

- Gibbon, J., Baldock, M. D., Locurto, C. M., Gold, L., & Terrace, H. S. (1977). Trial and intertrial durations in autoshaping. *Journal of Experimental Psychology: Animal Behavior Processes*, 3, 264-284.
- Glymour, C. & Cheng, P. W. (1998). Causal Mechanism and Probability: A Normative Approach. In M. Oaksford and N. Chater, eds., *Rational Models of Cognition*.
- Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T., & Danks D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, 111 (1), 1-31.
- Gopnik, A. & Glymour, C. (2002). Causal maps and Bayes nets: A cognitive and computational account of theory-formation. In P. Carruthers, S. Stich, M. Siegal,(Eds.) *The cognitive basis of science*. Cambridge: Cambridge University Press.
- 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.
- Gopnik, A. (2003). The theory theory as an alternative to the innateness hypothesis. In L. Antony and N. Hornstein (eds.) *Chomsky and his critics*. Blackwells, Oxford
- Griffiths, T. L. (2005). *Causes, Coincidence, & Theories*. Stanford University Doctoral Dissertation.
- Griffiths, T.L., Baraff, E.R., & Tenenbaum, J.B. (2004). *Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Society*.
- Griffiths, T.L., & Tenenbaum, J.B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, 51, 334-384.
- Hume, D. (1739/1978). A treatise of human nature. Oxford: Oxford University Press.

- Kehoe, E.J., Graham-Clarke, P., & Schreurs, B.G. (1989). Temporal patterns of the rabbit's nictitating membrane response to compound and component stimuli under mixed CS-US intervals. *Behavioral Neuroscience*, *103*(2), 283-95.
- Kemp, C., Griffiths, T. L., & Tenenbaum, J. B. (2004). Discovering latent classes in relational data. *MIT AI Memo 2004-019*.
- Kolata, G. (2005). Study Says Echinacea Has No Effect on Colds. *New York Times*, July 28, 2005.
- Lagnado, D. & Sloman, S.A. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 856-876.
- Novick, L. R., & Cheng, P. W. (2004). Assessing interactive causal influence. *Psychological Review*, *111*, 455-485.
- Pavlov, I.P., 1927. Conditioned reflexes. Oxford University Press, London.
- Reichenbach, H. (1956). *The Direction of Time*. University of California Press, Berkeley, CA
- Rescorla, R. A. (1967). Pavlovian conditioning and its proper control procedures. *Psychological Review*, *74*, 71-80.
- 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*, 64-69. New York: Appleton-Century-Crofts.
- Schulz, L. E. & Gopnik, A. (in submission) Preschoolers learn causal structure from conditional interventions.
- Schulz, L. E. & Gopnik, A. (2004). Causal learning across domains. *Developmental Psychology*, *40*(2), 162-176.

- Shanks, D. R. (1995). *The Psychology of Associative Learning*. Cambridge University Press, Cambridge, England.
- Shanks, D. R., Pearson, S. M., & Dickinson, A. (1989). Temporal Contiguity and the Judgment of Causality by Human Subjects. *The Quarterly Journal of Experimental Psychology*, *41B* (2), 139-159.
- Skinner, B. F. (1938). *The behavior of organisms*. New York: Appleton-Century-Crofts.
- Sloman, S.A., & Lagnado, D. (2005). Do we “do”? *Cognitive Science*, *29*, 5-39.
- Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children’s causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science*, *28*, 303-333.
- Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, Prediction, and Search*. The MIT Press, second edition.
- Tenenbaum, J. B., Griffiths, T. L., and Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309-318.
- Tenenbaum, J.B., Griffiths, T. L., and Niyogi, S. (in press). Intuitive theories as grammars for causal inference. To appear in Gopnik, A., & Schulz, L. (Eds.), *Causal learning: Psychology, philosophy, and computation*. Oxford University Press.
- Tenenbaum, J. B., & Niyogi, S. (2003). Learning causal laws. Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society.
- Waldmann, M. R. (2000). Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 53-76.

- Waldmann, M. R. (1996). Knowledge-based causal induction. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *The psychology of learning and motivation, Vol. 34: Causal learning*, 47-88. San Diego: Academic Press.
- Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General*, *121*, 222-236.
- White, P. A. (1995). *The understanding of causation and the production of action: from infancy to adulthood*. Hillsdale (USA) : L. Erlbaum Associates.

## **Part II. The role of prior knowledge in causal judgment**

**In which judgments under uncertainty can be explained as Bayesian inferences over causal models constructed from prior knowledge.**



## 5 Introduction to judgment under uncertainty

Everywhere in life, people are faced with situations that require intuitive judgments of probability. How likely is it that this person is trustworthy? That this meeting will end on time? That this pain in my side is a sign of a serious disease? Survival and success in the world depend on making judgments that are as accurate as possible given the limited amount of information that is often available. To explain how people make judgments under uncertainty, researchers typically invoke a computational framework to clarify the kinds of inputs, computations, and outputs that they expect people to use during judgment. We can view human judgments as approximations (sometimes better, sometimes worse) to modes of reasoning within a rational computational framework, where a computation is “rational” to the extent that it provides adaptive value in real-world tasks and environments. However, there is more than one rational framework for judgment under uncertainty, and behavior that looks irrational under one framework may look rational under a different framework. Because of this, evidence of “error-prone” behavior as judged by one framework may alternatively be viewed as evidence that a different rational framework is appropriate.

This paper considers the question of which computational framework best explains people’s judgments under uncertainty. To answer this, we must consider (1) what kinds of real-world tasks and environments people encounter, (2) which frameworks are best suited to these environments (i.e., which we should take to be normative), and (3) how well these frameworks predict people’s actual judgments under uncertainty (i.e., which framework offers the best descriptive model). We will propose that a causal Bayesian framework, in which Bayesian

inferences are made over causal models, represents a more appropriate normative standard and a more accurate descriptive model than previous frameworks for judgment under uncertainty.

The plan of the paper is as follows. We first review previous accounts of judgment under uncertainty, followed by the arguments for why a causal Bayesian framework provides a better normative standard for human judgment. We then present six experiments supporting the causal Bayesian framework as a descriptive model of people's judgments. Our experiments focus on the framework's ability to explain when and why people exhibit base-rate neglect, a well-known judgment phenomenon that has often been taken as a violation of classical Bayesian norms. Specifically, we test the hypotheses that people's judgments can be explained as approximations to Bayesian inference over appropriate causal models, and that base-rate neglect often occurs when experimenter-provided statistics do not map clearly onto parameters of the causal model participants are likely to invoke. In Experiments 1-4, we show that by clarifying the causal structure in these judgment scenarios and ensuring that experimenter-provided statistics map clearly onto the parameters of these causal models, we are able to substantially reduce -- and often, nearly eliminate -- base-rate neglect. In Experiments 5 and 6, we show that when the judgment prescribed by the causal Bayesian framework differs from classical (non-causal) Bayesian analyses, people's responses are more consistent with our framework. We conclude by discussing implications of the causal Bayesian framework for other phenomena in probabilistic reasoning, and for improving the teaching of statistical reasoning.

## 5.1 Statistical frameworks for judgment under uncertainty

Most previous accounts – whether arguing for or against human adherence to rationality – take some framework of statistical inference to be the normative standard (Anderson, 1990; Gigerenzer & Hoffrage, 1995; McKenzie, 2003; Oaksford & Chater, 1994; Peterson & Beach,

1967; Shepard, 1987; Tversky & Kahneman, 1974). Statistical inference frameworks generally approach the judgment of an uncertain variable, such as whether someone has a disease, by considering both the current data, such as the person’s symptoms, as well as past co-occurrences of the data and the uncertain variable, such as previous cases of patients with the same symptoms and various diseases. Because these frameworks focus on observations rather than knowledge, beliefs about the causal relationships between variables does not play a role in inference.

Using statistical inference frameworks as a rational standard, several hypotheses have been advanced to describe how people make judgments under uncertainty. Early studies of judgment suggested that people behaved as “intuitive statisticians” (Peterson & Beach, 1967), because their judgments corresponded closely to classical Bayesian statistical norms, which were presumed rational. Classical Bayesian norms explain how prior beliefs may be updated rationally in light of new data, via Bayes’ rule. To judge  $P(H | D)$ , the probability of an uncertain hypothesis  $H$  given some data  $D$ , Bayes’ rule prescribes a rational answer, as long as one knows (1)  $P(H)$ , the prior degree of belief in  $H$ , and (2)  $P(D | H)$  and  $P(D | \neg H)$ , the data expected if  $H$  were true and if  $H$  were false:

$$P(H | D) = \frac{P(H)P(D | H)}{P(D)} \quad (1)$$

where  $P(D) = P(H)P(D | H) + P(\neg H)P(D | \neg H)$ .

The intuitive statistician hypothesis did not reign for long. It was not able to account for a rapidly accumulating body of experimental evidence that people reliably violate Bayesian norms (Ajzen, 1977; Bar-Hillel, 1980; Eddy, 1982; Lyon & Slovic, 1976; Nisbett & Borgida, 1975; Tversky & Kahneman, 1974). For example, consider the “mammogram problem”, a Bayesian

diagnosis problem which even doctors commonly fail (Eddy, 1982). One well-tested version comes from Gigerenzer and Hoffrage (1995), adapted from Eddy (1982):

The probability of breast cancer is 1% for a woman at age forty who participates in a routine screening. If a woman has breast cancer, the probability is 80% that she will get a positive mammography. If a woman does not have breast cancer, the probability is 9.6% that she will also get a positive mammography. A woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer? \_\_\_\_ %

Participants often give answers of 70%-90% (Eddy, 1982; Gigerenzer & Hoffrage, 1995), while Bayes' theorem prescribes a much lower probability of 7.8%. In this case,  $H$  is "patient X has breast cancer",  $D$  is "patient X received a positive mammogram", and the required task is to judge  $P(H | D)$ , the probability that the patient has breast cancer given that she received a positive mammogram:

$$P(H | D) = \frac{P(H)P(D | H)}{P(H)P(D | H) + P(\neg H)P(D | \neg H)} = \frac{1\% \times 80\%}{1\% \times 80\% + 99\% \times 9.6\%} = 7.8\% \quad (2)$$

Kahneman and Tversky (1973) characterized the source of such errors as "neglect" of the base rate (in this case, the rate of cancer), which should be used to set  $P(H)$  in the above calculation of  $P(H | D)$  (in this case, the probability of cancer given a positive mammogram).<sup>2</sup>

---

<sup>2</sup> We should note that a popular explanation of the original mammogram problem suggests that people are confused by the given conditional probability, and think it means . This has been called the inverse fallacy (Villejoubert, & Mandel, 2002). We find this limited as an explanation, because people only seem to exhibit the inverse fallacy when they expect both probabilities to have roughly the same value. For instance, while people might agree that the probability of death in a plane crash is nearly 100% ( ), they surely would not agree to the inverse: that death is

The heuristics and biases view, which came to replace the intuitive statistician framework, sought to understand probabilistic judgments as heuristics, which approximate normative Bayesian statistical methods in many cases, but lead to systematic errors in others (Tversky & Kahneman, 1974). Given the focus of the heuristics and biases program on judgment errors, many concluded that people were ill-equipped to reason successfully under uncertainty. Slovic, Fischhoff, and Lichtenstein (1976) wrote: “It appears that people lack the correct programs for many important judgmental tasks.... it may be argued that we have not had the opportunity to evolve an intellect capable of dealing conceptually with uncertainty” (p. 174). Yet by the standards of engineered artificial intelligence systems, the human capacity for judgment under uncertainty is prodigious. People, but not computers, routinely make successful uncertain inferences on a wide and flexible range of complex real-world tasks. As Glymour (2001) memorably asked, “If we’re so dumb, how come we’re so smart?” (p. 8). Research in the heuristics and biases tradition generally did not address this question in a satisfying way.

One way to explain how people could be successful in the real world, despite their failings in laboratory tasks, is to argue that those tasks do not faithfully assess real-world reasoning. Perhaps the most influential such claim is the “natural frequency hypothesis” (Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995). According to this hypothesis, the statistics that naturally occur are not in the form of probabilities, thus only statistics presented as natural frequencies of events should activate our evolved capacity for judgment under uncertainty. Experiments show that people are better at statistical reasoning with natural frequencies, but there is an alternative explanation to the hypothesis that people routinely use

---

almost always a result of plane crashes ( ). Thus, it may be that people only confuse a conditional probability with its inverse when they expect the inverse to have the same value. It is this expectation, then, that needs to be explained.

natural frequencies for real-world judgment under uncertainty. Naturally frequency tasks are almost always arithmetically simpler than the corresponding probabilistic tasks, and this difference in complexity shows up in the expressions for Bayes' rule (equation 1) and natural frequency computation (equation 3) above. Thus, performance improvements alone are not convincing evidence that people are naturally better at reasoning about frequencies than probabilities, and do not address the question of whether the natural frequency algorithm represents a common method for making judgments in real-world environments.

All of these previous attempts to analyze judgment under uncertainty are fundamentally limited because they assume that purely statistical frameworks are an appropriate rational standard without seriously investigating whether they are in fact viable for inference in real-world environments. Purely statistical methods are best suited to reasoning about a small number of variables based on many observations of their patterns of co-occurrence – the typical situation in ideally controlled scientific experiments. However, real-world reasoning typically involves the opposite scenario: complex systems with many relevant variables and a relatively small number of opportunities for observing their co-occurrences. Because of this complexity, the amount of data required for reliable inference with purely statistical frameworks, which generally grows exponentially in the number of variables, is often not available in real-world environments. The conventional statistical paradigms developed for idealized scientific inquiry may thus be inappropriate as rational standards for human judgment in real-world tasks. Proposing heuristics as descriptive models to account for deviations from statistical norms only clouds the issue, as there is no way to tell whether an apparent deviation is a poor heuristic approximation to a presumed statistical norm, or a good approximation to some more adaptive approach.

We will propose that a *causal Bayesian framework* provides this more adaptive approach, and that it offers both a better normative standard than purely statistical methods and a better descriptive model than heuristic accounts. Like classical statistical norms, the framework we propose is Bayesian, but rather than making inferences from purely statistical data, inferences in our framework are made with respect to a causal model, and are subject to the constraints of causal domain knowledge. Our causal Bayesian framework is more adaptive than previous proposals, as it explains how rational judgments can be made with the relatively limited statistical data that is typically available in real-world environments. This approach also represents a better descriptive model than purely statistical norms or heuristics, which do not emphasize, or even have room to accommodate, the kinds of causal knowledge that seem to underlie much of people's real-world judgment.

The phenomenon of "base-rate neglect" is one of several judgment fallacies advanced by the heuristics & biases program as evidence that people reliably deviate from normative Bayesian judgment. Essential to their account is the notion that certain kinds of base rates are commonly neglected, and much of the literature focuses on which kinds of base rates people are most prone to neglecting. Some studies indicated that people more often neglected base rates that lacked causal relevance (Ajzen, 1977; Tversky & Kahneman, 1980). However, Bar-Hillel (1980) argued that the salience of the base rate determined whether people would use it, and that causal relevance was just one form of salience. Contrary to the conclusions of the heuristics & biases literature, we argue that for many well-known stimuli, the features of the base rate are not what lead people to exhibit apparent "base-rate neglect". We offer as evidence six experiments in which the description of the base rate is identical across two conditions, but people neglect the base rate in one condition and use it appropriately in the second condition. We further argue that

people in these experiments are not even neglecting the base rate at all, and we re-interpret incidents of purported “base-rate neglect” as cases in which the prescriptions of classical Bayesian norms are non-normative by the standards of causal Bayesian inference. Our experiments will show that when these prescriptive norms agree, people often use the given statistics normatively (by both standards), but when they disagree, people’s judgments more often adhere to the causal Bayesian standard than the classical Bayesian standard. Furthermore, when the problem makes clear which causal model should be used and how given statistics should be incorporated into that model, we find that people rarely neglect base rates.

We are not the first to propose that causal knowledge plays a role in base-rate neglect. Researchers in the heuristics and biases program investigated how causality influenced base-rate neglect, but their proposed models were vague and did not elucidate the rational role of causal reasoning in judgment under uncertainty. Ajzen (1977) proposed that a “causality heuristic” leads to neglect of information that has no apparent causal explanation. For example, if the base rate of students passing an exam is low, then a student’s performance will be causally influenced by the exam’s difficulty, and therefore the base rate is used to lower one’s judgment that a smart student passed the exam. However, if the exam was normal difficulty, but a psychologist selects mostly students who failed the exam for a post-exam interview (because the psychologist is mostly interested in students’ reaction to failure), the low base rate of passing the exam among interviewed students does not suggest a causal explanation for the students’ performance on the exam, and therefore this base rate is neglected when judging the probability that a smart student who happened to be interviewed passed the exam. While his account seems to explain participants’ judgments in his experiments, it does not seem to generalize well. For example, in the mammogram problem, the base rate of cancer seems to have a causal explanation, (e.g.,



environmental toxins, genetic factors, age, diet, etc.), hence the above account seems to predict that this base rate should be used. Furthermore, Ajzen (1977) did not explain how a focus on causal factors could lead to successful judgments in the real world, and thus did not address why people would have such a heuristic or why it should work the way that it does.

Following Ajzen (1977), Tversky and Kahneman (1980) proposed that “evidence that fits into a causal schema is utilized, whereas equally informative evidence which is not given a causal interpretation is neglected” (Tversky & Kahneman, 1980, p. 65). This was famously demonstrated by the “causal” variant of the “cab problem”, one of the earliest problems found to elicit base-rate neglect. The original version of the cab problem and the causal variant read as follows (from Tversky & Kahneman, 1980, p. 63):

#### *5.1.1.1 Original Cab Problem*

A cab was involved in a hit-and-run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data:

- (i) 85% of the cabs in the city are Green and 15% are Blue.
- (ii) A witness identified the cab as a Blue cab. The court tested his ability to identify cabs under the appropriate visibility conditions. When presented with a sample of cabs (half of which were Blue and half of which were Green) the witness made the correct identifications in 80% of the cases and erred in 20% of the cases.

*Question:* What is the probability that the cab involved in the accident was Blue rather than Green?

### 5.1.1.2 Causal Cab Problem

A cab was involved in a hit-and-run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data:

(i) Although the two companies are roughly equal in size, 85% of cab accidents in the city involve Green cabs, and 15% involve Blue cabs.

(ii) As in original problem above.

*Question:* What is the probability that the cab involved in the accident was Blue rather than Green?

The median and modal response to the original cab problem was 80% (reflecting neglect of the base-rate of 15% Blue cabs), while the median response to the causal cab problem was

60%, which is closer to the normative solution of  $41\% \left( \frac{15\% \times 80\%}{15\% \times 80\% + 85\% \times 20\%} = 41\% \right)$ . This

reflects improved, but not necessarily normative, use of the base rate. Tversky and Kahneman (1980) argued that the population base rate in the original problem is neglected because it does not fit into a causal schema; i.e., nothing causes there to be more green cabs. In contrast, they argued, the base rate in the causal cab problem fits into a causal schema: the higher accident rate of green cabs might be caused by Green cab drivers being more reckless. One major difficulty with this proposal is that it is unclear what it means for a statistic to fit into a causal schema. In particular, Tversky and Kahneman make the assumption that population base rates do not fit into causal schemas, but there is good reason to doubt this assumption. For instance, when judging whether someone's cough is more likely to be caused by a common cold or by lung cancer, the base rate of these illnesses in the population is obviously essential. Furthermore, Bar-Hillel (1980) showed that even population base rates can elicit good performance in the cab problem

(using a third variant, the “intercom problem”), casting doubt on the ability of causal schemas to explain base rate neglect.

In the heuristics and biases literature, the notion that attention to causal structure is in any sense rational or adaptive has received little attention. Instead, using causal schemas was viewed as an intuitive, fuzzy form of reasoning that, to our detriment, tends to take precedence over normative statistical reasoning when given the chance. In contrast to ill-defined causal *schemas*, however, inference over causal *models* (based on Bayesian networks) is a well-defined, rational, and adaptive method for judgment under uncertainty, which can succeed in real world tasks on which statistical methods fail. Therefore, it makes sense to revisit the idea that people’s judgments can in fact be both causally constrained *and* rational.

## 6 A causal Bayesian framework for judgment under uncertainty

Causal reasoning enables one to combine available statistics with knowledge of causal relationships, resulting in more reliable judgments, with less data required than purely statistical methods. It is becoming clear from research in artificial intelligence (Pearl, 2000), associative learning (Cheng, 1997; Glymour, 2001; Gopnik & Glymour, 2002; Gopnik & Sobel, 2000; Waldmann, 1996), and categorization (Ahn, 1999; Rehder, 2003) that causal reasoning methods are much better suited than purely statistical methods for learning and inference in real-world environments. In this section, we will present a framework for making inferences over causal models based on Bayesian networks, and then argue that it represents a better normative standard for judgment under uncertainty.

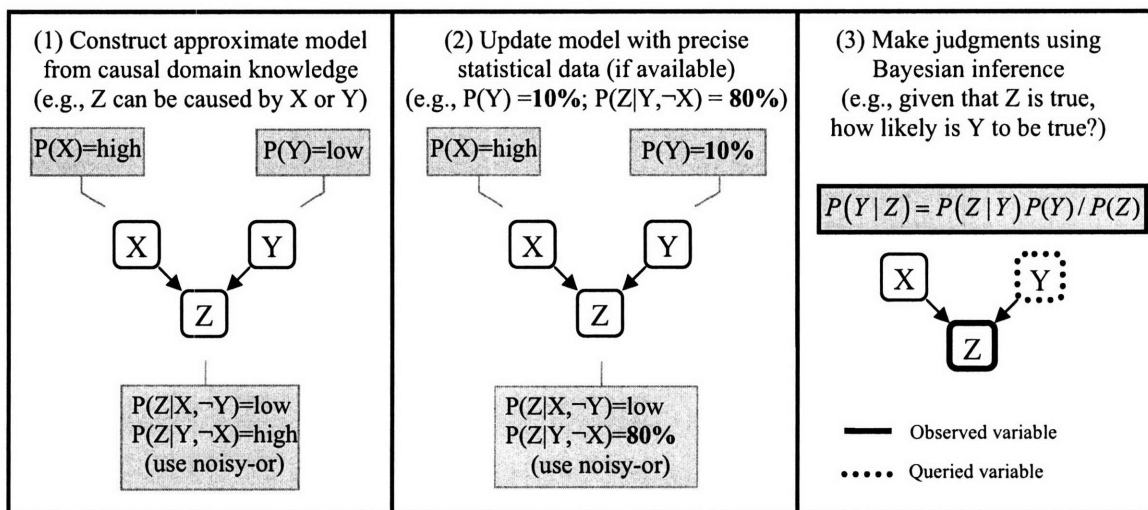
Motivation for this framework derives from two recent innovations in artificial intelligence for probabilistic reasoning about complex causal systems. First, probabilistic graphical models, including Bayesian networks, enable one to exploit independence and conditional independence relations between variables to dramatically reduce the amount of data required for successful inference in any one situation, or successful learning across situations (Pearl, 1988, Russell & Norvig, 2003). These systems have led to significant advances in machine learning and reasoning, with many successful real-world applications (e.g. Friedman, Linial, Nachman, and Pe'er, 2000; Oatley & Ewart, 2003; Spiegelhalter, Dawid, Lauritzen, and Cowell, 1993). Yet probabilistic graphical models require that independence and conditional independence relations either be known in advance or learned from data, and the relevant data-

driven learning algorithms (Jordan, 1999; Pearl, 1988; Spirtes, Glymour, & Scheines, 1993) require much larger statistical samples than are typically available to people in real-world tasks (Tenenbaum & Griffiths, 2003; Wasserman, 2004). More recently, causal graphical models (Pearl, 2000) offer to make Bayesian networks more practical by exploiting prior knowledge of causal structure to determine independence relations between variables. Causal Bayesian networks have been proposed as tools for understanding how people intuitively learn and reason about causal systems (e.g. Glymour & Cheng, 1998; Gopnik, et al., 2004; Griffiths & Tenenbaum, 2005; Sloman & Lagnado, in press; Steyvers, Tenenbaum, Wagenmakers & Blum, 2003; Tenenbaum & Griffiths, 2001, 2003; Waldmann, 2001), but their implications for more general phenomena of judgment under uncertainty have not been systematically explored.

## 6.1 Causal Bayesian inference as a rational method of judgment under uncertainty

The basis of judgment within our causal Bayesian framework is Bayesian inference over causal models derived from prior knowledge. Figure 25 depicts the three ideal phases of judgment in causal Bayesian inference: (1) constructing a causal model (2) updating the model's parameters with available statistical data to make them more precise; (3) inferring probabilities of target variables via Bayesian inference over the model. These three phases are not necessarily distinct and ordered in this way – this is more of an “ideal observer” model that provides a useful way to think about how causal reasoning guides people's judgments of probability. Expert systems based on Bayesian networks (Pearl, 1988) have often been built in just this way: an expert provides the initial qualitative causal model, objectively measured statistics determine the

model's parameters, and inference over the resulting Bayesian network model automates the expert's judgments about unobserved events in novel cases.



**Figure 25: In an “ideal observer” model for causal Bayesian inference, judgment under uncertainty divides into three phases: (1) causal domain knowledge is used to construct a causal model, with a prior belief distribution over parameters. (2) if statistical data are available they may be used to set more precise parameters values. (3) judgments are made by computing Bayesian inferences over the parameterized causal model.**

Our proposal differs in emphasis from other uses of causal Bayesian networks in the psychological literature (e.g., Glymour, 2001). We emphasize that people come to a judgment problem with prior knowledge that can be used to construct a causal model. The causal models that people construct are not just directed graphs with generic conditional probability tables, but rather associate individual edges with meaningful parameters that function in specific ways. To construct a causal model, a reasoner must utilize prior knowledge to decide whether a given variable is a cause or an effect of another (e.g., prior knowledge suggests that cancer causes a tumor, and a tumor causes a lump). Given a set of variables, knowledge of these relations can be

used to construct a causal model, which represents the causal relations among the variables.<sup>3</sup> Each variable has parameters that characterize how it depends on its causes, and the model is *fully-specified* when the parameters for all variables are specified. First, we will explain how computations over a fully-specified causal model can yield judgments under uncertainty, and then we will discuss how statistics, when available, can be used to update the parameters of a causal model prior to making a judgment.

When making causal Bayesian inferences, one is concerned with the same judgment as in classical Bayesian inferences: computing  $P(H | D)$ , the probability of a hypothesis,  $H$ , given some data,  $D$ . Rather than using Bayes' rule directly, however, computing  $P(H | D)$  in a causal Bayesian framework involves making inferences over a causal model relating the variables of  $H$  and  $D$ . In our framework, a causal model is formally represented as a causal Bayesian network (Glymour, 2003; Pearl, 2000), a directed acyclic graph in which the nodes represent variables, the arrows represent causal influences pointing from cause to effect, and a probability distribution is defined over the variables in a way that reflects the structure of the graph. Each variable  $X_i$  in a Bayesian network is conditionally independent of all other variables given the values of its direct causes, or *parents* – the variables corresponding to nodes that point to  $X_i$  in the graph. This *causal Markov condition* enables one to factor the full probability distribution over the variables into a product of conditional probabilities, one for each variable:

$$P(X_1 \dots X_n) = P(X_1 | \text{par}(X_1)) \times \dots \times P(X_n | \text{par}(X_n)) \quad (3)$$

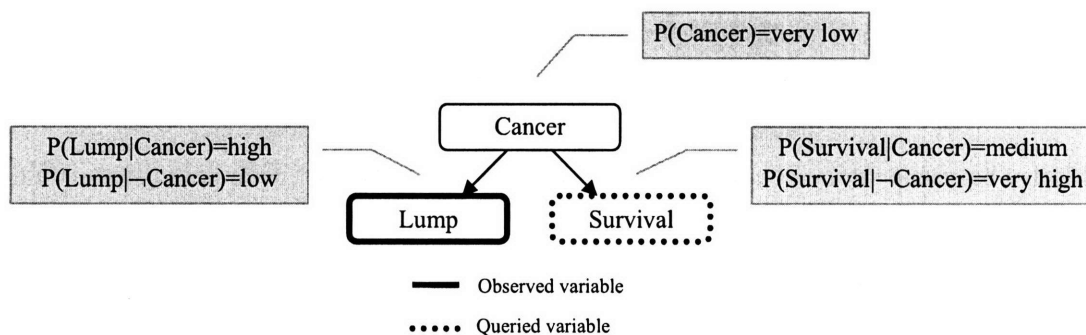
where  $\text{par}(X_i)$  is the set of parents of the variable  $X_i$ . One specifies  $P(X_i | \text{par}(X_i))$  by assigning to each variable  $X_i$  a conditional probability table, or CPT, which defines a probability

---

<sup>3</sup> The process of causal model construction itself is an important topic but is beyond the scope of this paper.

distribution over the variable's states given all possible states of its parents. Inferences, or judgments about the probable values of a set of variables, can be made over a causal model by fixing the values of all variables that have been "observed" ( $D$ ), and then computing the new probability distribution over the variable(s) of interest ( $H$ ), conditioned on the observed variables ( $D$ ) ("observing" a variable simply means knowing its value; the value could be obtained from communicating with others, reading about it, or other means).

These concepts can be understood better with an example. Suppose a person finds a possibly cancerous lump, and wants to estimate the probability of surviving. A simple causal model for this situation is depicted in Figure 26, with three binary (true/false) variables representing attributes of the person: *Cancer* (whether the person has cancer), *Lump* (whether the person has found a lump), and *Survival* (whether the person will survive).



**Figure 26: Example causal model that a typical person might use to make inferences about cancer. In a Bayesian network, the CPTs contain exact probabilities. In this case, the CPTs contain qualitative probabilities which reflect a typical person's rough estimates of these numbers.**

In this case, we wish to judge  $P(H | D)$  where  $D$  corresponds to  $Lump=true$  and  $H$  corresponds to  $Survival=true$ . To specify a complete probability distribution over the variables, three CPTs are needed,  $P(C)$ ,  $P(L | C)$ , and  $P(S | C)$ . Each CPT specifies the probability distribution of the variable conditioned on its parents (in Figure 26 we depict these probabilities with rough



estimates that could be generated from a typical person’s domain knowledge). Using the causal Markov condition, the joint distribution is factored as follows:

$$P(C, L, S) = P(C | par(C))P(L | par(L))P(S | par(S)) = P(C)P(L | C)P(S | C) \quad (4)$$

For the task in question, we wish to infer the probability of survival given that a lump was discovered. Inference would proceed by “observing” *lump* to be true, and then computing the new probability of *survival*:

$$P(S = true | L = true) = \sum_{x=true, false} \frac{P(C = x)P(S = true | C = x)P(L = true | C = x)}{P(L = true)} \quad (5)$$

(where summation over  $x$  means considering both  $C=true$  and  $C=false$ ).

Suppose we know that the rate of cancer in the population is 2%, the probability of finding a lump if you have cancer is 75%, the probability of finding a lump if you don’t have cancer is 10%, and the probability of surviving if you have cancer is 50%. The correct answer, then, is:

$$P(S = true | L = true) = \frac{2\% \times 50\% \times 75\% + 98\% \times 100\% \times 10\%}{2\% \times 75\% + 98\% \times 10\%} = \frac{10.55\%}{11.3\%} = 93.3\% \quad (6)$$

## 6.2 Causal Bayesian inference as a new normative standard

Leading accounts of judgment under uncertainty analyze people’s judgments as better or worse approximations to classical, non-causal statistical norms. Here, we argue that causal Bayesian inference represents a better normative standard for judgment under uncertainty than the classical Bayesian norm. Since both frameworks are rational in that they are grounded in Bayesian probability theory, the better normative standard should be the one with greater adaptive value (Anderson, 1990), which is the one that enables making successful inferences using the evidence that is most available in real world environments. Our argument for causal Bayesian inference as a new normative standard is based on its superior flexibility in leveraging

available causal knowledge to make appropriate inferences when sufficient observational data is not available.

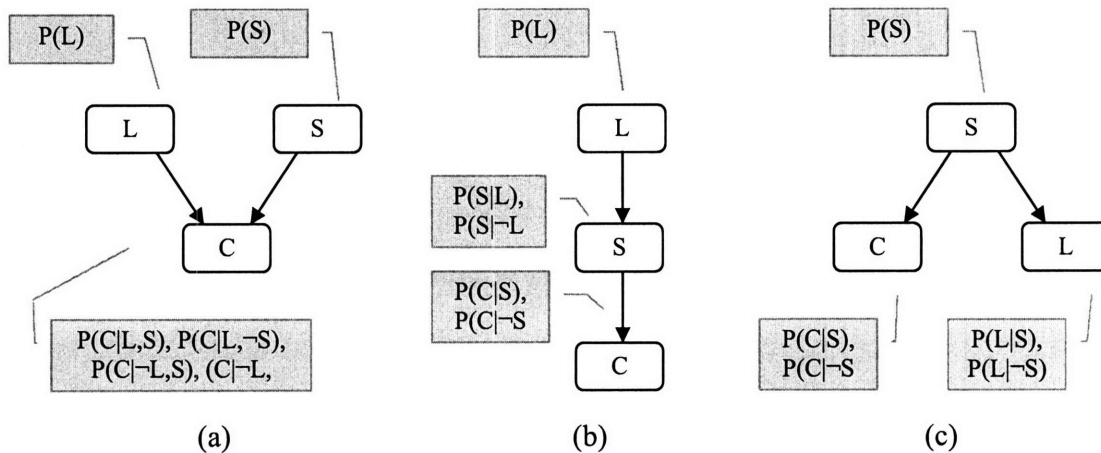
To make inferences using traditional statistical frameworks without knowledge of causal structure, one must have enough data to uniquely determine the full joint probability distribution or joint occurrence frequency of all relevant variables. But the number of observations required to fill the co-occurrence matrix is exponential in the number of variables, which often makes it difficult to obtain sufficient statistics for successful judgment from observations alone. In real world environments, one generally does not have time to wait for enough observations to completely fill a co-occurrence matrix. Causal Bayesian inference provides a way to go beyond the available data, by using causal domain knowledge to fill in the gaps when statistical data is insufficient.

To get around the problem of massive data requirements, statistical methods tend to restrict data collection to just a few variables. However, applying these limited data to a novel situation can be problematic. Consider, for example, the statistics in the mammogram problem presented earlier, and suppose a 45-year-old woman receives a positive mammogram. Can she use the given statistics, which are about 40-year-olds, to determine her chances of having breast cancer? What if we also know that her aunt had breast cancer and she is a vegetarian? Intuitively, the more information we know about a particular case, the better our inference should be, but to get this benefit using purely statistical methods we would have to collect new statistical data on the specific sub-population of 45-year-old vegetarian women whose aunts had cancer. People, on the other hand, are able to apply data appropriately when additional variables are introduced. We can reason that because cancer risk increases with age, a 45-year-old woman has a slightly higher chance of cancer than a 40-year-old. So too does a woman whose aunt had cancer, because genes

influence cancer risk. We may believe that vegetarians have a reduced chance of cancer, but only if we believe cancer is causally influenced by diet. It is knowledge of such causal relationships that enables us to make so many successful judgments with so limited data. Causal knowledge is also what differentiates everyday judgments from those of scientific researchers, who, because they do not wish prior causal intuitions to cloud their judgments, must laboriously collect data, sometimes over several months, just to make a single statistical inference.

When one has limited statistical data, if one has knowledge of the true causal structure generating the data, one can make more accurate judgments than with the data alone. For example, suppose we have statistical data indicating that the risk of breast cancer ( $C$ ) is increased by two lifestyle factors, being childless ( $L$ ) and having high stress levels ( $S$ ), but we do not have the full joint distribution among the three variables. Suppose we know that stress causes cancer, but it is not clear how being childless is causally related to cancer. If we wish to judge the likelihood of cancer in a childless woman, the correct judgment depends on the actual causal relationships that exist between these variables. If being childless causes breast cancer (Figure 27a), then the risk of cancer in a childless woman is increased, regardless of whether she has high stress levels (assuming for simplicity that these factors do not interact). However, it could be the case that being childless causes women to develop high stress levels (Figure 27b), but does not directly cause cancer. In this case, the risk of cancer in a childless woman is still increased, but we can ignore the fact that she is childless if we know her level of stress. Finally, it might be the case that having high stress levels causes women not to have children (Figure 27c). In this case, we should again ignore the fact that a woman is childless if we already know the woman's level of stress. The notion that it is rational to use statistical data differently for different causal structures is beyond the scope of classical statistical norms. The causal Bayesian

standard is able to prescribe the appropriate use limited data, by making use of the conditional independence relations determined by causal structure.



**Figure 27: Three causal structures that could be responsible for the same statistics. The correct judgment for  $P(C|A,B)$  depends on which causal structure was actually responsible for generating the observed statistics.**

The structure of a causal model influences not just how judgments should be made from provided statistics, but also the way the statistics should be interpreted. Consider the statistic for the probability of high stress levels given that one is childless,  $P(S|L)$ . For Figure 27b, the given statistic corresponds directly to a model parameter, thus it can be directly assigned to that parameter. For Figure 27c, there is no model parameter corresponding to  $P(S|L)$ , but one can straightforwardly account for it by assigning to  $P(L|S)$  the formula  $P(S|L)P(L)/P(S)$ . For Figure 27a,  $P(S|L)$  does not correspond directly to a parameter of the model, and there is no straightforward way to account for it, which means there is no single prescription for how such a statistic will influence future judgments from this model.

Psychological studies of judgment under uncertainty typically provide people with statistical data and then ask them to make judgments using the provided statistics. But if there are several different possible causal models that could have generated the statistics, then under

causal Bayesian inference, the normatively correct inference depends both on the particular model used and on how the statistics are assigned to the parameters of the model. Therefore, it is not possible to prescribe a single correct answer using the causal Bayesian framework unless (1) the model structure is known, (2) the provided statistics map unambiguously onto model parameters, and (3) no free parameters remain after assigning the provided statistics to the problem. Usually this means the provided statistics must directly correspond to a model parameter, but not always; there can still be an unambiguous mapping even if several parameters must be adjusted, provided there is only one way to adjust them. The issue of how provided statistics are used to update the parameters of a model, and how they are then used in subsequent inferences, plays a central role in our experiments. By providing statistics that clearly map onto an unambiguous causal model, we test in the following six experiments the extent to which people's judgments conform to the prescriptions of causal Bayesian inference.

## 7 Experiments

Although previous accounts present different descriptive models of the computations people perform in making judgments, they all adopt a classical Bayesian norm as the prescriptive standard for rational judgment under uncertainty. In contrast, we have proposed a new normative standard for judgment, causal Bayesian inference, which sometimes differs from classical Bayesian norms. The following six experiments test whether our framework for causal Bayesian inference provides a better descriptive model of judgment under uncertainty than previous descriptive accounts such as the natural frequency hypothesis and the heuristics & biases view.

In each experiment we created a pair of scenarios and randomly assigned participants to one of them. The formal statistical structures of the two scenarios were always identical from the

point of view of the classical Bayesian norm, thus the judgment prescribed by this norm is the same for the two scenarios. Furthermore, all other factors previously identified as playing a role in base-rate neglect (such as salience or causal relevance of the base rate) were held constant, thus the heuristics & biases view would predict that people exhibit identical levels of base rate neglect in the two scenarios. Crucially, however, the two scenarios always differ in their causal structure, such that the correct answers prescribed by our new causal Bayesian norm differ across scenarios. Since we hypothesize that people's judgments adhere to the principles of the causal Bayesian norm, our hypothesis, and only our hypothesis, predicts that people's judgments will differ between the two scenarios.

In Experiments 1-4, the first of the two scenarios is taken from a classic problem in the base-rate neglect literature. Studies from this literature have generally assumed that only a single answer could be "correct" based on a normative analysis of the given statistics in these problems. But in the causal Bayesian framework, the correct solution depends how the variables are causally related in a causal model, and how the statistics are assigned to the model's parameters. Since these classic scenarios are arguably consistent with multiple possible causal structures, it is problematic to prescribe a single solution. Moreover, if one adopts the structure most consistent with common causal knowledge, there is often no clear way to assign the statistics to parameters of the model. The second of the two scenarios in each experiment is a variant we designed to have a clear, unambiguous causal structure, and statistics that clearly map onto its parameters. The fully-specified causal models for these newly developed scenarios are unambiguous; therefore, it is possible to prescribe a single correct answer under our causal Bayesian norm.

The primary goal of the experiments is to test whether people's judgments are consistent with solutions prescribed by the causal Bayesian norm. However, because it is not appropriate to

prescribe a “correct” judgment (under this standard) unless it is clear which causal model should be used and how it should be updated, in Experiments 1-4 we prescribe a correct solution only to our newly developed scenario. We will employ the original scenario as a control condition to show that our participants’ solutions to these scenarios are often consistent with base-rate neglect, hence any better performance on the new scenarios cannot be explained by our population being more statistically adept than previously studied populations. To the extent that people’s judgments match the solutions prescribed by the causal Bayesian norm to our newly developed scenarios, it supports our hypothesis that people have the cognitive capacity to make causal Bayesian inferences. It also challenges the frequentist hypothesis claim that people lack a cognitive engine for making judgments from probabilities or relative frequencies (expressed as percentages), the data formats used in all scenarios in our studies.

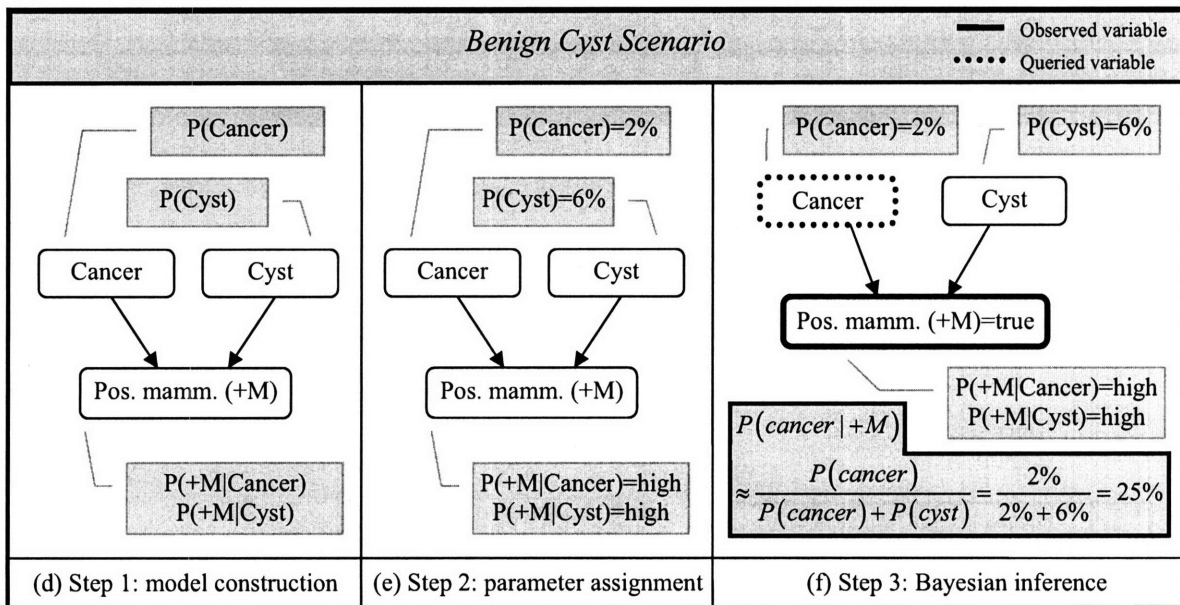
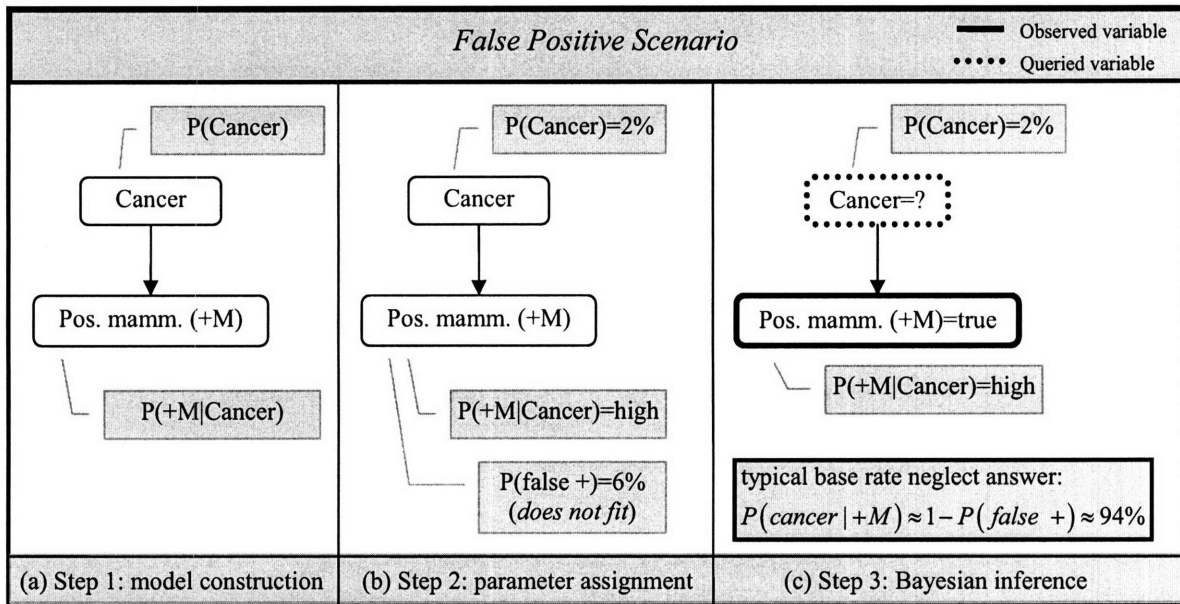
A secondary goal of the experiments is to challenge the heuristics and biases view that base rates are often neglected, and the more general view that people’s judgments under uncertainty are often flawed. In each of the first four experiments, the heuristics and biases view predicts high rates of base-rate neglect for both scenarios. To challenge these predictions, we sought to reduce base-rate neglect and improve correct response rates (by the standards of classical Bayesian norms) on the new scenario. This was done by carefully designing the causal structure of the newly developed scenarios such that both classical and causal norms would prescribe the same solution. If our hypothesis is correct, then people’s performance on the new scenarios will be *improved* by classical Bayesian standards over the original scenario. The heuristics and biases view would be unable to explain why people would do better on a problem that does not differ in any previously identified factor from one that produces base-rate neglect.

The experiments also serve a third purpose: to challenge the classical Bayesian norm as a rational standard for judgment. In Experiments 5 and 6, we provide two scenarios that clearly and compellingly require different solutions, yet the classical Bayesian framework cannot distinguish between them. Only the causal Bayesian framework prescribes a difference in correct answers between the two scenarios: that the base rate should rationally be ignored in one scenario, but not in the other. If the predicted difference is found in people's judgments, then the causal Bayesian framework would seem to provide not just a better descriptive model, but also a better normative standard for rational judgment.

## 7.1 Experiment 1

In the mammogram problem (Eddy, 1982; Gigerenzer & Hoffrage, 1995), presented earlier, the base rate of cancer in the population often seems to be neglected in people's judgments of  $P(\text{cancer} \mid +M)$ . We can construct a causal model of this scenario from common knowledge, which tells us that *cancer* is a cause of *positive mammograms*, as depicted in Figure 28a. In this model, the variable *cancer* has no parents, therefore the CPT for *cancer* contains just one parameter:  $P(\text{cancer})$ , which directly corresponds to the base rate provided in the problem. Because there is only one way to assign the base rate to this model parameter, the base rate should influence judgments by causal Bayesian standards. Since we hypothesize that people's judgments adhere to causal Bayesian standards, our hypothesis conflicts with previous conclusions that people neglect this base rate. In this experiment, we demonstrate empirically that people do not neglect this base rate in a newly developed scenario that differs only in causal structure, and we argue that the real difficulty people have is with the false-positive statistic.





**Figure 28: The three phases of causal Bayesian inference for the Experiment 1. (a-c) depict one possible interpretation of the causal structure of the false positive scenario, in which positive mammograms are generated only by cancer. (d-f) depict the causal structure described by the benign cyst scenario, in which positive mammograms are generated by cancer or benign cysts.**

In the causal Bayesian framework, one must first construct a causal model from domain knowledge prior to updating it with provided statistics, after which inferences can be computed. Figure 28(a-c) depict these three phases of inference for the mammogram scenario. We hypothesize that people may adopt a causal model for which the false-positive statistic,  $P(+M | -cancer) = 9.6\%$ , does not correspond to a model parameter. This statistic seems to describe the probability that positive mammograms would occur without being caused. If people believe positive mammograms to be the kind of thing that can be generated by a cause, but that cannot spontaneously occur, then the causal model most consistent with common knowledge may not have a parameter corresponding to the rate at which positive mammograms occur without cause. Figure 28b shows that the base rate of cancer fits into the model, but the false positive statistic does not. Since the statistic seems relevant, but does not fit into the model, people may be guessing a formula, such as subtracting the false-positive from the true positive rate, which produces responses consistent with base-rate neglect.

To test our hypothesis that people use base rates properly in causal inference, we developed a new scenario in which the causal model is clear and all the statistics clearly map onto parameters of the model. To clarify the causal structure, we provided an explicit alternative cause for positive mammograms in women who do not have cancer: benign cysts (see Figure 28d). We replaced the false-positive rate in the original problem with the base rate of dense but harmless cysts, and described the mechanism by which these cysts generate positive mammograms. This new statistic, the base rate of benign cysts in the population, directly maps onto the parameter for  $P(cyst)$  in the model (see Figure 28e).

### 7.1.1 Method

*Participants.* The 157 participants in this experiment were comprised of 78 airport passengers and 79 MIT undergraduate and graduate students (majors were not recorded, but were likely randomly distributed). Airport passengers were approached while waiting in line and were not compensated. MIT students were approached on campus, and were given token compensation.

*Materials.* Participants were randomly assigned to receive one of two paper and pen versions of Gigerenzer and Hoffrage's (1995) probabilistic mammogram question (adapted from Eddy, 1982) in a between-subjects design. The *false positive scenario* was similar to the original question, while the *benign cyst scenario* gave the base rate of benign cysts in the population rather than the rate of false positives. We chose not to include the true-positive rate,  $P(+M | cancer) = 0.8$ , but instead stated, "most women with breast cancer will receive a positive mammogram." This was done to encourage participants to provide answers based on their intuition rather than memorized mathematical formulas. Both scenarios required the exact same mathematical formula to calculate the answer, so there was no difference in arithmetic difficulty.

#### 7.1.1.1 False Positive Scenario

The following statistics are known about women at age 60 who participate in a routine mammogram screening, an X-ray of the breast tissue that detects tumors:

About 2% have breast cancer at the time of the screening. Most of those with breast cancer will receive a positive mammogram.

There is about a 6% chance that a woman without cancer will receive a positive mammogram.

Suppose a woman at age 60 participates in a routine mammogram screening and receives a positive mammogram. Please estimate the chance that she actually has breast cancer.

#### 7.1.1.2 Benign Cyst Scenario

The following statistics are known about women at age 60 who participate in a routine mammogram screening, an X-ray of the breast tissue that detects tumors:

About 2% have breast cancer at the time of the screening. Most of those with breast cancer will receive a positive mammogram.

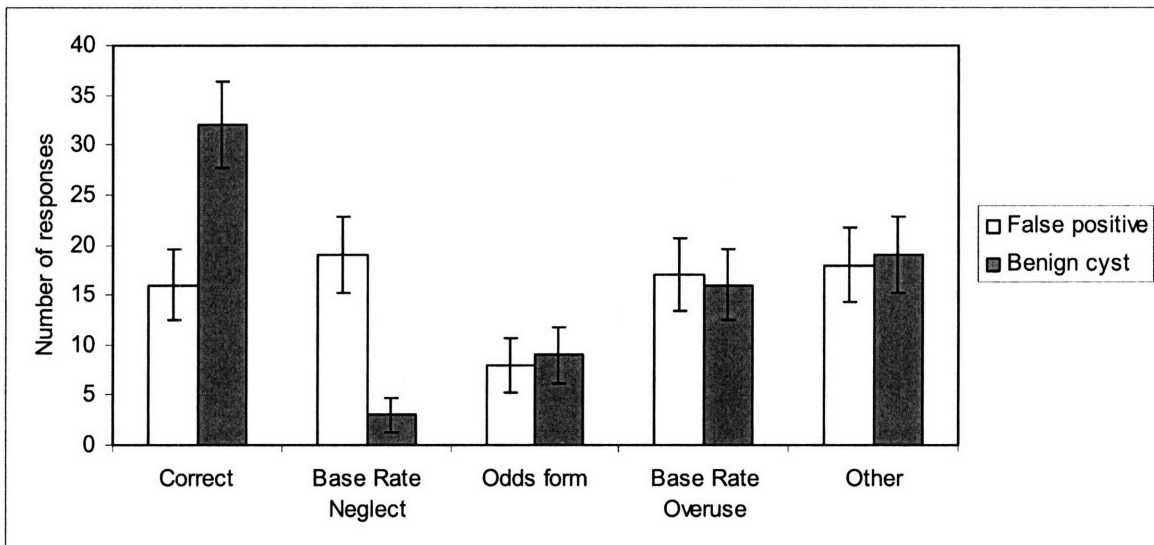
About 6% of those without cancer have a dense but harmless cyst, which looks like a cancerous tumor on the X-ray and thereby results in a positive mammogram.

Suppose a woman at age 60 participates in a routine mammogram screening and receives a positive mammogram. Please estimate the chance that she actually has breast cancer.

### 7.1.2 Results

The approximately correct answer to both scenarios (by the standards of classical Bayesian inference) was:  $\frac{2\%}{2\% + 98\% \times 6\%} \approx \frac{2\%}{2\% + 6\%} = 25\%$ . Preliminary analyses showed no significant differences between the responses of MIT students and airport passengers, so the two groups were collapsed for the remaining analyses. To maintain consistency with past research, we classified as *base-rate neglect* any answer greater than or equal to 69% (we assumed that “most” could be interpreted as 75% or higher, and a typical base-rate neglect answer is  $P(D|H) - P(D|\neg H)$ , for a lower bound on base-rate neglect of about 69%). We classified as *correct*, answers of the exact correct ratio or percentage, or a number up to 25% below it (in this

case, answers of 20% - 25%, again to accommodate the fact that most, but not all, women with cancer receive a positive result). There were two other types of answers that occurred frequently: *odds form* answers giving the odds of cancer vs. not cancer (in this case, 2/6 or 33%), which are technically incorrect but worth highlighting because they reflect correct use of the base rate, and *base-rate overuse* answers in which participants simply gave the base rate (in this case, 2%) as the answer. All other answers were classified as *other*, and were distributed across the spectrum with no discernable pattern. The results show that the benign cyst scenario significantly reduced *base-rate neglect* and significantly increased *correct* answer rates relative to the false positive scenario ( $\chi^2(4) = 17.07, p < .005$ ) (see Figure 29).



**Figure 29: Histogram of responses to Experiment 1. The correct answer was 25%. Responses were classified as correct (20%-25%), base-rate neglect ( $\geq 75\%$ ), odds form (33%), base rate overuse (2%), and other. A significant difference was found between the false positive and benign cyst conditions ( $\chi^2(4) = 17.07, p < .005$ ). Error bars represent the standard error of the normal approximation to the binomial distribution.**

### 7.1.3 Discussion

These results support our hypothesis that people are adept at making rational judgments from statistics that unambiguously map onto parameters of a clear causal model. The modal response to the benign cyst scenario was the *correct* answer (31 of 79 participants), while only 3 of 79 participants gave answers consistent with *base-rate neglect*. We consider this performance to be quite good, especially since half the participants were passengers at the airport, of all educational backgrounds, who completed the problem while waiting in line. Using the false positive scenario as a control, we replicated previous findings of frequent *base-rate neglect* (19 of 76 participants) and relatively low rates of *correct* answers (by classical Bayesian standards) (16 of 76 participants), demonstrating that our participants were not merely more statistically adept than the populations of previous studies.

Neither the natural frequency hypothesis nor the heuristics and biases view can account for our results, which, by classical Bayesian standards, show increased correct responses and few responses consistent with base-rate neglect on our newly developed benign cyst scenario. Tversky and Kahneman (1980), in explaining the results of their causal cab problem as well as Ajzen's (1977) study, state that "base-rate data that are given a causal interpretation affect judgments, while base-rates that do not fit into a causal schema are dominated by causally relevant data" (p. 50). Since the base rate of cancer itself is presented identically in the two scenarios, the heuristics and biases view cannot explain why it is more often used properly in the benign cyst scenario while so often "neglected" in the false positive scenario. The natural frequency hypothesis, in contrast, predicts poor performance on any statistical problem that does not involve natural frequencies, hence it cannot account for these results either.

In addition to gaining insight into how causality influences Bayesian inference tasks, we see implications for understanding the role of statistics in causal inference. The difficulty people have with the false positive statistic suggests that for situations in which an effect is generated by its causes, it may be difficult to interpret a statistic indicating the frequent occurrence of an effect in the absence of known causes. In this case, people may mistakenly interpret the statistic to be an error rate indicating the percentage of positive results that are wrong, in this case  $P(\neg cancer | +M)$ . Indeed, the majority of the responses we classified as *base-rate neglect* were answers of 94%, which would be correct if the false-positive statistic took the form  $P(\neg cancer | +M) = 6\%$  rather than  $P(+M | \neg cancer) = 6\%$ . One possible explanation is that people are more comfortable thinking in terms of  $P(\neg C | E)$ , the proportion of effect occurrences that are uncaused, rather than  $P(E | \neg C)$ , the rate of uncaused effect occurrences, and thus are likely to confuse  $P(E | \neg C)$  with  $P(\neg C | E)$ . Perhaps additional experiments could explore further the statistical format that best characterizes how people think about uncaused effects.

## 7.2 Experiment 2

In Experiment 1, we demonstrated that people often make accurate judgments about the uncertainty of a given cause being responsible for an observed effect. However, the mechanisms involved were described in essentially deterministic terms. In Experiment 2, we introduce a second source of uncertainty: probabilistic mechanisms. We again created a version of the mammogram problem (Eddy, 1982; Gigerenzer & Hoffrage, 1995), but this time we included the true positive rate,  $P(+M | cancer)$ , as well as the propensity of a benign cyst to cause a positive mammogram,  $P(+M | benign\ cyst)$ . This enabled us to test whether people reason appropriately

about the uncertainty introduced by probabilistic mechanisms. It also made the difficulty of the questions more similar to prior research, and provided a more rigorous test of people's judgment capabilities.

We also attempted to eliminate some potential confounds with the previous experiment. In Experiment 1, the salience and descriptive detail of the causal mechanism may have enabled people to pay more attention or think more clearly about the benign cyst scenario, which could account for their improved performance by classical Bayesian standards. For Experiment 2, both questions were written with only dry statistical information; instead of the description of the causal mechanism by which a benign cyst can lead to a positive mammogram, we provided only statistical data concerning the rate of benign cysts in the population and the likelihood of them causing a positive mammogram. We also made the two scenarios more similar to each other by specifying a relative frequency for the false positive rate in the original scenario, rather than a probability (we replaced "there is an X% chance that a woman without cancer will test positive" with "X% of women without cancer test positive"), which eliminates the potential confound between probabilities in one question vs. relative frequencies in the other.

### **7.2.1 Method**

*Participants.* The participants in this experiment were 59 MIT undergraduates. They were recruited during a study break in an undergraduate dormitory, and were given token compensation.

*Materials.* We again randomly assigned participants to one of two scenarios based on Gigerenzer and Hoffrage's (1995) probabilistic mammogram question (adapted from Eddy, 1982). The *false positive scenario* was similar to the original question, while the *benign cyst scenario* gave the base rate of benign cysts in the population, as well as a likelihood of a positive



mammogram given a benign cyst, rather than the rate of false positives. Under the classical Bayesian norm, the two calculations were equally difficult, ensuring that any improvement on the benign cyst scenario over the false positive scenario would not be due to an easier computation. The correct Bayesian calculation for the false positive scenario was:

$$P(H | D) = \frac{P(H) * P(D | H)}{P(H) * P(D | H) + P(-H) * P(D | -H)}$$

while the (approximately) correct Bayesian calculation for the benign cyst scenario was:

$$P(H | D) = \frac{P(H) * P(D | H)}{P(H) * P(D | H) + P(A) * P(D | A)} \quad (\text{where } A \text{ is the alternate cause, benign cysts})$$

#### 7.2.1.1 False Positive Scenario

Doctors often encourage women at age 50 to participate in a routine mammography screening for breast cancer. From past statistics, the following is known:

1% of the women had breast cancer at the time of the screening

Of those with breast cancer, 80% received a positive result on the mammogram

Of those without breast cancer, 15% received a positive result on the mammogram

All others received a negative result

Suppose a woman gets a positive result during a routine mammogram screening.

Without knowing any other symptoms, what are the chances she has breast cancer?

#### 7.2.1.2 Benign Cyst Scenario

Doctors often encourage women at age 50 to participate in a routine mammography screening for breast cancer. From past statistics, the following is known:

1% of the women had breast cancer at the time of the screening

Of those with breast cancer, 80% received a positive result on the mammogram

30% of the women had a benign cyst at the time of the screening

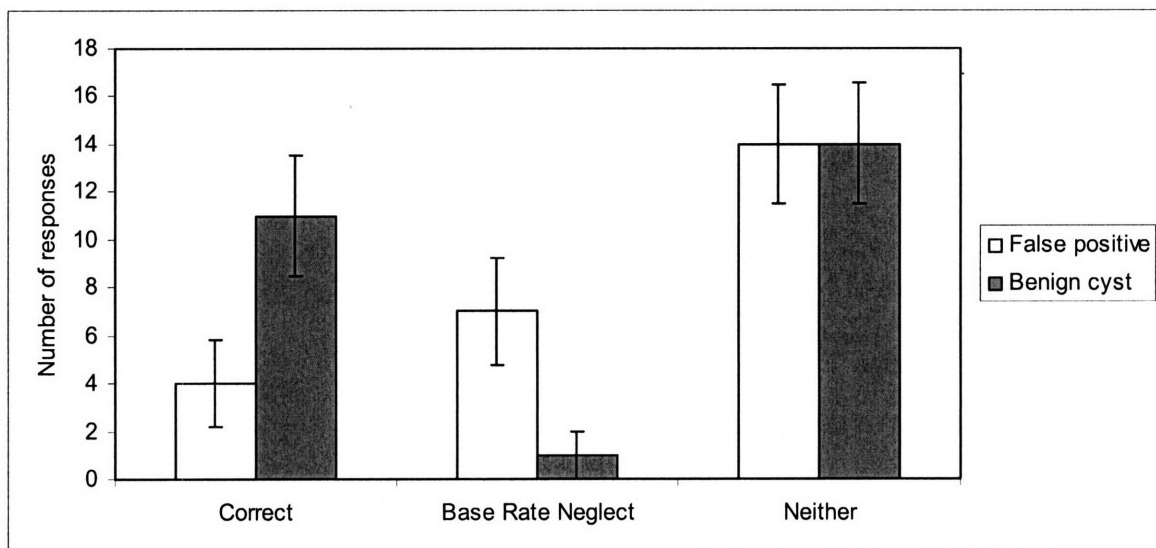
Of those with a benign cyst, 50% received a positive result on the mammogram

All others received a negative result

Suppose a woman gets a positive result during a routine mammogram screening.

Without knowing any other symptoms, what are the chances she has breast cancer?

### 7.2.2 Results



**Figure 30: Histogram of responses to Experiment 2. The correct answer was 5.1%. Responses were classified as correct (5.1%), base-rate neglect (>=65%), and other. A significant difference was found between false positive and benign cyst scenarios (Fisher’s Exact Test,  $p < .05$ ). Error bars represent the standard error of the normal approximation to the binomial distribution.**

The correct answer to the both scenarios is  $\frac{1\% \times 80\%}{1\% \times 80\% + 15\%} \approx 5.1\%$  (approximately). We

classified as *correct* only exactly correct answers, and we again classified as *base-rate neglect* any answer over  $P(D|H) - P(D|\neg H)$ , the lower bound of typical base-rate neglect answers (this lower bound was 65% in this problem). All other answers were classified as *neither*, and

were distributed across the spectrum with no discernable pattern. Our results show that *base-rate neglect* was significantly lower and *correct* answer rates were significantly higher on the benign cyst scenario (Fisher's Exact Test<sup>4</sup>,  $p < .05$ , see Figure 30). The size of the effect was also considerable, with *correct* responses more than doubled (11 of 30 on the benign cyst scenario vs. 4 of 29 on the false positive scenario), and responses consistent with *base-rate neglect* nearly eliminated (1 of 30 on the benign cyst scenario vs. 7 of 29 on the false positive scenario).

### 7.2.3 Discussion

The results of Experiment 2 again support our hypothesis that people generally make judgments consistent with the causal Bayesian framework. The benign cyst scenario in this case provided two sources of uncertainty: (1) multiple causes of an observed effect, and (2) probabilistic mechanisms by which those causes produce the effect. Although the arithmetic complexity involved prevented many participants from providing exactly correct responses, performance was relatively good overall, suggesting that people are often able to interpret and reason appropriately about statistics describing multiple probabilistic mechanisms.

Experiment 2 also addressed a number of potential confounds with Experiment 1, the most important of which was that the benign cyst scenario was much more descriptive of the mammogram mechanism than the false positive scenario, and hence potentially more salient. In Experiment 2, we did not describe the mechanism, and neither did we explicitly state that benign cysts were an alternative cause for positive mammograms. Merely by structuring the statistics in terms of an alternate cause, as opposed to a rate of false positives, we were able significantly

---

<sup>4</sup> The results are also significant by a  $\chi^2$  test, but because one of the cells of the expected distribution was less than 5, the  $\chi^2$  test is considered unreliable and therefore Fisher's Exact Test should be used.

reduce responses consistent with *base-rate neglect* and significantly increase *correct* answers (according to classical Bayesian norms). This suggests that if the existence of a mechanism seems plausible, such as a mechanism by which benign cysts can generate positive mammograms, one need not understand the mechanism entirely to reason about it appropriately.

### 7.3 Experiment 3

In Experiment 2, we demonstrated performance improvements on the mammogram problem even when the salient description of the causal mechanism was removed. However, there are two other important ways in which the false positive scenario of Experiment 1 may have been less salient than the benign cyst scenario. First, the statistics in the benign cyst scenario had *parallel construction*, in that both sentences began with “About X%”, which may have provided a cue to the correct idea that the two statistics should be compared to each other. Secondly, the false positive statistic might be less believable than the benign cyst statistic, which might lead people to disregard it or replace it with their own estimate of the false positive rate.

In Experiment 3 we attempted to eliminate these potential confounds. We removed the parallel construction in the benign cyst scenario, making the scenarios more similar by replacing the relative frequency in this scenario (“About 6% of those without cancer have a dense but harmless cyst”) with a probability (“There is a 6% chance that a woman without cancer will have a benign cyst”). We also asked participants to rate the believability of the statistics given, to determine whether people believe the benign cyst statistic more than the false positive statistic.

### 7.3.1 Method

*Participants.* The participants in this experiment were 60 MIT undergraduate and graduate students (majors were not recorded, but were likely randomly distributed). They were recruited in a main corridor on campus and given token compensation.

*Materials.* We again randomly assigned participants to one of two paper and pen scenarios based on those of Experiment 1. The *false positive scenario* was almost identical to that of Experiment 1, while the *benign cyst scenario* was made more similar to the *false positive scenario*, giving the probability of having a benign cyst, rather than the rate of benign cysts in the population. We also removed the word “about” in both conditions to indicate that the statistics were exact rather than estimated. Believability ratings were requested on the page following the question.

#### 7.3.1.1 False Positive Scenario

The following statistics are known about women at age 60 who participate in a routine mammogram screening, an X-ray of the breast tissue that detects tumors:

2% of women have breast cancer at the time of the screening. Most of them will receive a positive result on the mammogram.

There is a 6% chance that a woman without breast cancer will receive a positive result on the mammogram.

Suppose a woman at age 60 gets a positive result during a routine mammogram screening. Without knowing any other symptoms, what are the chances she has breast cancer?

### 7.3.1.2 *Benign Cyst Scenario*

The following statistics are known about women at age 60 who participate in a routine mammogram screening, an X-ray of the breast tissue that detects tumors:

2% of women have breast cancer at the time of the screening. Most of them will receive a positive result on the mammogram.

There is a 6% chance that a woman without breast cancer will have a dense but harmless cyst that looks like a cancerous tumor and causes a positive result on the mammogram.

Suppose a woman at age 60 gets a positive result during a routine mammogram screening. Without knowing any other symptoms, what are the chances she has breast cancer?

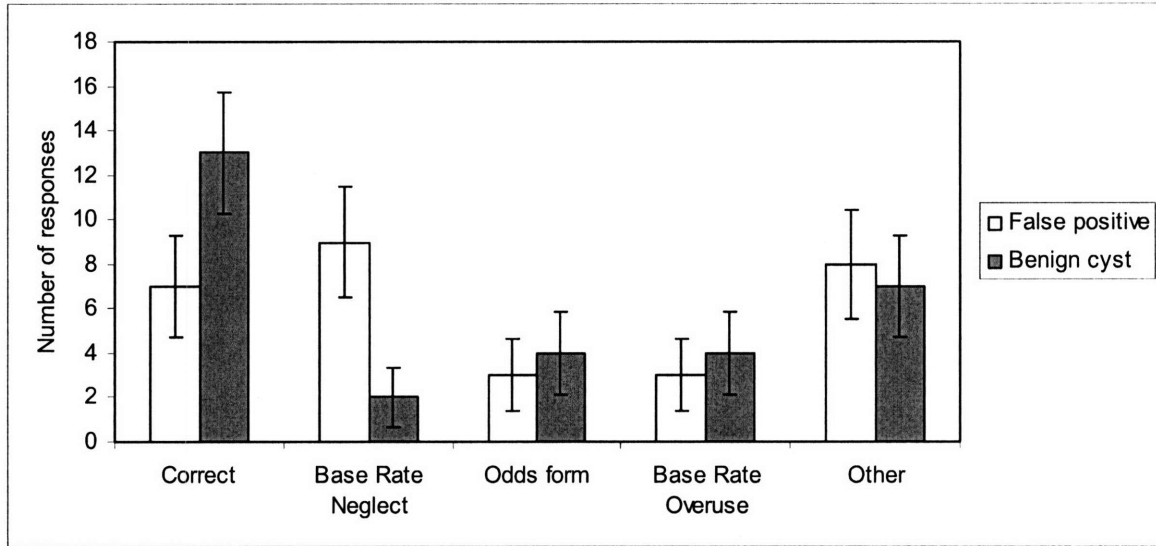
### 7.3.1.3 *Believability ratings*

The problem on the previous page gave several statistics about breast cancer and mammogram tests. To what extent do you think the given statistics accurately reflect the real-world probabilities of these events?

Answer on a scale of 1 to 7 (circle one):

1	2	3	4	5	6	7
Very inaccurate			Not sure			Very accurate

### 7.3.2 Results



**Figure 31: Histogram of responses to Experiment 3. The correct answer was 25%. Responses were classified as correct (20%-25%), base-rate neglect ( $\geq 75\%$ ), odds form (33%), base rate overuse (2%), and other. For analysis purposes, responses of odds form and base rate overuse were grouped with other. A significant difference was found between false positive and benign cyst scenarios ( $\chi^2(2) = 6.28, p < .05$ ). Error bars represent the standard error of the normal approximation to the binomial distribution.**

We again obtained significantly improved performance on the benign cyst scenario. Responses consistent with *base-rate neglect* were significantly lower and *correct* answer rates were significantly higher on the benign cyst scenario ( $\chi^2(2) = 6.29, p < .05$ , see Figure 31). The approximately correct answer to both scenarios (by the standards of classical Bayesian inference) was:  $\frac{2\%}{2\% + 98\% \times 6\%} \approx \frac{2\%}{2\% + 6\%} = 25\%$ . We classified as *correct* answers of exactly 25%, and classified as *base-rate neglect* any answer over 80% (however there were no responses between 80% and 90%). We also found several answers of *odds form* (answers of 2/6 or 33%) and *base rate overuse* (answers of 2%). All remaining responses were classified as *other* (answers of *odds*

*form* and *base rate overuse* are shown in Figure 31, but were collapsed with *other* for the statistical analysis). The results also indicate no significant difference in believability ratings between the two scenarios, with a trend towards the statistics of the benign cyst scenario being *less* believable (4.03 average rating for the benign cyst scenario vs. 4.37 average rating for the false positive scenario). From these results, we can draw a conclusion at the  $\alpha=0.05$  significance level that the benign cyst scenario is no more than 3% more believable than the false-positive scenario.

### 7.3.3 Discussion

The results of Experiment 3 confirm that the results of Experiment 1 were not due merely to differences in the believability of the statistics or the salience of the parallel statistical structure. We found that the believability of the false positive scenario was no different than that of the benign cyst scenario, thus it is unlikely that people might be replacing the false positive statistic with their own estimate of the false positive rate. Furthermore, we replicated the results of Experiment 1 using identical (non-parallel) statistical structure in the two conditions.

## 7.4 Experiment 4

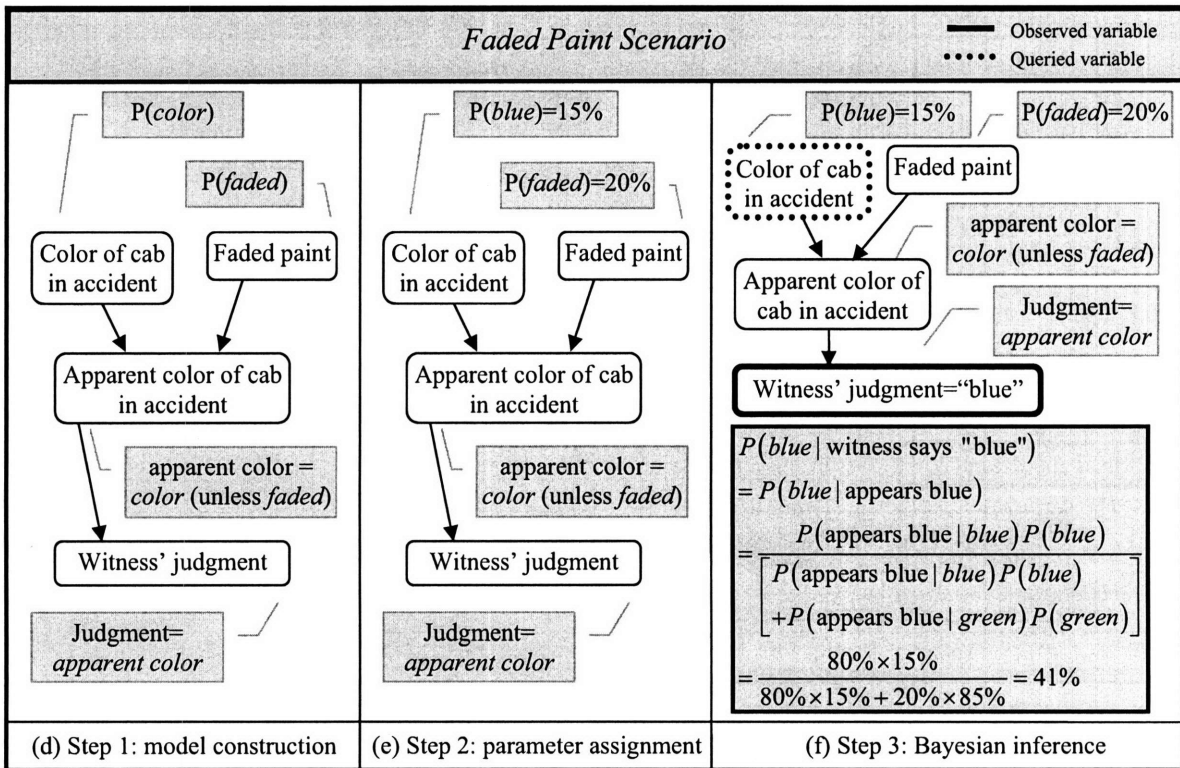
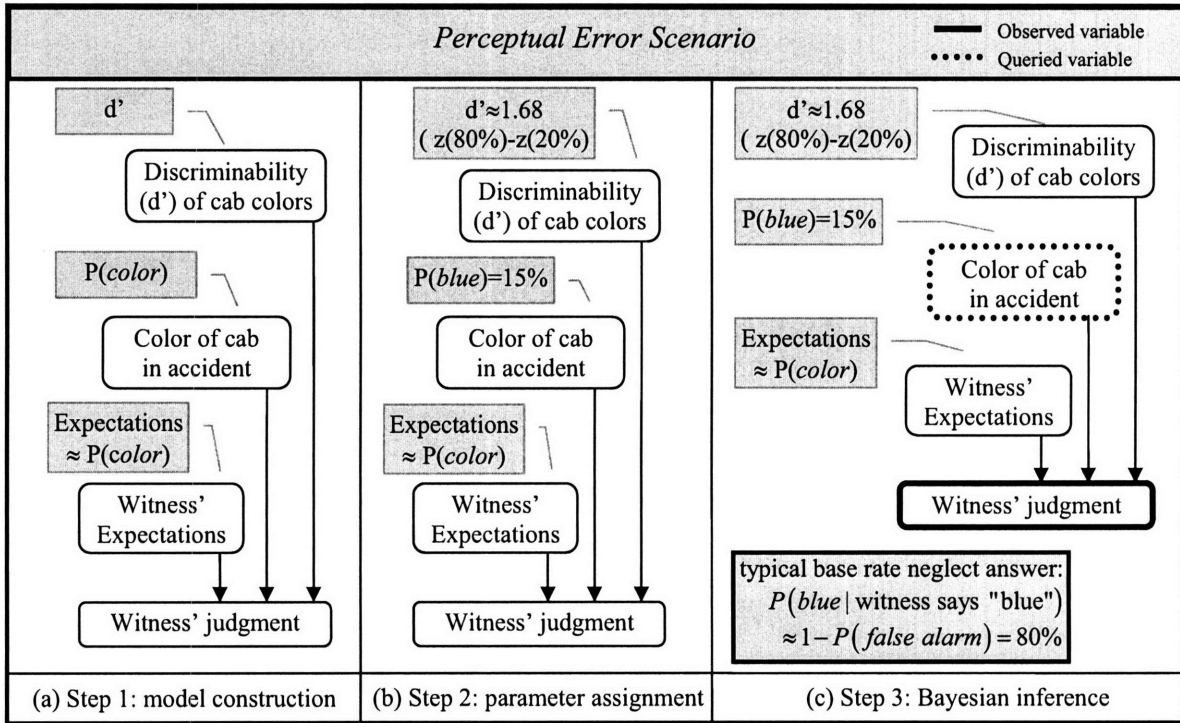
Experiments 1, 2 and 3 demonstrated that statistics that can be unambiguously assigned to causal model parameters are often used appropriately. This finding bears some resemblance to previous appeals to “causal relevance” in explaining base-rate neglect (Ajzen, 1977; Tversky & Kahneman, 1980), but our causal Bayesian account is importantly different. Under previous “causal relevance” accounts, Bayes’ rule is the prescribed method of judgment, and a statistic is used or not depending on whether or not it seems “causally relevant”. In contrast, the prescribed method of judgment in our framework is Bayesian inference over causal models, and statistics



are primarily used to update parameters of the model. Sometimes a statistic may be appropriately used in updating the model, but may not be required in the final judgment, reflecting apparent neglect of the statistic. This is what we propose happens when people seem to neglect the base rate in the “cab problem” (Kahneman & Tversky, 1972).

In the causal Bayesian framework, one must first construct a causal model from domain knowledge prior to updating it with provided statistics. In this case, the causal model may be more complex than anticipated by previous researchers. It is common knowledge that perceptual discriminations are strongly influenced not only by the intrinsic discriminability of objects in the world but also by prior knowledge, such as stereotypes or more mundane expectations. For instance, if you feel a rough oblong object in a fruit bowl you might guess it is a pear, whereas if you feel the same object in a vegetable drawer you might guess it is a potato. More formally, signal detection theory captures the contributions of both discriminability ( $d'$ ) and prior expectations (through criterion shifts of the area under the occurrence distributions). This results in three causes of the witness' judgment (see Figure 32b), and many more parameters than there are statistics provided. The statistics given in the problem are only sufficient to describe a much simpler causal model with a single binary cause (the color of the cab).

Since prior expectations influence people's judgments, unexpected judgments may be even more accurate than expected ones. Someone who thinks he found a pear in a vegetable drawer probably inspected it more closely than someone who thinks he found a potato in that drawer. According to Birnbaum (1983), ignoring the base rate of cabs may be justified if the witness' judgment can be characterized by signal detection theory. Birnbaum's (1983) analysis suggests that for an optimizing witness whose expectations match the true base rate, the probability of the witness being correct remains at approximately 80% across a large range of



**Figure 32: The three phases of causal Bayesian inference for Experiment 4. (a-c) depict one possible causal interpretation of the perceptual error scenario, in which the witness' judgment obeys signal detection theory, and the witness' error rate could be used to infer the discriminability of the colors. (d-f) depict the causal structure described by the faded paint scenario, in which the given statistics map more clearly onto parameters of the model.**

base rates. This means ignoring the base rate may actually be appropriate, because the witness has essentially already taken it into account. While this explanation has been offered before, it has not been used to explain findings that participants performed better on the version with a “causal” base rate (Tversky & Kahneman, 1980). The intuition on this problem seems to be that one must weigh the witness' testimony that the cab was blue against the countervailing evidence that green cabs are more reckless. This reflects an assumption that the witness did not take this countervailing evidence into account when making his judgment, thus participants may assume that the witness' expectations (50% blue, according to the population base rate) do not match the base rate of cabs in accidents (15% blue). The mismatch should reduce the witness' accuracy under a signal detection analysis, as participants' responses indicated. Although this account may be plausible, it is essentially speculative because we do not know what causal model people have in mind when solving these problems. It could be that people really have a signal detection model in mind, and their judgments are accurate by that model. Alternatively, people could merely have the intuition that reliability in perception should not vary across base rates.

In this experiment, we again seek to demonstrate that base rates that have been previously found to be neglected are actually generally used as prescribed by the causal Bayesian norm. As we argued earlier, the notion that population base rates do not fit into causal schemas is unjustified. Uncaused variables (those with no parents) in a causal Bayes net require a prior probability, and population base rates are an excellent way to set these priors. In this experiment,

we demonstrate that even the base rate in the original cab problem, previously labeled “non-causal”, can be easily interpretable and used properly. In our new *faded paint scenario*, 20% of green cabs have faded paint, making them appear blue, and 20% of blue cabs have faded paint, making them appear green, while the witness accurately reports the apparent color of the cab. This results in an identical computation (under classical Bayesian inference) to the original cab problem, but since it is clear how the model should be structured and how the given statistic should be assigned to its parameters, a correct solution can be prescribed using the causal Bayesian framework. The three phases of causal Bayesian inference for this scenario are depicted in Figure 32(d-f). Like the previous experiments, we tested people’s performance on our new scenario while using the original *perceptual error scenario* as a control.

#### **7.4.1 Method**

*Participants.* The 47 participants in this experiment were MIT undergraduate and graduate students (majors were not recorded, but were likely randomly distributed). They were approached in a student center, and given token compensation.

*Materials.* Participants were randomly assigned to receive one of two variants of Kahneman and Tversky’s cab problem in a between-subjects design. The *perceptual error scenario*, modeled after the original cab problem, attributed the witness’ mistakes to perceptual error. The *faded paint scenario* attributed the witness’ mistakes to faded paint. Both questions require the exact same calculation, so they are equally difficult to solve with classical Bayesian methods. The questions follow:

#### *7.4.1.1 Perceptual Error Scenario*

A cab was involved in a hit and run accident at night. Two cab companies operate in the city, the Green Co. and the Blue Co. (according to the color of the cab they run).

You know that:

(a) 85% of the cabs belong to the Green Co., and the remaining 15% belong to the Blue Co.

(b) A witness later identified the cab as a Blue Co. cab. When asked how he made the identification, he said the cab appeared blue in color.

(c) The court tested the witness' ability to identify cabs under similar visibility conditions. When presented with a sample of cabs, the witness identified only 80% of the Blue Co. cabs correctly, and only 80% of Green Co. cabs correctly. 20% of Blue Co. cabs were mistaken for Green Co. cabs, and 20% of Green Co. cabs were mistaken for Blue Co. cabs.

Even though the witness identified the cab as belonging to the Blue Co., it could have belonged to the Green Co. What do you think the chances are that the witness was right and the cab really belonged to the Blue Co.?

#### *7.4.1.2 Faded Paint Scenario*

A cab was involved in a hit and run accident at night. Two cab companies operate in the city, the Green Co. and the Blue Co. (according to the color of the cab they run).

You know that:

(a) 85% of the cabs belong to the Green Co., and the remaining 15% belong to the Blue Co.

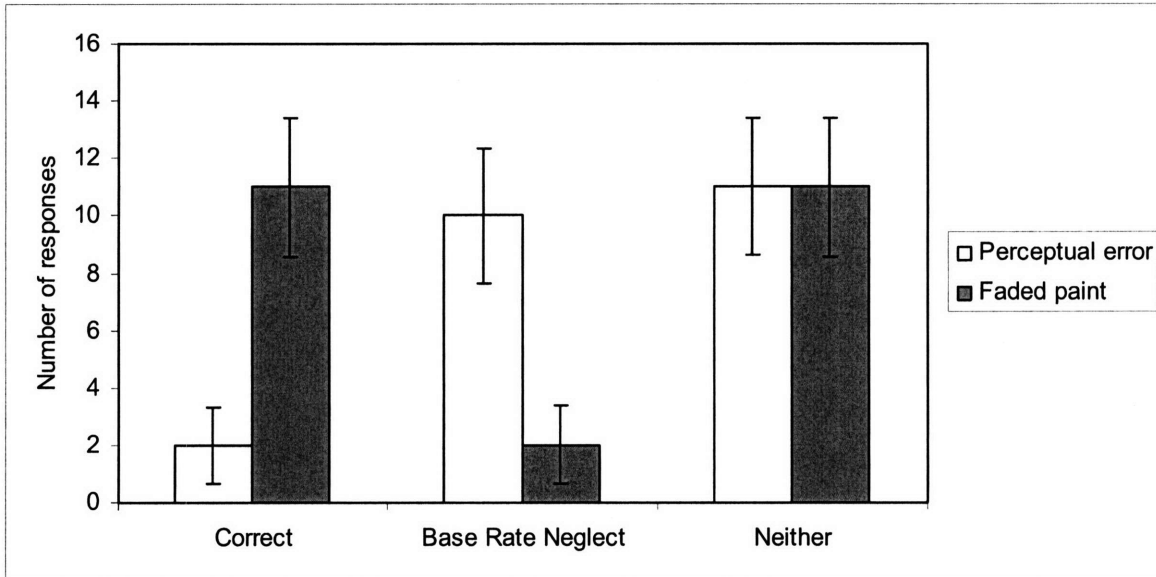
(b) A witness later identified the cab as a Blue Co. cab. When asked how he made the identification, he said the cab appeared blue in color.

(c) The court tested the appearance of cabs under similar visibility conditions. When testing a sample of cabs, only 80% of the Blue Co. cabs appeared blue in color, and only 80% of the Green Co. cabs appeared green in color. Due to faded paint, 20% of Blue Co. cabs appeared green in color, and 20% of Green Co. cabs appeared blue in color.

Even though the witness identified the cab as belonging to the Blue Co., it could have belonged to the Green Co. What do you think the chances are that the witness was right and the cab really belonged to the Blue Co.?

#### 7.4.2 Results

We classified as *correct* only answers that matched the normative solution by classical Bayesian standards:  $\frac{15\% \times 80\%}{15\% \times 80\% + 85\% \times 20\%} = \frac{12\%}{29\%} = 41\%$ . Consistent with prior studies, we classified as *base-rate neglect* any response of 80% or above. All other answers were classified as *neither*, and were distributed across the spectrum with no discernable pattern. The results show a significant reduction in *base-rate neglect* and a significant increase in *correct* responses for the faded paint scenario ( $\chi^2(2) = 11.55, p < .005$ ) (see Figure 33). The effect was very large, with *correct* responses dramatically improved (11 of 24 on the faded paint scenario vs. 2 of 23 on the perceptual error scenario), and responses consistent with *base-rate neglect* nearly eliminated (2 of 24 on the faded paint scenario vs. 10 of 23 on the perceptual error scenario).



**Figure 33: Histogram of responses to Experiment 4. Error bars represent the standard error of the normal approximation to the binomial distribution. The difference between conditions was large and highly significant ( $\chi^2(2) = 11.55, p < .005$ ).**

### 7.4.3 Discussion

The results show generally strong performance on the faded paint scenario, with nearly half the participants giving the exact correct solution on a mathematically difficult task. Using the original scenario as a control, we found that our participants are not merely more statistically adept than other populations, as they exhibited the classic pattern of responses consistent with base-rate neglect on the original scenario. These results are not predicted by the heuristics and biases view, which predicts that people use base rates that seem causally relevant, but neglect those that do not; in this case, the base rates are identical, yet people perform much better on the faded paint scenario. The results also cannot be accounted for by the natural frequency hypothesis, which predicts universally poor performance on any problem involving probabilities or relative frequencies (statistics expressed as percentages, but lacking a sample size).

Like in the previous experiments, we see important implications not just for the role of causality in judgments tasks, but also for the role of statistics in judgments from causal models. The cab problem is controversial in part because perception is a much more complicated process than can be summed up in two conditional probabilities. People's neglect of the base rate suggests that people may have a special model for perception in which accuracy remains relatively constant across base rates, but varies with the confusability of the items to be perceived, as would be expected if people are optimal signal detectors. This perceptual model may be one of many such models that people have for a variety of specific domains that enable them to make reasonably accurate qualitative judgments about complex mechanisms.

## 7.5 Experiment 5

In this experiment, we demonstrate more directly that causal structure should influence whether base rates will be used, and that the classical Bayesian framework, which ignores causal structure, can sometimes fail to provide an appropriate standard for judgment. The previous four experiments all compared scenarios that differed considerably in their descriptions, introducing possible imbalances in salience or plausibility that could be implicated in our results. In this experiment, we adopted stimuli that are nearly identical, except for a manipulation of base rate and requested judgment. All the scenarios were newly developed, modeled after social judgment tasks such as Kahneman and Tversky's (1973) famous lawyer-engineer problem, which asked participants to predict a man's career from his personality. One of the stimuli from that experiment follows:

A panel of psychologists have interviewed and administered personality tests to 30 [70] engineers and 70 [30] lawyers, all successful in their respective fields. On the basis of this information, thumbnail descriptions of the 30 [70] engineers and 70 [30]



lawyers have been written. You will find on your forms five descriptions, chosen at random from the 100 available descriptions. For each description, please indicate your probability that the person described is an engineer, on a scale from 0 to 100.

One of the descriptions follows: Jack is a forty-five year old man. He is married and has four children. He is generally conservative, careful, and ambitious. He shows no interest in political and social issues and spends most of his time on his many hobbies, which include carpentry, sailing, and mathematical puzzles.

The probability that Jack is one of the 30 engineers in the sample of 100 is \_\_\_%.

Kahneman and Tversky's (1973) findings were that people gave nearly the same response regardless of whether the personality test was administered to 30 engineers and 70 lawyers, or to 70 engineers and 30 lawyers. This violates classical Bayesian principles, because the base rate of engineers,  $P(\text{Engineer}) = 30\%$  or  $P(\text{Engineer}) = 70\%$ , should influence judgments of  $P(\text{Engineer} | \text{Personality})$ . In this experiment we investigated two sets of scenarios in which the base rate is varied from 30% to 70%, as it is in the lawyer-engineer problem.

The first set of scenarios involved college classes, while the second set of scenarios involved college sports teams. The class scenarios were a poetry class, which appealed mostly to women, and an electrical engineering class, which appealed mostly to men. Because women are shorter than men, the students in the poetry class were mostly 5'7" or under, while the students in the electrical engineering were mostly over 5'7". The sports teams were a horse racing team, which selected mostly people 5'7" or under, and a volleyball team, which selected mostly people over 5'7". Because it was easier to find short women than short men, the horse racing team had

mostly women, and because it was easier to find tall men than tall women, the volleyball team had mostly men.

For each set of scenarios, two different tasks were posed (each participant received only one task). The first task was to judge the probability that a male student randomly selected from the class or team is over 5'7":  $P(\text{height} > 5'7" | \text{gender} = \text{male})$ . According to the causal Bayesian framework, in the judgment of height from gender, one should ignore the base rate of heights in the class scenarios, but not in the team scenarios. Intuitively, this makes sense: a male student in the poetry class is no more likely to be short than a male student in the electrical engineering class, even if the base rate of people over 5'7" is 30% in the poetry class and 70% in the electrical engineering class. However, a male member of the horse racing team is much more likely to be short than a male member of the volleyball team, thus the difference in base rate of heights matters. This distinction between classes and teams derives from their causal structures, depicted in Figure 34(a) and (d). Given the causal model Figure 34(a), learning the gender of a student renders the fact that the student is in the poetry class irrelevant to judging the student's height, because the path from *InPoetryClass* to *Height* is blocked by *Gender* (technically, the causal markov condition of Bayesian networks specifies that *InPoetryClass* is conditionally independent of *Height* given *Gender*). Because it is irrelevant that the student is in the poetry class, the base rate of heights in the poetry class is also irrelevant (and similarly for the electrical engineering class).

The second task was to judge the probability that a randomly selected student over 5'7" is male:  $P(\text{gender} = \text{male} | \text{height} > 5'7")$ . By reversing which variable was observed and which was judged, we also reversed which scenario required ignoring the base rate. Intuitively, because the poetry class has more females than the electrical engineering class, more of the tall poets are

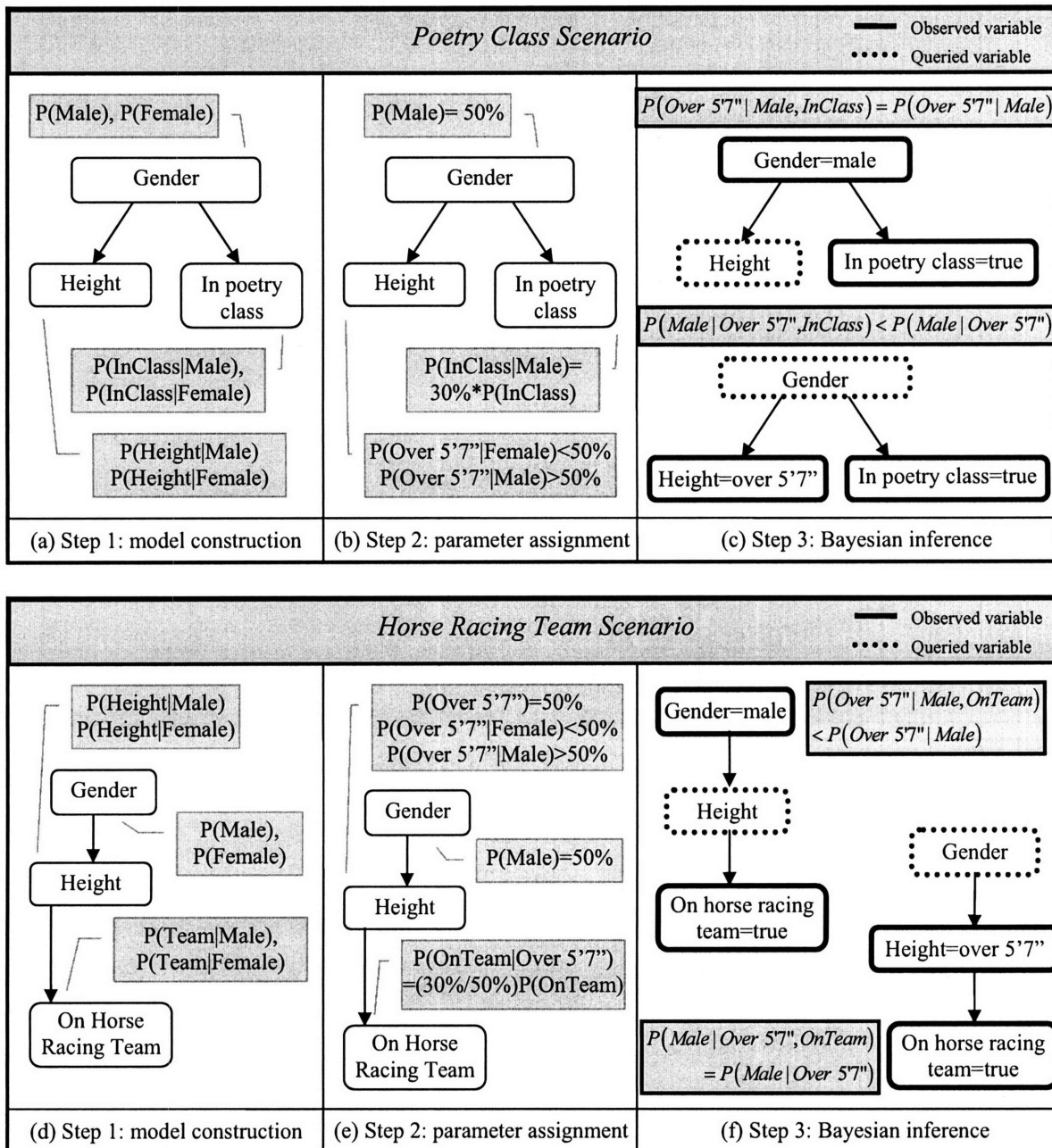


Figure 34: The three phases of causal Bayesian inference for Experiment 5. (a-c) depict the poetry class scenario, while (d-f) depict the horse-racing team scenario.

female than the tall engineers, thus the base rate of gender in the class matters. Conversely, a tall person on the horse racing team is no more likely to be male than a tall person on the volleyball team, assuming that gender does not influence selection for the team. Formally, the path from

*Gender* to *InPoetryClass* is not blocked by height, therefore the fact that the tall person is in the poetry class is relevant to judging the person's gender, and the base rate of gender in the class is relevant. In contrast, the path from *Gender* to *OnHorseRacingTeam* is blocked by *Height*, therefore the fact that the tall person is on the team is irrelevant for judging the person's gender, and the base rate of heights on the team should be ignored.<sup>5</sup>

With Experiment 5, we used the gender-height scenarios to test whether people use or ignore the base rate appropriately according to the causal Bayesian framework. This framework prescribes that people should ignore the base rate when the observed variable blocks the path between the variable to be judged and the variable representing being in the class or on the team, and should use the base rate when the variable to be judged falls between the observed variable and the variable representing being in the class or on the team.

### 7.5.1 Method

*Participants.* The 120 participants in this experiment were MIT undergraduates and graduate students (majors and years were not recorded, but were likely randomly distributed). They were approached in a student center, and were given token compensation.

*Materials.* The materials were modeled after the lawyer-engineer problem (Kahneman & Tversky, 1973), which compared judgments of  $P(\text{engineer} | \text{personality})$  in conditions of low vs. high base rate. We compared people's judgments of  $P(\text{height} | \text{gender})$  and  $P(\text{gender} | \text{height})$  in low base rate vs. high base rate conditions for two scenarios: *class* and *team*. The *class* scenario compared the low base rate poetry class (few male, thus few over 5'7")

---

<sup>5</sup> This is only true if enrollment on the horse racing team is not causally influenced by gender. Since sports teams usually have gender biases, this may not match everyone's intuition, as we discuss in our results.

to the high base rate engineering class (mostly male, thus mostly over 5'7"). The *team* scenario compared the low base rate horse racing team (few over 5'7", thus few male) to the high base rate volleyball team (mostly over 5'7", thus mostly male). Participants were randomly assigned to one of eight different conditions, as we crossed three factors: (1) *class* vs. *team*, (2) *low base rate* vs. *high base rate*, and (3) judging  $P(\text{height} | \text{gender})$  vs.  $P(\text{gender} | \text{height})$ . For judgments of  $P(\text{height} | \text{gender})$ , the given base rate was 30% over 5'7" (low base rate) or 70% over 5'7" (high base rate). For judgments of  $P(\text{gender} | \text{height})$ , the given base rate was 30% male (low base rate) or 70% male (high base rate). The materials follow: [brackets separate low base rate condition from high base rate condition] and {curly braces separate  $P(\text{height} | \text{gender})$  condition from  $P(\text{gender} | \text{height})$  condition}.

*1. Class scenario:*

A college [poetry | engineering] class has {mostly | 70%} [women | men].

Because it has mostly [women | men], {70% | most} of the people in the class are [5'7" or under | over 5'7"]. Suppose you meet someone from the class who is {male | over 5'7"}.

What do you think the chances are that {he is over 5'7" | this person is male}?

2. *Team scenario*:<sup>6</sup>

A college has a co-ed [horse-racing | volleyball] team. The team tries to recruit [short | tall] people because [shorter | taller] people make better [horse-racers | volleyball players]. In fact, {70% | almost all} of the people on the team are [5'7" or under | over 5'7"]. Since it's easier to find [short women | tall men] than [short men | tall women], {there are slightly more [female racers than male | male players than female] | 70% of the [racers are female | players are male]}. Suppose you meet someone from the team who is {male | over 5'7"}. What do you think the chances are that {he is over 5'7" | this person is male}?

In responding, we asked participants to select one of five qualitative probabilities:

Very Low      Low      Medium      High      Very High

---

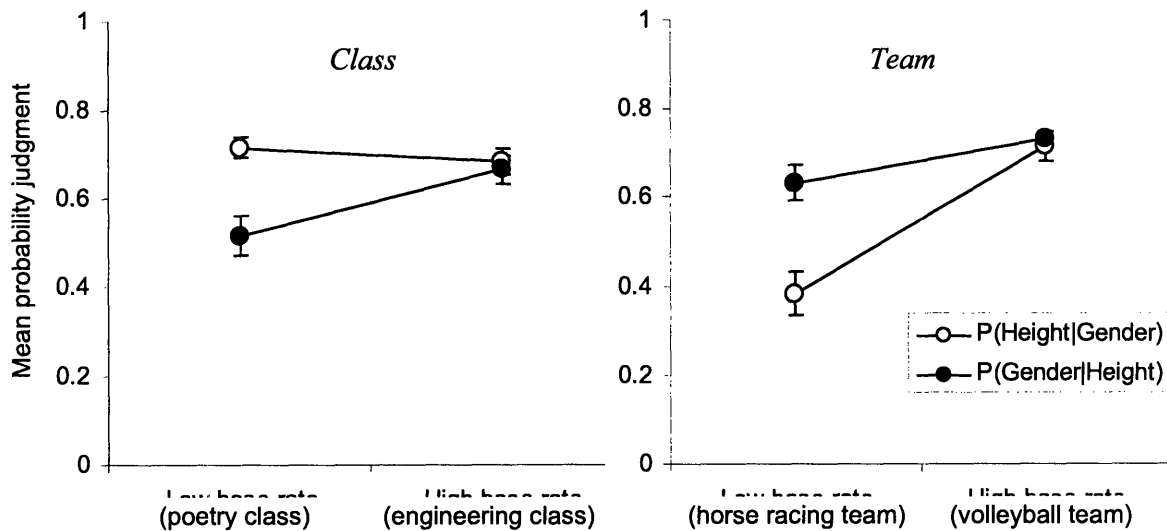
<sup>6</sup> Despite attempts to encourage the use of the model in Figure 36(a) by highlighting the causal influences, we inferred from pilot studies that in the sports team condition participants entertained the more natural model of Figure 36 (b), in which a link exists directly from gender to being on the team. This model is more natural because sports teams are rarely gender-blind (even co-ed teams often try to balance the number of male and female players). To discourage this model, we adjusted some language to make the statistic for which the base rate was not given (either gender or height) consistent with the model in Figure 36(b); whereas in pilot studies we had use the word “most”, in the experimental study we said the team had “slightly more” women than men in the case where the base rate of heights on the team is 70% 5'7" or under, and we said the team was “almost all” under 5'7" in the case where the base rate of gender is 70% female. This fine tuning of the non-base rate statistic was not necessary in the poetry/engineering class question (where we used the word “most” in both cases) because there is no reason to believe a causal link exists between one’s height and whether one is in the class, so there was no confusion about which model to use.

## 7.5.2 Results

The results strongly indicate that people use or ignore base rates according to the prescriptions of our proposed causal Bayesian norm. For ANOVA purposes, we coded the qualitative probabilities as follows: Very Low: 10%, Low: 25%, Medium: 50%, High: 75%, Very High: 90%. In a two-way ANOVA focusing just on the *class* scenario, a significant interaction was found between the *judgment* and the *base rate* ( $F(1, 56)=7.56, p<.01$ ). This indicates that the base rate manipulation in the *class* scenario affected judgments of  $P(\text{height} | \text{gender})$  differently than judgments of  $P(\text{gender} | \text{height})$ . These differences followed our predictions. Although the *poetry class* and *engineering class* had opposite base rates of heights, the judgment of  $P(\text{height} > 5'7" | \text{gender} = \text{male})$  did not differ significantly between the two conditions (Mann-Whitney,  $p=0.27$ , see Figure 35), indicating that people correctly ignored the base rate of heights. In contrast, judgments of  $P(\text{gender} = \text{male} | \text{height} > 5'7")$  in the *poetry class* condition were significantly lower than in the *engineering class* condition (Mann-Whitney,  $p<.05$ , see Figure 35), indicating people's judgments were correctly influenced by the base rates.

In a two-way ANOVA focusing just on the *team* scenario, a significant interaction was found between the *judgment* and the *base rate* ( $F(1,56)=10.09, p<0.005$ ). This indicates that the base rate manipulation in the *team* scenario affected judgments of  $P(\text{height} | \text{gender})$  differently than judgments of  $P(\text{gender} | \text{height})$ . These differences again followed our predictions. The *horse racing team* and the *volleyball team* had opposite base rates of heights, and judgments of  $P(\text{height} > 5'7" | \text{gender} = \text{male})$  in the *horse racing team* condition were significantly lower than in the *volleyball team* condition (Mann-Whitney,  $p<.001$ , see Figure 35), indicating people's

judgments were correctly influenced by the base rates. Differences in judgments of  $P(\text{gender} = \text{male} | \text{height} > 5'7")$  were not significantly different, but were marginally significant (Mann-Whitney,  $p=0.06$ , see Figure 35), indicating that participants partially ignored, but may not have fully ignored the base rate of genders as prescribed by the causal Bayesian norm.



**Figure 35: Experiment 5 results across all conditions. The 3-way interaction graph shows how base rate use in judgments of  $P(\text{male} | \text{over } 5'7")$  and  $P(\text{over } 5'7" | \text{male})$  differs across class and team scenarios. A significant 3-way interaction was found ( $F(1,112)=17.64, p<0.0001$ ).**

We suspect that base rates were not fully ignored in the *team* condition for judgments of  $P(\text{gender} = \text{male} | \text{height} > 5'7")$  because some participants intuitively considered causal the model depicted in Figure 36(b), which contains a causal link from gender to team membership, despite our attempts to assert that gender does not influence selection for the team. This model was likely more intuitive than that of Figure 36(a) because in the real world, gender usually influences sports team membership (e.g., a team with mostly women but some men suggests that

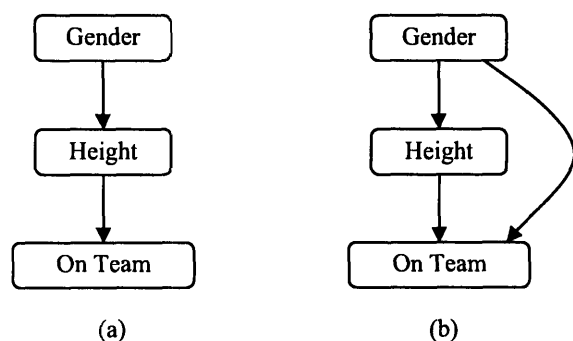


the sport appeals more to women than to men, or that women perform better than men). Since the volleyball team has 70% men whereas the horse racing team has 70% women, those participants who assume gender influences team membership should judge

$P(\text{male} | \text{over } 5'7", \text{ on volleyball team})$  to be higher than

$P(\text{male} | \text{over } 5'7", \text{ on horse racing team})$ , which would account for the marginally significant difference between the *low base rate* and *high base rate* conditions for the *team* scenario.

Experiment 6 was designed in part to eliminate this potential bias.



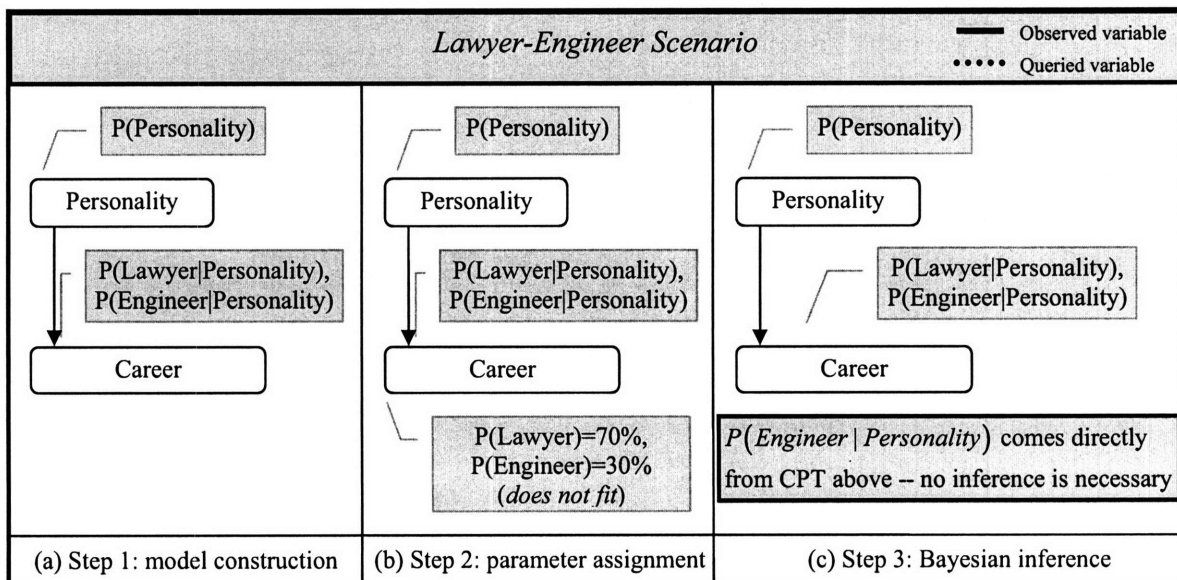
**Figure 36: Possible causal models representing the team scenario. The materials were designed to encourage participants to adopt the model of (a), although the model in (b) may better represent people’s existing knowledge.**

A 3-way ANOVA comparing all eight conditions revealed a highly significant 3-way interaction ( $F(1,112)=17.64, p<0.0001$ , see Figure 35). This indicates that the way base rate use differed between judgments of  $P(\text{height} | \text{gender})$  vs.  $P(\text{gender} | \text{height})$  varied across the scenarios of *class* vs. *team*. Base rates were ignored in judging  $P(\text{height} | \text{gender})$  for the *class* condition but used in the *team* condition, whereas base rates were used in judging  $P(\text{gender} | \text{height})$  for the *class* condition but largely ignored for the *team* condition. The only significant main effect in the 3-way ANOVA was for the base rate (low vs. high), indicating that

nothing about the scenario (class vs. team) or the judgment (gender vs. height) determines whether it should be ignored; rather, it is the interaction of the judgment with the scenario that determines whether the base rate is relevant.

### 7.5.3 Discussion

In several comparisons, we found strong effects of causal structure on judgments, demonstrating that people's use or neglect of the base rate varies according to what is rational under our proposed causal Bayesian norm. These results are not predicted by the heuristics and biases view, which proposed that people neglect the base rate when the individuating information seems more salient or more causally relevant than the base rate. In our questions, the base rate (e.g., 70% under 5'7") and the individuating information (e.g., you meet someone who is male) remain constant across conditions of *class* vs. *team*, yet we were able to induce use or neglect of the base rate by varying causal structure across the conditions. Appealing to the "causal relevance" of the base rate cannot explain these effects because in both cases the base rate is causally relevant: for the *team* scenarios, *height* causally influences selection for the team, while for the *class* scenarios, *height* is causally influenced by *gender*, which itself causally influences enrollment in the class. The results also cast doubt on the hypothesis that people use a representativeness heuristic (Kahneman & Tversky, 1973) in social judgment tasks. A representativeness heuristic would judge  $P(\text{male} | \text{over } 5'7")$  by the degree to a person over 5'7" represents a typical male, ignoring the base rate of males. Our results show that people use the low base rate of males in the *poetry class* scenario appropriately, judging a person over 5'7" in the poetry class as approximately equally likely to be male or female, contrary to the predictions of representativeness.



**Figure 37: Causal model for the lawyer-engineer problem. It is not clear how the interviewed variable should be connected causally to the rest of the model.**

The interaction of causal structure with base rate use may be responsible for a host of base-rate neglect findings on social judgment types of tasks, such as in Kahneman and Tversky's (1973) lawyer-engineer problem, described in the introduction to this experiment. Assuming that personality is a causal influence on the career choice of lawyers and engineers, the appropriate causal model for this scenario, including the three phases of causal Bayesian inference, is depicted in Figure 37. We are told that the base rate of lawyers and engineers interviewed is  $P(\text{lawyer} | \text{interviewed}) = 70\%$ ;  $P(\text{engineer} | \text{interviewed}) = 30\%$  (in one variant), but this base rate cannot be used to update the CPT for the *career* variable, because it has a parent in the model. The CPT for *career* only contains parameters for  $P(\text{career} | \text{personality})$ , not  $P(\text{career})$ . One could accommodate this base rate by including a variable for *interviewed*. However, we are not told how the lawyers and engineers were selected for interview. It is therefore not clear how a variable representing *interviewed* should be connected causally to the

rest of the model. If the *interviewed* variable is not causally connected, it is understandable that participants may ignore the base rate, assuming that whatever the causal connection is, learning the man's *personality*, a direct cause of *career*, renders the fact that the man was interviewed irrelevant to judging the man's *career*, just as learning the student's *gender* rendered the fact that the student was in a poetry class irrelevant to judging the student's *height*.

## 7.6 Experiment 6

In Experiment 5 we showed that whether people ignore or use base rates depends on causal structure, but causal structure was not the only difference between the scenarios. To reinforce the difference in causal structure between the conditions, we also relied on people's intuitive ideas about *poetry classes, engineering classes, horse racing teams, and volleyball teams*. This raises a potential confound: people's prior knowledge of the difference between these classes and teams could have been responsible for the effect. It could be the case, for instance, that people expect jockeys to be shorter than poets, and this alone explains why a male member of the horse racing team is judged shorter than a male member of a poetry class. Another issue with the materials of Experiment 5 is that they relied on people's intuitions about the heights of men and women, which could have led people to answer differently depending on, for instance, their intuition for the probability that a man is under 5'7".

We developed Experiment 6 to eliminate the potential confound of team and class names by keeping the type of team the same across conditions. This was done by describing a single scenario in which a CIA special operations team selected its members according to certain criteria. In one condition, part of the mission required female agents (like the poetry class), while in another condition, part of the mission required short agents (like the horse racing team). Causal models for these two scenarios are depicted in Figure 38. We also fashioned the question

to ask more directly about whether the base rate should influence the requested judgment, rather than relying on people's intuitions about the heights of men and women.

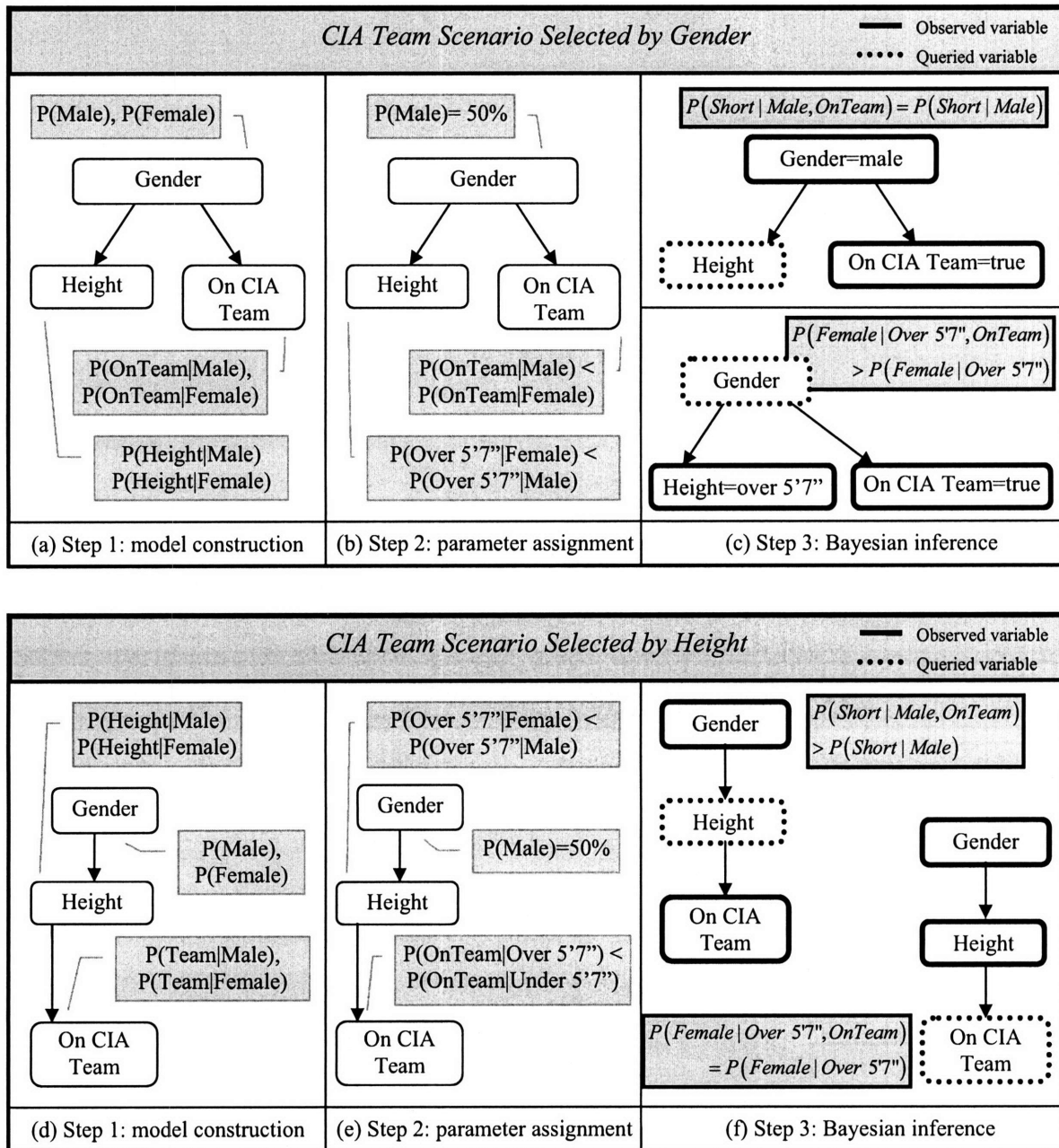


Figure 38: The three phases of causal Bayesian inference for the Experiment 6. (a-c) depict the CIA special operations team selected by gender, while (d-f) depict the CIA special operations team selected by height.

## 7.7 Method

*Participants.* The 152 participants in this experiment were MIT undergraduates and graduate students (majors were not recorded, but were likely randomly distributed). They were approached in a main corridor on campus, and were given token compensation.

*Materials:* Participants were assigned randomly to one of four conditions, as we crossed two factors: whether the team members were selected by gender or height (the causal structure), and whether the judgment was to infer gender from height or height from gender (the judgment). The cover story changed depending on the causal structure, while the question changed depending on the judgment. The materials follow, with the cover story shown separately from the question:

### *A. Cover Story*

#### *Causal structure 1: Selected by Gender*

The CIA's department of special operations recruited agents for a secret mission. Part of the mission required female agents, so agents were more likely to be selected if they were women.

Most of the final team members were women. Because most women are short, the team ended up being mostly under 5'7".

#### *Causal Structure 2: Selected by Height*

The CIA's department of special operations recruited agents for a secret mission. Part of the mission required short agents, so agents were more likely to be selected if they were short.

Most of the final team members were under 5'7". Because it was easier to find short women than short men, the team ended up being mostly women.

*B. Question*

*Judgment 1: Inferring Height from Gender*

There were several men on the team. Do you expect them to be shorter, taller, or no different than the average man?

Shorter

Taller

No different

*Judgment 2: Inferring Gender from Height*

There were several agents over 5'7" on the team. Compared to CIA agents over 5'7" on other teams, are these agents more likely to be female, more likely to be male, or about the same?

More likely female

More likely male

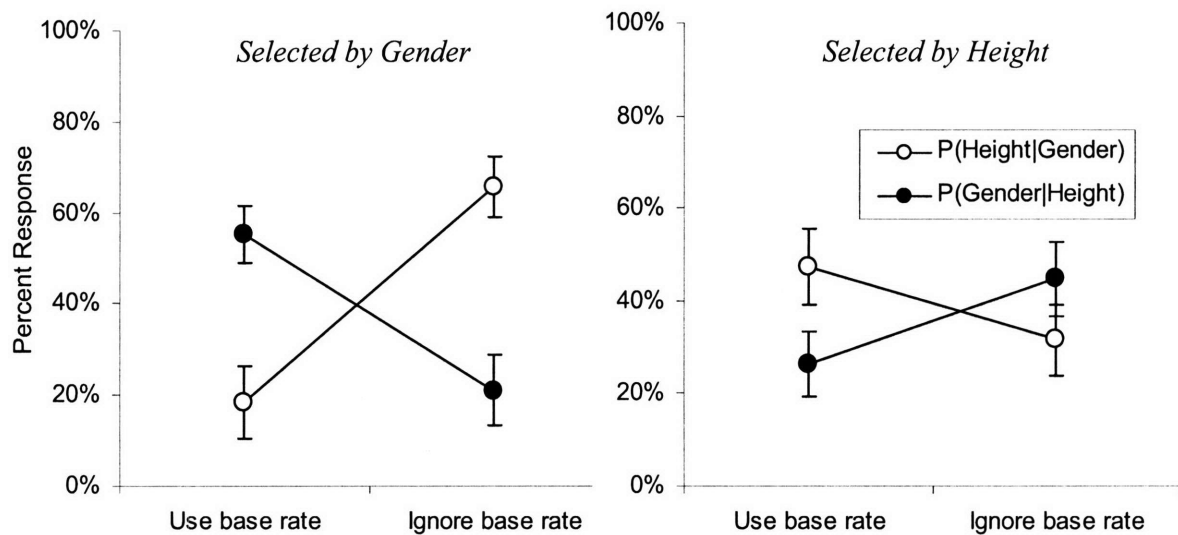
About the same

## 7.8 Results

The modal response in all conditions matched the predictions of our causal Bayesian hypothesis. For judgments of height from gender, responses of "no different" (indicating base rate ignorance) were significantly higher for the *selected by gender* causal structure than the *selected by height* causal structure ( $\chi^2(2)=9.7, p<.001$ ). This is consistent with the prescription of the causal Bayesian norm that when selected by gender, knowing the agent's *gender* renders *height* independent of being *on the team* (see Figure 38c). For judgments of gender from height, in contrast, responses of "about the same" (indicating base rate ignorance) were significantly higher for the *selected by height* causal structure than the *selected by gender* causal structure

( $\chi^2(2)=7.3, p<.05$ ). This is consistent with the prescription of the causal Bayesian norm that when selected by height, knowing the agent's *height* renders *gender* independent of being *on the team* (see Figure 38f).

We also analyzed the two factors separately, as well as their interaction. We found no main effect of causal structure or and no main effect of judgment, but a highly significant interaction between them ( $\chi^2(2)=16.44, p<0.0005$ ). Consistent with our predictions, base rate ignorance were significantly higher when the observed variable was situated between the variable to be judged and *on the team* in the causal model, whereas base rate use was significantly higher when the variable to be judged was situated between the observed variable *on the team* in the causal model (see Figure 39).



**Figure 39: Histogram of Experiment 6 results showing levels of base rate use compared to base rate neglect across gender vs. height scenarios. A two-way analysis of variance shows a significant interaction between the type of team (selected by gender vs. selected by height) and the judgment required (inferring gender from height vs. inferring height from gender) ( $\chi^2(2)=9.7, p<.001$ ).**



## 7.9 Discussion

The results of Experiment 6 confirmed the results of Experiment 5, while eliminating possible confounds in the stimulus materials. The demonstrated interaction between the requested judgment and the causal structure demonstrates that people often use or ignore the base rate exactly as prescribed by our proposed causal Bayesian norm. Ignoring the base-rate may often be a natural consequence of reasoning causally.

# 8 General Discussion

Our experimental results suggest that people's judgments under uncertainty may be best understood in terms of causal Bayesian inference: approximately rational statistical inferences over mentally represented causal models. Results from six experiments support three distinct claims: (1) people's judgments under uncertainty vary depending on the causal structure they believe to be underlying given statistics, and generally correspond well to the prescriptions of causal Bayesian inference; (2) when given a clear causal model people typically use base rates appropriately, which includes ignoring base rates when a causal Bayesian analysis suggests they should be ignored; and (3) people make approximately rational judgments which cannot be made using simple classical Bayesian norms (e.g., people rationally ignore the base rate of heights when judging height from gender in the poetry class scenario). In contrast to both the heuristics and biases view and the natural frequentist view, which cast people's ability to reason from probabilistic information as highly suspect, all six experiments found that participants' modal judgments are rational by the standards of causal Bayesian inference.

Our findings are not accounted for by earlier theories of judgment under uncertainty. The natural frequency hypothesis predicts uniformly poor performance on probability questions requiring the use of Bayes' rule, while the heuristics and biases view predicts that "causal" (or otherwise salient) statistics will dominate non-causal (or non-salient) base rates. The predictions of the heuristics and biases view may seem close to our view, but they run directly counter to our main finding: supposedly "non-causal" and "non-salient" base rates (such as the rate of breast cancer and the proportion of cabs that are blue) are more likely to be used correctly when the other statistics given are more "causal" (i.e., when statistics that fit poorly into a causal model, such as a false alarm rate or a perceptual error rate, are replaced with statistics that clearly map onto parameters of the causal model, such as the base rate of an alternative cause). Of the descriptive accounts reviewed, only our causal Bayesian hypothesis is consistent with the results presented in our experiments.

The results presented here support our causal Bayesian hypothesis, but there are other possible interpretations. Our manipulations in Experiments 1-4 often make the newly developed versions of judgment problems seem easier, more salient or more engaging, which may lead to better performance. Even if our newly provided statistics are naturally more salient, our results are inconsistent with the notion that mere salience of statistics drives attention which drives usage. For instance, people used the false-positive statistic in Experiment 1 as often as they used the benign cyst statistic, they just tended to misinterpret it as  $P(-cancer | +M)$  instead of  $P(+M | -cancer)$ . We would argue that increased intelligibility of our new scenarios comes as a result of the change to a more natural causal structure, and we have attempted to control for confounds that could have led incidentally to better performance on these versions. Furthermore, in Experiments 5 and 6 both versions were descriptively equivalent, hence arguably equally

salient, therefore these experiments confirm that causal structure influences judgments independently of salience. Another account of our results could hold that judgment is guided by causal reasoning heuristics, which might often approximate causal Bayesian inference but in some cases fall short of that normative framework's full capacity. We do not see the use of heuristics as inconsistent with our view. Rather, we treat causal Bayesian reasoning as a rational method of inference that, like any computation, can be approximated with heuristics. Thus, although future studies may provide evidence for a more heuristic level of judgment, we expect that these heuristics will be better characterized as approximations to causal Bayesian inference than to classical statistical methods.

## 8.1 Relation to the heuristics and biases view

The heuristics and biases view previously identified causality as playing an important role in base-rate neglect (Ajzen, 1977; Tversky & Kahneman, 1980), but did not address the crucial rational function of causal schemas in real-world judgment under uncertainty. In contrast to this view, which treats causal reasoning as a potentially distracting heuristic leading to judgment errors, we view causal inference one of the foundations of people's intuitive judgment ability, and we interpret effects of causality on judgment as core evidence for understanding how judgment works so well. We have attempted to explain how causal reasoning constitutes a rational system for making everyday judgments under uncertainty – in an adaptive sense, more rational than classical statistical norms – and we have shown how people's judgments under uncertainty, and deviations from normative statistical answers, may reflect sophisticated causal reasoning abilities. Rather than trying to identify the factors that induce or reduce errors such as base-rate neglect, as in Bar-Hillel (1980), we have explained how a rational inference engine can

yield seemingly irrational judgments when people assume a different causal structure from the experimenters or are presented with statistical data that do not correspond to model parameters.

## 8.2 Relation to the natural frequency hypothesis

The natural frequentist hypothesis does not address how people reason with explicit probabilities. By claiming that natural frequencies are the only statistical data that can be handled by the cognitive engine people have evolved for statistical inference, this view cannot account for results that show good performance on problems involving probabilities or relative frequencies (expressed as percentages). Since people have been shown to be adept at reasoning with probabilities and percentages under the right circumstances, in the experiments presented here and in other studies (e.g., Bar-Hillel, 1980; Peterson & Beach, 1967) the natural frequentist hypothesis does not seem to provide a general account of when and why judgments succeed or fail. We have also argued that the natural frequency hypothesis has significant limitations in its ability to account for real-world judgment under uncertainty, where there are often far too many possible factors and far too few prior observations to make valid judgments by natural frequencies alone.

In arguing against the natural frequentist hypothesis, we do not mean to imply that natural frequencies are not useful. On the contrary, they are an extremely important source of input that can be used for updating parameters of causal models or for computing proportions when only two variables are of interest. We also do not object to the claim that people are skilled at reasoning with natural frequencies; the data clearly show they are. But rather than conclude that evolution has only equipped us with a natural frequency engine, we would argue that people do better on these tasks because the natural frequency format makes the task simpler; it highlights nested sets that can be used in a calculation of proportions. Unlike most natural

frequency experiments, our experiments were carefully controlled such that both conditions required equally complex calculations. Thus we can be more confident that the differences in performance between the conditions are due to the causal content of the scenarios rather than their mathematical form.

### 8.3 Explaining other apparent errors of judgment under uncertainty

We anticipate that a number of other “fallacies” in the judgment literature may be artifacts of attempting to analyze people’s causal judgments as approximations to traditional statistical methods, and may be more productively explained in terms of causal Bayesian reasoning. We do not claim this view can account for all cases where intuitive judgments appear to depart from classical statistical norms, or even for all cases of base-rate neglect, but there are several important classes of judgments where it appears to offer some useful insights.

One such class of judgments are “causal asymmetries”, where people more readily infer effects from causes, as in judgments of  $P(E|C)$ , than causes from effects, as in judgments of  $P(C|E)$ . For example, Tversky and Kahneman (1980) report that people expressed more confidence in judging a man’s weight from his height than a man’s height from his weight. In the context of causal Bayesian reasoning, this asymmetry is understandable:  $P(E|C)$  is a parameter of a causal model, which should be available directly from causal domain knowledge, whereas  $P(C|E)$  requires a Bayesian computation that depends on knowledge of  $P(E|C)$ , and thus should be more difficult to judge. Intuitively, when judging  $P(C|E)$ , one must often consider many potential causes of  $E$  and integrate over the possible states of these causes. No such complexity is involved in judging  $P(E|C)$ , which can be obtained directly from the CPT of a causal model.

Another judgment phenomenon that our framework addresses is the difficulty people have with Simpson's paradox. An important version of this paradox is characterized by  $P(E|C, K) \geq P(E|\neg C, K)$  for all sub-populations ( $K$ ), but  $P(E|C) < P(E|\neg C)$  when the populations are combined into one; in each of the subpopulations,  $C$  seems to cause  $E$ , but overall,  $C$  seems to prevent  $E$ . Waldmann & Hagmayer (2001) showed that if people believe the co-factor ( $K$ ) to be a cause of  $E$ , they correctly condition on  $K$  when inferring the strength and direction of the contingency (in this case, people who believe  $K$  is a cause of  $E$  would conclude that  $C$  causes, rather than prevents,  $E$ , because for a given  $K$ ,  $C$  makes  $E$  more likely). However, the study did not address what factors account for the degree to which the situation seems paradoxical.

Our proposal suggests that Simpson's paradox seems paradoxical because people generally interpret statistical data as parameters of a causal model. If people believe a causal link exists between  $C$  and  $E$ , people will interpret the contingency statistic,  $P(E|C) < P(E|\neg C)$ , to be describing the parameter of that single causal link. For example, consider the following statistics: overall, those who use sunscreen more often are more likely to get skin cancer, but for both sunbathers and non-sunbathers, those who use sunscreen more often are less likely to get skin cancer. Since we know that a causal link exists between sunscreen and skin cancer, we naturally interpret the statistic as a parameter describing the causal power of that link. In this case, the sense of paradox is driven by the impression that the power of sunscreen to cause cancer is reversed (from preventive to generative) when the populations are combined. This paradox does not occur for the following statistics: overall, those who wear sunglasses more often are more likely to get skin cancer, but for both sunbathers and non-sunbathers, those who wear sunglasses more often are no more likely to get skin cancer. Because we know that

sunglasses cannot causally influence skin cancer, we do not interpret the contingency to be describing the causal link. There is no paradox because the power of sunglasses to cause cancer is not being changed when the population is combined; rather, it is clear that the apparent contingency results from a common cause (sun exposure). It is the prior knowledge of the preventive link between sunscreen and skin cancer that differentiates these two examples; when a causal or preventive link is known to exist between two variables, people naturally interpret the statistic to be describing the power of that causal link. Simpson's paradox becomes strongest when the causal link between the two variables is well known, but the common cause is not readily apparent. For example, people who eat vegetables regularly are more likely to get cancer, yet for every age group, people who eat vegetables regularly are less likely to get cancer. The paradox is dissolved by discovering that older people eat more vegetables, thus old age is a common cause of both cancer and eating vegetables regularly.

## 8.4 Learning Structure from Statistical Data

Causal Bayesian inference represents a method of combining statistical data with prior knowledge to make judgments. In the related area of learning causal structure, an active debate currently exists between views that emphasize the importance of statistical data versus prior knowledge. Glymour and Cheng's (1998) approach to causal structure induction seeks to formalize methods that enable both causal structure and parameters to be learned from statistical data alone. Waldmann (1996) argues that statistics alone are not enough to explain learning, and demonstrates that prior knowledge of causal directionality and causal relevance can affect learning causal structure from data. Taking this idea further, Ahn and Kalish (2000) argue that prior mechanism knowledge, including knowledge of intermediate causes in a chain, is crucial for learning causal structure, and especially for resolving ambiguous correlations. Tenenbaum

and Griffiths (2001, 2003) propose a model that synthesizes the roles of prior causal knowledge and statistical data, in which knowledge serves to constrain the space of possible causal structures, and that data can then be used to favor one structure over another. Our approach to judgment also calls for a synthesis between prior knowledge and statistics, but like Ahn and Kalish (2000), our experiments suggest that understanding how a causal mechanism works may be crucial to interpreting statistics that describe it.

## 8.5 Deterministic Mechanisms and Randomly Occurring Causes

Another way to interpret our results is in terms of a bias towards deterministic mechanisms. In the original mammogram and cab problems, the statistics given implied that the mechanisms were fundamentally stochastic, randomly generating positive mammograms 9.6% of the time with no cause, or randomly causing the witness to make a mistake 20% of the time. A number of studies have cast doubt on people's abilities to comprehend randomness, including such well-known phenomena as the gambler's fallacy, the hot-hand fallacy (Gilovich, Vallone, & Tversky, 1985), and the law of small numbers (Tversky & Kahneman, 1971). In experiments 1-3, as part of clarifying the causal structure of the scenario, we moved the main source of randomness from the efficacy of the mechanism to the presence or absence of other causal variables. It could be that people are good at reasoning about nondeterministic scenarios when the main source of nondeterminism is in the random occurrence of causal variables, but they find it less natural to reason about mechanisms randomly failing (unless those failures can be attributed to some external factor, at which point the source of randomness becomes the presence of a cause that deterministically disables the mechanism). The notion that causal reasoning may be accomplished by modeling deterministic mechanisms, with indeterminism introduced through uncertainty about the presence of hidden causal variables, has recently been proposed in both the



artificial intelligence (Pearl, 2000) and psychological literatures (Schulz, Sommerville, & Gopnik, in press; Luhmann & Ahn, in press).

## 8.6 Making Statistics Easy

People clearly have natural abilities to interpret statistics, yet they are often poor at interpreting published statistical data. Previous researchers have argued that we can leverage people's known natural abilities to teach them how to interpret published statistics. Sedlmeier and Gigerenzer (2001), for instance, present evidence that people can be taught to recast probabilities or relative frequencies, expressed as percentages, as natural frequencies, expressed as whole numbers, which improves correct response rates. However, if one does not hypothesize that people have a specialized cognitive engine for using natural frequencies, this effect could be seen as the result of mathematically simplifying a difficult problem. Just as one can break down a multi-digit multiplication problem into several single-digit multiplications added together, one can break down a probability question by considering nested subsets of a large number of individuals. The problem certainly becomes easier, but one need not hypothesize a special cognitive engine to explain it.

Our causal Bayesian hypothesis suggests a different approach to making statistics easy: replacing statistics that fit poorly into a causal model with statistics that correspond directly to causal model parameters. Furthermore, by embedding the statistics within a causal model, people may be able to understand the world better than they would with statistics alone, regardless of whether they are probabilities or natural frequencies. Consider the contrast in insight between the false-positive and benign cyst mammogram scenarios of Experiment 1 if one attempts to generalize beyond the very restricted circumstances given in the problem setup. Imagine a woman who gets a positive mammogram, but hears that it can be unreliable so she decides to get

a second mammogram. If that second mammogram also comes back positive, how much more confident should we be that she has cancer?

- The statistics in the false-positive problem suggest that she should be much more concerned after the second mammogram comes back positive: 6% of women without cancer will receive a positive mammogram, therefore 0.36% will test positive twice, assuming the second mammogram result is independent of the first. The chance of having cancer given two positive mammograms, then, is  $\frac{2\%}{2\% + 98\% \times 0.36\%} = 85\%$ .

- In contrast, the causal structure described by the benign cyst mammogram scenario of Experiment 1 suggests that the unreliability of the mammogram apparent in the false-positive scenario is an illusion. The mammogram reliably detects cancer, but it also reliably detects benign cysts, and if a woman has a benign cyst she will get a positive mammogram the second time as well. In this scenario, two positive mammograms is no more diagnostic of cancer than one positive mammogram. The answer remains  $\frac{2\%}{2\% + 98\% \times 6\%} = 25\%$ .

It may seem strange that supposedly equivalent statistics lead to greatly different inferences. This occurs because both sets of statistics are mere reflections of a much more complex underlying generative process. For instance, you may believe that any benign cyst has a 50% chance of being detected, or you may believe that only 50% of benign cysts are dense enough to be detected, but those that are dense enough will be detected every time. This second situation can be modeled causally by adding variables to represent the size and density of the cyst or tumor, and then specifying a threshold at which it is large enough to be detected deterministically. Of the views we have considered, we believe that only the causal Bayesian hypothesis can account

for how people extrapolate meaning by going beyond the statistics to represent a causal model, and using the model to make new inferences that are under-determined with statistics alone.

The most intuitively valuable statistics are those that correspond transparently to parameters of known causal relationships. However, for some situations, the true causal structure may not be obvious to people. In these cases, one should explain to people the true causal structure, as well as provide statistics that map onto that structure. For example, consider the statistic that patients are more likely to survive after being treated by doctor B than by doctor A (from Bar-Hillel, 1990). One could easily get the impression that doctor A provides inferior care, unless one is specifically informed that doctor A specializes in life-threatening diseases, and doctor B does not. This new information invokes a causal structure in which a person's disease state causally influences both the choice of doctor and the likelihood of survival. With this new structure, it is easy to see that the low survival rate of doctor A's patients may be due solely to a higher base rate of patients with life-threatening diseases, and hence the quality of care may be the same or even better than that of doctor B. But with the wrong causal structure in mind, people could easily and understandably jump to false and dangerous conclusions.

# Conclusion

The need to make intuitive statistical judgments is a pervasive fact of life in human society. But if we relied only on purely statistical information, we would be in dire straights, as the remarkable flexibility, success, and inductive potential of common sense would be impossible. Fortunately, our physical, biological and social environments are causally structured, and our intuitive theories of the world are often – but not always – sufficient to capture the most relevant structures for enabling appropriate causal Bayesian inferences. In experimental studies, if we present participants with a clear causal structure and statistics that clearly map onto that structure, we can nearly eliminate traditional judgment errors such as base-rate neglect and dramatically boost the incidence of correct Bayesian reasoning. Those who have a stake in improving statistical reasoning in complex, everyday settings – scientists, educators, doctors, advertisers, politicians, and many others – could do well to follow the same approach in communicating their questions, their data, and their conclusions to the lay public.

# References

- Ahn, W., & Kalish, C. (2000). The role of covariation vs. mechanism information in causal attribution. In R. Wilson, & F. Keil (Eds.) *Cognition and explanation*. Cambridge, MA: MIT Press.
- Ahn, W. (1999). Effect of Causal Structure on Category Construction. *Memory & Cognition*, 27, 1008-1023.
- Anderson, J. (1990). *The Adaptive Character of Thought*. Hillsdale, NJ : Erlbaum Associates.
- Ajzen, I. (1977). Intuitive Theories of Events and the Effects of Base-Rate Information on Prediction. *Journal of Personality and Social Psychology*, 35 (5), 303-314.
- Bar-Hillel, M. (1990). Back to Base Rates. In RM Hogarth (Ed), *Insights in Decision Making: A Tribute to Hillel J. Einhorn*, 200–216. Chicago: University of Chicago Press.
- Bar-Hillel, M. (1980) The base-rate fallacy in probability judgments. *Acta Psychologica*, 44, 211-233.
- Birnbaum, M.H. (1983). Base Rates in Bayesian Inference: Signal Detection Analysis of the Cab Problem. *American Journal of Psychology*, 96, 85-94.
- Cheng, PW (1997). From covariation to causation: A causal power theory. *Psychological Review*, 104(2), 367-405.
- Cosmides, L. and Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, 58, 1-73.
- Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds), *Judgment under uncertainty: Heuristics and biases*, 249-267. Cambridge: Cambridge University Press.

- Friedman, N., Linial, M., Nachman, I., and Pe'er D. (2000). Using Bayesian Networks to Analyze Expression Data. *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology (RECOMB 2000)*, 127-135.
- Gigerenzer, G. & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological review*, *102* (4), 684-704.
- Gilovich, T., Vallone R., & Tversky, A. (1985). The Hot Hand in Basketball: On the Misperception of Random Sequences. *Cognitive Psychology*, *17*, 295-314.
- Glymour, C. (2003). Learning, prediction and causal Bayes nets. *Trends in Cognitive Sciences*, *7*, 43-48.
- Glymour, C. (2001). *The Mind's Arrows: Bayes Nets and Graphical Causal Models in Psychology*. Cambridge, MA: The MIT Press.
- Glymour, C., & Cheng, P. (1998). Causal Mechanism and Probability: A Normative Approach. in M. Oaksford and N. Chater (Eds.), *Rational Models of Cognition*. Oxford: Oxford University Press.
- Gopnik, A., Glymour, C., Sobel D., Schulz L., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes-Nets. *Psychological Review*, *111* (1), 3-32.
- Gopnik, A., & Glymour, C. (2002). Causal maps and Bayes nets: A cognitive and computational account of theory-formation, in S. Stich, P. Carruthers (Eds.), *The Cognitive Basis of Science*, 117-132. New York: Cambridge University Press.
- Gopnik, A., & Sobel, D. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child development*, *71* (5), 1205-1222.

- Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology* 51(4), 285-386.
- Jordan, M. I. (Ed.) (1999). *Learning in Graphical Models*. Cambridge, MA: MIT Press.
- Kahneman, D. & Tversky, A. (1973). On the psychology of prediction. *Psychological Review* 80 (4), 237-251.
- Kahneman, D. & Tversky, A. (1972). On prediction and judgment. *Oregon Research Institute Bulletin* 12 (4).
- Luhmann, C. C., & Ahn, W. (in press). The meaning and computation of causal power: A critique of Cheng (1997) and Novick and Cheng (2004). *Psychological Review*.
- Lyon, D. & Slovic, P. (1976). Dominance of accuracy information and neglect of base rates in probability estimation. *Acta Psychologica*, 40, 287-298.
- McKenzie, C. R. M. (2003). Rational models as theories -- not standards -- of behavior. *Trends in Cognitive Sciences*, 7, 403-406.
- Nisbett, R. E. & Borgida, E. (1975). Attribution and the psychology of prediction. *Journal of Personality and Social Psychology*, 32, 932-943.
- Oaksford, M. & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, 101, 608-631.
- Oatley, G. & Ewart, B. (2003). Crimes analysis software: 'pins in maps', clustering and Bayes net prediction. *Expert Systems with Applications*, 25 (4), 569-588.
- Pearl, J. (2000). *Causality: models, reasoning, and inference*. New York: Cambridge University Press.
- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco: Morgan Kaufmann.

- Peterson, C. & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, 68 (1), 29-46.
- Rehder, B. (2003). Categorization as causal reasoning. *Cognitive Science*, 27, 709-748.
- Russell, S. J. & Norvig, P. (2003). *Artificial Intelligence: A Modern Approach, 2nd Edition*. Englewood Cliffs, NJ: Prentice-Hall.
- Schulz, L. E., Sommerville, J., & Gopnik, A. (in submission). God does not play dice: Causal determinism and children's inferences about unobserved causes. *Child Development*.
- Sedlmeier, P. & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General*, 130, 380-400.
- Shepard, R. N. (1987). Towards a universal law of generalization for psychological science. *Science*, 237, 1317-1323.
- Sloman, S.A., & Lagnado, D. (2005). Do we “do”? *Cognitive Science*, 29, 5-39.
- Slovic, P., Fischhoff, B., and Lichtenstein, S. (1976). Cognitive processes and societal risk taking. In J. S. Carol and J. W. Payne (Eds.), *Cognition and Social Behavior*. Hillsdale, NJ: Erlbaum.
- Spiegelhalter, D., Dawid, P., Lauritzen, S., Cowell, R. (1993). Bayesian analysis in expert systems. *Statistical Science*, 8, 219-282.
- Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction, and Search*. New York: Springer-Verlag.
- Steyvers, M., Tenenbaum, J., Wagenmakers, E.J., Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, 27, 453-489.

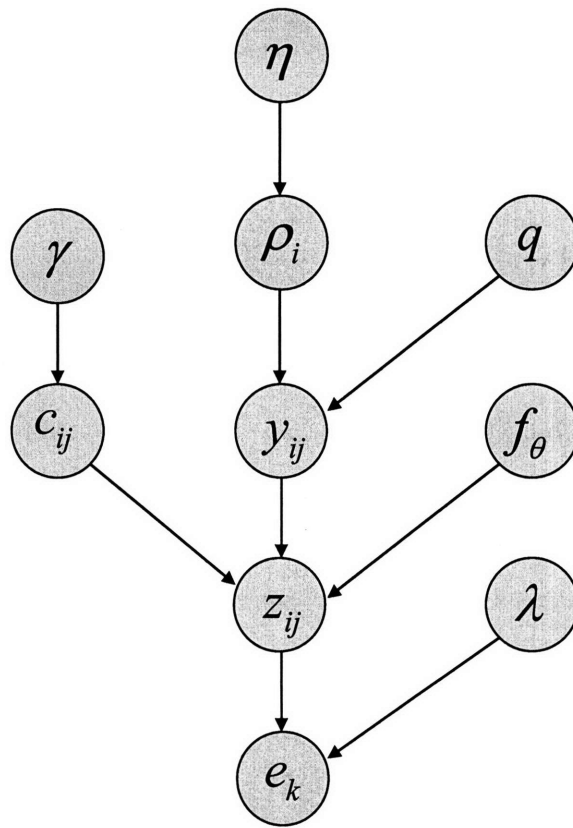


- Tenenbaum, J. B. & Griffiths, T. L. (2001) Structure learning in human causal induction. *Advances in Neural Information Processing Systems*. Leen, T., Dietterich, T., and Tresp, V., Cambridge, MIT Press, 2001, 59-65.
- Tenenbaum, J. B. & Griffiths, T. L. (2003). Theory-based causal inference. *Advances in Neural Information Processing Systems 15*. Becker, S., Thrun, S., & Obermayer. (Eds). Cambridge, MIT Press, 35-42.
- Tversky, A. & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, 1971, Vol. 76, No. 2. 105-110.
- Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124-1131.
- Tversky, A. & Kahneman, D. (1980). Causal schemas in judgments under uncertainty. *Progress in social psychology*, 49-72, M. Fishbein (Ed.). Erlbaum.
- Villejoubert, G. & Mandel, D. R. (2002). The inverse fallacy: An account of deviations from Bayes's theorem and the additivity principle. *Memory and Cognition* 30 (2), 171-178.
- Waldmann, M. R. (1996). Knowledge-based causal induction. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *The psychology of learning and motivation, Vol. 34: Causal learning*, 47-88. San Diego: Academic Press.
- Waldmann, M. R. (2001). Predictive versus diagnostic causal learning: Evidence from an overshadowing paradigm. *Psychonomic Bulletin & Review*, 8, 600-608.
- Waldmann, M. R., & Hagmayer, Y. (2001). Estimating causal strength: The role of structural knowledge and processing effort. *Cognition*, 82, 27-58.
- Wasserman, L. (2004). *All of Statistics : A Concise Course in Statistical Inference*. Springer.

# Appendix A: A hierarchical generative model for event-based causal learning

This appendix details the formal specification of the computational model. For those less interested in technical detail, Section 2 provides the important points for understanding the model's predictions for the experiments. The hierarchical generative model specifies a probability distribution over causal structure among objects, as well as event data generated by that structure. At the first level, causal powers are generated probabilistically among the cause objects. At the second level, cause and effect events are generated via Poisson processes and causal processes. Several simplifying assumptions are made in this model, making it applicable to our experiments, but potentially inapplicable to more complex stimuli. These assumptions can of course be relaxed, and future work could be done to test such a model's predictions.

In this simplified model, there is only one kind of causal power, and there is only one effect object that can be affected by this causal power. There is a specific set of cause objects, some of which have the causal power, but whether or not an object has the causal power is unobservable. Cause objects make contact with the effect object to form cause events at specific times. When a cause event happens, the causal powers in the cause object can produce effect events in the effect object after some temporal delay. Once the effect occurs, the cause events are no longer active. Thus, two cause events occurring in succession can only produce a maximum of one effect event. This generative model is depicted in Figure 40.



**Figure 40: a hierarchical generative model for causal events.  $c_{ij}$  and  $e_k$  are observable data, while the other nodes must be inferred.**

### 8.6.1 Specification of the generative model

Parameters for generating causal structure:

$\mu_N, \sigma_N$  : the mean and variance of the number of cause objects

$\eta$  : probability of each cause object having the causal power

$q$  : the causal strength (the probability that a cause will produce the effect on any given occurrence)

$f$  : the type of distribution over the delay (e.g., delta, uniform, gamma, normal)

$\theta$  : the parameters of the delay distribution (e.g., mean and variance)

Generated causal structure:

$N \sim \text{Normal}(\mu_N, \sigma_N)$ : the number of cause objects.

$\rho_i \sim \text{Bernoulli}(\eta)$ : whether object  $i$  has the causal power (true or false)

Parameters for generating events:

$\gamma$  : the rate at which a cause occurs when idle

$\lambda$  : the rate at which the effect occurs spontaneously

Generated events:

$c_{ij} \sim \text{Poisson process}(\gamma)$ : the time of the  $j^{\text{th}}$  occurrence of cause  $i$ .

$y_{ij} \sim \text{bernoulli}(q(\rho_i))$ : 1 if the  $j^{\text{th}}$  occurrence of cause  $i$  successfully produced the effect (or would have had the effect not been produced by another cause earlier), else 0 if the  $j^{\text{th}}$  occurrence of cause  $i$  failed to produce the effect.

$z_{ij} \sim y_{ij} f_{\theta}(x - c_{ij})$ : the time of the effect occurrence produced by the  $j^{\text{th}}$  occurrence of cause  $i$ , or 0 if it failed to produce the effect.

$\psi_k$  : the index of the cause that produced the effect, or 0 if the effect occurred spontaneously.

$e_k \sim \text{Poisson}(\lambda + \delta(z))$ : the time of the  $k^{\text{th}}$  occurrence of the effect.

### 8.6.2 Inference of causal powers

The goal of inference is to compute  $P(\rho | D)$ , the posterior probability of a particular assignment of causal powers to the cause objects, given the observable data.  $D$  represents all the data available that can inform our inference. Next, we formally specify the data:

Data ( $D$ ):

$N_{obs}$  : the number of observable cause objects

$C = \{c_{ij}\}$  : the set of all cause occurrences

$E = \{e_k\}$  : the set of all effect occurrences

The generative model as defined so far is not complete until we specify actual values for the parameters. But in most cases, we do not know what these values are. Nevertheless, we can still make inferences by specifying a distribution over the parameters of the generative model, representing our prior belief in what those parameters might be. By integrating over all possible values of the parameters, weighted by this prior, we can flexibly consider many potential causal domains, without making the model too rigidly adapted to any particular domain.

Priors over parameters:

$$\eta \sim \text{beta}(1,1)$$

$$\phi_\lambda \sim \text{exp}(1): \text{ the expected rate of the effect to occur spontaneously}$$

$$\lambda \sim \text{gamma}(2, \phi_\lambda)$$

$$\phi_\gamma \sim \text{exp}(1): \text{ the expected rate of the causes to occur spontaneously}$$

$$\gamma \sim \text{gamma}(2, \phi_\gamma)$$

$$q \sim \text{uniform}(0,1)$$

$$f = \text{gamma}(\alpha, \beta)$$

$$\theta : \alpha, \beta \sim \text{uniform}$$

Inference:

The probability of the causal power vector given the data can now be written as:

$$P(\rho | D) \propto P(\rho)P(D | \rho) = \eta^{\sum p_i} (1 - \eta)^{N - \sum p_i} P(C)P(E | C, \rho) \quad \text{Eq}$$

14

$P(C)$  is defined simply by the Poisson distribution, where  $n_i$  is the number of occurrences of cause  $i$  and  $T$  is the total amount of time observed.

$$P(C) = \prod_{i=1}^N \frac{(\lambda T)^{n_i} e^{-\lambda T}}{n_i!} \quad \text{Eq 15}$$

For simplification, we have assumed that once an effect occurs, all previous cause occurrences can no longer produce the effect. We define the likelihood of the effects using intermediate variables  $y$  and  $z$ .

$$P(y_{ij} | \rho_i, q) = \begin{cases} q\delta_{y_{ij},1} + (1-q)\delta_{y_{ij},0} & \text{if } \rho_i = 1 \\ \delta_{y_{ij},0} & \text{otherwise} \end{cases} \quad (\delta_{a,b} \text{ is 1 when } a=b, \text{ else 0})$$

$z_{ij}$  is the time at which the  $j^{\text{th}}$  occurrence of the  $i^{\text{th}}$  cause produces the effect. The probability density function for  $z_{ij}$  depends on  $y_{ij}$ . If the cause successfully produced the effect ( $y_{ij}=1$ ), then  $z_{ij}$  is determined by the distribution over temporal delay,  $f_\theta$ . Otherwise,  $z_{ij}$  is defined to be 0.  $z_{ij}$  cannot be less than  $c_{ij}$ , and cannot be greater than  $c_{i,j+1}$ .

$$p(z_{ij} | C, y_{ij}) = \begin{cases} \delta(z_{ij}) & \text{if } y_{ij} = 0 \\ f_\theta(z_{ij} - c_{ij}) & \text{if } y_{ij} = 1 \text{ and } c_{ij} < z_{ij} < c_{i,j+1} \\ 0 & \text{otherwise} \end{cases} \quad (\delta(a) = \infty \text{ when } a=0, \text{ else 0})$$

The likelihood of the  $k^{\text{th}}$  occurrence of the effect occurring at time  $e_k$  depends on whether there is a  $z_{ij}$  equal to  $e_k$ . If so, it is equal to the probability that no other cause, including the background, produced the effect at an earlier time. If not, it is equal to the probability that the background produced the effect at that time, and no cause produced the effect at an earlier time.

$$P(e_k | e_{k-1}, Z, \lambda) = \begin{cases} e^{-\lambda(e_{k+1}-e_{k-1})} \prod_{c_{xv} \in C: e_{k-1} < c_{xv} < e_k, x \neq i} P(z_{xv} = 0) + P(z_{xv} > e_k) & \text{if } \exists i, j : z_{ij} = e_k \wedge c_{ij} > e_{k-1} \\ \lambda e^{-\lambda(e_{k+1}-e_{k-1})} \prod_{c_{xv} \in C: e_{k-1} < c_{xv} < e_k} P(z_{xv} = 0) + P(z_{xv} > e_k) & \text{otherwise} \end{cases}$$

The likelihood of the full set of effect occurrences is equal to the product of the probabilities of each occurrence:

$$P(E | Z, \lambda) = \prod_{e_k \in E} P(e_k | e_{k-1}, Z, \lambda)$$

Finally, the probability of the effects given the causes is given by the generative model:

$$P(E | C, \rho, q, \lambda, f_\theta) = P(Y | \rho, q)P(Z | C, Y, f_\theta)P(E | Z, \lambda)$$

To fully compute  $P(E | C, \rho)P(E|C)$ , we must integrate over the parameters:

$$P(E | C, \rho) = \int \int \int P(E | C, \rho, q, \lambda, f_\theta).$$