CAUSES, COINCIDENCES, AND THEORIES

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF PSYCHOLOGY
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Thomas L. Griffiths
December 2004

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Joshua B. Tenenbaum
(Principal Co-Advisor)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Gordon H. Bower
(Principal Co-Advisor)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Ewart A. C. Thomas

Approved for the University Committee on Graduate Studies.

Preface

Sir Edmond Halley was an astronomer, sailor, and arguably the first statistician. In 1695, Halley was computing the orbits of a set of comets for inclusion in Newton's *Principia mathematica* when he noticed a surprising regularity: the comets of 1531, 1607, and 1682 took remarkably similar paths across the sky, and visited the Earth approximately 76 years apart. In the first edition of the *Principia*, Newton had argued that comets, like planets, follow orbits corresponding to conic sections – parabolas, hyperbolas, and ellipses – although he concluded that "as to the transverse diameters of their orbits, and the periodic times of their revolutions I leave them to be determined by comparing comets together which after long intervals of time return again on the same orbit" (Newton, 1687/1962, p. 532). Halley inferred that the comets of 1531, 1607, and 1682 were not three separate events, but three consequences of a single common cause: a comet that had visited the Earth three times, travelling in an elliptical orbit. He went on to state that "if according to what we have already said it should return again about the year 1758, candid posterity will not refuse to acknowledge that this was first discovered by an Englishman" (Halley, 1752, p. Ssss3).[1] The comet returned in December 1758, as predicted, and has continued to visit the Earth approximately every 76 years since, providing a sensational confirmation of Newton's physics.

The discovery of Halley's comet is a compelling example of *causal induction*: inferring causal structure from data. In modern scientific practice, causal relationships are identified through careful experimentation and statistical analysis. While the development of statistical methods for designing and analyzing experiments has greatly streamlined scientific

[1] Halley gave the more precise prediction of "about the end of the year 1758, or the beginning of the next" (1752, p. Rrrr2) earlier in the text. Making this prediction was by no means trivial, as it had to take into account the perturbation of the orbit of the comet by Saturn and Jupiter. Detailed accounts of Halley's discovery are provided by Cook (1998), Hughes (1990) and Yeomans (1991).

argument, science was possible before statistics: in many cases, the causal relationships between variables that are critical to understanding our world could be discovered without any need to perform explicit calculations. Science was possible because people have an intuitive capacity for causal induction, using data to assess the plausibility of causal relationships. This capacity is sufficiently accurate as to have resulted in genuine scientific discoveries, and allows us to construct the intuitive theories that express our knowledge about the world. The same notions about causality that allowed Halley to discover his comet guide our more mundane everyday inferences, from evaluating whether drinking coffee improves our productivity to working out which programs crash a computer.

Explaining Halley's discovery requires appealing to two factors: prior knowledge, and intuitive statistical inference. The prior knowledge that guided Halley was the physical theory laid out by Newton in the *Principia mathematica*. Newton's theory identified the entities and properties relevant to understanding a physical system, formalizing notions like velocity and acceleration, and characterized the relations that can hold among these entities. This physical theory allowed Halley to generate a set of hypotheses about the causal structure responsible for his astronomical observations: they could have been produced by three different comets, each travelling in a parabolic orbit, or by one comet, travelling in an elliptical orbit. Choosing between these hypotheses required the use of statistical inference. While Halley made no formal computations of the probabilities involved, the similarity in the paths of the comets and the fixed interval between observations convinced him that "it was highly probable, not to say demonstrative, that these were but one and the same Comet" (from the Journal Book of the Royal Society, July 1696, reproduced in Hughes, 1990, p. 353).

Scientific inferences like that of Halley can provide clues about how people solve problems of causal induction in the course of everyday life. The capacity to reason about the causes of events is an essential part of cognition from early childhood, whether it concerns the forces involved in physical systems (e.g., Shultz, 1982b), the essential properties of natural kinds (e.g., Gelman & Wellman, 1991), or the mental states of others (e.g., Perner, 1991). Often, these causal relationships need to be inferred from data. Explaining how people make these inferences is not just a matter of accounting for how causation is identified from correlation, but of accounting for how complex causal structure is inferred in the *absence* of (statistically significant) correlation. Halley's discovery illustrates that people can infer causal relationships from samples too small for any statistical test to produce significant

results – in this case, three observations – and solve problems like inferring hidden causal structure that still pose a major challenge for statisticians and computer scientists. In this thesis, I will argue that both of the factors involved in Halley's discovery – prior knowledge, in the form of a causal theory, and statistical inference – are necessary to make such inferences possible, and present a computational framework that indicates how they are combined.

In the computational framework that forms the heart of this thesis, *theory-based causal induction*, prior knowledge is formalized as a theory which generates a hypothesis space of causal models that could have produced data. These hypotheses are evaluated by Bayesian inference, resulting in a tight coupling between prior knowledge, in the form of a causal theory, and statistical learning. I will argue that there are three aspects of prior knowledge that are central to causal induction – the ontology of entities, properties, and relations that organizes a domain, the plausibility of specific causal relationships, and the functional form of those relationships – and that these three aspects are the key constituents of causal theories. This account identifies the limitations of existing algorithms for causal induction, makes it possible to explore the relationship between causes and coincidences, identifies which aspects of causal induction should be domain-sensitive, clarifies how causal mechanism knowledge is involved in human causal induction (c.f. Ahn & Kalish, 2000; Glymour & Cheng, 1998; White, 1995), and suggests how statistical learning might be extended from evaluating single causal relationships to evaluating entire causal theories.

The plan of the thesis is as follows. Chapter 1 introduces the key issues that I will address, discussing how prior knowledge can influence causal induction. The analysis of a scientific example indicates which aspects of prior knowledge are relevant to the assessment of new causal relationships, and I connect these aspects of prior knowledge with intuitive theories. The major challenge of the thesis is explaining how such theories can be integrated with statistical learning.

Chapter 2 takes up this challenge, summarizing the background behind the computational tools used throughout the thesis. The chapter introduces the fundamental ideas behind causal graphical models, which will be used as a basic representation of causal relationships, and highlights the ways in which generic algorithms for learning causal structure are inadequate for explaining human inferences.

Chapter 3 presents the theory-based causal induction framework. The chapter discusses how the aspects of intuitive theories relevant to causal induction can be formalized, and how

they can be incorporated with statistical inference. The notion of a theory as a hypothesis space generator is introduced, and used to explain how top-down and bottom-up information interact in causal induction.

Chapter 4 applies the theory-based approach to the problem of learning from contingency data. Causal graphical models are used to interpret two leading models of human judgments in this task, and to illustrate that they only evaluate the strength of a causal relationship, not asking the question of whether a causal relationship actually exists. A simple causal theory is used to develop an alternative model, causal support, that addresses this structural question. Causal support predicts several phenomena that cannot be explained by other models, and provides the best account of several existing datasets.

Chapter 5 examines how people learn about causal relationships in physical systems where events take place over a series of discrete trials. Here the problem is accounting for how people learn so much from so little, in some cases identifying underlying causal structure from just a single observation. The theory-based approach explains these inferences as the result of strong constraints imposed by a simple physical theory. The framework is used to account for people's inferences about two kinds of physical systems: detectors and machines.

The models discussed in Chapters 4 and 5 both concern events that occur on discrete trials. Chapter 6 extends the causal graphical model framework to cover events that occur in continuous time. This makes it possible to model systems that have complex dynamics, and to explain how people infer causal structure from the rates and times at which events occur. The remainder of the chapter uses this extended framework to analyze people's inferences about two dynamic systems: particle emissions and explosions.

Chapter 7 discusses the relationship between causes and coincidences. Psychologists and philosophers differ strongly in their treatment of coincidences, with the former focussing on the irrationality of human reasoning about chance, and the latter noting the close relationship between coincidences and scientific discovery. The chapter explores this tension, providing a formal definition of coincidences as events that provide support for a hypothesis that one ultimately decides is false. This illuminates the role of coincidences in scientific discovery, as well as how they can lead us astray, and provides some clues about the process of theory change.

Chapter 8 considers the implications of the theory-based approach to causal induction for the notion of an intuitive theory. The first part of the chapter examines the extent to which the aspects of intuitive theories relevant to causal induction should be domain-sensitive,

and explores the consequences of using cross-domain causes. This analysis motivates a discussion of the role of causal mechanism knowledge in causal induction. The chapter also considers the question of how theories might be acquired, synthesizing some of the results from the previous chapters.

Chapter 9 concludes the thesis.

Acknowledgements

Over the last few years, I have had the privilege to meet and work with some wonderful people. I would like to thank everybody who has played a role in the development of the ideas presented in this thesis:

I met Josh Tenenbaum back when there was only one hypothesis space, and theories were not even a glint in his eye. I think we have both benefitted from sharing an unusually productive collaborative relationship, going beyond the usual mode of interaction between student and advisor, and I am looking forward to continuing to share ideas with him as a colleague. In addition to the substantive outcomes of this collaboration, I have enjoyed his camaraderie and his ability to recognize the most interesting aspects of everything. By holding me to the same rigorous standards he sets for himself, he helped me discover both personal and intellectual capacities that I was not aware I possessed.

Gordon Bower and Ewart Thomas, the other members of my reading committee, have both enriched my time at Stanford, and tempered our youthful enthusiasm with their wisdom. Persi Diaconis and Lee Ross both made me think about this work in a new light, although that process is not yet complete. Persi was also responsible for introducing me to coincidences.

Mark Steyvers and Dave Blei have both been great collaborators, giving me the opportunity to work on a variety of projects that have nothing to do with this thesis, which was at times a great relief.

My lab- and class-mates have made graduate school both intellectually and socially stimulating. Thanks to the Computational Cognitive Scientists at MIT – Chris Baker, Liz Baraff, Charles Kemp, Konrad Koerding, Tevye Krynski, Amy Perfors, Lauren Schmidt, and Pat Shafto – and my friends at Stanford – Nick Davidenko, Phil Goff, Julie Heiser, Julie Turchin (née McGuire), Danny Oppenheimer, and Kelly Wilson. Liz and Danny deserve to be singled out, for helping to run what must be hundreds of subjects over the last few

years. I will continue to pay Danny in icecream, but Liz gets the less tangible reward of my sincere gratitude. Ronnie Bryan, Onny Chatterjee, Anne Chin, Carrie Niziolek, and Davie Yoon were also all part of the lab experience at Stanford and MIT, and contributed ideas and data.

Being an International Exchange Scholar involved certain challenges, and I thank Lorie Langdon, Denise Heintze, Rolando Villalobos, and Pat Cook for resolving those challenges with efficiency and a smile. While working on this thesis, I was supported by a Hackett Studentship from the University of Western Australia and a Burt and Deedee McMurtry Stanford Graduate Fellowship, for which I am extremely grateful.

Finally, I would like to thank my family, which has expanded along with my intellectual horizons. My parents, Rod and Judy Griffiths, have always supported me in achieving my goals, and we continue to discover parallels in our interests through lengthy international telephone conversations. My brother, Simon Griffiths, inspires me with his travels, maturity, and independent spirit. Tsing and Keith Bardin took the idea of a "host family" to heart, making me feel welcome in America time after time. Enrique and Viviana Lombrozo have also provided me with another place to consider home, and have warmly tolerated the fact that I am always doing something urgent whenever I visit. They have also generously shared their daughter, Tania, with me. My work has been enriched by her intellectual clarity, and my life has been enriched by her love.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The fundamental problem in explaining how people infer causal structure from data is understanding how we learn so much from so little. Halley inferred a common cause from three observations. In some of the experiments discussed in this thesis, people do likewise from only a single observation. The study of causal induction has a long history in philosophy (e.g., Hume, 1739/1978), statistics (e.g., Pearson, 1911), and psychology (e.g., Piaget, 1930).[1] However, this history sheds little light on people's remarkable capacity for causal induction. Hume (1748) emphasized the importance of large samples in inferring causal relationships, stating that "Even after one instance or experiment, where we have observed a particular event to follow upon another, we are not entitled to form a general rule, or foretell what will happen in like cases; it being justly esteemed an unpardonable temerity to judge the whole course of nature from one single experiment, however accurate or certain" (p. 50). Similarly, the statistical tests that scientists use to evaluate causal claims, and which are at the heart of many contemporary algorithms for identifying causal structure (e.g., Pearl, 2000; Spirtes et al., 1993), require large samples to produce results.

People learning a lot from a little is a familiar problem in cognitive science, and has a familiar solution. The strength of the conclusion reached in any inductive inference is a function of both data and prior knowledge. Strong conclusions from sparse data must thus be a result of strong prior knowledge. This principle appears most prominently in Chomsky's (1965) notorious "poverty of the stimulus" argument for the role of innate knowledge in language acqusition. However, it need not be connected to nativism, and has implications that go beyond the study of language. In fact, it suggests a general strategy for studying

[1] Reviews of some of this history are provided by Shultz (1982b), White (1990; 1995) and Pearl (2000).

the mind: in any setting where people learn a lot from a little, we can explore the prior knowledge that informs their inferences by examining the conclusions that they draw from different patterns of data. In the case of causal induction, examining what people infer from a few observations can begin to tell us about the principles by which human knowledge about causality is organized.

The claim that causal induction is guided by prior knowledge is not novel. Several cognitive scientists have proposed that human causal learning is best thought of as a knowledge-based, theory-based or top-down process (e.g., Waldmann, 1996; Lagnado & Sloman, 2004). However, these proposals have been strictly qualitative and informal. The goal of this thesis is to provide a *computational* account of human causal induction, in the sense introduced by Marr (1982). This involves answering three questions: "what is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out" (Marr, 1982, p. 25). Answering these questions requires being precise about the nature of the prior knowledge that informs causal induction, and explaining how it is integrated with statistical inference. Having developed such a computational account, we can begin to investigate the content of the knowledge that is necessary to explain human inferences about causal relationships in different settings. A first step in this process is identifying the ways in which prior knowledge influences human causal induction.

1.1 The role of knowledge in causal induction

The nature of the prior knowledge that allowed Halley to make his causal inference is very clear – it was Newton's mathematical theory of physics. Newton's theory identifies a set of observable and unobservable entities, specifies the plausible relationships among these entities, and defines the functional form of those relationships. In most cases of causal induction, the nature of the prior knowledge involved is less apparent. A central claim of this thesis is that this knowledge generally has the same structure as that used by Halley: a causal theory. In this section, I will present a scientific example more representative of everyday causal induction, illustrating how prior knowledge can influence our expectations about causal relationships.

Many empirical studies of causal induction (e.g., Buehner & Cheng, 1997; Buehner, Cheng, & Clifford, 2003; Lober & Shanks, 2000) use medical cover stories, such as evaluating the influence of chemicals on outcomes like gene expression. These studies typically examine

learning about a single causal relationship, such as whether injecting a chemical results in gene expression, and pay little attention to the effects of prior knowledge. I will use a slightly more complex medical scenario to examine the influence of prior knowledge on causal induction. This scenario is based upon a real scientific problem of causal induction, using the results of a study published in the journal *Toxicological Sciences* by Hamadeh, Bushel, Jayadev, Martin, DiSorbo, Sieber, Bennett, Tennant, Stoll, Barrett, Blanchard, Paules, and Afshari (2002). Hamadeh et al. (2002) used gene microarray analysis to assess causal relationships between chemicals and gene expression in mice. The methods they used to discover these relationships amount to little more than the assessment of covariation, but the resulting structure leads to strong expectations about what other causal relationships might be observed.

I will use this example to argue that human causal induction is influenced by three aspects of prior knowledge: information about the types of entities, properties, and relations that arise in a domain (which I will refer to as an *ontology*), constraints on the plausible relations among these entities, and constraints on the functional form of such relations. Each of these aspects of prior knowledge has previously been identified by psychologists as playing a role in causal induction, although no previous work has considered them simultaneously, or provided a detailed account of how they interact with statistical learning. I will highlight some of the ways in which these ideas have arisen elsewhere as I discuss each aspect of prior knowledge.

1.1.1 Ontology

Among several other results, Hamadeh et al. (2002) reported the set of causal relationships among six variables shown in Figure 1.1. Four of these variables refer to the injection of chemicals – clofibrate, Wyeth 14,643, gemfibrozil, and phenobarbital – and the other two indicate expression of a particular gene. Imagine that another variable, X, was involved in the study. Without knowing anything more about X, it would be hard to predict what causal relationships X might participate in. However, if you knew what variable X referred to – injection of a chemical or expression of a gene – you would probably have quite strong expectations about the causal relationships in which X would participate. In particular, discovering that X represents the injection of a chemical would probably lead you to believe that if X did participate in any causal relationships, it is likely that it would cause one of the genes to be expressed, or perhaps be affected by the expression of one of the genes.

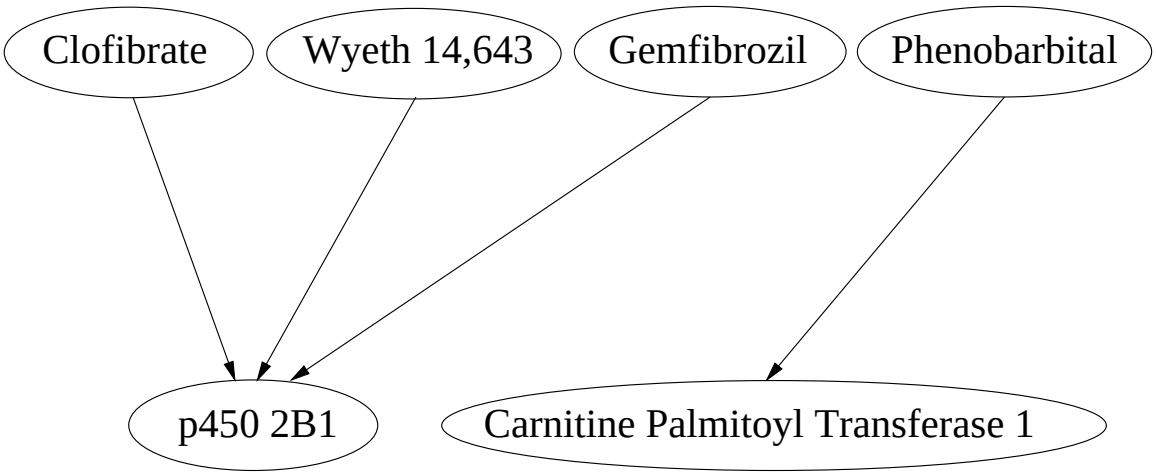

Figure 1.1: Causal structure relating four chemicals, shown at the top of the figure, to the expression of two genes, shown at the bottom of the figure, as reported by Hamadeh et al. (2002).

Chemicals and genes differ in the properties that apply to them, and the causal relationships in which they participate.

The question of how entities are differentiated based upon their causal properties has been thoroughly explored in developmental psychology, through consideration of the ontological commitments reflected in the behavior of infants and young children. Both infants and young children have strong expectations about the behavior of physical objects, and these expectations are quite different from those for intentional agents (e.g., Opfer & Gelman, 2001; Saxe, Tenenbaum, & Carey, in press; Shultz, 1982a; Spelke, Phillips, & Woodward, 1995). Similarly, children have different expectations about the properties of biological and non-biological entities (e.g., Springer & Keil, 1991). Gopnik, Sobel, Schulz, and Glymour (2001) have shown that children use the causal properties of entities to determine whether they belong to a novel type – objects that differed in appearance but both activated a "detector" were more likely to both be considered "blickets" than objects with similar appearance that differed in their causal properties (these studies will be discussed in more detail in Chapter 5). Research with adults has also examined how the types of entities influence causal inferences. For example, Lien and Cheng (2000) examined the circumstances under which causal properties are generalized across the members of a category.

1.1.2 Plausible relations

Knowledge of the types of entities can provide quite specific information about the *plausibility* of causal relationships. The expectation that chemicals influence genes, and not vice versa, is one illustration of type influencing plausibility, but carving up the types of entities at a finer scale can produce even stronger expectations. For example, the chemicals clofibrate, Wyeth 14,643, and gemfibrozil are all peroxisome proliferators, while phenobarbital is an enzyme inducer. If you were told that the chemical represented by X was a peroxisome proliferator, the relationship between the other peroxisome proliferators and gene expression might lead you to expect that X would influence expression of the gene p450 2B1, but not carnitine palmitoyl transferase I.

There is little controversy as to whether expectations about the plausibility of causal relations influence causal induction – even those who endorse covariation-based views recognize the role of "top-down knowledge" (e.g., Cheng, 1993; 1997). The key issue is how such knowledge is integrated with other sources of evidence (e.g., Alloy & Tabachnik, 1984). This issue is compounded when the goal is not just to learn a single causal relationship, but to simultaneously learn about multiple relationships. Waldmann (1996; 2000; Waldmann & Holyoak, 1992; Waldmann, Holyoak, & Fratianne, 1995) has shown that the way that people evaluate the strength of relationships among a set of variables is dramatically affected by their expectations about the underlying causal structure.

1.1.3 Functional form

Type information can also affect expectations about the *functional form* of causal relationships: whether they are positive or negative, and whether multiple causes interact or are independent. For example, discovering that the three other peroxisome proliferators all increase expression of p450 2B1 would probably lead you to believe that the chemical represented by X would also increase expression of this gene. Furthermore, finding that the other peroxisome proliferators each have an independent effect on the gene would suggest that X would combine with the variables representing other chemicals in a similar fashion.

A number of psychological studies have dealt with issues related to functional form, although often doing so obliquely. Many theories of animal learning assume that multiple causes of a single effect combine additively, each making a constant contribution to the

effect (e.g. Rescorla & Wagner, 1972). A number of researchers, including Shanks, Wasserman, and their colleagues, have advocated these linear models as accounts of human causal learning (e.g., Shanks, 1995a; 1995b; López, Cobos, Caño, & Shanks, 1998; Shanks & Dickinson, 1987; Wasserman, Elek, Chatlosh, & Baker, 1993). Waldmann (1996) described a study that shows that the validity of the linearity assumption depends upon people's beliefs about how the cause influences the effect. Cheng (1997) has argued that these linear models result in purely "covariational" measures, and people's causal inferences result from the innate assumption of a particular non-linear functional form, which is a probabilistic generalization of a logical OR gate (Glymour, 1998; Griffiths & Tenenbaum, in press; Tenenbaum & Griffiths, 2001). Kelley (1973) made a comprehensive attempt to spell out the functional forms of causal relationships, suggesting that causal induction from small numbers of observations may be guided by "causal schemas," such as sufficiency, necessity, and compensatory causation.

In addition to determining whether or not an effect occurs, the functional form of a causal relationship can also determine when that effect occurs. The time between the occurrence of a potential cause and the occurrence of an effect is a critical variable in many instances of causal induction. Several studies have explored covariation and temporal proximity as cues to causality in children, typically finding that the event that immediately precedes an effect is most likely to be perceived as the cause, even if there is covariational evidence to the contrary (e.g., Shultz, Fisher, Pratt, & Rulf, 1986). Hagmayer and Waldmann (2002) presented an elegant series of studies that showed that different assumptions about the delay between cause and effect could lead to different interpretation of the same set of events, determining which events were assumed to be related. Anderson (1990) provided a computational analysis of data involving the interaction between spatial separation and temporal contiguity in causal induction.

1.2 Causal induction is guided by causal theories

The toxicology example introduced above illustrates how causal induction can be influenced by three aspects of causal knowledge – ontological assumptions, constraints on plausibility, and constraints on functional form. These three aspects of prior knowledge can support strong expectations about possible causal relationships. Having an ontology, knowing the plausibility of relationships among the entities identified within that ontology, and knowing

the functional form of those relationships provides information that makes it possible to generalize about the causal relationships among new variables. For example, if X refers to a peroxisome proliferator, and it is known that clofibrate, another peroxisome proliferator, increases expression of a newly discovered gene, represented by Y, then we might expect that X would also increase Y. This kind of generalization is central to causal induction, making it possible for known causal relationships to influence our expectations about new relationships. However, it is not something that has been addressed by previous accounts of causal induction: inferring this relationship between X and Y does not require covariation between the two variables, or knowledge of the mechanism by which X influences Y.

I will argue that these three aspects of prior knowledge reflect the role of intuitive theories in guiding causal induction. Even in settings where we might not have the kind of explicit theory that was available to Halley, we often have an implicit theory that guides our inferences. In the case of the toxicology example, our theory concerns the interactions of genes and chemicals. Our limited understanding of this domain is sufficient to tell us that whether something is a chemical or a gene (or a peroxisome proliferator) is an important determinant of its causal properties. In particular, it affects our expectations about which causal relationships are plausible, and the form that those relationships might take.

Many cognitive scientists have suggested that human cognition and cognitive development can be understood by viewing knowledge as organized into intuitive theories, with a structure analogous to scientific theories (Carey, 1985a; Gopnik & Meltzoff, 1997; Karmiloff-Smith, 1988; Keil, 1989; Murphy & Medin, 1985). This approach has been used to explain people's intuitions in the biological (Atran, 1995; Inagaki & Hatano 2002; Medin & Atran, 1999), physical (McCloskey, 1983) and social (Nichols & Stich, 2003; Wellman, 1990) domains, and suggests some deep and interesting connections between issues in cognitive development and the philosophy of science (Carey, 1985a; Gopnik, 1996).

While there are no formal accounts of intuitive theories, there is a consensus on what kind of knowledge they incorporate: an ontology, indicating the types of entities that can be encountered in a given domain, and a set of causal laws expressing the relations that hold among these entities. For example, Carey (1985b) states that:

> A theory consists of three interrelated components: a set of phenomena that are in its domain, the causal laws and other explanatory mechanisms in terms of which the phenomena are accounted for, and the concepts in terms of which the phenomena and explanatory apparatus are expressed. (p. 394)

When discussing causal theories, it is often productive to distinguish among different levels at which a theory might operate. In a philosophical work that has inspired much of the treatment of theories in cognitive development, Laudan (1977) made such a distinction, separating everyday scientific theory from higher level "research traditions." He characterizes a research tradition as consisting of

> ... an ontology which specifies, in a general way, the types of fundamental entities which exist in the domain or domains within which the research tradition is embedded. ... Moreover, the research tradition outlines the different modes by which these entities can interact. (p. 79)

This distinction between these different levels of theory has been carried over into research on cognitive development, where Wellman (1990) and Wellman and Gelman (1992) distinguished between "specific" and "framework" theories:

> Specific theories are detailed scientific formulations about a delimited set of phenomena ... framework theories outline the ontology and the basic causal devices for their specific theories, thereby defining a coherent form of reasoning about a particular set of phenomena. (p. 341)

All of these definitions draw upon the same elements – ontologies and causal laws.

The three aspects of prior knowledge that influence causal induction map loosely onto the content of intuitive theories identified in these definitions. The division of the entities in a domain into a set of different types is the role of an ontology, and causal laws identify which relationships are plausible, and what form they take. It thus seems reasonable to assert that the form of the causal knowledge that guides causal induction is that of a causal theory. In particular, it is a theory that plays the role of a framework theory, providing a set of constraints that are used in discovering the causal structure of a system (the analogue of a specific theory). Despite the widespread use of intuitive theories in explaining cognition and cognitive development, there exist no formal accounts of the content of intuitive theories, or their role in guiding human cognition. In the remainder of the thesis, I will develop such an account for the case of causal induction.

Table 1.1: Key issues in causal induction

Causal graphical models
Can causal graphical models represent the content of human causal knowledge?
Can standard algorithms for learning causal structure explain human inferences?
Causes and coincidences
What makes a coincidence?
What is the role of coincidences in theory change?
Domain specificity
Which aspects of causal induction are domain-sensitive?
Theories and mechanisms
What role does mechanism knowledge play in causal induction?
Cognitive development and everyday learning
Do cognitive development and everyday learning act upon the same representations?

1.3 Key issues in causal induction

Developing a computational account of the role of intuitive theories in causal induction has the potential to shed light on a number of questions about how people assess causal relationships. These issues will arise throughout the thesis, and are summarized in the form of questions in Table 1.1.

1.3.1 Causal graphical models

Causal graphical models (also known as Bayesian networks) are a language for representing and reasoning about causal relationships that has been developed in computer science and statistics (Pearl, 2000; Spirtes, Glymour, & Schienes, 1993), and has begun to be used in psychology (e.g., Danks & McKenzie, under revision; Gopnik, Glymour, Sobel, Schulz, Kushnir, & Danks, 2004; Glymour, 1998; 2001; Griffiths & Tenenbaum, in press; Lagnado & Sloman, 2002; Rehder, 2003; Steyvers, Wagenmakers, Blum, & Tenenbaum, 2003; Tenenbaum & Griffiths, 2001, 2003; Waldmann & Martignon, 1998). I will provide a formal introduction to causal graphical models in the next chapter, but the basic idea is that the causal relationships among a set of variables can be represented in a graph in which variables are nodes and causation is indicated with arrows.

By providing a computational framework for addressing issues of causality, causal graphical models make it tempting to conclude that simple graphical representations are going to be sufficient to capture the causal knowledge that people have about the world. Indeed,

Gopnik and Glymour (2002; Gopnik et al., 2004) argue that the "causal maps" provided by such graphical representations can capture the knowledge contained in intuitive theories. Gopnik, Glymour, and their colleagues have also argued that the standard algorithms developed for learning causal graphical models might provide a domain-general account of human causal induction (e.g., Glymour, 2001; Gopnik & Glymour, 2002; Gopnik et al., 2004). These algorithms (e.g., Cooper & Herskovits, 1992; Pearl, 2000; Spirtes et al., 1993) are primarily data-driven, making little use of prior knowledge. This property of these algorithms has been criticized by statisticians (e.g., Humphreys & Freedman, 1996), and suggests that such algorithms may not be ideal for explaining human inferences. This leaves us with two questions: whether causal graphical models are the appropriate representation for human causal knowledge, and whether the algorithms for causal learning developed in computer science and statistics can shed light on the process by which people identify causal relationships.

1.3.2 Causes and coincidences

Coincidences receive quite different treatment from psychologists and philosophers. Psychologists typically use the conclusions that people draw from coincidences to support the argument that human beings reason poorly about chance (e.g., Gilovich, 1993; Plous, 1993). In contrast, philosophers of science have emphasized the connection between causes and coincidences, recognizing that coincidences are not merely improbable events, but events that provide support for a causal relationship (e.g., Horwich, 1982; Owens, 1992). Recognizing this connection begins to explain how it is that noticing suspicious coincidences often leads people to make significant scientific discoveries. Answering the question of what makes an event a coincidence has the potential to shed light on a deeper issue: since coincidences are often involved in scientific discoveries, what might they tell us about the processes by which intuitive theories change?

1.3.3 Domain specificity

One of the central questions in the investigation of causal induction in children is the extent to which such abilities are domain-specific. Different domains, such as physics, biology, and psychology, employ different ontologies and different kinds of causal relationships. The early manifestation of domain-specific causal inferences, such as knowledge of the causal properties of objects (e.g., Spelke, Breinlinger, Macomber, & Jacobson, 1992) has led to

claims that these inferences are the result of distinct and specialized cognitive modules (e.g., Leslie, 1994). It is these claims of domain-specificity that Gopnik and Glymour (2002) argue against when they appeal to the kind of domain-general learning mechanisms provided by causal graphical models. Precisely formulating the interaction between theory and evidence provides the opportunity to establish which aspects of causal induction should be domain-sensitive.

1.3.4 Theories and mechanisms

Historically, psychological theories about causal induction have fallen into two camps (Newsome, 2003): covariation-based approaches characterize human causal induction as the consequence of a domain-general statistical sensitivity to covariation between cause and effect (e.g., Cheng & Novick, 1990; 1992; Shanks & Dickinson, 1987), while mechanism-based approaches focus on the role of prior knowledge about the mechanisms by which causal force can be transferred (e.g., Ahn & Kalish, 2000; Shultz, 1982b; White, 1995).

Neither of these accounts is satisfactory. While sensitivity to covariation is an important aspect of causal learning, what counts as evidence for a causal relationship and how much evidence is required to conclude that such a relationship exists are both determined by prior knowledge. Appealing to causal mechanism recognizes the importance of this knowledge, but suffers from the vagueness of the notion of "mechanism." Various ideas have been appealed to by advocates of this view (e.g., Ahn & Kalish, 2000; Bullock, Gelman, & Baillargeon 1982; Shultz, 1982b; White, 1995) and different definitions are used in formal approaches to causality (e.g. Glymour & Cheng, 1998; Pearl, 1996). A complete understanding of the mechanisms mediating between cause and effect is clearly not necessary for causal induction – if one possessed such knowledge there would be nothing to learn, and recent results suggest that people have quite a limited understanding of the mechanisms involved in causal systems (e.g., Rozenblit & Keil, 2002). This leaves open the questions of which aspects of causal mechanism knowledge are involved in causal induction.

1.3.5 Cognitive development and everyday learning

Many of the most extensive studies of causal induction have been performed by developmental psychologists (e.g., Bullock et al., 1982; Gopnik et al., 2004; Shultz, 1982b). Explaining cognitive development poses a challenging problem: accounting for how children

acquire the detailed understanding of the causal structure of their environment exhibited by adults. Causal induction is at the heart of this problem, being the process by which children come to identify new causal relationships. However, developmental psychologists have also emphasized the importance of large-scale shifts in children's causal understanding over the course of cognitive development (e.g., Carey, 1985a; Gopnik & Meltzoff, 1997; Piaget, 1930). Instead of concerning just a single causal relationship, such changes are described as concerning entire causal theories.

Computational accounts of causal induction have implications for whether cognitive development and everyday causal induction act upon the same representations. Gopnik and Glymour (2002) propose that causal graphical models provide an appropriate representation for the causal knowledge of children, and that algorithms for learning causal graphical models can describe how this knowledge changes. This proposal suggests that there is a single representational substrate that is modified by both the long-term process of cognitive development and the short-term process of learning a new causal relationship (which is the more common form of causal induction as experienced by adults). However, this need not be the case: if causal graphical models are not the only representational format for human causal knowledge, then cognitive development and causal learning might operate at different levels of representation.

1.4 Summary

Understanding human causal induction requires explaining how people can infer causal relationships from small amounts of data. Such inferences are the result of strong constraints from prior knowledge. By formalizing how prior knowledge and statistical inference interact in causal induction, we can examine what kind of knowledge is needed to explain human inferences across different contexts. The kind of knowledge that influences causal induction is that expressed in intuitive theories: ontological assumptions about the types of entities and the properties and relations that apply to those entities, and causal laws that state the plausibility and form of causal relationships. Developing a computational account of the role of theories in causal induction has the potential to provide insight into a number of questions about how people assess causal relationships. I begin to explore these questions in the next chapter, introducing causal graphical models and using them to formulate the computational problem of causal induction that will be the focus of this thesis.

Chapter 2

Causal graphical models

Causality was excluded from the subject matter of statistics for much of the 20th century, largely as a consequence of the fervent arguments against causality made by Karl Pearson (e.g., Pearson, 1911). Techniques like structural equation modeling (Wright, 1921) were intended to make it possible to evaluate causal models, but their use for this purpose remains controversial (e.g., Freedman, 1991). Significant advances towards understanding the circumstances under which causal relationships could be identified were made in the 1970s and 1980s, resulting in several different formal treatments of causality (Rubin, 1974; Robins, 1986; 1987; see Rubin, 1990, and Holland, 1986, for reviews). Causal graphical models synthesize many of these advances into a single intuitive framework, providing a means of representing the causal relationships among a set of variables (Pearl, 2000; Spirtes et al., 1993). Recent work in computer science has focused on developing algorithms for learning causal structure from data, a project that has drawn some of the same criticisms as traditional structural equation modeling (e.g., Humphreys & Freedman, 1996).

Causal graphical models, also known as Bayesian networks or Bayes nets, have recently begun to be used in psychological research on causality (e.g., Danks & McKenzie, under revision; Gopnik, Glymour, Sobel, Schulz, Kushnir, & Danks, 2004; Glymour, 1998; 2001; Griffiths & Tenenbaum, in press; Lagnado & Sloman, 2002; Rehder, 2003; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003; Tenenbaum & Griffiths, 2001, 2003; Waldmann & Martignon, 1998). Several authors, most notably Gopnik, Glymour, and their colleagues, have argued that causal graphical models provide a means of representing the content of intuitive theories, and that the algorithms developed for learning causal graphical models might provide a domain-general account of how these theories are learned (e.g., Glymour,

2001; Gopnik & Glymour, 2002; Gopnik et al., 2004). I will argue against this view, claiming that representing intuitive theories requires going beyond the capacity of causal graphical models, and that, by exploiting the knowledge contained within these theories, human causal induction goes beyond generic algorithms for causal learning.

The plan of this chapter is as follows. First, I will briefly summarize how a causal graphical model is defined, and how such models can be used to make inferences about observations and the consequences of interventions. I will then use this formalism to state precisely the computational problem of causal induction, and argue that the standard algorithms that are used to solve this problem are inadequate as an account of human causal induction, failing to incorporate the kind of prior knowledge discussed in the previous chapter. This raises the question of how such knowledge might be incorporated into algorithms for learning causal structure, which provides some first clues about how intuitive theories might be formalized. A detailed consideration of some of the criticisms that have been leveled at causal graphical models and their relevance to the project at hand is provided in Appendix A.

2.1 Defining a causal graphical model

A causal graphical model has three components: a set of variables, a causal structure defined over those variables, and a set of assumptions about the functional form of the relationships indicated by this structure. I will describe these three components in turn.[1]

2.1.1 Variables

The first step in defining a causal graphical model is to identify the variables that will be used to encode events in the domain. The causal graphical model will define a probability distribution over these variables which can be used to answer questions about what patterns we might expect to observe, and the consequences of actions on this system. For instance, in a simplified version of the toxicology example discussed above, we might imagine that

[1]The various proposals for using graphical models to represent causal relationships differ slightly in their details. For example, Pearl (2000) assumes that the functions relating variables are deterministic, adding a stochastic component to the system through the presence of independent exogeneous noise variables. This makes it easier to address many of the questions one might have about a causal system, such as counterfactuals, at the cost of some notational complexity. Here, I will take the simpler route of defining causal graphical models as generic Bayesian networks (c.f. Pearl, 1988) supplemented with a calculus for interventions.

we are interested in whether a particular chemical causes expression of a particular gene in mice. For any given mouse, we know whether that mouse was treated with the chemical, and whether that mouse expressed the gene.

Many of the causal graphical models I consider in this thesis will be defined on binary random variables representing the truth of atomic sentences in a simple logical language.[2] Like predicate logic, this language will involve predicates, constants, variables, and quantifiers. For example, in our toxicology setting, we might have a single mouse `m`, a single chemical `c`, and a single gene `g`, allowing us to define atomic sentences such as $\texttt{Injected}(\texttt{c},\texttt{m})$, indicating that `m` was injected with chemical `c`, and $\texttt{Expressed}(\texttt{g},\texttt{m})$, indicating that `m` expresses gene `g`. We could then define binary random variables C and E that indicate the truth of these two predicates. For example, we might say that C assumes the value c^+ when $\texttt{Injected}(\texttt{c},\texttt{m})$ is true, and c^- when it is false. Treating C and E as random variables, we might ask questions like "Given that mouse `m` expresses the gene, what is the probability that the mouse was treated with the chemical?" which would be answered by computing $P(c^+|e^+)$.

In some cases, it will be more efficient to collapse many binary variables into a single variables that can take multiple values. For example, if a mouse `m` was injected with a chemical `c` at time `t`, then we could encode this information with a set of variables $\texttt{Injected}(\texttt{c},\texttt{m},\texttt{T})$ where `T` ranges over all possible times, and is only true at time `t`. Alternatively, we could represent this information with a variable $\texttt{InjectionTime}(\texttt{c},\texttt{m})$ which has values corresponding to possible times. If $\texttt{InjectionTime}(\texttt{c},\texttt{m}) = \texttt{t}$, then $\texttt{Injected}(\texttt{c},\texttt{m},\texttt{T})$ is true for $T = t$ and false otherwise. I will indicate whenever variables take on values other than simple Boolean truth or falsehood.

It is important to discriminate logical statements from their probabilistic counterparts. Throughout the thesis, I will use `typewriter` font to indicate logical statements, with predicates such as `Injected(c,m,T)` being capitalized, constants such as `m` being lower-case, and variables such as `T` being upper-case. Likewise, I will use *italic* font to indicate statements about random variables, with variables such as C being upper-case, and the values of those variables such as c^+ being lower-case.

[2]The principles behind graphical models can be applied to random variables of any kind. In focusing on logical statements, I am motivated by the use of graphical models for probabilistic logical reasoning in artificial intelligence (e.g., Russell & Norvig, 2002), rather than their more general use for representing structured probability distributions in machine learning and statistics (e.g., Jordan, 1998).

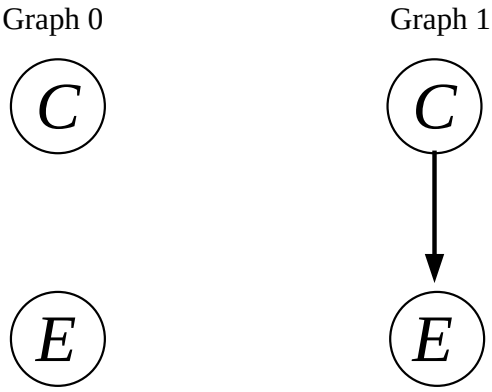

Figure 2.1: Directed graphs involving two variables, C and E. C is a potential cause, and E the effect of interest. In Graph 0, the two variables are independent, while Graph 1 depicts a causal relationship.

2.1.2 Causal structure

A causal graphical model uses a directed graph to represent the causal relationships among a set of variables. Nodes in the graph represent variables, and directed edges represent direct causal connections between those variables (Glymour, 1998; Glymour & Cooper, 1999; Pearl, 2000; Spirtes et al. 2001). The direction of the edges is illustrated using arrows, with "parent" nodes having arrows to their "children." The graph shown in Figure 1.1 thus represents the causal structure among a set of variables in the fashion used in causal graphical models. Two simpler examples are the directed graphs denoted Graph 0 and Graph 1 in Figure 2.1, which I will later use in Chapter 4 when explaining human inferences from contingency data. Both graphs are defined over two binary variables, C and E, representing a potential cause and its effect respectively. Each graph represents a hypothesis about the causal relations that could hold among these variables. In Graph 0, C and E are independent. In Graph 1, C causes E.

The causal structure used in a causal graphical model has implications for the dependencies that manifest in the probability distribution associated with that model. Any causal graphical model with variables $\{X_1, \ldots, X_n\}$ implies a probability distribution of the form $P(X_1, \ldots, X_n) = \prod_i P(X_i|\mathrm{Pa}(X_i))$, where $\mathrm{Pa}(X_i)$ is the set of parents of the node associated with X_i. This factorization of the probability distribution follows from the assumption that each variable X_i is independent of all of its non-descendants in the graph when conditioned upon its causes, $\mathrm{Pa}(X_i)$. This assumption is called the causal Markov condition, and

the relationship between statistical dependence and causation that it implies forms the basis of many algorithms for learning causal structure (e.g., Pearl, 2000; Spirtes et al., 1993).

This idea can be illustrated for Graph 0 and Graph 1, from Figure 2.1. In general, we can decompose any probability distribution into a product of terms for each variable, such as $P(E, C) = P(E|C)P(C)$. The causal structure represented by Graph 0 indicates that C and E are independent of one another. As a consequence, we can simplify this expression, to obtain $P(E, C) = P(E)P(C)$. Under this factorization, the joint probability of E and C is obtained by multiplying the probability of E by the probability of C. In contrast, Graph 1 indicates a dependency between the two variables, consistent with the factorization $P(E, C) = P(E|C)P(C)$. Under this factorization, the joint probability of E and C is obtained by multiplying the probability of C with a different probability distribution over E for each value of C.

2.1.3 Parameterization

The graphical structure of a causal graphical model identifies the causal relationships among variables, but says nothing about the precise nature of these relationships – they could be deterministic or probabilistic, and multiple causes of an effect could act independently or interact strongly. This information is captured by the *parameterization* of a causal graphical model. The parameterization specifies a conditional probability distribution for each variable, conditioned upon its parents in the graph. For a set of variables $X_1, \ldots, X_n$, these conditional probabilities are $P(X_i|\mathrm{Pa}(X_i))$. By the causal Markov condition, multiplying these conditional probabilities together provides the joint distribution over all X_i, which can be used to predict which values those variables are likely to take on.

In some cases, the parameterization of a model is trivial – for example, in Graph 0, we just need to specify $P(E)$ and $P(C)$. This can be done using a single numerical parameter w_0 for each node, providing the probability that that variable takes a positive value. For example, we could specify $P(E)$ by $P(e^+; w_0) = w_0$. However, when a node has parents, there are many different ways in which the relationship between causes and effects could be defined. For example, in Graph 1 we need to account for how the cause C influences the effect E.

The simplest means of parameterizing a variable – what I will call the *generic* parameterization – is to use a separate parameters to define the conditional probability distribution for a variable conditioned on each state of its parents. For example, in Graph 1, E has one

parent, C, with two states. We could thus specify the conditional probability $P(e^+|c)$ using two parameters, defining $P(e^+|c^-) = w_0$ and $P(e^+|c^+) = w_1$. Since these parameters can take on any value in $[0, 1]$, we can choose them in a way that allows us to capture any possible pattern of dependency between cause and effect.

The generic parameterization is extremely flexible, but the number of parameters required to specify the conditional probability of a variable increases exponentially with the number of its parents. Other parameterizations specify all of the conditional probability distributions using a simple function with a small number of parameters. The conditional probability distribution associated with a node can be any probabilistically sound function of its parents, including a deterministic function. Different assumptions about the kinds of mechanisms in a domain naturally lead to different parameterizations, so there will not be a single parameterization that can be used to characterize all settings in which causal learning takes place. Here, I will describe four simple parameterizations: noisy-OR, noisy-AND-NOT, logistic, and linear. I will show how these parameterizations can be applied to Graph 1.

The *noisy-OR* parameterization (Pearl, 1988) results from a natural set of assumptions about the relationship between cause and effect. For Graph 1, these assumptions are that C is a *generative* cause, increasing the probability of the effect; that the probability of E as a result of factors other than C is w_0; and that C influences E independently of these other factors. This gives

$$P(e^+|c; w_0, w_1) = 1 - (1 - w_0)(1 - w_1)^c. \tag{2.1}$$

where w_1 is a parameter associated with the strength of C and $c^+ = 1, c^- = 0$ for the purpose of arithmetic operations. This expression gives w_0 for the probability of E in the absence of C, and $w_0 + w_1 - w_0 w_1$ for the probability of E in the presence of C. This parameterization is called a noisy-OR because if w_0 and w_1 are both 1, Equation 2.1 reduces to the logical OR function: the effect occurs if and only if either some background factor or C is present. With w_0 and w_1 in the range $(0, 1)$ it generalizes this function to allow probabilistic causal relationships. If E had multiple parents $X_1, \ldots, X_n$, we could associate a separate strength w_i with each parent, and the noisy-OR parameterization would give

$$P(e^+|x_1, \ldots, x_n; w_0, w_1, \ldots, w_n) = 1 - (1 - w_0) \prod_i (1 - w_i)^{x_i} \tag{2.2}$$

where again $x_i = 0$ if X_i is absent, and 1 if X_i is present.

A parameterization for preventive causes can be derived from a set of assumptions similar to those made in the noisy-OR. In the case of Graph 1, these assumptions are that E occurs in the absence of C with probability w_0, and C independently prevents E from occurring with probability w_1. The resulting *noisy-AND-NOT* generalizes the logical statement that E will occur if the background factors are present and not C, allowing the influence of these factors to be probabilistic. The conditional probability can be written as

$$P(e^+|c; w_0, w_1) = w_0(1 - w_1)^c, \tag{2.3}$$

which gives w_0 for the probability of E in the absence of C and $w_0(1 - w_1)$ when C is present. As with the noisy-OR, both w_0 and w_1 are constrained to lie in the range $[0, 1]$, and the function can be generalized to accommodate the influence of multiple parents.

Finally, a *linear* parameterization of Graph 1 assumes that the probability of E occuring is a linear function of background factors and C. This corresponds to assuming that the presence of a cause simply increases the probability of an effect by a constant amount, regardless of any other causes that might be present. As with the logistic parameterization, there is no distinction between generative and preventive causes. The result is

$$P(e^+|c; w_0, w_1) = w_0 + w_1 \cdot c. \tag{2.4}$$

This parameterization requires that we constrain $w_0 + w_1$ to lie between 0 and 1 to ensure that Equation 2.4 results in a legal probability distribution. Because of this dependence between parameters, the linear parameterization is not normally used for causal graphical models, but I introduce it as it will prove useful in Chapter 4, when explaining the rational basis of some existing accounts of causal induction.

2.2 Observations and interventions

So far, I have described how causal graphical models can be used to make predictions about observed data: the causal structure implies a particular factorization of the probability distribution over a set of variables, and the parameterization specifies the conditional probabilities involved in this factorization. However, causal induction can make use of data other than those gathered by passive observation. Human learners often know the consequences

of actions upon a system, whether they be formal experiments or merely exploratory manipulation.

Predicting the values that variables will take on when a system is observed only requires knowing the correlations among those variables, and not the causal relationships responsible for those correlations. In contrast, predicting the consequences of actions upon a system requires knowing the causal relationships themselves. In the literature on causal graphical models, actions that fix the values of variables are referred to as *interventions*. Intervening on a variable renders all other causes of that variable redundant. Consequently, knowing that a variable took a particular value as the result of an intervention should provide no information about the variables that would normally act as its causes. Interventions are thus dealt with by performing "surgery" on a graph, removing the incoming edges from the variable that was manipulated (Pearl, 2000). The consequences of an intervention can be evaluated by performing probabilistic inference using the modified graph, treating it just like any other graphical model.

I will illustrate the difference between observations and interventions using Graph 1. Returning to our toxicology example, C indicates the truth of `Injected(c,m)` – whether mouse `m` was injected with chemical `c` – and E indicates the truth of `Expressed(g,m)` – whether mouse `m` expressed gene `g`. Assume that half of the mice in the experiment treated with the chemical, $P(c^+) = 0.5$, and that the gene is likely to be expressed if the mice are treated, $P(e^+|c^+) = 0.7$, but unlikely otherwise $P(e^+|c^-) = 0.1$. Upon *observing* that `m` expresses the gene (e^+), we might ask how likely it is that `m` was treated with the chemical (c^+). Answering this question involves computing $P(c^+|e^+)$, which can be done using the probability calculus:

$$\begin{aligned} P(c^+|e^+) &= \frac{P(e^+|c^+)P(c^+)}{P(e^+|c^+)P(c^+) + P(e^+|c^-)P(c^-)} \\ &= \frac{0.7 \times 0.5}{0.7 \times 0.5 + 0.1 \times 0.5} \\ &= 0.875. \end{aligned}$$

Observing that the gene is expressed thus increases the probability that `m` was treated with the chemical.

Now imagine that we have access to genetic engineering equipment that allows us to directly manipulate the gene in question. We can use this equipment to produce gene

expression (e^+) in m, and then ask how likely it is that m was treated with the chemical (c^+). To indicate that e^+ took its value as the result of an intervention, I will denote this probability $P(c^+|\text{do}(e^+))$, following the notation introduced by Pearl (2000). Under the procedure outlined above, this probability is evaluated by removing the edges from C to E, and reasoning with the resulting graph. In this case, C and E are rendered independent, so the value of E has no influence on the value of C. Consequently, $P(c^+|\text{do}(e^+)) = P(c^+) = 0.5$. If m expresses the gene as the consequence of an intervention, we gain no information about whether m was treated with the chemical.

Reasoning about interventions requires that the edges in a graphical model reflect the causal relationships among a set of variables, and not just correlations. Formally, this requirement arises because graph surgery treats the causes and effects of the manipulated variable differently. Causality is thus treated as a primitive, used in evaluating the consequences of intervention. Recent work in philosophy has pursued this idea from the opposite direction, starting with intervention as a primitive and using this as the basis for a definition of causality (Woodward, 2003). The ability to address interventions is the key innovation that extends the graphical models used to represent structured probability distributions in artificial intelligence and statistics (e.g., Pearl, 1988) into a framework for reasoning about causality (e.g., Pearl, 2000).

2.3 Quantification and plates

In my presentation of causal graphical models so far, I have assumed that the random variables corresponding to the nodes of the model indicate the truth of logical atomic sentences, such as `Injected(c,m)`. However, it will often be convenient to talk about causal relationships that hold over all instantiations of a particular logical variable. In the language of first-order logic, these are relationships that hold when we *quantify* over a variable (in this case, applying universal quantification). For example, imagine we had four mice, denoted $\texttt{m}_\texttt{i}$ for $i = 1, \ldots, 4$, and believed that a causal relationship held between `Injected(c,M)` and `Expressed(g,M)` for all `M` (i.e. $\texttt{M} \in \{\texttt{m}_1, \ldots, \texttt{m}_4\}$). Using the random variable C_i to indicate the truth of $\texttt{Injected(c,m}_i\texttt{)}$, and E_i to indicate the truth of $\texttt{Expressed(g,m}_i\texttt{)}$, the causal graphical model describing the relationships among these variables will consist of four copies of exactly the same causal structure, as shown in Figure 2.2 (a).

Causal relationships that hold under quantification introduce redundancies into the

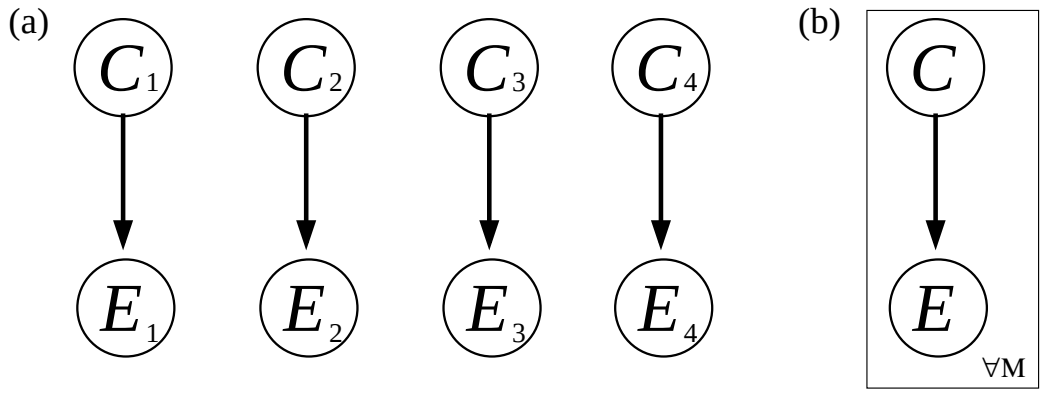

Figure 2.2: Plate notation for causal graphical models. (a) A causal relationship that holds over all instantiations of a logical variable produces causal graphical models with redundant structure. In this case, C_i indicates `Injected(c,`$\mathtt{m}_i$`)`, and E_i indicates `Expressed(g,`$\mathtt{m}_i$`)` for mice $\mathtt{m}_1, \ldots, \mathtt{m}_4$. (b) Quantification can be expressed efficiently using plates. Here C indicates `Injected(c,M)` and E indicates `Expressed(g,M)`, while the plate indicates that the relationship holds for all mice `M`.

causal structure among a set of variables and the parameters that are used to define a probability distribution over those variables. These redundancies can be exploited by developing an efficient notation for quantification. I will use a variant of *plates* (Buntine, 1994) for this purpose. Plate notation augments a graph with a set of boxes (called plates) surrounding subgraphs, indicating how many times that subgraph should be replicated. This formalism is typically used to capture redundancies in the structure and parameters of graphical models that result from generating independent samples. Independent sampling is an implicit form of quantification, where the dependencies exhibited in the graphical model are assumed to hold over all samples.

Figure 2.2 (b) shows how a single graph, together with a plate, can be used to represent a causal relationship that holds for all mice. The variables C and E indicate `Injected(c,M)` and `Expressed(g,M)` respectively. The plate shows that the relationship between these variables holds for all instantiations of the logical variable `M`, using the symbol for universal quantification, $\forall$. The causal graphical model shown in Figure 2.2 (a) can be obtained by substituting $\mathtt{m}_1, \ldots, \mathtt{m}_4$ for `M`. It should be emphasized that the structure depicted in Figure 2.2 (b) is primarily a notational innovation, and that the real underlying causal graphical model for representing the structure that holds among the variables and their probabilities is that shown in Figure 2.2 (a). However, using plates in this fashion will make it possible

for us to efficiently represent causal graphical models for large numbers of variables, and to understand when parameters are shared across variables. This becomes particularly important in Chapter 6, where the number of variables under consideration is effectively infinite.

2.4 The problem of causal induction

Causal graphical models provide us with the tools to take the first step towards a computational account of human causal induction, in the sense introduced by Marr (1982): they allow us to give a precise definition of the underlying computational problem. In this thesis, I will treat the problem of causal induction as that of identifying the causal graphical model responsible for generating a dataset $\mathcal{D}$. This problem has been extensively explored in the literature on causal graphical models in computer science and statistics, where it is known as the problem of *structure learning* (e.g., Heckerman, 1998).[3] Learning the causal structure that relates a large number of variables is a difficult computational problem, as the number of possible structures is a super-exponential function of the number of variables. Research in computer science and statistics has focussed on two strategies for solving this problem. *Constraint-based* methods attempt to identify causal structure on the basis of the patterns of dependency exhibited by a set of variables, while *Bayesian* methods evaluate the probability that a particular structure generated the observed data. I will discuss these strategies in turn, and consider their adequacy as accounts of human causal induction.

2.4.1 Constraint-based algorithms

Constraint-based algorithms for structure learning (e.g., Pearl, 2000; Spirtes et al., 1993) proceed in two steps. First, standard statistical tests like Pearson's χ^2 test are used to identify which variables are dependent and independent. Since the causal Markov condition implies that different causal structures should result in different patterns of dependency among variables, the observed dependencies provide constraints on the set of possible causal structures. The second step of the algorithms identifies this set, reasoning deductively from

[3]The other key problem in learning causal graphical models is *parameter estimation*. Structure learning involves identifying the topology of the causal graph, while parameter estimation involves determining the parameters of the functional relationships between causes and effects for a given causal structure. Parameter estimation will make an appearance in Chapter 4, when I consider alternative accounts of human inferences from contingency data.

the pattern of dependencies. The result is one or more causal structures that are consistent with the statistically significant dependencies exhibited by the data.

The use of statistical dependence between variables as the basis for structure learning means that constraint-based algorithms make very weak commitments as to the nature of any causal relationship. In particular, these algorithms make no assumptions about functional form. This may be appropriate in some of the scientific applications that inspired these methods, where the spirit of frequentist statistics favors minimal assumptions. By not making any commitments about the consequences of causal relationships other than statistical dependency, constraint-based algorithms provide a general-purpose tool for causal induction that can be applied easily across many domains. This generality is part of the appeal of these algorithms for developmental psychologists seeking to explain the acquisition of causal knowledge without recourse to domain-specific learning mechanisms (Gopnik & Glymour, 2002; Gopnik et al., 2004).

Throughout this thesis, I will argue that constraint-based algorithms cannot be used to explain human causal induction. This argument is based upon the data-driven, bottom-up approach to causal induction embodied in these algorithms, which results in two major problems. First, constraint-based algorithms do not account for the ways in which prior knowledge influences human inferences that I identified in the previous chapter. As these algorithms are defined, they use only a weak form of prior knowledge – the knowledge that particular causal relationships do or do not exist (e.g., Spirtes et al., 1993). They do not use prior knowledge concerning the underlying ontology, the plausibility of relationships, or their functional form. This insensitivity to prior knowledge has previously been pointed out by some critics of constraint-based algorithms (Humphreys & Freedman, 1996; Korb & Wallace, 1997). Prior knowledge provides essential constraints on human inferences, making it possible to infer causal relationships from very small samples. Without it, constraint-based algorithms require relatively large amounts of data in order to detect a causal relationship.

Second, constraint-based algorithms cannot combine weak sources of evidence, or maintain graded degrees of belief. This is a direct consequence of the policy of first conducting statistical tests, then reasoning deductively from the results. Statistical tests impose an arbitrary threshold on the evidence that data provide for a causal relationship. Using such a threshold is a violation of what Marr (1982) termed the "principle of least commitment":

> This principle requires not doing something that may later have to be undone, and I believe that it applies to all situations in which performance is fluent.

> It states that algorithms that are constructed according to a hypothesize-and-test strategy should be avoided because there is probably a better method. My experience has been that if the principle of least commitment has to be disobeyed, one is either doing something wrong or something very difficult. (p. 106)

Causal induction with a large number of variables (and without constraints from prior knowledge) *is* very difficult, and could be used to justify violation of the principle of least commitment. However, when making inferences about small numbers of variables, it is unnecessary and creates problems. In particular, thresholding evidence makes it hard to combine multiple weak sources of evidence for a causal relationship. The binarization of evidence is carried forward by deductively reasoning from the observed patterns of dependency. Such a process means that a particular causal structure can only be identified as consistent or inconsistent with the data, admitting no graded degrees of belief that might be updated through the acquistion of further evidence.

2.4.2 Bayesian structure learning

The Bayesian approach to structure learning (Cooper & Herskovits, 1992; see Heckerman, 1998) treats causal induction as a special case of the more general statistical problem of identifying the statistical model most likely to have generated an observed dataset. Bayesian inference provides a solution to this problem. The heart of this solution is Bayes' rule, which can be used to evaluate the probability that a hypothetical model h was responsible for generating data $\mathcal{D}$. The *posterior* distribution, $P(h|\mathcal{D})$, is evaluated by combining *prior* beliefs about the probability that h might generate any dataset with the probability of $\mathcal{D}$ under the model h, typically referred to as the *likelihood*. Bayes' rule stipulates how these probabilities should be combined, giving

$$P(h|\mathcal{D}) = \frac{P(\mathcal{D}|h)P(h)}{\sum_{h' \in \mathcal{H}} P(\mathcal{D}|h')P(h')} \tag{2.5}$$

where $\mathcal{H}$ is the *hypothesis space*, the set of all models that could possibly have produced $\mathcal{D}$.

As with any Bayesian inference, Bayesian structure learning requires specifying a prior probability and a likelihood for every hypothesis within a precisely delimited hypothesis space. In typical applications of this method, $\mathcal{H}$ consists of all directed graphs defined over the available variables. I will index these hypotheses as Graph i, since our primary concern

is with causal structure. The data $\mathcal{D}$ consist of the values that those variables assume as the result of observation and intervention. Standard Bayesian structure learning algorithms evaluate $P(\mathcal{D}|\text{Graph } i)$ by assuming the generic parameterization, defining a prior over the parameters, and then integrating over the specific values of those parameters (e.g., Cooper & Herskovits, 1992). This makes it possible to compute the probability of the data given a particular graphical structure without committing to a particular choice of parameter values. The prior over graph structures, $P(\text{Graph } i)$, is typically either uniform (giving equal probability to all graphs), or gives lower probability to more complex structures. Algorithms that use these principles differ in whether they then proceed by searching the space of structures to find that with the highest posterior probability (Friedman, 1997), or evaluate particular causal relationships by integrating over the posterior distribution over graphs (Friedman & Koller, 2000).

While Bayesian structure learning can deal with weak evidence and graded degrees of belief, the standard assumptions about priors, likelihoods, and hypothesis spaces mean that this approach is just as limited in its treatment of prior knowledge as the constraint-based algorithms described above. However, the Bayesian approach can be extended to incorporate the kinds of prior knowledge that influence causal induction. Different assumptions about the functional form of causal relationships can be captured by including models with different parameterizations in the hypothesis space, and the plausibility of causal relationships can be used in defining the prior probability of different graph structures. Recent work in computer science has begun to explore methods that use more complex ontologies, with each type of entities being characterized by a particular pattern of causal relationships with a particular functional form (e.g., Segal, Pe'er, Regev, Koller, & Friedman, 2003). This work is motivated by problems in bioinformatics that, as in many of the settings for human causal induction, require learning complex structures from limited data (e.g., Segal, Shapira, Regev, Pe'er, Botstein, Koller, & Friedman, 2003).

2.5 Beyond causal graphical models

Formulating the problem of causal induction as a Bayesian decision as to which causal graphical model generated a dataset provides a precise specification of how prior knowledge guides this inference. Knowledge about the ontology, plausibility, and functional form should define the prior, likelihood, and hypothesis space for Bayesian inference. However,

expressing this knowledge requires going beyond the representational capacities of causal graphical models. While this knowledge can be *instantiated* in a causal graphical model, it generalizes over a set of such models, and thus cannot be *expressed* in any one model.

A motivating intuition for this distinction, and for many of the ideas in the remainder of the thesis, can be obtained by considering an analogy between language and causality. Language comprehension and causal learning both center on the same fundamental problem of inductive inference: inferring the structure most likely to have generated some data. In language, the data take the form of a spoken or written sentence, and the structure to be recovered is a syntactic structure, such as a parse tree. In causal learning, the data concern the states of variables in a causal system, and the structure to be recovered is a causal graphical model. Both inductive problems are made particularly challenging by two salient features: the data severely underconstrain the underlying structure, and the set of structures that could have generated that data is effectively infinite.

To explain how syntactic structures can be inferred from sentences, linguists posit a generative grammar as a separate level of linguistic knowledge more abstract than any specific syntactic structure. The grammar generates a strongly constrained space of syntactic structures that could result in sentences in the particular language. The constraints supplied by the grammar are sufficient to allow the identification of a relatively small number of syntactic structures for any particular sentence. These syntactic structures are identified by parsing a sentence with respect to the grammar. Probabilistic grammars (see Charniak, 1993; Jurafsky & Martin, 2000; Manning & Shutze, 1999) augment the deterministic rules of traditional grammars with probabilities. This approach allows parsing to be formulated as a problem of Bayesian inference, with the grammar defining the hypothesis space of parse trees, together with prior probabilities and likelihoods for each of these structures.

Just as a grammar cannot be expressed in a single parse tree, the prior knowledge that constrains causal learning cannot be expressed in a single causal graphical model. This is because of an inherent limitation in the expressive capacity of graphical models. Causal graphical models are formally equivalent to a probabilistic form of propositional logic (e.g., Russell & Norvig, 2002). A causal graphical model can be used to encode any probabilistic logical rule that refers to the properties of specific entities in the domain. An example of such a rule might be that presence of chemical `c` causes expression of gene `g` in mouse `m` with probability 0.8. The addition of plates, as discussed in Section 2.3, introduces a limited form of quantification and makes it possible to define a causal relationship that holds over all

entities. For example, we could use causal graphical models with plates to represent about the relationship between `c` and `g` for all mice. However, causal graphical models cannot encode rules that state *conditions* on causality that generalize over entities, such as the rule that a causal relationship between any chemical and any gene exists with probability 0.8.

More generally, causal graphical models cannot capture the fact that there are different types of entities, or the way that the types of entities involved in a potential relationship influence our expectations about the plausibility and functional form of that relationship. Such notions require going beyond causal graphical models, and considering richer probabilistic logics. The development of probabilistic predicate logic remains an open problem in artificial intelligence research (Friedman, Getoor, Koller, & Pfeffer, 1999; Kersting & DeRaedt, 2000; Koller & Pfeffer, 1997; Milch, Marthi, & Russell, 2004; Muggleton, 1997). In the next chapter, I consider how we can use some of the ideas behind this research to develop a different level of representation for causal knowledge: a set of principles that can be used to guide inferences about the causal structure that was most likely to have generated a dataset. This level of representation is that of intuitive theories.

Chapter 3

Theory-based causal induction

I have argued that the problem of causal induction can be formulated as a Bayesian decision in which the hypotheses are causal graphical models. This requires specifying a hypothesis space, and a prior and likelihood for every hypothesis in that space. In this chapter, I will argue that the problems of specifying these components of Bayesian inference and of explaining how intuitive theories guide causal induction can both be solved by thinking of theories as *hypothesis space generators*. I will explain what this means, introduce a simple formalism for defining theories that can play this role, and demonstrate this formalism by developing a simple theory applicable to the toxicology example introduced in Chapter 1.

3.1 Theories as hypothesis space generators

The problem of specifying a hypothesis space and a prior and likelihood for each hypothesis in that space can be solved by defining a probabilistic procedure for generating causal graphical models. Such a procedure needs to specify probability distributions from which the variables, structure, and parameterization of causal graphical models are drawn. The hypothesis space is the set of causal graphical models that can be generated by sampling from these distributions, the prior is the probability with which a given model is generated by this process, and the likelihood is determined by the parameterization of that model. By limiting which causal structures and parameterizations can be generated, it is possible to impose strong constraints on the hypotheses considered when reasoning about a causal system.

The central claim of this thesis is that intuitive theories play exactly this role, generating

hypothesis spaces for causal induction. The commitments and consequences of this claim can be understood by extending the analogy between language comprehension and causal induction introduced in the previous section, where I suggested that the prior knowledge that guides causal induction is not just specified at the same abstract level of representation with respect to a causal graphical models as a grammar is to a parse tree. Extending the analogy, a theory plays the same role in solving the problem of causal induction that a grammar plays in language comprehension: like a grammar, a theory *generates* the hypotheses used in induction.[1] A schematic illustration of the correspondence between these two problems is shown in Figure 3.1. Under this view, the solution to the inductive problem of causal learning has the same character as identifying the syntactic structure of sentences: just as grammars generate a space of possible phrase structures, theories generate a space of possible causal graphical models. Causal learning is thus a problem of "parsing" the states of the variables in a system with respect to a causal theory. If the theory provides strong enough constraints, such parsing can be done swiftly and easily, picking out the causal structure that is most likely to have generated the data.

3.1.1 Formalizing intuitive theories

The first step in developing this account is to identify the basic elements of intuitive theories – the equivalents of terminals, non-terminals, and rewrite rules for a context-free grammar – and explain how these are used to generate causal graphical models. When cognitive scientists appeal to an intuitive theory to explain the inferences that people make in a given domain, they typically mean a structured representation with causal content, similar in spirit to a scientific theory (e.g., Carey, 1985a). As discussed above, accounts in philosophy of science and cognitive development are more precise about the structure and content of such theories, seeing them as constructed from an ontology and causal laws (Carey, 1985b; Gopnik & Meltzoff, 1997; Wellman, 1990; Wellman & Gelman, 1992). Providing a formal treatment of causal theories that captures their richness and complexity, as well as

[1] This equation of theories and grammars is foregrounded in Chomsky's early writing on language:

> The grammar of a language can be viewed as a theory of the structure of this language. Any scientific theory is based on a certain finite set of observations and, by establishing general laws stated in terms of certain hypothetical constructs, it attempts to account for these observations... Similarly, a grammar is based on a finite number of observed sentences... and it "projects" this set to an infinite set of grammatical sentences by establishing general "laws" ... [stated in terms of] phonemes, words, phrases, and so on. (Chomsky, 1956, p. 113)

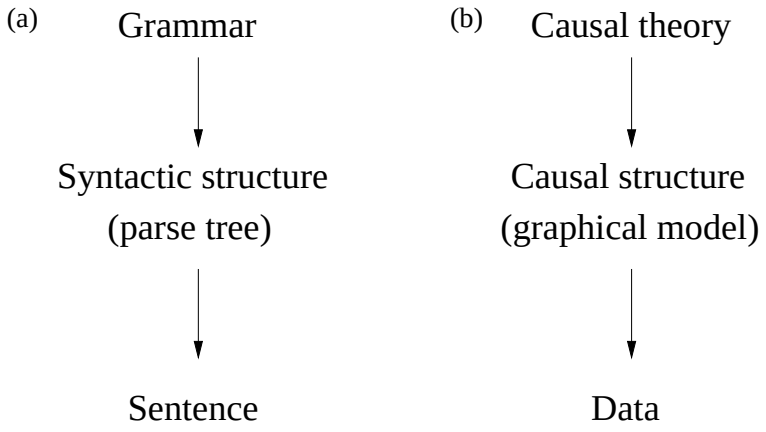

Figure 3.1: Three levels of representation in (a) language comprehension and (b) causal induction. Each level generates the level below, and language comprehension and causal induction both involve inferring the middle level based upon data below and constraints from above.

the breadth of inferences that they are supposed to support, is a task that goes beyond the scope of this thesis. I will formalize just the aspects of intuitive theories relevant to generating hypothesis spaces for causal induction.

The intuitive theories that I present in this thesis will have three components, corresponding to the three aspect of prior knowledge that influence causal induction identified in Chapter 1 and the three elements of the definition of a causal graphical model introduced in Chapter 2. These three components are an ontology, a set of principles that identify plausible relations, and a statement of the functional form of those relations. These three components of a theory each generate one part of a causal graphical model, being the variables, the causal structure, and the parameterization respectively.

The first component of a theory, the ontology, identifies the types of entities that exist in a domain,[2] the number of entities of each type (or a distribution over this number), and the predicates that can be used to describe these entities. In the toxicology example, there are three types of entities: `Chemical`, `Gene`, and `Mouse`. Any entity in the domain must belong to one of these three types. The number of entities of each type can either be stipulated, or treated as a random variable drawn from a specific distribution. For example, I might state that the number of Chemicals, N_C, the number of Genes, N_G, and

[2]The term "type" is used here in the technical sense assoiated with a typed or many-sorted logic (see Enderton, 1972). Types restrict quantifiers and the application of predicates, with each predicate only being applicable to entities of particular types.

the number of Mice, N_M are drawn independently from distributions P_C, P_G, and P_M respectively, but leave these distributions undefined – in many applications, N_C, N_G and N_M will be apparent, and we need not be concerned about generating them. The predicates defined on these types state which properties and relations can take arguments of particular types, and what values those predicates can take on. In the toxicology case, these would include `Injected`(`Chemical`, `Mouse`), indicating that a particular chemical was injected into a particular mouse, and `Expressed`(`Gene`, `Mouse`), indicating that a particular gene was expressed by a particular mouse. Both of these predicates are Boolean, being either true or false. This ontology is summarized in Figure 3.2.

The ontology required for this example is relatively simple, but the kind of knowledge that people have in other situations may be much more complex. For example, an ontology could be hierarchical, with objects belonging to types at multiple levels and predicates applying based upon the type at each of those levels. A theory that captured the causal structure of our original toxicology example would have such a hierarchical ontology, breaking entities first into `Chemical` and `Gene`, and then further dividing `Chemical` into `PeroxisomeProliferator` and `EnzymeInducer` and `Gene` into the subtypes influenced by these kinds of chemical. The number of entities that are of these subtypes could be determined by probabilistically allocating the known set of chemicals and genes into the two types. This procedure can express the fact that it might be more likely that a chemical is a `PeroxisomeProliferator` than an `EnzymeInducer`.

The second component of an intuitive theory is a set of rules that determine which causal relationships are plausible. These rules can be based upon the types of the entities involved, or the predicates that apply to them. In the cases I consider, the rules will be based purely on types.[3] In the toxicology example, we know that the structure of the problem is such that injections do not cause other injections, and nor does gene expression. It might be possible that the expression of one gene influences the expression of another, but I will assume that this is not the case for the moment. The only relationships with which we will concern ourselves are those between chemicals and genes. Figure 3.2 states a rule by

[3]Defining the rules based purely on type results in simpler theories. More generally, we could allow predicates to play a role in determining whether causal relationships are plausible. In fact, this is done implicitly even when only type is used, since a typed logic can be reduced to standard first-order logic by introducing predicates that indicate type (see Enderton, 1972). Pursuing this strategy requires distinguishing between predicates that participate in causal relationships and predicates that are just used to determine the plausibility of those relationships. The former are used to generate the variables of the causal graphical models, while the latter define the prior probability of each model (see Tenenbaum & Griffiths, in prep, for an example of this).

Ontology:

Types	Number
`Chemical`	$N_C \sim P_C$
`Gene`	$N_G \sim P_G$
`Mouse`	$N_M \sim P_M$

Predicates	Values	
`Injected(Chemical,Mouse)`	`Boolean:`	{`T`,`F`}
`Expressed(Gene,Mouse)`	`Boolean:`	{`T`,`F`}

Plausible relations:
`Injected(C,M)` $\rightarrow$ `Expressed(G,M)`
True for all `M` with probability p for each `C`, `G` pair

Functional form:

`Injected(C,M)` $\sim$ Bernoulli($\cdot$)

`Expressed(G,M)` $\sim$ Bernoulli(ν) for ν from a noisy-OR:

Cause	Strength
(Background)	$w_0 \sim \text{Uniform}(0,1)$
`Injected(C,M)`	$w_i \sim \text{Uniform}(0,1)$

Figure 3.2: Theory for causal induction from contingency data in a medical setting.

which the plausibility of such relationships might be expressed, assigning a probability p to the existence of a causal relationship between a particular chemical and a particular gene, regardless of the mouse involved. All other causal relationships have probability 0.

The final component of an intuitive theory is a statement of the functional form that causal relationships are expected to possess. This requires specifying a parameterization (or distribution over parameterizations) for each predicate identified in the ontology. For the toxicology example, we need to define the probability that a particular mouse receives an injection of a particular chemical. This probability will not influence any of our subsequent analyses, and thus is not specified: the theory indicate that this is a Bernoulli event, being true with some probability, but does not give the probability. In contrast, `Expressed(G,M)` is identified as a Bernoulli event with parameter ν, where ν is computed using the noisy-OR parameterization (Equation 2.2), allowing each cause – in this case `Injected(C,M)` for some `C` – has an independent opportunity to influence the effect with probability w_i. The parameters w_i are all assumed to be drawn from a uniform distribution, reflecting a lack of expectations about the strengths of the causes.

3.1.2 Generating a hypothesis space

The process by which a causal graphical model is generated from a theory is as follows:

1. *Generate variables.* Sample the number of entities of each type from the distribution specified in the **Ontology**. Generate the complete set of *grounded predicates* for these entities. This is the set of variables that form the nodes of the graph.
2. *Generate structure.* Sample links between nodes using the probabilistic procedure stated in the **Plausible relations** component of the theory.
3. *Generate parameterization.* For each node, sample a parameterization as specified in the **Functional form** component of the theory.

This generative process defines a hypothesis space, together with a prior probability and, by specifying the parameterization, a likelihood for each model in that space.

I will illustrate how this generative process works by using the theory given in Equation 3.2. I will assume that the number of chemicals, genes, and mice involved in a particular experiment is known, and implicitly condition on this information. For example, we might have a single chemical `c`, a single gene `g`, and N_M mice $\mathtt{m}_1, \ldots, \mathtt{m}_{N_M}$. The set of grounded predicates is constructed by substituting all possible entities for the arguments of each predicate in the ontology. In our case, this set consists of N_M statements indicating whether `Injected(c,`$\mathtt{m}_i$`)` holds of mouse $\mathtt{m}_i$, and N_M statements indicating whether `Expressed(g,`$\mathtt{m}_i$`)` holds of mouse $\mathtt{m}_i$. We then have to consider possible causal structures on these $2N_M$ variables. Since the constraints on plausible relations are such that if `Injected(c,`$\mathtt{m}_i$`)` causes `Expressed(g,`$\mathtt{m}_i$`)` for some mouse $\mathtt{m}_i$, then it does so for all mice, we can use plate notation (Section 2.3) to efficiently summarize the causal structures in the hypothesis space, representing the relationship between `Injected(c,M)` and `Expressed(g,M)`, quantifying over all mice with a logical variable `M`. The constraints on plausible relations imply that the only possible causal relationship in this graphical model is that from `Injected(c,`$\mathtt{m}_i$`)` to `Expressed(g,`$\mathtt{m}_i$`)`, and that this relationship holds with probability p. The hypothesis space $\mathcal{H}$ thus consists of two causal graphical models: one in which `Injected(c,M)` causes `Expressed(G,M)`, which has prior probability p, and one in which `Injected(C,M)` does not cause `Expressed(G,M)`, which has prior probability $1 - p$. These are Graph 1 and Graph 0, shown at the top of Figure 3.3, taking C to stand for `Injected(C,M)`, E to stand for `Expressed(G,M)`, and using plates to quantify over `M`.

The same procedure can be used to generate a hypothesis space of causal graphical models for other numbers of entities. For example, with two chemicals, $\mathtt{c}_1$ and $\mathtt{c}_2$, and two genes, $\mathtt{g}_1$ and $\mathtt{g}_2$, the hypothesis space contains sixteen causal graphical models, with the

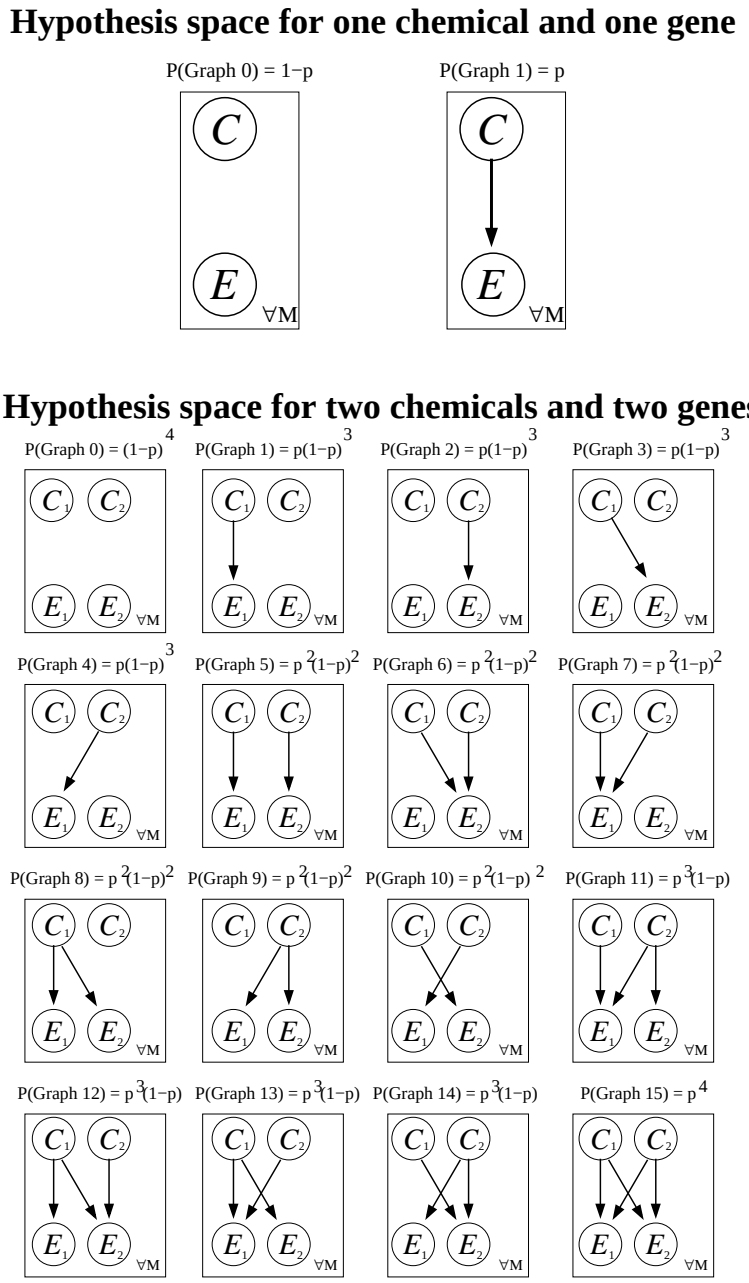

Figure 3.3: Hypothesis spaces generated by the theory shown in Figure 3.1. The top of the figure shows the hypothesis space for one chemical and one gene, which includes only two causal structures. With two chemicals and two genes, the hypothesis space includes sixteen causal structures, as shown in the lower portion of the figure. In the graphs, C corresponds to `Injected(c,M)` for `Chemical c` and E corresponds to `Expressed(g,M)` for `Gene g`. C_1, C_2, E_1, and E_2 should be interpreted similarly. `M` is a logical variable, and the plates indicate that these relationships hold for all mice `M`.

prior probabilities determined by the number of causal relationships expressed in the graph. This hypothesis space is shown in the lower portion of Figure 3.3. The same theory can be used to define a hypothesis space for five chemicals and ten genes, or fifty chemicals and a thousand genes – the theory provides abstract principles that can be used to construct a hypothesis space for any set of objects, just as a grammar can be used to construct all possible parses of a sentence.

3.2 Top-down and bottom-up information

The theory-based causal induction framework combines the formalization of theories developed in the previous section with the principles of Bayesian structure learning. This framework provides a precise account of how top-down and bottom-up information are combined in causal learning. A theory T supplies a hypothesis space of causal graphical models $\mathcal{H}_T$, each having a prior probability $P(\text{Graph } i|T)$. Each graph has a parameterization, defining a likelihood, $P(\mathcal{D}|\text{Graph } i, T)$. These probabilities can be combined via Bayes' rule (Equation 2.5) to compute a posterior probability for all of the models in the hypothesis space, $P(\text{Graph } i|\mathcal{D}, T)$. These posterior probabilities indicate how likely it is that each graphical structure was responsible for generating the data $\mathcal{D}$. The posterior probability of Graph i given data $\mathcal{D}$ and theory T is

$$P(\text{Graph } i|\mathcal{D}, T) = \frac{P(\mathcal{D}|\text{Graph } i, T)P(\text{Graph } i|T)}{P(\mathcal{D}|T)} \tag{3.1}$$

where $P(\mathcal{D}|T) = \sum_{i=1}^{|\mathcal{H}_T|} P(\mathcal{D}|\text{Graph } i, T)P(\text{Graph } i|T)$ and $|\mathcal{H}_T|$ is the size of the hypothesis space.

Equation 3.1 indicates how the top-down influence of prior knowledge is combined with the bottom-up influence of data to identify a particular causal structure. The top-down influence is introduced by the prior, $P(\text{Graph } i|T)$, while the bottom-up influence is introduced by the likelihood, $P(\mathcal{D}|\text{Graph } i, T)$. The posterior distribution over causal structures is proportional to the product of these two terms, making it tempting to believe that they exert equal influence on the result. This "additive combination" view of the interaction between prior knowledge and data is fairly widespread in the causal induction literature (e.g., Alloy & Tabachnik, 1984; Koslowski, 1996), but it misses some important subtletles of Equation 3.1.

The theory T does not simply add evidence in favor of a particular causal structure: it determines what counts as data, how that data is interpreted, and what other hypotheses are considered. By generating the variables of the causal graphical models, the theory establishes what properties of the world can have bearing on possible causal structures. $\mathcal{D}$ consists of the values that these variables take on. Furthermore, $P(\mathcal{D}|\text{Graph } i, T)$ is determined by the theory T as the causal structure of the graph, due to the dependence of the parameterization upon the theory. Different assumptions about the functional form of causal relationships will justify different conclusions from the same data. Finally, the normalizing constant $P(\mathcal{D}|T)$ depends upon the other hypotheses generated by T. The posterior probability of any particular graph will depend upon the number of other causal structures generated by T that could explain $\mathcal{D}$. These three factors are all neglected by conceiving of causal induction as the result of an additive combination of top-down and bottom-up information.

3.3 Descriptive and explanatory goals

Just as a grammar generates the set of syntactic structures considered in parsing a sentence, a theory generates the set of causal models considered in causal induction. Pursuing this analogy with linguistics suggests two goals to be pursued in studying human induction: a "descriptive" goal of characterizing the theories used in different settings, and an "explanatory" goal of accounting for their origins (c.f. Chomsky, 1965). I will pursue the descriptive goal in Chapters 4 to 6, using the theory-based framework developed in this chapter to identify the assumptions that are necessary to explain people's inferences about different causal systems. Chapter 4 analyses the case of causal learning from contingency data, discussing experiments that use a cover story very similar to the toxicology example, and showing that people make assumptions about functional form even in this abstract setting. I then examine people's inferences across a variety of different causal systems: Chapter 5 concerns simple physical systems that can be modeled in discrete time, such as detectors and other machines, while Chapter 6 focuses on systems that operate in continuous time, such as particle emissions and explosions. Considering this broad range of systems makes it possible to address several important issues: learning from small samples, use of observations and interventions, inferences about hidden causes, reasoning about dynamical systems, and the effects of domain on causal induction.

Just as examining the properties of different languages provides insights into their common grammatical structures, examining the kind of knowledge necessary to explain people's inferences about the different systems discussed in Chapter 4 to 6 reveals commonalities in the way that this knowledge is organized, and provides hints as to how it might be acquired. In Chapters 7 to 9 I return to the explanatory goal, considering how theories might be acquired. Chapter 7 argues that coincidences play a key role in the process of theory change, Chapter 8 discusses theory acquisition more broadly, and Chapter 9 connects these ideas back to the key issues with which the thesis began. However, as with language acquisition, these questions are deep and difficult, and I will only scratch their surface here.

In each of the following chapters, I will use specific formal theories like that shown in Figure 3.2 to develop computational models of people's inferences. These theories should not be interpreted as an attempt to definitively state the knowledge that people have about a particular a domain, or as claims about mental representations. What is important about these theories is the set of constraints that they imply for causal structures. There are many ways of specifying these theories that differ in scope and terminology but imply the same constraints, and the theories that I present are intended to describe these constraints as concisely as possible. These theories represent working assumptions about the knowledge necessary to explain people's inferences in particular settings. Making such assumptions is necessary in order to demonstrate the utility of the central idea that motivates this framework: that the interaction between prior knowledge and statistical inference in causal induction can be understood by conceiving of theories as hypothesis space generators.

Chapter 4

Contingency data

> The contagion spread rapidly and before its progress could be arrested, sixteen persons were affected of which two died. Of these sixteen, eight were under my care. On this occasion I used for the first time the affusion of cold water, in the manner described by Dr. Wright. It was first tried in two cases... The effects corresponded exactly with those mentioned by him to have occurred in his own case and thus encouraged the remedy was employed in five other cases. It was repeated daily, and of these seven patients, the whole recovered.
>
> James Currie (1798/1960, p. 430)

The most basic problem of causal induction is learning that a relationship exists between a single cause and effect. Psychologists have extensively investigated how people infer causal relationships from contingency data, which indicate the frequency with which cause and effect co-occur. As the epigraph illustrates, people are quite capable of inferring causal relationships from such data. Its author, Dr. James Currie, was an eighteenth-century ship's surgeon who later went into practice in Liverpool. After having heard a Dr. William Wright give an account of the efficacy of being doused with cold water in treating an extended fever, Currie conducted his own experiment, with results described above. He was sufficiently encouraged that he went on to use the treatment with hundreds of other patients, publishing a detailed treatise on the matter (Currie, 1798/1960). Washing the skin of the patient is still used to ease fevers, although modern medicine cautions against using water cold enough to induce shivering.

A number of mathematical models have been proposed to explain how people use contingency data to evaluate causal relationships (e.g., Anderson, 1990; Anderson & Sheu, 1995; Allan, 1980; Cheng, 1997; Jenkins & Ward, 1965; López, Cobos, Caño, & Shanks, 1998; Cheng & Novick, 1990; 1992; Shanks, 1995b). These models tend to downplay the role of prior knowledge, assuming that such knowledge serves to provide a set of candidate causes, but contingency data are used to evaluate those causes. Indeed, Cheng (1997) goes so far as to suggest that "the assumption that causal induction and the influence of domain-specific prior causal knowledge are separable processes is justified by numerous experiments in which the influence of such knowledge can be largely ignored" (p. 370).

I will provide an account of human causal induction from contingency data within the theory-based framework developed in the previous chapter. Most experiments using contingency data select candidate causes and effects for which causal relationships are plausible. This uniformity of plausibility underlies claims about the separability of causal induction and prior knowledge, and means that the aspect of causal theories that determines the plausibility of relationships will not be as relevant here as in other settings discussed in later chapters. However, as the theory-based framework emphasizes, prior knowledge is not restricted to plausibility: it also determines assumptions about functional form. This framework thus makes two claims about causal learning from contingency data: that variation in the assumed functional form should produce variation in human judgments, and that causal inferences can be understood as statistical inferences. In this chapter, I will test these claims by using the framework to define a new model of causal judgments, called "causal support". In the process, I will also attempt to clarify the assumptions behind previous rational models, such as ΔP (Allan, 1980; Jenkins & Ward, 1965; López et al., 1998) and causal power (Cheng, 1997),.

Causal support predicts several phenomena that are problematic for other rational models. My presentation will be organized around these phenomena. The first phenomenon I will describe is the interaction between covariation, measured by ΔP, and the base-rate probability of the effect in the absence of the cause in determining human judgments. This interaction manifests in two curious phenomena, the "frequency illusion" – a decrease in causal judgments as the base-rate decreases when $\Delta P = 0$ (Allan & Jenkins, 1983; Buehner, Cheng, & Clifford, 2003; Shanks, López, Darby, & Dickinson, 1996) – and non-monotonic effects of changes in base-rate at other values of ΔP (Lober & Shanks, 2000). I will also discuss effects of sample size (White, 1998; 2002; 2003b) and inferences from incomplete

Table 4.1: Contingency Table Representation used in Causal Induction

	Effect Present (e^+)	Effect Absent (e^-)
Cause Present (c^+)	$N(e^+c^+)$	$N(e^-c^+)$
Cause Absent (c^-)	$N(e^+c^-)$	$N(e^-c^-)$

Note: $N(\cdot)$ denotes the frequency of a particular event.

contingency tables. No other rational model of can explain all of these phenomena, or fit as wide a range of datasets as causal support.

The plan of the chapter is as follows. First I will outline the problem of causal induction from contingency data in more detail, describing the experimental paradigms that are the focus of my investigation, the two leading rational models, ΔP and causal power, and some of the data that has been gathered in support of them. Then, I will use the theory-based framework described in the previous chapter to analyze this problem, and to derive causal support. The body of the chapter discusses the phenomena predicted by causal support but not by other models, explaining the statistical origins of these predictions.

4.1 Causal induction from contingency data

Much psychological research on causal induction has focused upon the problem of learning a single causal relationship from contingency data: given a candidate cause, C, and a candidate effect, E, and information about the frequency with which the effect occurs in the presence and absence of the cause, represented by the numbers $N(e^+, c^+), N(e^-, c^-)$ and so forth in Table 4.1, people are asked to assess the extent to which C causes E. In the toxicology example introduced in Chapter 1, C might be injecting a chemical into a mouse, and E the expression of a particular gene. For this case, $N(e^+, c^+)$ would be the number of injected mice expressing the gene, while $N(e^-, c^-)$ would be the number of uninjected mice not expressing the gene.

This contingency information is usually presented to participants in one of three modes. Early experiments on causal induction would either explicitly provide participants with the numbers contained in the contingency table (e.g., Jenkins & Ward, 1965), which I will refer to as a "summary" format, or present individual cases one by one, with the appropriate frequencies (e.g., Ward & Jenkins, 1965), which I will refer to as an "online" format. Some more recent experiments use a mode of presentation between these two extremes, showing

a list of all individual cases simultaneously (e.g., Buehner, Cheng, & Clifford, 2003; White, 2003b), which I will refer to as a "list" format.

Experiments also differ in the questions that are asked of the participants. Participants can be asked to rate the strength of the causal relationship, the probability of a causal relationship, or their confidence that a causal relationship exists. Understanding the effects of question wording is an ongoing task (e.g., White, 2003a), but one variable that has been shown to have a strong effect is asking counterfactual questions, such as "What is the probability that a mouse not expressing the gene before being injected will express it after being injected with the chemical?" (Buehner et al., 2003; Collins & Shanks, submitted).

Causal induction tasks also vary in their treatment of the valence of the potential cause, and the nature of the rating scale used for responses. Causes can be either "generative," increasing the probability of an outcome (as in our mouse gene example), or "preventive," reducing its probability (as in the case of Dr. Currie's cold water treatment). Some experiments use exclusively generative or exclusively preventive causes and ask for judgments on a nonnegative scale (e.g., 0 to 100), while others mix generative and preventive causes and ask for judgments on a scale that has both positive and negative ends (e.g., -100 to 100).

Given the many ways in which experiments on causal judgment can differ, it is important to identify the scope of the present analysis. I will discuss experiments that use all three modes of presentation, as each mode captures an aspect of causal induction that is important for the development of rational models: the summary format removes memory demands and allows a deliberative inference, the online format taps intuitions about causality that are engaged by direct interaction with data, and the list format falls between these extremes. I will focus on experiments that require participants to make judgments about potential causes of a single kind, generative or predictive. Most of the critical datasets in the current debate about rational models of causal induction are of this form (e.g., Buehner & Cheng, 1997; Lober & Shanks, 2000).

4.1.1 Rational models

Recent work has focused on connecting the judgments people make in causal induction tasks to some rational standard, following the same approach as that taken in this thesis. I will describe two leading rational models of causal induction which are at the center of a debate about modeling causal judgments: ΔP and causal power.

ΔP and associative strength

One common approach to modeling judgments about causal relationships is to combine the frequencies from a contingency table in the form

$$\Delta P = \frac{N(e^+, c^+)}{N(e^+, c^+) + N(e^-, c^+)} - \frac{N(e^+, c^-)}{N(e^+, c^-) + N(e^-, c^-)} = P(e^+|c^+) - P(e^+|c^-), \quad (4.1)$$

where $P(e^+|c^+)$ is the empirical conditional probability of the effect given the presence of the cause, estimated from the contingency table counts $N(\cdot)$. ΔP thus reflects the change in the probability of the effect occuring as a consequence of the occurence of the cause. This measure was first suggested by Jenkins and Ward (1965), subsequently explored by Allan (1980; 1993; Allan & Jenkins, 1983), and has appeared in various forms in both psychology and philosophy (Cheng & Holyoak, 1995; Cheng & Novick, 1990; 1992; Melz, Cheng, Holyoak & Waldman, 1993; Salmon, 1980). One argument for the appropriateness of ΔP as a normative model uses the fact that it is the asymptotic value of the weight given to the cause C when the causal induction task is modeled with a linear associator trained using the Rescorla-Wagner (Rescorla & Wagner, 1972) learning rule (Cheng, 1997; Cheng & Holyoak, 1995; Chapman & Robbins, 1990; Danks, 2003; Wasserman, Elek, Chatlosh & Baker, 1993).

The Power PC theory and causal power

Cheng (1997) rejected ΔP as a measure of causal strength because it is a measure of covariation, not causality. According to Cheng (1997; Novick & Cheng, 2004), human judgments reflect a set of assumptions about causality that differ from those of purely "covariational" measures such as ΔP and conventional statistics. Cheng's (1997) Power PC theory attempts to make these assumptions explicit, providing an axiomatic characterization of causality and proposing that human causal judgments correspond to "causal power," the probability that C produces E in the absence of all other causes. Causal power for a generative cause can be estimated from contingency data, with Cheng (1997) giving the expression:

$$\text{power} = \frac{\Delta P}{1 - P(e^+|c^-)}. \quad (4.2)$$

Causal power takes ΔP as a component, but predicts that ΔP will have a greater effect when $P(e^+|c^-)$ is large. Causal power can also be evaluated for preventive causes, following

from a similar set of assumptions about the nature of such causes. The causal power for a preventive cause is:

$$\text{power} = \frac{-\Delta P}{P(e^+|c^-)}, \tag{4.3}$$

For preventive causes, the effect of $P(e^+|c^+)$ on causal power is reversed, with ΔP having a greater influence when $P(e^+|c^+)$ is small.

The measure of causal power in Equation 4.2 can be derived from a counterfactual treatment of "sufficient cause" (Pearl, 2000). Causal power corresponds to the probability that, for a case in which C was not present and E did not occur, E would occur if C was introduced. This probability depends upon ΔP, corresponding to the raw increase in occurrences of E, but has to be normalized by the proportion of the cases in which C could actually have influenced E. If some of the cases already show the effect, then C had no opportunity to influence those cases and they should not be taken into account when evaluating the strength of C. The requirement of normalization introduces $P(e^-|c^-) = 1 - P(e^+|c^-)$ in the denominator.

To illustrate the difference between causal power and ΔP, consider the problem of establishing whether injecting chemicals into mice results in gene expression. Two groups of 60 mice are used in two experiments, evaluating the effect of different chemicals on different genes. In each experiment, one group is injected with the chemical, and the other group receives no injection. In the first experiment, 30 of the uninjected mice express the gene, $P(e^+|c^-) = 0.5$, and 36 of the injected mice express it, $P(e^+|c^+) = 0.6$. In the second experiment, 54 of the uninjected mice express the gene, $P(e^+|c^-) = 0.9$, and all 60 of the injected mice express it, $P(e^+|c^+) = 1$. In each case $\Delta P = 0.1$, but the second set of results seem to provide more evidence for a relationship between the chemical and gene expression. In particular, if we imagine that the frequency of gene expression among the uninjected mice would be reproduced exactly in the other group of mice prior to injection, it seems that the first chemical produces gene expression in only six of the thirty mice who would not have otherwise expressed the gene, while *all* of the mice not expressing the gene in the second experiment have their fates altered by the injection. This difference is reflected in causal power, which is 0.2 in the first case and 1 in the second.

4.1.2 The debate over ΔP and causal power

ΔP and causal power make different predictions about the strength of causal relationships, and several experiments have been conducted with the aim of determining which model gives a better account of human data (e.g., Buehner & Cheng, 1997; Collins & Shanks, submitted; Lober & Shanks, 2000; Perales & Shanks, 2003; Shanks, 2002; Vallee Tourangeau, Murphy, Drew, & Baker, 1998). Each model captures some of the trends identified in these experiments, but there are several results that are predicted by only one of the models, as well as phenomena that are predicted by neither. These negative results are almost equally distributed between the two models, and suggest that there may be some basic factor missing from both. The problem can be illustrated by considering two sets of experiments: those conducted by Buehner and Cheng (1997) and Lober and Shanks (2000).

The experiments conducted by Buehner and Cheng (1997; Buehner et al., 2003) explored how judgments of the strength of a causal relationship vary when ΔP is held constant. This was done using an experimental design adapted from Wasserman et al. (1993), giving 15 sets of contingencies expressing all possible combinations of $P(e^+|c^-)$ and ΔP in increments of 0.25. Experiments were conducted with both generative causes, for which C potentially increases the frequency of E as in the cases described above, and preventive causes, for which C potentially decreases the frequency of E, and with both online and summary formats. For the moment, I will focus on the online study with generative causes (Buehner & Cheng, 1997, Experiment 1B), where a total of 16 trials gave the contingency information. The results of this experiment showed that at constant values of ΔP, people made judgments that were sensitive to the value of $P(e^+|c^-)$. Furthermore, this sensitivity was consistent with the role of $P(e^+|c^-)$ in causal power.

However, as was pointed out by Lober and Shanks (2000), the results also proved problematic for the Power PC theory. The design used by Buehner and Cheng (1997) provides several situations in which sets of contingencies give the same value of causal power. The data are shown in Figure 4.1, together with the values of ΔP and causal power. ΔP and causal power gave r scores of 0.889 and 0.881 respectively, with scaling parameters $\gamma = 0.98, 1.05$.[1] As can be seen from the figure, both ΔP and causal power predict

[1]Throughout this chapter, in each case where I have fit a computational model to empirical data, I have used a scaling transformation to account for the possibility of non-linearities in the rating scale used by participants. This is not typical in the literature, but is necessary to separate the quantitative predictions from a dependency on the linearity of the judgment scale – an issue that arises in any numerical judgment task. I use the transformation $y = \text{sign}(x)\text{abs}(x)^\gamma$, where y are the transformed predictions, x the raw

important trends in the data, but since these trends are orthogonal, neither model provides a full account of human performance. The only sets of contingencies for which the two models agree are those where ΔP is zero. For these cases, both models predict negligible judgments of the strength of the causal relationship. In contrast to these predictions, people give judgments that seem to decrease systematically as $P(e^+|c^-)$ decreases. Similar effects with $\Delta P = 0$ have been observed in other studies (e.g., Allan & Jenkins, 1983; Shanks et al., 1996), where the phenomenon is referred to as the "frequency illusion".

In a further test of the two theories, Lober and Shanks (2000) conducted a series of experiments in which either causal power or ΔP was held constant while the other varied. These experiments used both online (Experiments 1-3) and summary (Experiments 4-6) formats. The results showed systematic variation in judgments of the strength of the causal relationship at constant values of causal power, in a fashion consistent with ΔP. The results of Experiments 4-6 are shown in Figure 4.2, together with the values of ΔP and causal power. The models gave r scores of 0.980 and 0.581 respectively, with $\gamma = 0.8, 1.1$. While ΔP gave a good fit to these data, the human judgments for contingencies with $\{P(e^+|c^+), P(e^+, c^-)\}$ of $\{30/30, 18/30\}$, $\{24/30, 12/30\}$, $\{12/30, 0/30\}$ are not consistent with ΔP: they show a non-monotonic trend, with smaller judgments for $\{24/30, 12/30\}$ than for either of the extreme cases. The quadratic trend over these three sets of contingencies was statistically significant, but Lober and Shanks (2000) stated that "...because the effect was non-linear, it probably should not be given undue weight" (p. 209). For the purposes of Lober and Shanks, this effect was not important because it provided no basis for discrimination between ΔP and causal power: neither of these theories can predict a non-monotonic change in causal judgments as a function of the base-rate probability $P(e^+|c^-)$.

The results of Buehner and Cheng (1997) and Lober and Shanks (2000) illustrate that neither ΔP nor causal power provides a full account of people's judgments in causal induction tasks. These are not isolated results: ΔP and causal power cannot explain several other phenomena of human causal induction. One of these phenomena is the effect of sample size: both ΔP and causal power are defined using the conditional probabilities $P(e|c)$, and are thus insensitive to the number of observations expressing those probabilities. However, human judgments change as the number of observations contributing to a contingency table

predictions, and γ a scaling parameter selected to maximize the linear correlation between the transformed predictions and the data. This power law transformation accommodates a range of non-linearities.

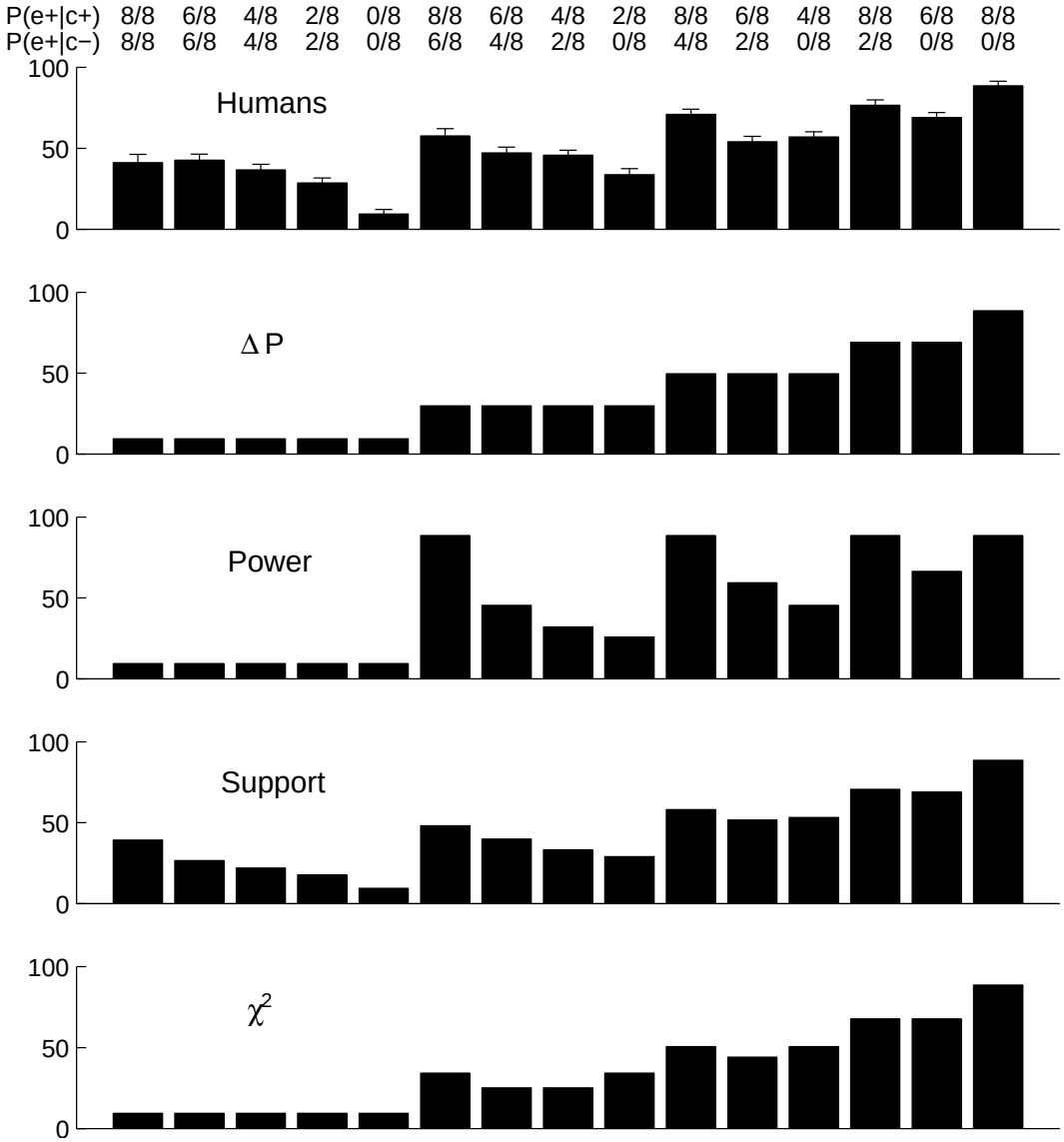

Figure 4.1: Predictions of rational models compared with the performance of human participants from Buehner and Cheng (1997, Experiment 1B). Numbers along the top of the figure show stimulus contingencies, error bars indicate one standard error.

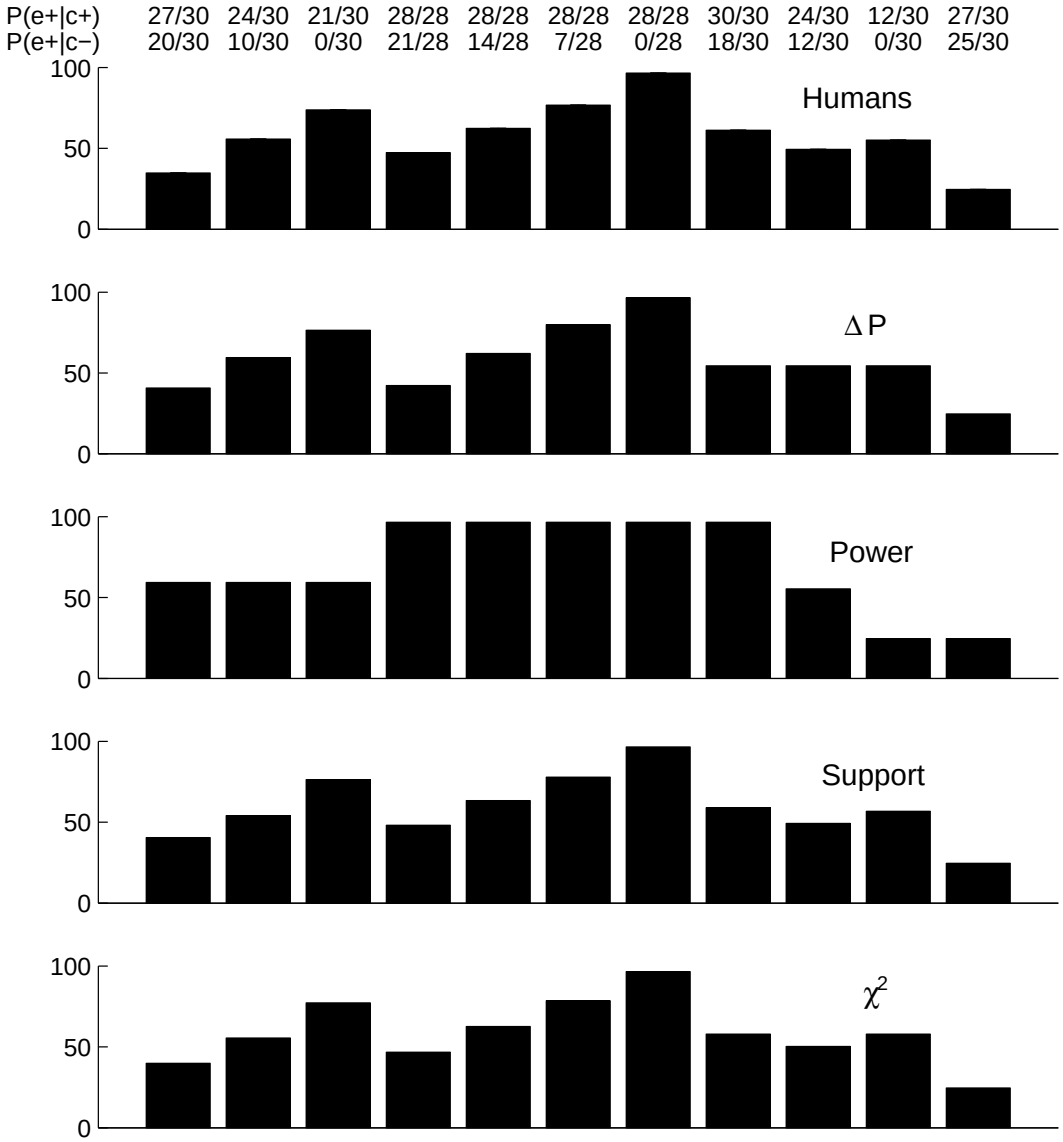

Figure 4.2: Predictions of rational models compared with the performance of participants from Lober and Shanks (2000, Experiments 4-6). Numbers along the top of the figure show stimulus contingencies.

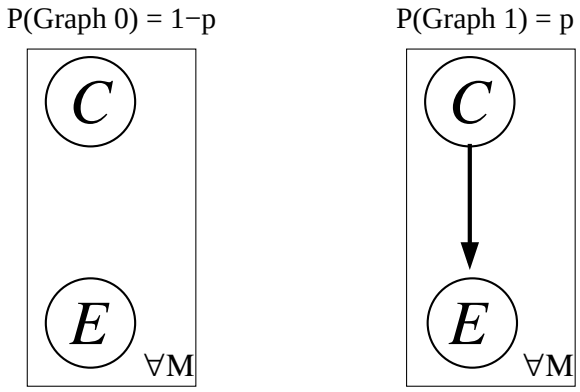

Figure 4.3: Hypothesis space for causal induction from contingency data. C corresponds to `Injected(c,M)` for `Chemical c` and E corresponds to `Expressed(g,M)` for `Gene g`. The plates indicate that these relationships hold for all mice `M`.

varies (White, 1998; 2002; 2003b). Another is inferences from incomplete data: people can assess causal relationships in circumstances where there is not enough information to compute the conditional probabilities of the effect in the presence and the absence of the cause. In the early stages of both everyday and scientific inferences, we might be presented with an incomplete contingency table. Neither ΔP nor causal power can explain the judgments that people make from such data. In the remainder of the chapter, I will use the theory-based framework to define a new model of causal induction from contingency data, and to provide insight into the problems of ΔP and causal power.

4.2 Theory-based causal induction

The experiments conducted by Buehner and Cheng (1997) and Lober and Shanks (2000) used cover stories with content similar to our toxicology example, asking people to evaluate causal relationships between chemicals and genes. The theory given in Figure 3.2 can be used to generate a hypothesis space of causal models expressing the different kinds of structure that might explain the contingency data yielded by the fictitious experiments described in these cover stories. We have a single chemical and a single gene, so the hypothesis space $\mathcal{H}$ contains the two models Graph 0 and Graph 1 shown at the top of Figure 3.3. For ease of reference, these models are reproduced in Figure 4.3. According to the theory, these models should use the Noisy-OR parameterization, with each cause independently having the chance to influence the effect.

Under this hypothesis space, the probability that a causal relationship exists is the probability that Graph 1 was responsible for generating the data, $\mathcal{D}$, being the frequencies with which cause and effect co-occurred. We could compute the posterior probability of this causal structure by applying Bayes' rule, as in Equation 3.1. Alternatively, we could choose to evaluate the *evidence* that $\mathcal{D}$ provides for this structure using the log likelihood ratio, also known as the "Bayes factor" (Kass & Raftery, 1995). I will use this measure to define "causal support," the evidence a data set $\mathcal{D}$ provides in favor of Graph 1 over Graph 0:

$$\text{support} = \log \frac{P(\mathcal{D}|\text{Graph 1})}{P(\mathcal{D}|\text{Graph 0})}. \tag{4.4}$$

To evaluate causal support, it is necessary to compute $P(\mathcal{D}|\text{Graph 1})$ and $P(\mathcal{D}|\text{Graph 0})$. Using a procedure described in Appendix B, it is possible to compute these probabilities without committing to particular values of the parameters w_0 and w_1 by integrating over all possible values these parameters could assume.

As illustrated in Figure 4.4, the major determinant of causal support is the extent to which the posterior distribution over w_1 places its mass away from zero. The contingency data for the top three cases shown in the figure all result in the same estimate of causal power (approximately the peak of the posterior distribution on w_1), but increasing the number of observations contributing to these contingencies decreases uncertainty about the value of w_1. It thus becomes more apparent that w_1 has a value greater than zero, and causal support increases. However, higher certainty does not always result in an increase in causal support, as shown by the next three cases in the figure. Causal power is zero for all three cases, and once again the posterior distribution shows higher certainty when the number of observations is large. Greater confidence that w_1 should be zero now results in a *decrease* in causal support, although the effect is weaker than in the previous case. The last three cases illustrate how causal support can differ from causal power. The contingencies $\{30/30, 18/30\}$ suggest a high value for w_1, with relatively high certainty, and consequently strong causal support; $\{24/30, 12/30\}$ suggest a lower value of w_1, with less certainty, and less causal support; and $\{12/30, 0/30\}$ produces an even lower value of w_1, but the higher certainty that this value is greater than zero results in more causal support.

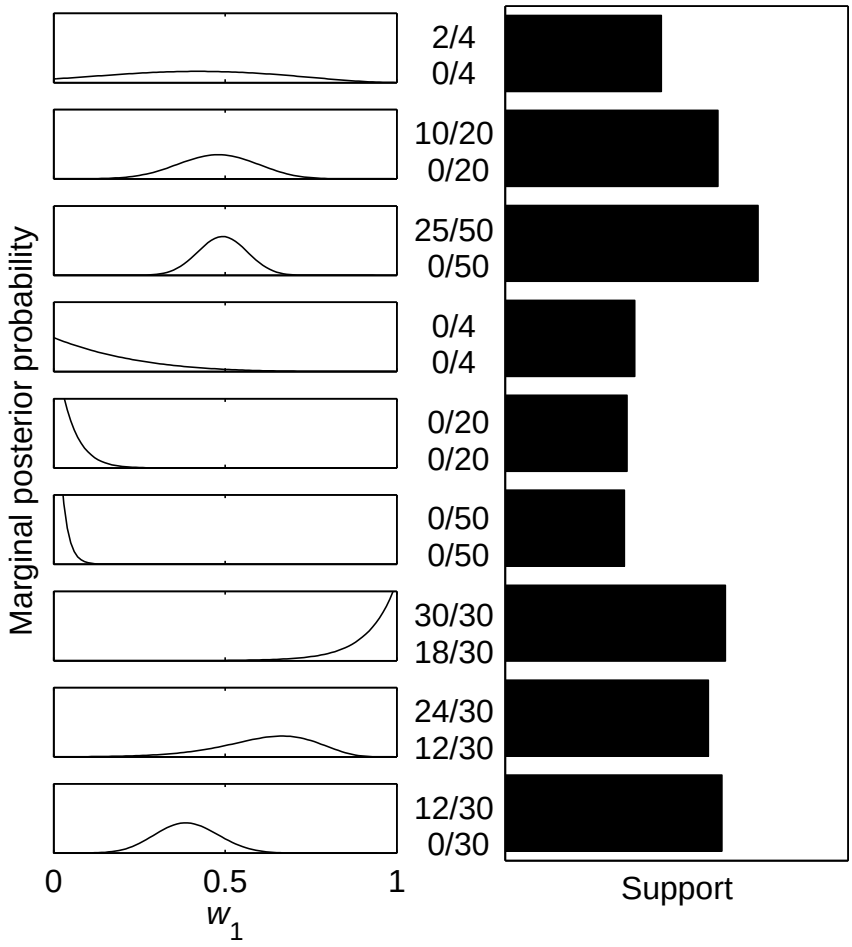

Figure 4.4: Marginal posterior distributions on w_1 and values of causal support for six different sets of contingencies. The first three sets of contingencies result in the same estimates of ΔP and causal power, but different values of causal support. The change in causal support is due to the increase in sample size, which reduces uncertainty about the value of w_1. As it becomes clear that w_1 takes on a value other than zero, the evidence for Graph 1 increases, indicated by the increase in causal support. The second set of three contingencies shows that increasing sample size does not always result in increased causal support, with greater certainty that w_1 is zero producing a mild decrease in causal support. The third set of three contingencies illustrates how causal support and causal power can differ. While the peak of the distribution over w_1, which will be close to the value of causal power, decreases across the three examples, causal support changes in a non-monotonic fashion.

4.3 Alternative accounts

We can gain insight into the critical features of this account of causal induction from contingency data by comparing its predictions to those of other models. The theory-based account assumes that people approach causal induction as a decision between different causal structures, and that their intuitive theory of the interactions between chemicals and genes involves a functional form equivalent to the noisy-OR. By exploring the consequences of modifying these assumptions, we can evaluate what role they play in explaining people's inferences. I will describe two classes of models that each result from modifying one of these assumptions. The first class makes assumptions about functional form, but does not treat causal induction as a problem of structure learning. The two models in this class are the psychological models of causal induction from contingency data introduced above – ΔP and causal power. The second class treats causal induction as structure learning, but does not make assumptions about functional form, and includes the χ^2 test used in constraint-based algorithms.

4.3.1 Functional form without structure learning

My discussion of learning causal graphical models in Chapter 2 focused on the problem of structure learning, which in this case amounts to asking whether or not a causal relationship exists. However, there is a second problem that arises when learning such models, the problem of parameter estimation. This involves assuming that a relationship exists, and evaluating its strength. There are several approaches to parameter estimation in graphical models (Heckerman, 1998). The simplest approach is maximum-likelihood estimation. For a particular graphical structure and parameterization, the likelihood of a set of parameters for data $\mathcal{D}$ is the probability of $\mathcal{D}$ under the distribution specified by that structure, parameterization, and choice of parameter values. Given Graph 1 with a noisy-OR parameterization, maximum-likelihood parameter estimation would involve choosing the values for w_0, w_1 that maximize $P(\mathcal{D}|w_0, w_1, \text{Graph 1})$. A Bayesian alternative is maximum a posteriori parameter estimation, in which a prior is defined on w_0 and w_1, and parameter values are chosen to maximize the product of this prior and the likelihood.

The two rational models at the heart of the current theoretical debate about elemental causal induction address the same component of the underlying computational problem in fundamentally similar ways. Both ΔP and causal power are maximum-likelihood estimates

of the causal strength parameter w_1 in Graph 1, but under different parameterizations (Tenenbaum & Griffiths, 2001). As shown in Appendix B, ΔP corresponds to the linear parameterization (Equation 2.4), whereas causal power for generative causes corresponds to the noisy-OR parameterization (Equation 2.1) and for preventive causes corresponds to the noisy-AND-NOT parameterization (Equation 2.3). Glymour (1998) also showed that causal power corresponds to the strength parameter in a noisy-OR. Since they are estimates of the parameters of a fixed graphical structure, ΔP and causal power both measure the strength of a causal relationship, based upon the assumption that the relationship exists.

As point estimates of a parameter, ΔP and causal power share several properties. Firstly, they do not answer the question of whether or not a causal relationship exists. Large values of ΔP or causal power are likely to be associated with genuine causal relationships, but small values are non-diagnostic. Shanks (1995a, p. 260) notes this, pointing out that while ΔP for the effect of smoking on lung cancer is small, around 0.00083, "no one would deny that the relationship between smoking and lung cancer is an important one". This relates to a second property of both ΔP and causal power: these measures contain no information about uncertainty in the estimates involved. This uncertainty is crucial to deciding whether a causal relationship actually exists. Even a small effect can provide strong evidence for a causal relationship, provided its value is estimated with high certainty. A major factor influencing certainty in a parameter estimate is the number of observations contributing to the estimate, highlighting a third property of both ΔP and causal power: neither is affected by the size of the sample from which their values are computed. One of the reasons we are able to conclude that smoking causes lung cancer, despite its relatively low strength, is that thousands of datapoints have contributed to this estimate of the strength of the cause, and we are quite sure that it is greater than zero.

Identifying both ΔP and causal power as maximum-likelihood parameter estimates also helps to illustrate how these measures differ: they make different assumptions about the functional form of a causal relationship. The linear relationship assumed by ΔP seems less consistent with the intuitions people express about causality than the noisy-OR, an important insight which is embodied in Cheng's (1997) Power PC theory. Cheng's (1997) distinction between "causal" and "covariational" measures turns on this fact: she views the noisy-OR parameterization as resulting from the correct set of assumptions about the nature of causality, and it is the use of this parameterization that distinguishes the Power PC theory from covariation-based accounts. Under the theory-based account, the appropriate

parameterization for the relationship between a cause and its effects will depend upon an individual's prior knowledge: in some settings, other parameterizations may be more appropriate than the noisy-OR.

4.3.2 Structure learning without functional form

The first step of constraint-based algorithms for learning causal structure is to evaluate the dependencies among a set of variables. This is typically done using the standard frequentist analysis of contingency tables, Pearson's (1904/1948) χ^2 test for independence. The use of the χ^2 test as a model for human causal judgment was suggested in the psychological literature (e.g., Allan, 1980), but was rejected on the grounds that it neglects the kind of asymmetry that is inherent in causal relationships, providing information solely about the existence of statistical dependency between the variables (Allan, 1980; López et al., 1998; Shanks, 1995b). χ^2 also makes no commitment about the functional form of the relationship between cause and effect: it simply detects any kind of statistical dependency between C and E.

4.3.3 Comparing the models

These three models – ΔP, causal power, and χ^2 – can be used to examine which of the assumptions behind the theory-based account are relevant to the explanation of particular phenomena in causal induction. In the remainder of the chapter, I will show that causal support can provide an explanation for phenomena that are problematic for existing psychological models of causal induction, and cannot be accounted for by using standard statistical tools for assessing causal relationships. I will consider four such phenomena in turn: the interaction between ΔP and $P(e^+|c^-)$, non-monotonic effects of $P(e^+|c^-)$, sample size effects, and inferences from incomplete contingency tables.

4.4 Interaction between ΔP and $P(e^+|c^-)$

In section 4.1.2 I discussed results from Experiment 1B of Buehner and Cheng (1997; later published in Buehner et al., 2003), which are shown in Figure 4.1. This experiment used an online format, and produced trends that were independently problematic for both ΔP and causal power, as well as a trend that neither model could predict: people's judgments at

$\Delta P = 0$ decrease as the base-rate probability of the effect, $P(e^+|c^-)$, decreases. The fundamental problem in explaining the results of Buehner and Cheng (1997) is accounting for the interaction between ΔP and the base-rate probability in producing human judgments. ΔP predicts no interaction, and causal power predicts the simple relationship given in Equation 4.2, but neither of these predictions matches human judgments. I will show that causal support is able to correctly predict this interaction, demonstrating that the model can capture the trends found by Buehner and Cheng (1997), and examine why it makes this prediction.

Figure 4.1 shows the data from Buehner and Cheng (1997, Experiment 1B), together with predictions of four models: ΔP, causal power, causal support, and χ^2. As noted in section 4.1.2, both ΔP and causal power capture some trends in the data, but miss others, resulting in correlations of $r = 0.889$ and $r = 0.881$. Causal support provides the best quantitative account of this dataset, $r = 0.968$, $\gamma = 0.668$, and accounts for all of the major trends in human judgments, including those at $\Delta P = 0$. χ^2 gives a correlation of $r = 0.889$, $\gamma = 0.596$.

Causal support correctly predicts the interaction between ΔP and $P(e^+|c^-)$ in influencing people's judgments. In particular, it is the only model that predicts human judgments at $\Delta P = 0$. The explanation for these predictions is not that there is decreasing evidence *for* a causal relationship as $P(e^+|c^-)$ decreases, but rather that there is no evidence for or against a causal relationship when $P(e^+|c^-) = 1$, and increasing evidence *against* a causal relationship as $P(e^+|c^-)$ decreases. This account depends on the assumption that the causal relationship – if it exists – is generative (increasing the probability of the effect, rather than preventing it). At one extreme, when $\{P(e^+|c^+), P(e^+|c^-)\} = \{8/8, 8/8\}$, all mice expressed the gene irrespective of treatment, and it is clear that there is no evidence for a causal relationship. But there can also be no evidence against a (generative) causal relationship, because of a complete "ceiling" effect: it is impossible for the cause to increase the probability of E occuring above its baseline value when $P(e^+|c^-) = 1$. This uncertainty in causal judgment when $P(e^+|c^-) = 1$ and $\Delta P = 0$ is predicted by both causal support, which is essentially 0, and also (as Cheng, 1997, points out) by causal power, which is undefined there.

Only causal support, however, predicts the gradient of judgments as $P(e^+|c^-)$ decreases. Causal support becomes increasingly negative as the ceiling effect weakens and the observation that $\Delta P = 0$ provides increasing evidence against a generative causal relationship.

At the other extreme, when $P(e^+|c^-) = 0/8$, no untreated mice expressed the gene, and there are eight opportunities for a causal relationship to manifest itself in the treated mice if such a relationship in fact exists. The fact that the effect does not appear in any treated mice, $P(e^+|c^+) = 0/8$, suggests that the drug does *not* cause gene expression. The intermediate cases provide intermediate evidence against a causal relationship. The contingencies $\{2/8, 2/8\}$ offer six chances for the treatment to have an effect, and the fact that it never does so is slightly weaker evidence against a relationship than in the $\{0/8, 0/8\}$ case, but more compelling than for $\{6/8, 6/8\}$, where the cause only has two chances to manifest itself and the observation that $\Delta P = 0$ could easily be a coincidence. This gradient of uncertainty shapes the Bayesian structural inference underlying causal support, but it does not impact the maximum-likelihood parameter estimates underlying causal power or ΔP.

Figure 4.5 reveals why causal support is sensitive to the change in $P(e^+|c^-)$, showing the posterior distribution on w_1 for each set of contingencies. Greater certainty in the value of w_1 is reflected in a more peaked distribution, and causal support becomes larger as it becomes more apparent that w_1 is greater than zero. The top five plots show the cases where $\Delta P = 0$. Despite the fact that ΔP is the same in these five cases, the posterior distributions on w_1 look quite different. Four of the distributions have a maximum at $w_1 = 0$, consistent with the estimate of causal power for these contingencies, but they differ in the certainty of the estimate, reflected in the breadth of the posterior about this point.[2] As $P(e^+|c^-)$ increases, fewer observations contribute to the estimate of w_1, and the posterior becomes more diffuse, being almost uniform for $P(e^+|c^-) = 1$.

This explanation depends upon two aspects of this theory-based account: the general assumption that causal induction is a statistical inference comparing causal structures, and the specific assumption that the appropriate functional form is the noisy-OR. The importance of these assumptions is supported by the failure of causal power, an estimate of the strength of a causal relationship under the noisy-OR, and the generic Bayesian structure-learning model. One prediction of this account is that people should alter their judgments in circumstances where a different functional form is appropriate. An opportunity to test this prediction comes from preventive causes. Since the noisy-OR only allows causes to

[2]The distribution for w_1 when $P(e^+|c^+) = P(e^+|c^-) = 1$ is mostly flat but has a very slight peak at $w_1 = 1$, despite causal power being undefined for this case. This is because there is also uncertainty in the value of w_0. If w_0 actually takes on any value less than 1, and the large number of occurrences of the effect in the absence of the cause is just a coincidence, then the large number of occurrences of the effect in the presence of the cause still needs to be explained, and the best explanation is that w_1 is high.

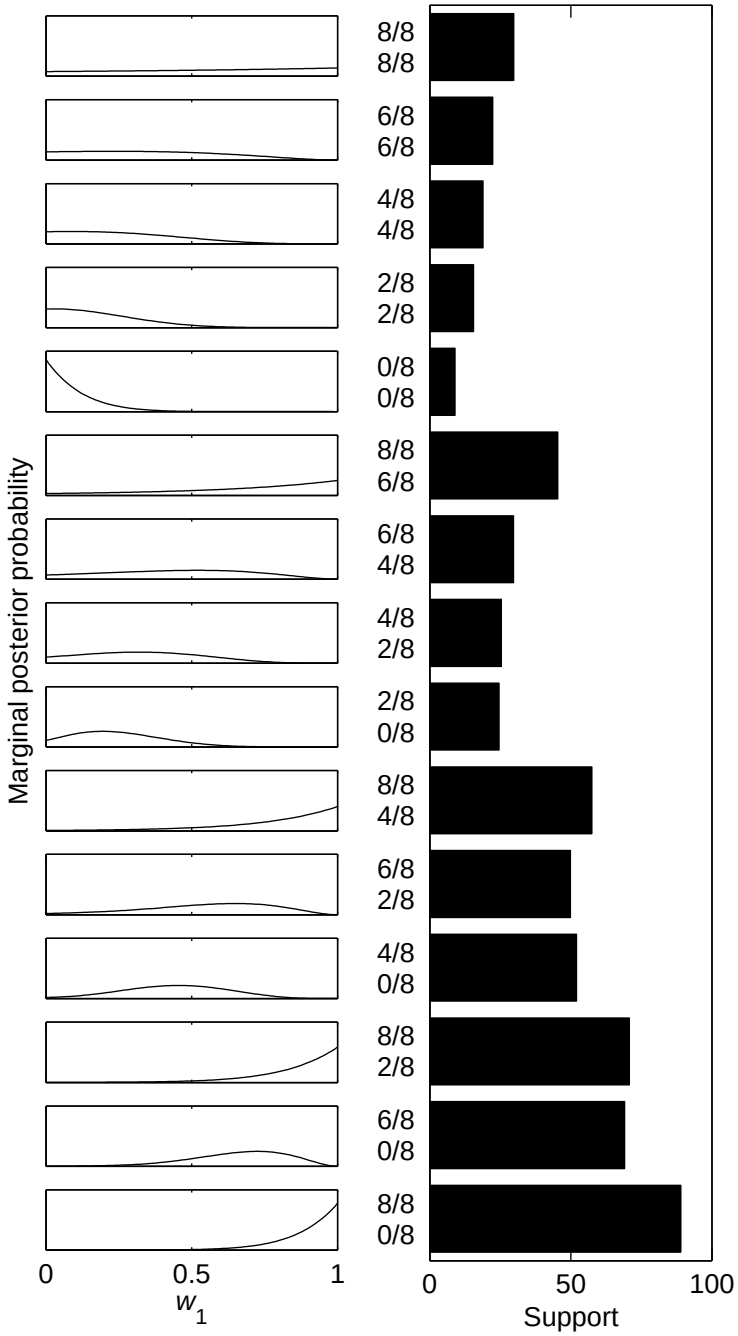

Figure 4.5: Marginal posterior distributions on w_1 and values of causal support for the contingencies used in Buehner and Cheng (1997, Experiment 1B).

increase the probability of their effects, a different parameterization is required to capture the properties of causes which decrease the probability of their effects. Buehner and Cheng (1997, Experiment 1A) used a design similar to that already described for generative causes to assess people's judgments about preventive causes. The results of this experiment are shown in Figure 4.6. From the figure, it can be seen that the trend at $\Delta P = 0$ is reversed for preventive causes: a decrease in $P(e^+|c^-)$ results in an increase in people's judgments.

Modeling this data requires making an assumption about the functional form of preventive causes. A simple theory of this kind would have the same content as the theory for generative causes shown in Figure 3.2, except for the assumptions about functional form, replacing the noisy-OR with the noisy-AND-NOT parameterization (Equation 2.3. Under a theory using the noisy-AND-NOT, the evidence for Graph 1 can still be evaluated via Equation 4.4, but the different parameterization results in different probabilities for $P(\mathcal{D}|\text{Graph 1})$ and $P(\mathcal{D}|\text{Graph 0})$. First, it should be noted that comparing structures with the noisy-AND-NOT parameterization provides a poor account of human judgments for generative causes, being strongly anti-correlated with human judgments for generative causes. However, the noisy-AND-NOT model gives a good account of human judgments for preventive causes, with $r = 0.922$, $\gamma = 0.537$, while the noisy-OR model produces a negative correlation. The explanation for the predictions of the noisy-AND-NOT model is similar to the generative case, although now the "ceiling effect" is a "floor effect": as $P(e^+|c^-)$ increases there is more opportunity for a non-zero value of w_1 to be demonstrated.

The alternative models discussed above can also be applied to this preventive setting. Under a linear parameterization, the appropriate measure of associative strength is -ΔP. Causal power for inhibitory causes is given in Equation 4.3. Glymour (1998) pointed out that this definition of causal power corresponds to w_1 under a noisy-AND-NOT parameterization, and the same argument as for the noisy-OR shows it to be a maximum-likelihood estimate of this parameter. χ^2 is insensitive to the distinction between generative and preventive causes, and can be applied just as in the generative case. Comparing these models to the data shows that causal power performs similarly to causal support with the noisy-OR parameterization, $r = 0.912$, $\gamma = 1.278$, while ΔP gives $r = 0.800$, $\gamma = 0.943$ and χ^2 gives $r = 0.790$, $\gamma = 0.566$. The predictions of these models are shown in Figure 4.6.

Inferences about generative causes are best captured using a noisy-OR parameterization, and inferences about preventive causes are best captured by the noisy-AND-NOT. This suggests that people make different assumptions about the functional form of generative

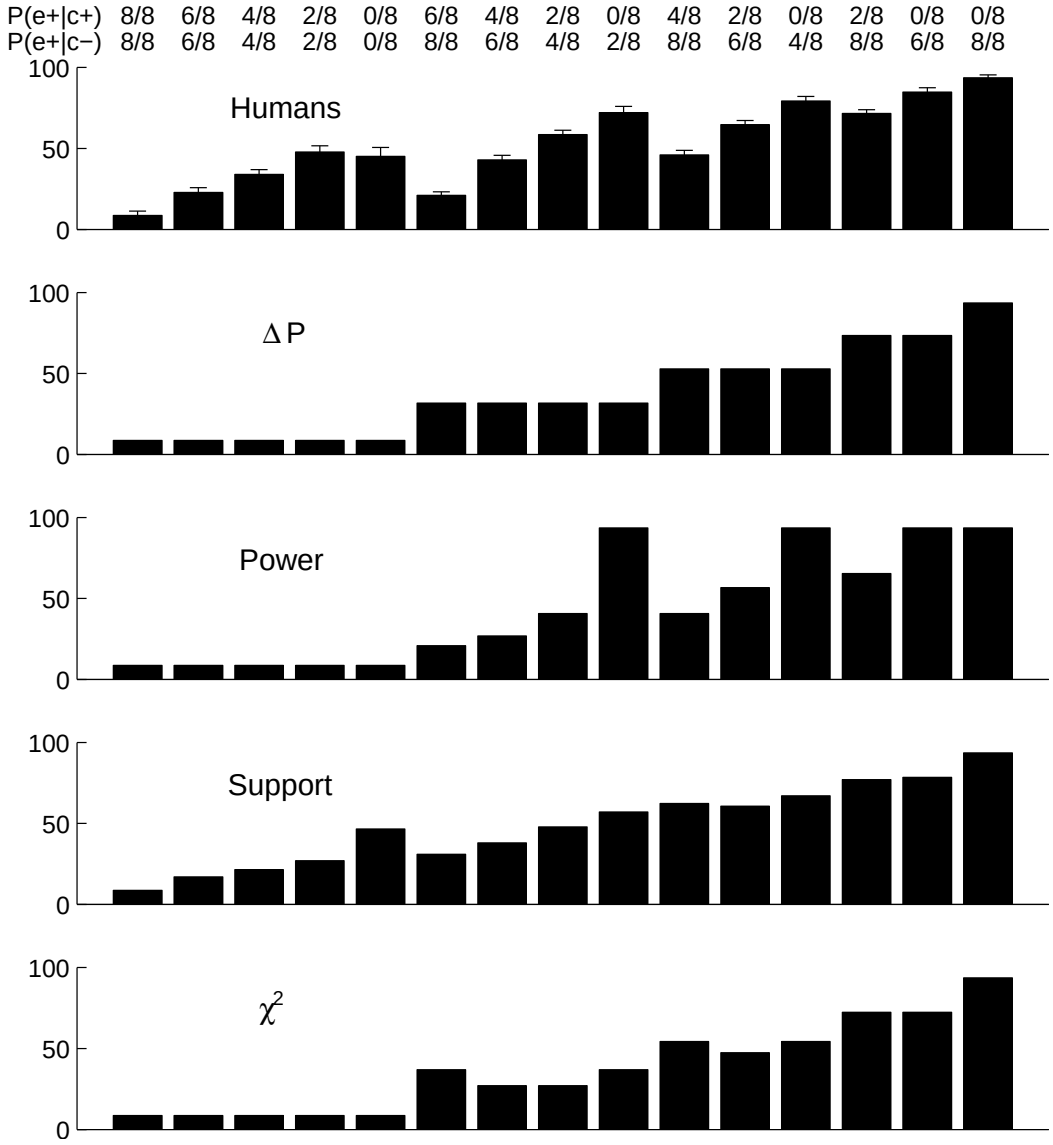

Figure 4.6: Predictions of rational models compared with the performance of human participants from Buehner and Cheng (1997, Experiment 1A). Numbers along the top of the figure show stimulus contingencies, error bars indicate one standard error.

and preventive causes, a fact that is consistent with the inclusion of functional form in the intuitive theories used in the theory-based framework. A further question raised by these results is whether there are any circumstances under which people's judgments will be best captured by other parameterizations. In particular, standard Bayesian structure learning algorithms use what I called the "generic" parameterization in Chapter 2, defining the probability of the effect in the presence and absence of the cause by separate parameters, with $P(e^+|c^-) = w_0$ and $P(e^+|c^+) = w_1$. The same parameterization is assumed by Anderson's (1990; Anderson & Sheu, 1995) rational model of causal induction (Griffiths & Tenenbaum, in press). This parameterization makes no assumptions about the nature of the causal relationship between C and E, postulating simply that the two variables are dependent.

The generic parameterization makes no assumptions about the nature of the relationship between cause and effect. Bayesian structure learning with this parameterization amounts to assessing whether there is any difference in the probability with which the effect occurs in the presence and absence of the cause. Consequently, we should expect that explicitly asking people to assess whether there is a difference in the probability with which the effect occurs under two different conditions will produce responses consistent with the generic parameterization. This hypothesis is tested in Experiment 4.1, which also aimed to replicate the results of Buehner and Cheng (1997) with generative and preventive causes.

4.4.1 Experiment 4.1: The effect of functional form

Method

Participants. Participants were 187 Stanford University undergraduates. There were 73 participants in the *generative* condition, 58 in the *difference* condition, and 67 in the *preventive* condition.

Stimuli. The three conditions used the same contingencies, presented in slightly different fashion. Participants were asked to evaluate a set of contingencies as either evidence for a *generative* causal relationship between a chemical and a gene, evidence for a *preventive* causal relationship between a chemical and a gene, or evidence for a *difference* in the probability of gene expression between two species of mice. The *generative* condition used the stimuli from Buehner and Cheng (1997, Experiment 1B), as given in Figure 4.1, the *preventive* condition used the stimuli from Buehner and Cheng (1997, Experiment 1A), as

given in Figure 4.6. These stimuli differ only in whether $P(e^+|c^+)$ is greater than $P(e^+|c^-)$ or vice versa. The *difference* condition used the same contingencies, randomizing whether the probability of the effect was higher for the first species or the second.

Procedure. The stimuli were presented in a one-page survey, which outlined the medical cover-story and then asked people to evaluate the evidence provided by the contingencies, stating that each set of contingencies indicated the results of a laboratory study. It was emphasized that each study concerned a different chemical (or pair of species) and a different gene. In the *generative* condition, participants were given the following instructions:

> For each study, write down a number between 0 and 100 representing the degree to which the chemical causes the gene to be expressed. A rating of 0 indicates that the chemical DOES NOT CAUSE the gene to be expressed at all, and a rating of 100 indicates that the chemical DOES CAUSE the gene to be expressed every time. Use intermediate ratings to indicate degrees of causation between these extremes.

In the *preventive* condition, the instructions read:

> For each study, write down a number between 0 and 100 representing the degree to which the chemical prevents a virus being caught. A rating of 0 indicates that the chemical DOES NOT PREVENT the virus at all, and a rating of 100 indicates that the chemical DOES PREVENT the virus every time. Use intermediate ratings to indicate degrees of prevention between these extremes.

Finally, the *difference* condition gave the instructions:

> For each study, write down a number between 0 and 100 representing how likely you think it is that the two species differ in their probability of expressing that gene. A rating of 0 indicates that the two species DEFINITELY have THE SAME probability of expressing the gene, while a rating of 100 indicates that the two species DEFINITELY have DIFFERENT probabilities of expressing the gene. Use intermediate ratings to indicate degrees of likelihood between these extremes.

Participants completed the survey as part of a booklet of unrelated surveys.

Table 4.2: Rank-Order Correlations for Different Rational Models

	Buehner & Cheng (1997)		Experiment 4.1		
Model	*Generative*	*Preventive*	*Generative*	*Difference*	*Preventive*
ΔP	0.883	0.769	0.968	0.968	0.946
Power	0.942	0.884	0.949	N/A	0.971
χ^2	0.880	0.761	0.966	0.958	0.940
Support:					
Noisy-OR	**0.961**	-0.814	**0.971**	0.850	-0.721
Generic	0.876	0.760	0.957	**0.975**	0.930
Noisy-AND-NOT	-0.868	**0.893**	-0.732	-0.336	**0.971**

Note: Boldface indicates highest correlation in each column.

Results and Discussion

People's judgments in the *generative* and *preventive* conditions replicated the results of Buehner and Cheng (1997), having a correlation of $r = 0.993$ and $r = 0.989$ with Experiments 1B and 1A respectively. The correlations of the six models with the three different conditions are shown in Table 4.2. As some of the models were negatively correlated with the data, the practice of finding the optimal non-linear scale transformation could not be applied here, since it would attempt to reduce negative correlations to zero correlations. Consequently, I report the rank-order correlation for each model. For comparison, the table also reports rank-order correlations for the experiments of Buehner and Cheng (1997).

Human judgments for the five stimuli for which $\Delta P = 0$ are shown in Figure 4.7. ΔP, causal power, and χ^2 all predict that there should be no variation in responses across these stimuli. Contrary to these predictions, the effect of $P(e^+|c^-)$ on judgments was statistically significant in all three conditions – *generative* ($F(4, 288) = 5.32, MSE = 216.92, p < 0.001$), *preventive* ($F(4, 264) = 2.63, MSE = 313.25, p < 0.05$), and *difference* ($F(4, 228) = 2.70, MSE = 359.19, p = 0.031$). There was also a statistically significant interaction between $P(e^+|c^-)$ and condition ($F(8, 780) = 3.57, MSE = 291.11, p < 0.001$). As can be seen from the figure, this variation was exactly as should be expected if people are performing Bayesian structure learning with the appropriate parameterization. Changing the context in which causal induction is performed thus seems to affect the functional form that people assume, a phenomenon that is completely consistent with the theory-based account.

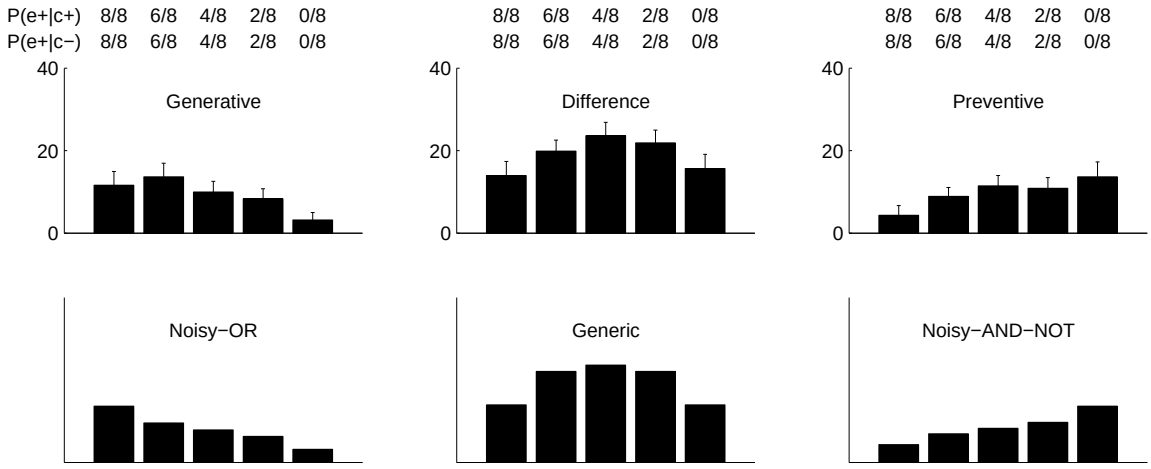

Figure 4.7: Effect of assumptions about functional form on causal induction. The top row shows people's judgments for a set of stimuli for which $\Delta P = 0$, under three different kinds of instructions, as described in the text. The bottom row shows the predictions of the theory-based account under three different assumptions about the functional form of a causal relationship. There appears to be a direct correspondence between task instructions and assumed functional form.

4.5 Non-monotonic effects of $P(e^+|c^-)$

Accounting for the interaction between ΔP and the base-rate probability, $P(e^+|c^-)$, is fundamental to explaining the results of Buehner and Cheng (1997). It is also important in explaining other phenomena of causal induction. The second dataset discussed in Section 4.1.2 was Experiments 4-6 from Lober and Shanks (2000), shown in Figure 4.2. ΔP accounts for these data quite well, reflected in the high correlation coefficient, $r = 0.980$, $\gamma = 0.797$, while causal power does poorly, $r = 0.581$, $\gamma = 1.157$. However, neither of these models can predict the non-monotonic effect of $P(e^+|c^-)$ seen with the $\{P(e^+|c^+), P(e^+, c^-)\}$ pairs $\{30/30, 18/30\}$, $\{24/30, 12/30\}$, $\{12/30, 0/30\}$. A similar, but weaker, trend can be seen in the online data of Buehner and Cheng (1997, Experiment 1B), shown in Figure 4.1, for the contingencies $\{8/8, 4/8\}$,$\{6/8, 2/8\}$,$\{4/8, 0/8\}$. These non-monotonic trends cannot even be predicted by models that form linear combinations of the entries in a contingency table, such as Anderson and Sheu (1995) and Schustack and Sternberg (1981), despite their many free parameters and great flexibility.

Causal support with the noisy-OR parameterization gives the best quantitative fit to this dataset, $r = 0.994$, $\gamma = 0.445$, with χ^2 performing similarly, $r = 0.993$, $\gamma = 0.502$. Both

causal support and χ^2 predict non-monotonic trends, as shown in Figure 4.2. The intuitive reasons for these predictions were mentioned when discussing Figure 4.4, which uses exactly the same set of contingencies: while $\{24/30, 12/30\}$ suggests a higher value of causal power than $\{12/30, 0/30\}$, such a difference in contingencies is more likely to arise by chance. As with the explanation of predictions for $\Delta P = 0$ given in Section 5, the high certainty in the value of w_1 for $\{12/30, 0/30\}$ results partly from the low value of $P(e^+|c^-)$.

The non-monotonic trend observed in Experiments 4-6 of Lober and Shanks (2000) did not appear in their Experiments 1-3, despite the use of the same contingencies, as shown in Figure 4.8. The only difference between these two sets of experiments was the presentation format, with an online format being used in Experiments 1-3, and a summary format in Experiments 4-6. This presents a challenge for the explanation based on causal support given above. However, I will argue that this discrepancy can be resolved through a finer-grained analysis of these experiments.

Catena, Maldonando, and Cándido (1998) and Collins and Shanks (2002) both found that people's judgments in online experiments are influenced by response frequency. Specifically, people seem to make judgments that are based upon the information presented since their last judgment, meaning that "primacy" or "recency" effects can be produced by using different response schedules (Collins & Shanks, 2002). Lober and Shanks (2000) used a procedure in which participants made judgments after blocks of trials, with 6 blocks of length 10 in Experiment 1, two blocks of length 18 and one of length 20 in Experiment 2, and 3 blocks of length 20 in Experiment 3. The actual trials presented in each block were selected at random. Thus, while the overall contingencies might match those used in Experiments 4-6, the contingencies contributing to any individual judgment varied. This may account for the difference between the results of the two experiments. In particular, the smaller sample sizes contributing to the contingencies may affect people's judgments.

To test this hypothesis, I used the records of the individual trials seen by the participants in these experiments to establish the contingencies each participant saw in each block, and evaluated the performance of different models in predicting the judgments of individual participants for each block from these contingencies.[3] The results of these analyses are given in Table 4.3. The correlations are lower than those reported elsewhere in the paper

[3]The raw data from Lober and Shanks (2000) was supplied by Klaus Melcher. The models compared in this section were fit using the same scaling parameter for all participants within the same experiment. Causal power was not computed for this comparison, as the presence of extreme contingencies in several cases resulted in undefined values of causal power, interfering with correlations and averaging.

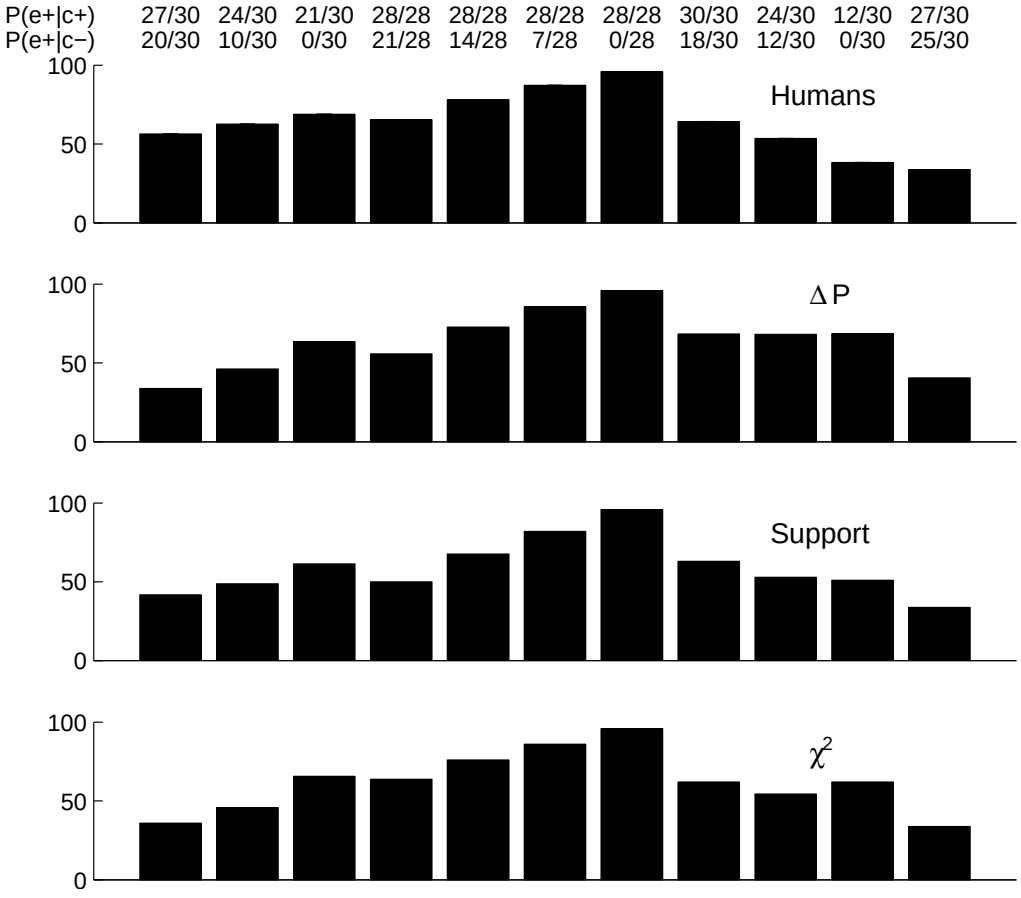

Figure 4.8: Predictions of rational models compared with the performance of participants from Lober and Shanks (2000, Experiments 1-3). Numbers along the top of the figure show stimulus contingencies, but the results are constructed by averaging over the blocks of trials seen by individual subjects, in which contingencies varied.

Table 4.3: Correlations of Rational Models with Lober and Shanks (2000)

	ΔP	Support	χ^2
Experiment 1	0.354	0.350	**0.354**
Experiment 2	0.462	0.465	**0.471**
Experiment 3	0.336	**0.382**	0.303
Overall means	0.695	**0.895**	0.829

Note: Boldface indicates highest correlation in each row.

because they concern the responses of individual participants rather than average scores. While all models did equally well in predicting the results of Experiments 1 and 2, causal support gave a better account of the results of Experiment 3, with a correlation of $r = 0.382$ as compared to $r = 0.336$ for ΔP, and $r = 0.303$ for χ^2. The model predictions, averaged across blocks and participants in the same fashion as the data, are shown in Figure 4.8. The mean values of causal support do not show any predicted non-monotonicity, in accord with the data. As shown in Table 4.3, the mean causal support correlates better with the mean human judgments than the mean predictions of any other model, with $r = 0.895$ for causal support, $r = 0.695$ for ΔP, and $r = 0.829$ for χ^2.

There are two reasons why causal support does not predict a non-monotonic trend for Experiments 1-3: smaller samples, and variation in observed contingencies. The effect of sample size is simple: causal support predicts a non-monotonic trend for contingencies derived from 30 trials in each condition, but this trend is almost completely gone when there are only 10 trials in each condition. Small samples provide weaker evidence that the strength of the cause is different from zero in all conditions, with $\{12/30, 0/30\}$ and $\{24/30, 12/30\}$ giving equivalently weak support for a causal relationship. The effect of variation in observed contingencies is more complex. Since both ΔP and causal power are estimated from the conditional probabilities $P(e|c)$, and the empirical probabilities give unbiased estimates of the true probabilities, averaging across many sets of contingencies generated according to those probabilities gives mean values that approximate the true ΔP and causal power. Causal support is a more complex function of observed contingencies, and averaging causal support across a set of samples produces different results from computing causal support from the average contingencies. In particular, variation in the contingencies within blocks results in some blocks providing very weak evidence for a causal relationship (for example, in the $\{12/30, 0/30\}$ condition, one participant saw a block in which the cause was present

on eight trials, but the effect occurred on only one of these trials). Such results have the greatest effect in the $\{12/30, 0/30\}$ condition, and the mean causal support in this condition consequently ends up slightly lower than causal support estimated from mean contingencies.

Although a non-monotonic effect of $P(e^+|c^-)$ appears in both Lober and Shanks (2000, Experiments 4-6) and Buehner and Cheng (1997), it has not been the principal target of any experiments. Since no existing model of elemental causal induction can predict this phenomenon, confirming its existence would provide strong evidence in favor of causal support. Experiment 4.2 was designed to explore this non-monotonic effect further.

4.5.1 Experiment 4.2: Testing for non-monotonicities

Method

Participants. 108 Stanford undergraduates participated for course credit.

Stimuli. The contingencies used in the experiment are shown in Figure 4.9. They included three sets of three contingencies at fixed values of ΔP but different values of $P(e^+|c^-)$, and several distractors. The sets of contingencies with fixed ΔP used $\Delta P = 0.40$, $\Delta P = 0.07$ and $\Delta P = 0.02$. ΔP predicts no effect of $P(e^+|c^-)$ within these sets, so any effect provides evidence against this model. Causal power predicts a monotonic increase in people's judgments as $P(e^+|c^-)$ increases, and causal support predicts a non-monotonic trend in the first two sets of contingencies, and a monotonic increase with $P(e^+|c^-)$ in the third. Finding non-monotonic effects in the first two sets of contingencies would thus provide evidence for causal support over causal power.

Procedure. The experiment was conducted in survey form. The instructions placed the problem of causal induction in a medical context:

> Imagine that you are working in a laboratory and you want to find out whether certain chemicals cause certain genes to be expressed in mice. Below, you can see laboratory records for a number of studies. In each study, a sample of mice were injected with a certain chemical and later examined to see if they expressed a particular gene. Each study investigated the effects of a **different** chemical on a **different** gene, so the results from different studies bear no relation to each other.
>
> Of course, these genes may sometimes be expressed in animals not injected with a chemical substance. Thus, a sample of mice who were not injected with any

> chemical were also checked to see if they expressed the same genes as the injected mice. Also, some chemicals may have a large effect on gene expression, some may have a small effect, and others, no effect.

Participants were then asked for ratings on a total of 14 different contingency structures. The instructions for producing the ratings were:

> For each study, write down a number between 0 and 20, where 0 indicates that the chemical DOES NOT CAUSE the gene to be expressed at all, and 20 indicates that the chemical DOES CAUSE the gene to be expressed every time.

Each participant completed the survey as part of a booklet of unrelated experiments.

Results and Discussion

The results are shown in Figure 4.9. As predicted, there was a statistically significant effect of $P(e^+|c^-)$ in all three sets of contingencies with fixed ΔP. For $\Delta P = 0.40$, $F(2, 214) = 6.61$, $MSE = 17.43$, $p < 0.005$, for $\Delta P = 0.07$, $F(2, 214) = 3.82$, $MSE = 26.93$, $p < 0.05$, for $\Delta P = 0.02$, $F(2, 214) = 6.06$, $MSE = 11.27$, $p < 0.005$. Since a quadratic trend analysis would only test deviation from linearity, and not non-monotonicity, the effect of $P(e^+|c^-)$ in each of these sets of contingencies was evaluated by testing each pair of means with neighboring values of $P(e^+|c^-)$. The response for $\{40/100, 0/100\}$ was significantly greater than that for $\{70/100, 30/100\}$, $t(107) = 3.00$, $p < 0.005$, and $\{70/100, 30/100\}$ was significantly less than $\{100/100, 60/100\}$, $t(107) = 3.81$, $p < 0.001$, indicating a non-monotonic trend at $\Delta P = 0.40$. The response for $\{7/100, 0/100\}$ was significantly greater than that for $\{53/100, 46/100\}$, $t(107) = 2.13$, $p < 0.05$, and $\{53/100, 46/100\}$ was significantly less than $\{100/100, 93/100\}$, $t(107) = 2.70$, $p < 0.01$, indicating a non-monotonic trend at $\Delta P = 0.07$. At $\Delta P = 0.02$, $\{2/100, 0/100\}$ was greater than $\{51/100, 49/100\}$, $t(107) = 4.05$, $p < 0.001$, but there was no significant difference between $\{51/100, 49/100\}$ and $\{100/100, 98/100\}$, $t(107) = 0.55$, $p = 0.58$, providing no evidence for non-monotonicity.

The results of this experiment suggest that the non-monotonicitic trend seen by Lober and Shanks (2000) is a robust aspect of human judgments, even though it may be a small effect. Such trends can be used to assess models of causal induction. Causal support, χ^2, and ΔP gave similarly high correlations with the experimental results, with $r = 0.952$, $\gamma = 0.604$, $r = 0.965$, $\gamma = 0.446$, and $r = 0.943$, $\gamma = 0.568$, respectively. Causal power performed far worse, with $r = 0.660$, $\gamma = 0.305$. The statistically significant effects of

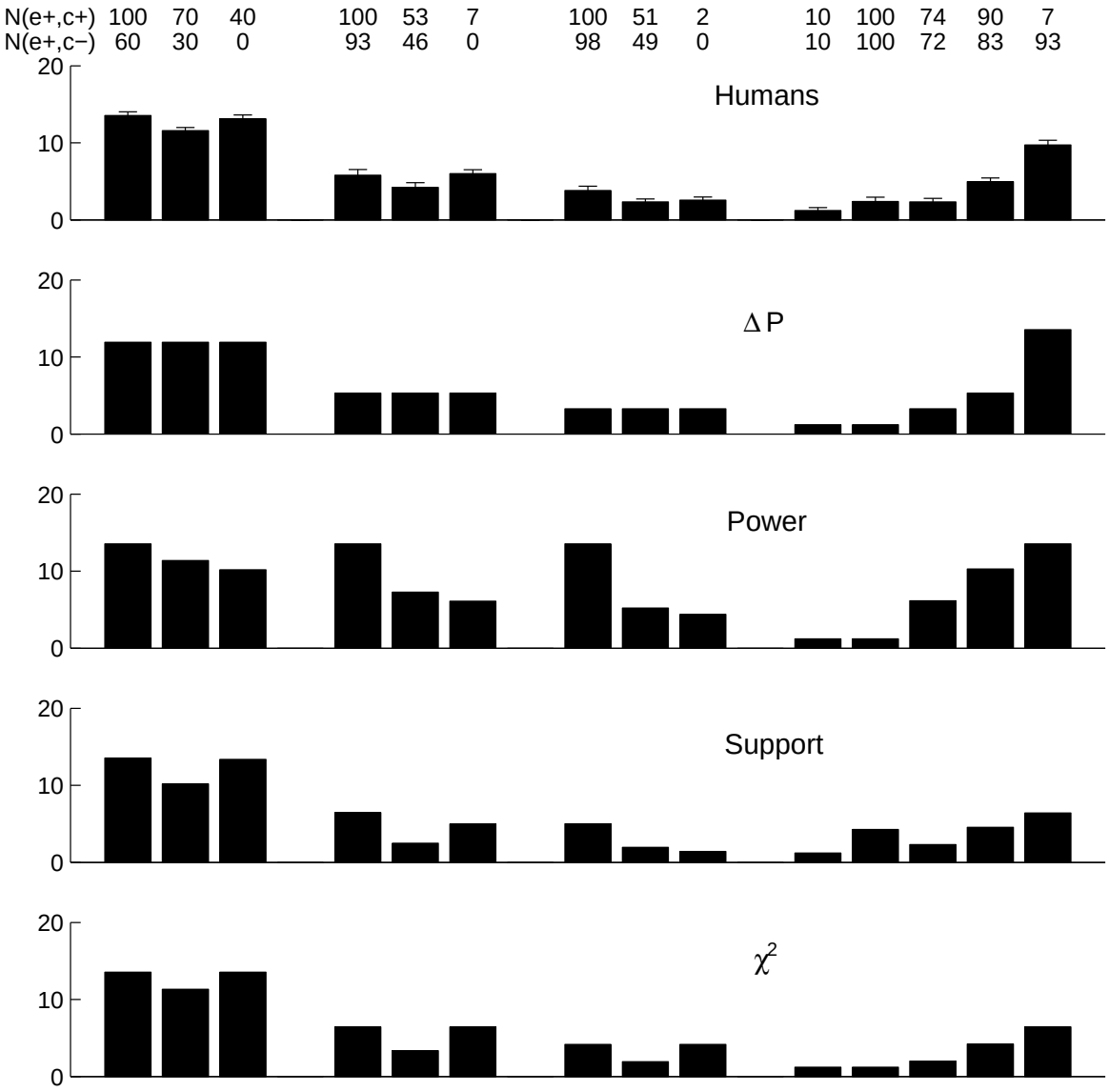

Figure 4.9: Predictions of rational models compared with results of Experiment 4.2. Numbers along the top of the figure show stimulus contingencies. These numbers give the number of times the effect was present out of 100 trials, for all except the last column, where the cause was present on 7 trials and absent on 193. The first three groups of contingencies are organized to display non-monotonicities in judgments, the last group contains distractor stimuli. Error bars indicate one standard error.

$P(e^+|c^-)$ in the three sets of contingencies with fixed ΔP are contrary to the predictions of ΔP. Only causal support and χ^2 predicted the observed non-monotonic trends.

4.6 Sample size effects

In explaining the results of Lober and Shanks (2000, Experiments 1-3), I touched upon the issue of sample size. Sample size is an important factor that affects structure learning but not parameter estimation: larger samples provide better grounds for assessing the evidence for a causal relationship, but do not affect parameter estimation. Both ΔP and causal power are computed using the conditional probabilities $P(e|c)$, rather than the number of observations contributing to these probabilities. Consequently, they predict no variation in judgments as the number of trials on which the cause was present or absent is varied. In contrast, both causal support and χ^2 are sensitive to sample size.

There are two dimensions along which sample size might be varied: the ratio of the number of trials on which the cause is present or absent ($N(c^+)$ and $N(c^-)$), and the total number of trials (N). Variation of the ratio of $N(c^+)$ to $N(c^-)$ has been explored extensively by White (1998; 2002; 2003b). These experiments revealed several effects of sample size, inconsistent with the predictions of ΔP and causal power, and were taken to support White's (2002) proportion of Confirming Instances (pCI) model, which differs from ΔP only when $N(c^+) \neq N(c^-)$. All four models – ΔP, causal power, causal support, and χ^2 – were applied to the results of these experiments, producing the correlations shown in Table 4.4. For several of these experiments, ΔP was constant for all stimuli, resulting in an undefined correlation between ΔP and human judgments, denoted N/A. I also present the results of the five models on Experiment 1 of Anderson and Sheu (1995), in which the entries in single cells of the contingency table were varied systematically, resulting in some sample size variation as well as other effects. The observed effects of sample size were broadly consistent with causal support, which gave either the best or close to the best account of all twelve datasets. There was no systematic relationship between the format in which contingency information was presented and the performance of the models, although causal support gave the best correlations with the three online datasets.

Variation of the total number of trials producing a set of contingencies, N, has been studied less extensively. White (2003b, Experiment 4) conducted an experiment in which this quanitity was varied, and found no statistically significant effect on human ratings.

Table 4.4: Correlations of Rational Models with Sample Size Experiments

Paper	Experiment	Format	ΔP	Power	Support	χ^2
White (1998)	3 (seeds)	summary(16)	0.907	**0.929**	0.924	0.865
	(contentless)	summary(16)	0.922	0.867	**0.935**	0.885
White (2002)	1	list(8)	**0.956**	0.765	0.938	0.936
	2	list(12)	0.772	0.852	**0.916**	0.830
	3	list(6)	N/A	0.760	**0.837**	0.146
White (2003b)	1	list(8)	0.200	0.389	**0.854**	0.791
	2	online(8)	0.070	0.409	**0.812**	0.640
	3	summary(8)	0.392	0.383	**0.677**	0.586
	4	list(8)	N/A	0.037	0.679	**0.729**
	5	list(8)	N/A	0.373	**0.803**	0.631
	6	online(4)	N/A	N/A	**0.676**	N/A
Anderson & Sheu (1995)	1	online(80)	0.884	0.816	**0.894**	0.329

Note: Boldface indicates highest correlation in each row. Number in parentheses in Format column indicates number of stimuli.

As shown in Figure 4.4, causal support makes clear predictions about the effect of N on the evidence for a causal relationship, and the demonstration of such an effect would provide evidence for the involvement of structure learning in human causal induction. One possible explanation for the lack of a sample size effect in White's (2003b) experiment is the use of ratings as a dependent measure: the effect of sample size might be concealed by allowing people the possibility of giving equal ratings to different stimuli. Consequently, it makes sense to explore this phenomenon further using a more sensitive response measure: Experiment 4.3 was designed to examine whether sample size affects people's assessment of causal relationships, using a rank ordering task.

4.6.1 Experiment 4.3: Sample size effects with ranking

Method

Participants. Participants were 20 members of the MIT community who took part in the experiment in exchange for candy.

Stimuli. Nine stimuli were used, each composed of different contingency data. The two critical sets of contingencies were a set for which $\Delta P = 0$, consisting of $\{0/4, 0/4\}$, $\{0/20, 0/20\}$, and $\{0/50, 0/50\}$, and a set for which $\Delta P = 0.5$, $\{2/4, 0/4\}$, $\{10/20, 0/20\}$,

and $\{25/50, 0/50\}$. ΔP, pCI, and causal power are constant for these sets of stimuli, and any ordering should thus be equally likely. χ^2 predicts an increase in judgments with sample size for the $\Delta P = 0.5$ set, but is constant for the $\Delta P = 0$ set. Causal support predicts sample size should result in an increase in judgments with $\Delta P = 0.5$, and a decrease with $\Delta P = 0$, as shown in Figure 4.4. The experiment also included three distractors, to conceal our manipulation: $\{3/4, 1/4\}$, $\{12/20, 8/20\}$, and $\{50/50, 0/50\}$.

Procedure. Participants read a description of an imaginary laboratory scenario, similar to that used in Experiment 1, and were shown nine cards that expressed the stimulus information described above. They were given the following instructions:

> Each of the cards in front of you summarizes the results of a different study. Look these summaries over carefully, and then place them in order from the study from which it seems LEAST LIKELY that the chemical causes the gene to be expressed, to the study in which it seems MOST LIKELY that the chemical causes the gene to be expressed.

The wording of the question in terms of likelihood followed the procedure reported by White (2003b). If participants asked whether cards could be ranked equally, they were told that they could order them randomly.

Results and Discussion

Analysis of the orderings produced by the participants showed that 17 out of 20 ordered the stimuli with $\Delta P = 0.5$ by increasing sample size (binomial test, $p < 0.001$), while 16 out of 20 ordered the stimuli with $\Delta P = 0$ by decreasing sample size (binomial test, $p < 0.001$). We computed rank-order correlations with the responses of individual participants for each of the five models. The rank-order correlations with the four models were computed for each participant, averaging these correlations to result in scores for each model.[4] Causal support and χ^2 performed equivalently, $\rho = 0.948$ and $\rho = 0.945$ respectively, followed by ΔP, $\rho = 0.905$, and causal power, $\rho = 0.859$. Causal support gave the highest correlation with the responses of eleven participants, causal power and χ^2 with four participants each, and ΔP with only one participant.

[4]Correlations were averaged using the Fisher z transformation. These results include only 19 of the 20 participants, since causal support perfectly predicted the ordering given by one participant, resulting in an infinite z score. The reported mean correlation is thus an underestimate for causal support. While the other models do not predict an ordering for the two critical sets, they do predict an ordering among the full set of nine stimuli, hence $\rho > 0$.

The results indicate that people are sensitive to sample size when making causal judgments. Specifically, increasing sample size increases judgments when effects are large, but decreases judgments for zero effects. Only causal support can explain this pattern of results. Sensitivity to sample size is a property of structure learning, not parameter estimation, and thus provides evidence that people approach problems of causal induction as structure learning problems.

4.7 Inferences from incomplete contingency tables

One theme of this thesis is the claim that everyday causal induction has several commonalities with the reasoning of early scientists. Among these commonalities is the need to make inferences from limited data. In many settings where people infer causal relationships, they do not have all of the information that is typically provided in causal induction tasks. Specifically, without a carefully designed experiment, we often do not know the frequency of the effect in the absence of the cause, leaving some of the cells in a contingency table empty. While the epigraph to this chapter indicates the attention that James Currie paid to the number of patients who recovered both with and without treatment, such reporting was the exception rather than the rule prior to the development of modern experimentation. Many early medical texts, such as Edward Jenner's (1798) famous treatise on the smallpox vaccine, consist of a description of a number of cases in which the treatment proved successful, providing only $N(e^+, c^+)$. In order to make an inference from such data, his readers had to use their expectations about the frequency of infections in the absence of treatment.

ΔP and causal power are both undefined when there are no trials on which the cause was absent, since $P(e^+|c^-)$ cannot be computed. This is a problem, as people readily make causal judgments under such circumstances. For example, suppose that a doctor claims to have invented a treatment that will cure a rare illness, Hopkins-Francis syndrome. He tells you that he has given this treatment to one patient with Hopkins-Francis syndrome, and after one month, all the patient's symptoms are gone. How much evidence does this provide for the treatment's effectiveness? It may provide some evidence, but not strong evidence, since we do not know how many patients would recover spontaneously in this interval.

A few months later, the doctor tells you that he has now given the treatment to three patients, and after one month all of their symptoms are gone. These data provide stronger evidence, but not that much stronger. The evidence is strengthened once more when, a

few months later, the doctor tells you that he has given the treatment to twenty patients, and after one month all of their symptoms are gone. Finally, the doctor tells you that he has also seen twenty patients over the same time period who received a placebo instead of the new treatment, and all people in this group still had symptoms after a month of observation. Moreover, the people who received the treatment or the placebo were chosen at random. Now this provides very strong evidence for the treatment's effectiveness.

We can identify five stages in the accumulation of evidence in this example. The first stage is the baseline, with no information, and the contingencies $\{0/0, 0/0\}$. After a single observation, we have $\{1/1, 0/0\}$. Two more observations provide $\{3/3, 0/0\}$, and 17 more successful treatments give $\{20/20, 0/0\}$. Finally, the control condition provides $\{20/20, 0/20\}$. ΔP and causal power can only be computed for this last case, where they both indicate strong evidence for a causal relationship, ΔP = power = 1.00. They are undefined for the other contingency tables, and thus cannot capture the weak but growing evidence these tables provide. Causal support is 0 for $\{0/0, 0/0\}$, reflecting the lack of evidence for or against a causal relationship (negative values of causal support indicate evidence for Graph 0, while positive values indicate evidence for Graph 1). Causal support then gradually increases as the observations accumulate, taking values of 0.41, 0.73, and 1.29, before jumping dramatically to 23.32 for when the control condition is added. Unlike ΔP or causal power, causal support thus predicts our intuitive ordering of the strength of evidence provided by these five stimuli. In Experiment 4.4, I examined whether this ordering matched the judgments of naive participants.

4.7.1 Experiment 4.4: Incomplete contingency tables

Method

Participants. Participants were 20 members of the MIT community who took part in the experiment in exchange for candy.

Stimuli. The five stimuli described above were used: $\{0/0, 0/0\}$, $\{1/1, 0/0\}$, $\{3/3, 0/0\}$, $\{20/20, 0/0\}$, and $\{20/20, 0/20\}$.

Procedure. The procedure was identical to that of Experiment 2, with the stimuli being presented on cards and participants being asked to provide an ordering.

Results and Discussion

Analysis of the orderings produced by the participants showed that 15 out of 20 perfectly reproduced the ordering predicted by causal support (binomial test, $p < 0.001$). The other five participants still showed some conformity to the predictions of causal support, with a mean correlation of $\rho = 0.51$. ΔP and causal power are undefined for all but one of these stimuli, preventing the computation of any correlations.

Causal support is the only model we have considered that is capable of capturing people's inferences from incomplete contingency tables. The ability to infer causal relationships from limited data is an extremely important part of causal induction in both everyday and scientific settings. Even today, many medical treatments are initially tested with small samples and incomplete contingency tables. Many doctors say they do not believe in a drug's effectiveness until it passes large-scale studies with appropriate controls, in part because standard statistical practice does not provide a rigorous way to evaluate treatment effectiveness with such limited data. However, the researchers who are actually coming up with and testing new treatments need to have some way of evaluating which treatments are promising and which are not, or they would never make any progress. Causal support provides an account of the rational basis of these intuitions.

4.8 Summary

Learning a single causal relationship from contingency data is arguably the most basic form of causal learning, and is certainly that which has inspired the most psychological research. The results discussed in this chapter suggest that causal theories play a subtle but important role in guiding causal induction from contingency data. Models developed using the theory-based framework outperform leading psychological models – ΔP and causal power – as well as models based upon standard algorithms for causal learning developed in computer science and statistics. The success of these models can be traced to two factors: the use of a causal theory that postulates the correct functional form, and the formulation of causal induction as structure learning. Varying the context of causal learning alters the functional form that people assume, resulting in variations in behavior that can be explained within this framework. Construing causal induction as a problem of structure learning leads to an explanation for several phenomena that are problematic for other models, including non-monotonic effects of $P(e^+|c^-)$ on human judgments, effects of sample

size, and inferences from incomplete contingency tables. This approach explains not only lay people's fundamental intuitions about cause and effect, but also the intuitions that drove discovery for early scientists, such as Dr. James Currie of the epigraph, and that continue to be important in the early stages of much contemporary scientific research.

Chapter 5

Discrete physical systems

Learning a single relationship from contingency data is the simplest case of causal induction. However, it does not provide the opportunity to demonstrate all of the contributions of the theory-based approach, as the effects of prior knowledge are relatively weak in most settings where such learning is performed. By studying inferences about other kinds of systems, we can see how this knowledge varies, and how it can provide strong constraints on causal induction. In this chapter, I will focus on physical systems that operate in discrete time. People have strong expectations about causality in physical systems, making it possible to draw inferences from small samples, and to identify complex hidden structure.

The human ability to infer causal relationships from small samples is at odds with both covariation-based accounts of causal induction, and standard algorithms for causal learning. Hume (1748) emphasized the importance of large samples in inferring causal relationships, stating that "Even after one instance or experiment, where we have observed a particular event to follow upon another, we are not entitled to form a general rule, or foretell what will happen in like cases; it being justly esteemed an unpardonable temerity to judge the whole course of nature from one single experiment, however accurate or certain" (p. 50). For Hume, causal induction required "many uniform instances" (1748, p. 52). Similarly, the statistical tests that scientists use to evaluate causal claims, and which are at the heart of constraint-based algorithms, require large samples to achieve significance.

Inferring hidden causal structure presents a similar challenge for computational models of causal induction. Everyday reasoning draws on notions that go far beyond the observable world, just as modern science draws upon theoretical constructs beyond the limits of measurement. The richness of our intuitive theories is a direct result of our ability to identify

hidden causal structure. The central role of hidden causes in intuitive theories makes the question of how people infer hidden causal structure fundamental to understanding human reasoning. Psychological research has shown that people can infer the existence of hidden causes from otherwise unexplained events (Ahn & Luhmann, 2003), and determine hidden causal structure from very little data (Kushnir et al., 2003). However, most existing algorithms for identifying hidden structure require strong evidence – such as a pattern of dependencies among variables that cannot be explained by any set of causal relationships among those variables – and large samples.

In the remainder of the chapter, I will discuss the inferences of children and adults about two kinds of physical systems: detectors and machines. In each case, I will use the theory-based account to explain how it is possible for people to learn so much from so little, and argue that this cannot be explained by other models.

5.1 Detectors

Gopnik and Sobel (2000) introduced a novel paradigm for investigating causal inference in children, in which participants are shown a number of blocks, along with a machine – the "blicket detector." The blicket detector "activates" – lights up and makes noise – whenever a "blicket" is placed on it. Some of the blocks are blickets, others are not, but their outward appearance is no guide. Participants observe a series of trials, on each of which one or more blocks are placed on the detector and the detector activates or not. They are then asked which blocks have the power to activate the machine.

Gopnik and Sobel have demonstrated various conditions under which children successfully infer the causal status of blocks from just one or a few observations (Gopnik et al., 2001; Sobel et al., in press). Two experiments of this kind are summarized in Table 5.1. In these experiments, children saw two blocks, `a` and `b`, placed on the detector either together or separately across a series of trials. On each trial the blicket detector either became active or remained silent. I will encode the placement of `a` and `b` on the detector with variables A and B respectively, and the response of the detector with the variable E. After seeing a series of trials, children were asked whether each object was a blicket. Table 5.1 gives the proportion of 4-year-olds who identified `a` and `b` as blickets after several different sequences of trials.

Table 5.1: Probability of Identifying Blocks as Blickets for 4-year-old Children

Condition	Stimuli	a	b
one cause	$e^+\|a^+b^-$ $e^-\|a^-b^+$ $2e^+\|a^+b^+$	0.91	0.16
two cause	$3e^+\|a^+b^-$ $2e^+\|a^-b^+$ $e^-\|a^-b^+$	0.97	0.78
indirect screening-off	$2e^+\|a^+b^+$ $e^-\|a^+b^-$	0.00	1.00
backwards blocking	$2e^+\|a^+b^+$ $e^+\|a^+b^-$	1.00	0.34
association	$e^+\|a^+b^-$ $2e^+\|a^-b^+$	0.94	1.00

Note: The *one cause* and *two cause* conditions are from Gopnik, Sobel, Schulz, and Glymour (2001, Experiment 1), *indirect screening-off*, *backwards blocking*, and *association* conditions are from Sobel, Tenenbaum, and Gopnik (in press, Experiment 2).

5.1.1 Theory-based causal induction

The inferences that both adults and children draw about blickets and blicket detectors will be explained with reference to a simple causal theory. Such a theory should reflect people's intuitive expectations about how machines (and detectors) work, and be informed by the instructions provided in the experiment. In the experiments I will discuss (Gopnik et al., 2001, Experiment 1; Sobel et al., in press, Experiment 2), children were introduced to the blicket detector by being told that it was a "blicket machine," and that "blickets make the machine go," and saw blocks that activated the machine being identified as blickets, and blocks that did not activate the machine being identified as non-blickets. A theory expressing this information is sketched in Figure 5.1.

As discussed in previous chapters, this theory has three parts: an ontology, a set of plausible relations, and the functional form of those relations. The ontology identifies the kinds of entities in our domain, which are divided into two types: `Block` and `Detector`. The number of entities of each type, N_B and N_D are specified by distributions P_B and P_D. This ontology is hierarchical, with `Block` being divided into `Blicket` and `NonBlicket`. The probability that a `Block` is a `Blicket` is set by a parameter p. The ontology also identifies

Ontology:

Types	Number
`Block`	$N_B \sim P_B$
`Blicket`	p
`NonBlicket`	$1-p$
`Detector`	$N_D \sim P_D$
`Trial`	$N_T \sim P_T$

Predicates	Values	
`Contact(Block,Detector,Trial)`	`Boolean:`	{`T`, `F`}
`Active(Detector,Trial)`	`Boolean:`	{`T`, `F`}

Plausible relations:
`Contact(B,D,T)` $\rightarrow$ `Active(D,T)`
Relation holds over all `T` for any `D` if `B` is a `Blicket`

Functional form:

`Contact(B,D,T)` $\sim$ Bernoulli($\cdot$)
`Active(D,T)` $\sim$ Bernoulli(ν) for ν from a noisy-OR:

Cause	Strength
(Background)	$w_0 = 0$
`Contact(B,D,T)`	$w_i = 1$

Figure 5.1: Theory for causal induction with deterministic blicket detectors.

a set of predicates that apply to these entities: `Contact(b,d,t)` indicates that `Block b` is on `Detector d` in `Trial t`, and `Active(d,t)` indicates that `Detector d` is active on `Trial t`.

The plausible relations state that only blickets can cause detectors to activate, and every blicket can activate every detector. The functional form gives the probabilities of different kinds of events, stating how causal relationships influence these probabilities. The theory indicates that contact between a block and a detector is a relatively rare event. While the specific probabilities given here will not have any impact on our analysis, they could be used to make predictions about other kinds of experiments. The critical piece of information supplied by the functional form concerns how activation of a detector is affected by its causes. The theory indicates that the probability of activation follows a noisy-OR distribution (Equation 2.1). w_i is the "causal power" of blicket i (c.f. Cheng, 1997) – the probability that blicket i will cause the detector to activate. w_0 represents the probability that the detector will activate without any blickets being placed upon it. The causal theory shown in Figure 5.1 makes two important assumptions through the setting of the w_i parameters. First, the detector cannot activate unless a blicket is in contact with it ($w_i = 0$). Second, the probability with which a blicket will activate the detector, is

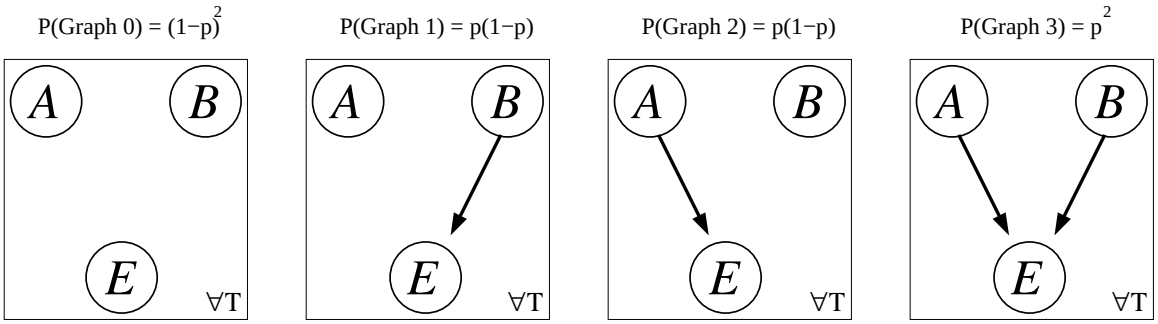

Figure 5.2: Causal structures generated by the theory for the blicket detector with two blocks, `a` and `b`, and one detector, `d`. A and B indicate the truth value of `Contact(a,d,T)` and `Contact(b,d,T)` for `Block a` and `b` and `Detector d`, while E indicates the truth value of `Active(d,T)`. The plates indicate that these causal relationships hold for all trials `T`.

$w_i = 1.00$. These two assumptions make this the *deterministic detector* theory, embodying a simple "activation law" (Sobel et al., in press): only blickets can activate the blicket detector, and they always do so.

In tasks involving the blicket detector, participants usually know the number of blocks, N_B, the number of detectors, N_D, and the number of trials, N_T. However, they usually do not know which blocks are blickets. The question of whether a block is a blicket comes down to whether that block causes the blicket detector to activate, since only blickets can cause activation of the detector. This question can be addressed via a Bayesian inference over causal networks: the posterior probability that a block is a blicket is simply the posterior probability that there is a causal relationship between placing that block on a detector and the activation of the detector.

The deterministic detector theory generates a hypothesis space, $\mathcal{H}$ of causal networks for any events involving blocks and detectors. Assuming that we know that we have two blocks, `a`, and `b`, and a single detector, `d`, $\mathcal{H}$ would consist of four graph structures, as shown in Figure 5.2. I will use the variables A and B to indicate `Contact(a,d,T)` and `Contact(b,d,T)` respectively, and E to indicate `Active(d,T)`, all for the same trial, indicated by the logical variable `T`. The prior probabilities of these models, $P(\text{Graph } i)$, are set by the causal theory. The likelihood of a set of trials under these models can be evaluated using the probabilities given by the noisy-OR. The posterior probability distribution over this set of causal models can be evaluated for each set of trials shown in Table 5.1, denoting the set of trials $\mathcal{D}$ and applying Bayes' rule. The probability that a particular block is a

blicket can be evaluated by summing the posterior probability of the models in which such a causal relationship exists. For example, to evaluate the probability that A causes E, we would add $P(\text{Graph } 2|\mathcal{D})$ and $P(\text{Graph } 3|\mathcal{D})$.

The predictions of this account are given in Table 5.2. These predictions provide a strong qualitative correspondence with the judgments of the children in the experiments. The most interesting case is that of *backwards blocking*, where the Bayesian model predicts that the probability that a causal relationship exists between B and E after the series of trials is p, the prior probability that such a relationship exists. The analysis of this experiment is as follows. After the $e^+|a^+b^+$ trials (which I will denote $\mathcal{D}_1$), at least one block must be a blicket: the consistent hypotheses are Graphs 1, 2, and 3. After the $e^+|a^+b^-$ trial (with the accumulated data being denoted $\mathcal{D}_2$), only Graphs 2 and 3 remain consistent. The nonzero posterior probabilities are then given as follows (all others are zero): $P(\text{Graph } 2|\mathcal{D}_1) = P(\text{Graph } 1|\mathcal{D}_1) = \frac{p(1-p)}{p^2+2p(1-p)}$, $P(\text{Graph } 3|\mathcal{D}_1) = \frac{p^2}{p^2+2p(1-p)}$, $P(\text{Graph } 2|\mathcal{D}_2) = \frac{p(1-p)}{p^2+p(1-p)} = 1-p$, and $P(\text{Graph } 3|\mathcal{D}_2) = \frac{p^2}{p^2+p(1-p)} = p$. Consequently, the probability that A causes E is 1, and that B causes E is p.

However, this deterministic detector theory cannot explain all of the inferences that children make about blickets. In particular, it cannot explain the *two cause* condition in Experiment 1 of Gopnik et al. (2004). This condition was used as a control for the *one cause* condition, demonstrating that children drew quite different inferences about the causal relationships among a set of objects when the same associative relations (the frequency with which cause and effect co-occurred) were maintained, but the structure of the trials manifesting those relations was modified. This control experiment involved showing children a block (b) which activated the detector on two out of three occasions. Such an event cannot be explained by our deterministic theory, under which a block either causes a detector to activate all the time, or never. A set of trials in which a block activates a detector on two out of three occasions has zero probability under all causal models. Consequently, the posterior distribution is undefined for this case, indicated by the question marks in Table 5.2.

The problem raised by the *two cause* condition can be addressed by relaxing one of the assumptions of the deterministic detector theory. If we allow blickets to activate detectors only some of the time, then inconsistent patterns of activation like that exhibited by block b are possible. We can make this change by altering

Table 5.2: Predictions of Probabilistic Theory and Alternative Models

	Deterministic		Probabilistic		Generic		Noisy-OR	
Condition	a	b	a	b	a	b	a	b
one cause	1.00	0.00	0.99	0.07	0.54	0.27	0.65	0.25
two cause	?	?	1.00	0.81	0.32	0.32	0.52	0.28
indirect screening-off	0.00	1.00	0.13	0.90	0.33	0.50	0.29	0.56
backwards blocking	1.00	p	0.93	0.41	0.33	0.25	0.49	0.40
association	1.00	1.00	0.82	0.98	0.27	0.27	0.38	0.43

Cause	Strength
(Background)	$w_0 = 0$
$\texttt{Contact}(\texttt{B}, \texttt{D}, \texttt{T})$	$w_i = 1$

to reflect the possibility of errors

Cause	Strength
(Background)	$w_0 = \epsilon$
$\texttt{Contact}(\texttt{B}, \texttt{D}, \texttt{T})$	$w_i = 1 - \epsilon$

where ϵ is a parameter of the theory reflecting the error rate of the detector. This *probabilistic detector* theory gives the same predictions as the deterministic detector theory in the limit as $\epsilon \to 0$, but also predicts that both a and b are blickets with probability 1.00 in the *two cause* condition. Different values of ϵ give different predictions. The predictions of this theory with $\epsilon = 0.1$ and $p = 1/3$ are shown in Table 5.2. This model captures some of the finer details of children's judgments that are not captured by the deterministic detector, such as the fact that b is judged less likely to be a blicket than a in the *two cause* condition.

5.1.2 Alternative accounts

As before, considering alternative accounts of these data provides insight into the assumptions that allow the theory-based approach to succeed. I will compare this account with three alternatives: constraint-based algorithms, Bayesian structure learning with the generic parameterization, and a Bayesian model in which the noisy-OR parameters w_0 and w_1 can take any value between 0 and 1, as was used in the account of causal induction from contingency data in Chapter 4. All of these alternatives approach human inferences as a decision between causal structures, but differ from the theory-based account in their assumptions about the functional form of the relationship between cause and effect. Examining these

alternative accounts reveals that the strong expectations about functional form embodied in our deterministic and probabilistic theories are necessary to explain how children can infer causal relationships with high certainty from small samples.

Constraint-based algorithms

Gopnik et al. (2004) argue that children's inferences about blicket detectors can be explained by standard constraint-based algorithms for learning causal graphical models. They point out that, given appropriate information about the dependencies between the variables A, B, and E, these algorithms will infer the appropriate causal structure – for example, that `a` is a blicket in the *one cause* condition. However, there are two significant problems with this account: inferring the dependencies, and using probabilistic prior knowledge.

The first step in applying a constraint-based algorithm is to identify the statistical dependencies that hold among a set of variables. Typically, this is done using frequentist tests such as the χ^2 test for independence. These tests impose no constraints on the functional form of the relationships between variables, and deciding that two variables are dependent requires imposing some criterion of statistical significance on the results of the tests. This raises a problem: the inferences that children make in these experiments are the result of only a handful of observations – far fewer than would be required to produce statistically significant results. Gopnik et al. (2004) address this issue by suggesting that 'the sample size is given a large fictitious multiplier' (p. 21). Introducing such fictitious multipliers is hazardous – if they are applied indiscriminately, we should expect human causal induction to result in a great many false alarms. Such a suggestion seems inconsistent with the accuracy exhibited by human inferences in both this and the other settings I consider here.

Allowing the sample size to be multiplied by some fictitious amount is a post-hoc solution to a fundamental problem raised by using standard statistical tests to evaluate all causal relationships. Under our account, the reason why small samples are so compelling in the case of the blicket detector is that children have strong expectations about the functional form of the relationship between placing blickets on the detector and the detector activating – namely that "blickets make the machine go," and that the machine does not go in the absence of blickets. The χ^2 test makes far weaker assumptions about functional form, and thus requires more information to identify a relationship. Fictitious multipliers thus act as a proxy for the prior knowledge that children are exploiting when making their inferences.

A second problem with explaining these results using constraint-based algorithms is that

these algorithms cannot exploit probabilistic prior knowledge. While particular structures can be ruled out on the basis of prior knowledge, it is difficult to see how the knowledge that the probability that a block is a blicket is p can be used by these algorithms. The theory-based account predicts that such knowledge should be useful in the *backwards blocking* condition, where the probability that `b` is a blicket (under the deterministic theory) is p. In the next section, I will describe some experiments that reveal the importance of these priors in people's inferences. Furthermore, by reasoning deductively from a pattern of dependencies, constraint-based algorithms cannot maintain degrees of uncertainty: a causal structure is either consistent or inconsistent with the data. Both the *backwards blocking* conditions and the experiment with ambiguous evidence discussed above illustrate that people exhibit graded degrees of belief in the existence of a causal relationship.

Bayesian structure learning with the generic parameterization

Standard Bayesian structure-learning algorithms assume that the relationship between A, B, and E can be expressed using the generic parameterization, using a separate parameter to define the probability of E for each combination of values of A and B. This approach faces the same problem as constraint-based algorithms in attempting to explain children's inferences: it makes weak assumptions about functional form, and consequently requires large samples to identify the existence of a causal relationship. This is illustrated in Table 5.2, which shows the predictions obtained by applying Bayesian structure learning using the generic parameterization (with a uniform distribution over parameters) to the stimuli given in Table 5.1. The assumptions used to generate the predictions are exactly those of the theory in Figure 5.1 (with $p = 1/3$), except for the functional form. The predictions often deviate from human judgments – for example, in the *two causes* condition, `a` and `b` have a probability of being blickets that is scarcely different from the prior, because there is no strong evidence that contact between `a` and `b` and the blicket detector affects the probability with which the detector activates. In the cases where predictions move in the same direction as children's inferences, the probability that any block will be identified as a blicket remains close to the prior probability in all cases. Both of these issues are consequences of using the generic parameterization: if identifying a causal relationship requires determining that two variables are dependent, small samples can only produce small changes in beliefs about causal relationships.

Relaxing the noisy-OR parameters

As a final comparison, we can try to explain these inferences by applying the theory developed for chemicals and genes in the previous section (Figure 3.2) to blicket detectors. This theory assumes that the functional form is a noisy-OR, but that the noisy-OR parameters (including the baseline probability w_0) are drawn from a uniform distribution on $[0, 1]$. Thus, like the deterministic and probabilistic detector theories outlined above, blickets can only increase the probability that the detector activates, but unlike those theories, blickets can vary dramatically in their strengths, and the detector can activate in the absence of any blickets. These are also the assumptions Cheng (1997) makes in deriving causal power. The predictions under this "Noisy-OR" theory are shown in Table 5.2. The assumption of generativity is not sufficient to explain children's inferences: just assuming a noisy-OR does not place sufficiently strong constraints on the functional form of the relationship between blickets and the activation of blicket detectors. The model gives predictions that are slightly more consistent with human judgments than the generic parameterization, but small samples still do not produce dramatic changes in the extent to which a block is believed to be a blicket.

5.1.3 Priors and ambiguous evidence

The theory-based account explains how children can infer causal relationships from small amounts of data, positing strong constraints on the relationships considered plausible and the functional form of those relationships. It also makes two further predictions about human performance on these tasks which discriminate it from alternative accounts such as constraint-based algorithms. The first prediction is that prior beliefs, in the form of expectations about the probability that a block is a blicket, the parameter p in the theory, should influence people's judgments. Specifically, in the *backwards blocking* condition, the posterior probability that `b` is a blicket is just p. The second prediction is that people should be able to maintain graded degrees of belief in the face of ambiguous evidence. Again, the *backwards blocking* experiment provides one example of this, but the theory-based account predicts that people should be able to infer that a block is a blicket despite never obtaining definitive evidence, such as seeing it light up the detector all on its own. A series of studies have been conducted that test these predictions with both adults and children (Sobel et al., in press; Tenenbaum, Sobel, & Gopnik, submitted). I will summarize the results of these

experiments, and show how they can be explained by the theory-based account. I will focus on the adult experiments, which provide data about the beliefs of the participants after each trial, but the same qualitative effects hold with 4-year-old children.

Tenenbaum et al. (submitted, Experiment 1) explored the extent to which people's judgments were affected by prior probabilities by conducting an analogue of the *backwards blocking* condition of Sobel et al. (in press, Experiment 1), but varying the probability of a causal relationship. The experiment was done with adults, using a "super-pencil" detector which functioned exactly like a blicket detector, but identified whether golf pencils contained a special kind of lead. Participants were randomly assigned to two conditions, which determined how they were introduced to the detector. In both conditions, participants saw 12 pencils placed on the detector, one after the other. In the *rare* condition, only two of these pencils caused the detector to activate. In the *common* condition, the detector activated for 10 of the 12 pencils. The two conditions were designed to establish different prior beliefs about the probability that a causal relationship existed.

The experiment had three stages. First, participants were shown two new pencils, `a` and `b`, and asked to rate the probability that they were super-pencils. They then saw `a` and `b` placed on the detector together, and the detector activating, and were again asked to rate the probability that they were super-pencils. Finally, just `a` was placed on the detector, and the detector activated. Once again, participants rated the probability that `a` and `b` were super-pencils. The mean ratings in the *rare* and *common* conditions are shown in Figure 5.3 (a) and (b) respectively. Manipulating the frequency with which pencils were identified as super-pencils had the expected effect on people's baseline judgments, indicating a difference in prior beliefs. It also affected the judgments that people made after each trial. As shown in the figure, the pattern of judgments is perfectly predicted by Bayesian inference guided by the deterministic detector theory (or the probabilistic detector theory with $\omega = 1 - \epsilon$ as $\epsilon \rightarrow 0$): the probability of `a` and `b` being super-pencils should increase after the first trial, and then the second trial should provide unequivocal evidence that `a` is a super-pencil while the probability that `b` is a super-pencil should return to the prior p. The model predictions shown in the figure were obtained by setting p to the baseline probability given by the participants. Similar results were obtained with 4-year-old children by Sobel et al. (in press, Experiment 3).

This experiment illustrates that people's causal inferences are affected by their prior beliefs in exactly the way the theory-based account predicts. Tenenbaum et al. (submitted,

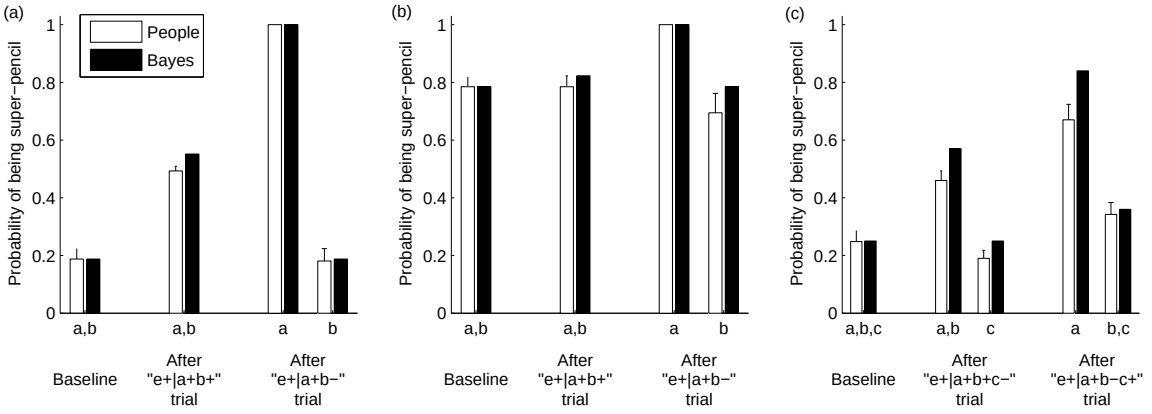

Figure 5.3: Adult judgments with "super-pencils," an analogue of the blicket detector task, from Tenenbaum, Sobel, & Gopnik (submitted). (a) and (b) show inferences from the same set of trials, but with different prior probabilities for super-pencils, being rare and common respectively. (c) Inferences from ambiguous evidence.

Experiment 2) also showed that people could make inferences from ambiguous evidence in a fashion consistent with a theory-based Bayesian inference. This experiment was also conducted with super-pencils, and people saw the detector activated by 2 out of 12 pencils before beginning the critical trials. They were shown three new pencils, `a`, `b`, and `c`, and were asked to rate the probability that these pencils were super-pencils. They then saw `a` and `b` placed on the detector together, causing the detector to activate, and gave new ratings. Finally, they saw `a` and `c` placed on the detector together, causing the detector to activate, and were asked to rate the probability that each of the pencils was a super-pencil. The mean ratings are shown in Figure 5.3 (c).

In this experiment, people received no unambiguous clues that a particular pencil was a super-pencil: there were no trials on which a single pencil caused the detector to activate. Nonetheless, people were able to infer that `a` was quite likely to be a super-pencil, while `b` and `c` were less likely to be super-pencils, but more likely than they had been at the start of the experiment. Similar results were obtained with 4-year-old children (Tenenbaum et al., submitted, Experiment 3). The hypothesis space generated by the deterministic detector theory with three blocks and one detector is shown in Figure 5.4. The predictions of a Bayesian model using this theory are shown in Figure 5.3 (c), setting p to the baseline probability given by the participants. The hypothesis space $\mathcal{H}$ generated by the theory consists of eight causal graphical models, indicating all possible sets of causal relationships

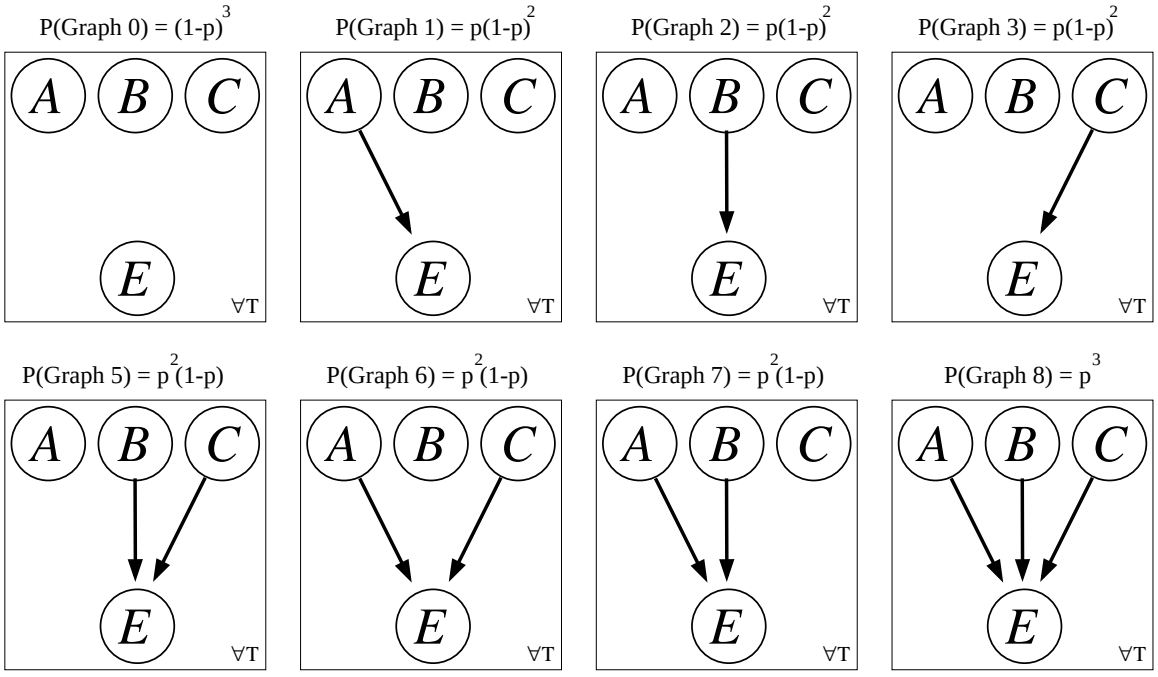

Figure 5.4: Causal structures generated by the theory for the blicket detector with three blocks, `a`, `b`, and `c`, and one detector, `d`. A and B and C indicate whether contact between the appropriate block and the detector occurred on a particular trial, while E indicates whether the detector activated. The plates indicate that these causal relationships hold for all trials `T`.

between the three blocks and the detector. Using this hypothesis space, the model predicts four quantitatively different levels of belief for different pencils at different points in the experiment: the baseline probability, the probability that `a` and `b` are super-pencils after the first trial, the probability that `a` is a super-pencil after the second trial, and the probability that `b` and `c` are super-pencils after the second trial. People also show these four levels of graded belief in the existence of a causal relationship.

5.1.4 Learning the right theory

I have outlined two different theories for the blicket detector – the deterministic detector theory, and the probabilistic detector theory. In some cases, such as the *one cause* and *two causes* conditions, it seems that the probabilistic detector theory provides a better characterization of children's inferences. However, the instructions the children received suggested that the deterministic theory might be more appropriate. This raises an important

question: how might a learner choose between different theories? This question returns to one of the most interesting aspects of the tale of Halley's comet: that the return of the comet provided an indication of the validity of Newton's theory, the theory which had made it possible for Halley to recognize the causal structure responsible for his observations.

The process of selecting between theories on the basis of evidence can be modeled naturally within the theory-based causal induction framework. Indeed, this is one of the great strengths of the framework: it begins to indicate how intuitive theories might be learned from data. In the case of the blicket detector, the problem is quite constrained, being a matter of choosing between the deterministic and the probabilistic theory. This decision can be made by using Bayes' rule, treating theories T as hypotheses

$$P(T|\mathcal{D}) = \frac{P(\mathcal{D}|T)P(T)}{P(\mathcal{D})}, \tag{5.1}$$

where $P(\mathcal{D}) = \sum_T P(\mathcal{D}|T)P(T)$. The critical probabilities in this expression are of the form $P(\mathcal{D}|T)$, being the probability of a dataset $\mathcal{D}$ under a theory T. These probabilities can be computed by summing over all causal graphical models generated by T, being the members of the hypothesis space $\mathcal{H}_T$. Thus we have

$$P(\mathcal{D}|T) = \sum_{i=1}^{|\mathcal{H}_T|} P(\mathcal{D}|\text{Graph } i)P(\text{Graph } i|T)$$

which can be computed using just probabilities defined above: the probability of the data under a particular causal graphical model, and the prior probability of such a model under the theory. Other probabilities, such as the probability of a particular causal structure, or that an object is a blicket, can be evaluated by summing over theories T.

Figure 5.5 illustrates how this process of selecting a causal theory can operate concurrently with inferring the causal properties of the entities in a domain. The figure shows the posterior distribution over the two theories – deterministic and probabilistic – and the probability that blocks `a` and `b` are blickets as data $\mathcal{D}$ accumulates. In this case, the data are the trials used in the *two cause* condition. The prior gives a probability of 0.99 to the deterministic theory, and 0.01 to the probabilistic theory, p is set to 0.3, and ϵ is set to 0.1. The first three datapoints are all $e^+|a^+b^-$, being events in which `a` is placed on the detector and the detector activates. This is sufficient to identify `a` as a blicket under either theory, and weakly favors the deterministic theory. The fourth datapoint is $e^-|a^-b^+$, with

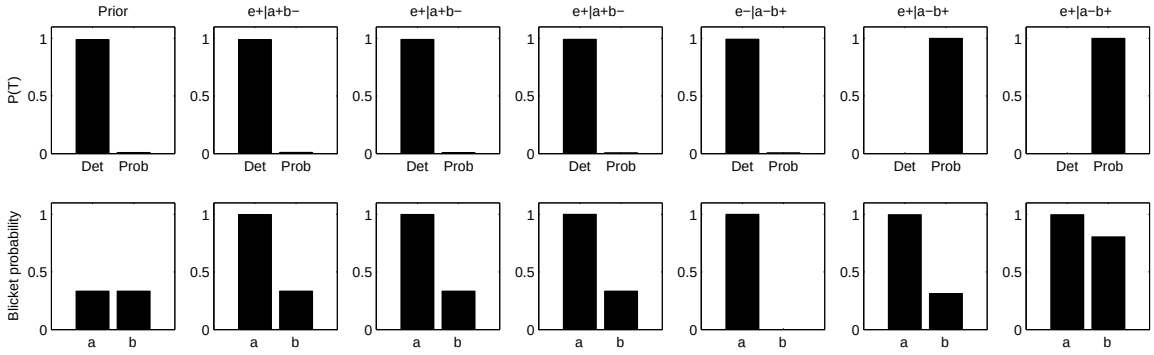

Figure 5.5: Choosing between two theories. The bar graphs along the top of the figure show the probabilities of the two theories, with "Det" indicating the deterministic detector theory, and "Prob" indicating the probabilistic detector theory. The bar graphs along the bottom show the probabilities that the blocks `a` and `b` are blickets. The probabilities after successive trials are shown from left to right.

`b` placed on the detector and the detector not activating. Under the deterministic theory, `b` is definitely not a blicket. Under the probabilistic theory, there remains a small chance that `b` is a blicket, and since the probabilistic theory is still viable, the probability that `b` is a blicket is non-zero. The fifth datapoint is $e^+|a^-b^+$, the activation of the detector when `b` is placed upon it. The fourth and fifth datapoints are mutually contradictory under the deterministic theory, and have a probability of zero. This event can only be explained by the probabilistic theory, in which `b` is definitely a blicket, and consequently the posterior probability of the probabilistic theory and of `b` being a blicket both become 1.00. By the end of the trials in the *two cause* condition, one should be firmly convinced that blicket detectors are probabilistic.

The selection of an appropriate causal theory based upon evidence provides a possible explanation for why children in the *one cause* and *two cause* conditions produced responses consistent with the probabilistic detector theory, while children in the other conditions acted in a fashion more consistent with the deterministic detector theory. Under the deterministic theory, children in the *one cause* condition should never say that `b` is a blicket, but children did so on 16% of trials. Since the *one cause* and *two cause* conditions were presented within-subjects, one possibility is that some of the children saw the *two cause* trials, inferred that the probabilistic detector theory was appropriate, and then used this theory when they subsequently experienced the *one cause* condition. Examination of the data of Gopnik et al. (2001) provides tentative support for this conclusion: *all* of the children who identified `b`

as a blicket in the *one cause* condition had seen the *two cause* condition beforehand.[1] The possibility that children changed their expectations about how blicket detectors operate in the course of the experiment deserves further investigation.

5.2 Machines

The blicket detector illustrates how people make inferences about a specific kind of physical system, which has predictable properties: a detector. Other kinds of machines follow more generic forms of mechanical causation. In this section, I will provide a case study in the learning of hidden causes, examining how people infer the causal structure that underlies the behavior of a a simple machine – the stick-ball machine (Gopnik et al., 2004; Kushnir et al., 2003). First, I will introduce this apparatus, and summarize the results of experiments that used the stick-ball machine to investigate whether children and adults could combine evidence from observations interventions, and if they could infer hidden causes (Gopnik et al., 2004; Kushnir et al., 2003). I will then present a causal theory that can be used to explain these inferences. As with the blicket detector, using this theory makes it possible to identify causal structure from only a small amount of data (observations or interventions). The theory extends my previous analyses by allowing for the possibility of hidden causal structure.

The stick-ball machine, also known as the puppet machine, is a physical system consisting of a number of colored balls mounted on sticks which can move up and down on a box (see Figure 5.6). The mechanical apparatus moving the balls is concealed by the box, keeping the actual causal relationship unknown. The balls can either move on their own, or be moved by the experimenter. Different patterns of observations and interventions lead people to believe in different underlying causal structures. Studying which structures are inferred for different stimuli provides the opportunity to understand how people make such inferences.

Gopnik et al. (2004) described a series of experiments using the stick-ball machine to assess causal induction in children. Table 5.3 summarizes the results of these experiments. In all cases, children were familiarized with the machine, and told that "special" balls caused other balls to move. I will discuss conditions in which children saw two balls, `a` and `b`, move in various patterns, using the variables A and B to indicate the motion of `a` and `b` on a

[1] I thank David Sobel for making these data available.

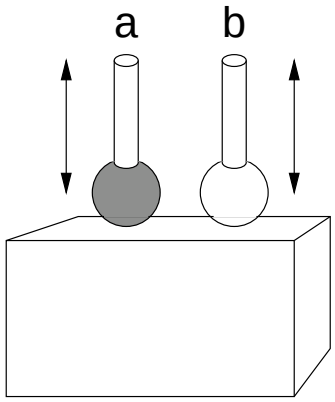

Figure 5.6: A two-ball stick-ball machine (Kushnir et al., 2003).

Table 5.3: Modal Inferences by Children and Bayes for Two-Ball Machines

Expt	Condition	Stimuli	Children	Bayes
1,2,3	*common effects*	$5a^+b^+$, $a^-\|\text{do}(b^+)$	`a` is special	`a` is special
2	*association*	$a^+\|\text{do}(b^+)$, $a^-\|\text{do}(b^+)$	`b` is special	`b` is special
3	*common cause*	$5a^+b^+$, $a^-\|\text{do}(b^+)$, $b^-\|\text{do}(a^+)$	hidden cause	hidden cause

Note: Experiment numbers and conditions refer to Gopnik et al. (2004).

given trial. In the *common effects* condition of Experiments 1-3, children saw `a` and `b` move together several times, $5a^+b^+$, then saw the experimenter intervene to move `b` without `a` moving, $a^-|\text{do}(b^+)$. Most children inferred that `a` was special, and causing `b` to move. In the *association* condition of Experiment 2, it was established that this inference made use of the difference between observations and interventions, with children seeing $5a^+|\text{do}(b^+)$ followed by $a^-|\text{do}(b^+)$. These stimuli differed from the *common effects* stimuli only in the use of intervention on the $a^+|\text{do}(b^+)$ trials, but produced quite different responses, with the majority of children favoring the hypothesis that `b` was special, and causing `a` to move. In the *common cause* condition of Experiment 3, children saw $5a^+b^+$, followed by two interventions: $a^-|\text{do}(b^+)$, and $b^-|\text{do}(a^+)$. They were asked why the balls were moving together, and the majority of the children referred to an unobserved variable as the cause of these events.[2]

[2]One condition (Experiment 1, *common cause*) is not included in the table. This condition examined inferences involving a three-ball machine, being a version of the *common cause* condition of Experiment 3 in which the common cause was observable. For three balls `a`, `b`, and `c`, the stimuli consisted of several $a^+b^+c^+$ trials, followed by $b^-c^-|\text{do}(a^+)$ and $a^-b^-|\text{do}(c^+)$. Children inferred that the motion of both `a` and `c` was caused by the motion of `b`. This inference can be explained by the theory-based account, under the assumption that $\alpha < \omega$, but I will not discuss it in detail here.

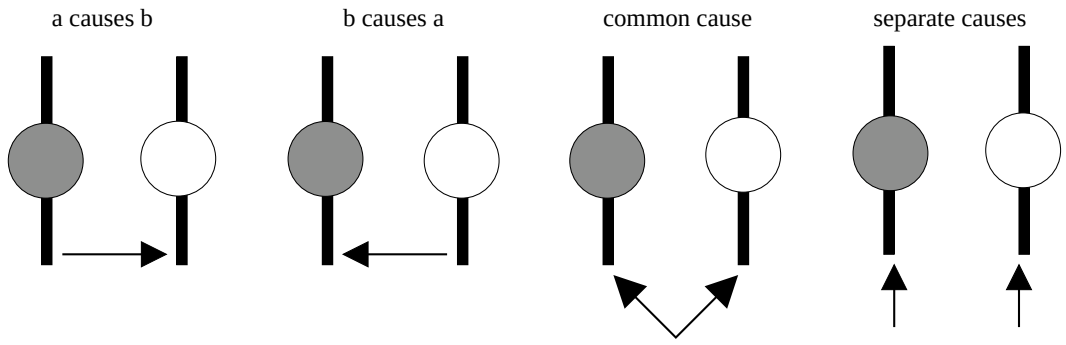

Figure 5.7: Schematic diagrams indicating possible causal structures for the stick-ball machine (after Kushnir, Gopnik, Schulz, & Danks, 2003).

The experiments reported by Gopnik et al. (2004) suggest that children discriminate between observations and interventions when assessing causal relationships, and that they are capable of recognizing the presence of hidden causes. Kushnir et al. (2003) conducted two experiments that extend these results to adults. In both experiments, participants were familiarized with the machine, told that if one ball caused the other to move it did so "almost always," and saw the two balls move together four times. There were three test conditions in Experiment 1, seen by all participants. In the *common unobserved cause* condition, participants saw $4a^+b^+$, then four trials in which the experimenter intervened, twice moving `a` with no effect on `b`, $2b^-|\text{do}(a^+)$, and twice moving `b` with no effect on `a`, $2a^-|\text{do}(b^+)$. In the *independent unobserved cause* condition, participants saw $2a^+b^-$, $2a^-b^+$, a^+b^+, $2a^-|\text{do}(b^+)$, and $2b^-|\text{do}(a^+)$. In the *one observed cause* condition, participants saw $4b^+|\text{do}(a^+)$ and $2b^-|\text{do}(a^+)$. Experiment 2 replicated the *common unobserved cause* condition, and compared this with a *pointing control* condition in which interventions were replaced with observations where the experimenter pointed at the moving ball ($4a^+b^+$, $2a^-b^+$, $2a^+b^-$). On each trial, participants identified the causal structure they thought responsible by indicating images similar to those shown in Figure 5.2. The results of both experiments are shown in Table 5.4. In each condition, the majority of people indicated a single structure – common cause in the *common unobserved cause* condition, separate causes in the *independent unobserved causes* condition, `a` causes `b` in the *one observed cause* condition, and separate causes in the *pointing control*.

Table 5.4: Probability of Choosing Different Causal Structures in Kushnir et al. (2003)

Condition	`a` causes `b`		`b` causes `a`		common cause		separate causes	
common unobserved cause	0.00	(0.12)	0.01	(0.12)	**0.65**	**(0.71)**	0.34	(0.05)
independent unobserved causes	0.00	(0.00)	0.00	(0.00)	0.04	(0.01)	**0.96**	**(0.99)**
one observed cause	**0.65**	**(0.67)**	0.06	(0.00)	0.08	(0.00)	0.21	(0.33)
pointing control	0.00	(0.04)	0.04	(0.04)	0.17	(0.16)	**0.79**	**(0.76)**

Note: Numbers in parentheses are predictions of Bayesian model. Boldface indicates majority.

5.2.1 Theory-based causal induction

Explaining the inferences of children and adults about the stick-ball machine requires addressing three challenges: accounting for the difference between observations and interventions, explaining how it is possible to identify hidden causes, and justifying the fact that so little data is required to identify relatively complex causal structures. These three challenges can be addressed by a theory-based account, using a causal theory like that shown in Figure 5.8.

The theory shown in Figure 5.8 differs from the theories considered in previous sections in incorporating a type of entity that is unobserved – the `HiddenCause`. The number of entities of this type is unbounded, representing the fact that there could be arbitrarily many such hidden causes. This is possible because hidden causes not connected to balls have no influence on the probability with which events involving those balls occur. The way that hidden causes are connected to balls, and to each other, is also unaffected by the fact that the number of such causes is unbounded.

In the case of a physical system like the stick-ball machine, specifying the plausible relations among a set of variables involves identifying the possible physical structures that could be responsible for the motion of the elements of the system. The ontology divides the components of these physical structures into two types: components of the type `HiddenCause` are the "prime movers" in the system, the source of the force that is ultimately responsible for any observed motion, while components of the type `Ball` are passive elements, which can transfer force but not generate it. A graph structure defined on the predicates `Moves` and `Active` applied to these components indicates how force flows through the system. The parameters p and q determine how likely it is that force is able to flow from one ball to another, and from a hidden cause to a ball.

The process by which the hidden cause connected to a ball is selected deserves some further explanation. If it is decided that a ball is connected to a hidden cause, then the

Ontology:

Types	Number
`Ball`	$N_B \sim P_B$
`HiddenCause`	$N_H = \infty$
`Trial`	$N_T \sim P_T$

Predicates	Values	
`Moves(Ball,Trial)`	`Boolean:`	$\{\mathtt{T}, \mathtt{F}\}$
`Active(HiddenCause,Trial)`	`Boolean:`	$\{\mathtt{T}, \mathtt{F}\}$

Plausible relations:

$\mathtt{Moves}(\mathtt{B}_1, \mathtt{T}) \rightarrow \mathtt{Moves}(\mathtt{B}_2, \mathtt{T})$
True for all `T` with probability p for each $\mathtt{B}_1 \neq \mathtt{B}_2$ pair

$\mathtt{Active}(\mathtt{H}, \mathtt{T}) \rightarrow \mathtt{Moves}(\mathtt{B}, \mathtt{T})$
Each `B` has an edge from some `H` with probability q. If such an edge exists, then the particular `H` is chosen based upon the number of existing edges:

$$P(\mathtt{Active}(\mathtt{H}, \mathtt{T}) \rightarrow \mathtt{Moves}(\mathtt{B}, \mathtt{T})) \propto \begin{cases} M_{\mathtt{H},i} & M_{\mathtt{H},i} > 0 \\ s & \mathtt{H} \text{ is new} \end{cases}$$

where $M_{\mathtt{H},i}$ is the number of edges from `H` when the edges are chosen for the ith ball.

Functional form:

$\mathtt{Active}(\mathtt{H}, \mathtt{T}) \sim$ Bernoulli(α)

$\mathtt{Moves}(\mathtt{B}_1, \mathtt{T}) \sim$ Bernoulli(ν) for ν from a noisy-OR:

Cause	Strength
(Background)	$w_0 = 0$
$\mathtt{Moves}(\mathtt{B}_2, \mathtt{T})$	$w_i = \omega$
$\mathtt{Active}(\mathtt{H}, \mathtt{T})$	$w_i = \omega$

Figure 5.8: Theory for causal induction with the stick-ball machine.

particular hidden cause is selected by sampling from a distribution in which each hidden cause `h` that is connected to at least one ball is chosen with probability proportional to the number of other balls to which `h` is connected, and a new hidden cause is chosen with probability proportional to a constant s. This procedure allows balls to share the same hidden causes, or to have independent hidden causes, and does not impose an upper bound on the number of hidden causes that appear in a physical system. The sampling scheme is that of the *Chinese restaurant process* (Aldous, 1985; Pitman, 2002), which is commonly used in non-parametric Bayesian models (e.g., Blei, Griffiths, Jordan, & Tenenbaum, 2004) and is formally equivalent to the system involving a "coupling probability" used in Anderson's (1990) rational model of categorization. The distribution that results from this process is *exchangeable*, meaning that the order in which the hidden causes are chosen does not affect the probability of a particular configuration of connections. When s is small, the scheme favors structures in which many balls have the same hidden cause. When s is large, it is more likely that balls will have independent hidden causes.[3]

The set of all structures defined on two balls that is generated by the theory is shown in Figure 5.9. This set includes all simple causal structures one might identify as possible descriptions of a physical system like the stick-ball machine. Graph 0 is a system in which balls are disconnected from hidden causes and from one another, and thus will never move. Graph 1 is a system in which moving `a` causes `b` to move, but neither `a` nor `b` will move on their own. Graphs 3, 7, 11, 15, and 19 all indicate a bidirectional causal relationship between A and B. Causal graphical models do not usually allow such relationships, being restricted to acyclic directed graphs. I discuss how these relationships are dealt with in Appendix C.

The functional form identified by the causal theory summarizes a set of expectations about the interactions between physical objects. It states that no object moves without a cause, and objects are likely to move when caused to do so. These two principles are simplified versions of Newton's first and second laws of motion: that an object in a uniform state of motion will remain that way unless influenced by an external force, and that the application of force results in acceleration. Such a functional form results in strong constraints upon the kind of data that one might expect to see under different causal structures. For

[3]The theory that I have chosen to use here allows a maximum of one hidden cause per ball. This simplifies the mathematical description of the theory, and results in a smaller hypothesis space. Similar theories can be defined that allow multiple causes per ball, using different non-parametric Bayesian priors (e.g., Griffiths & Ghahramani, in prep).

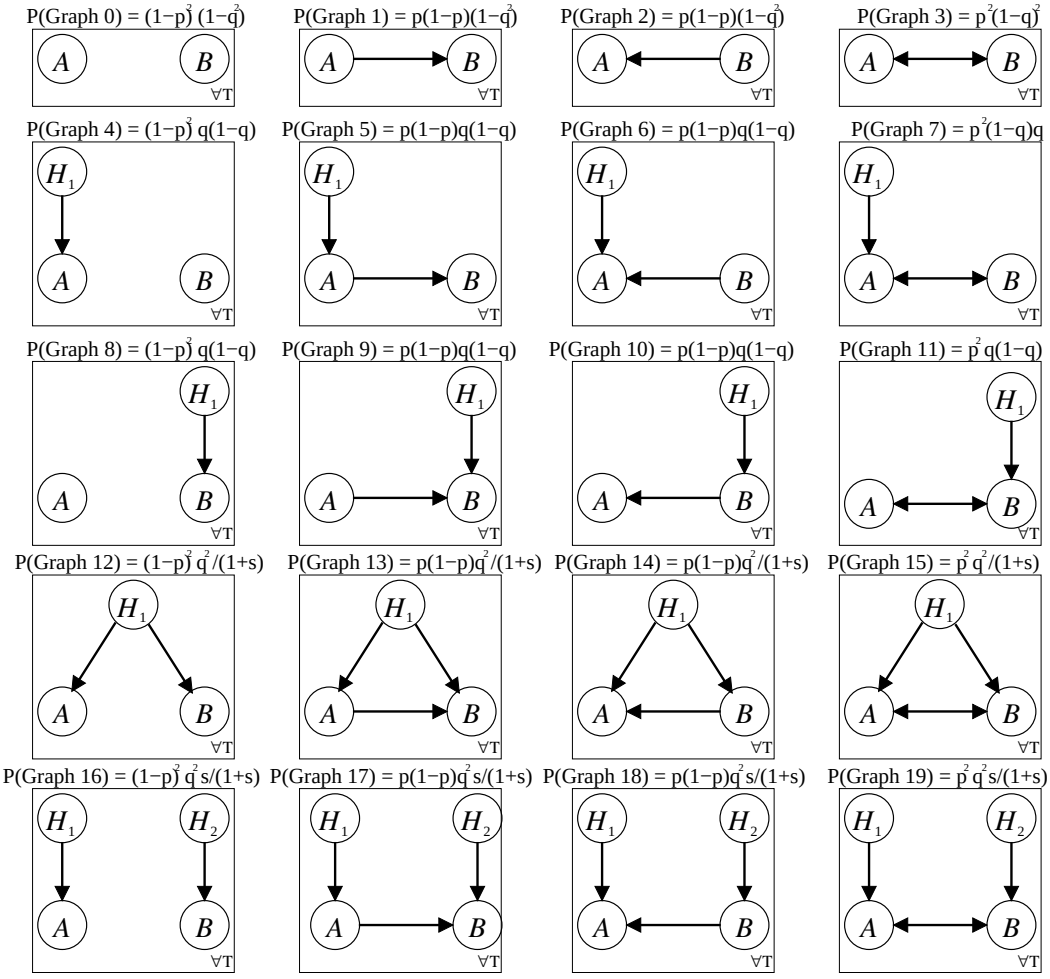

Figure 5.9: Hypothesis space for a two-ball stick-ball machine. A and B indicate `Moves(a,T)` and `Moves(b,T)` for `Ball` `a` and `b` respectively, while H_i indicates `Active(`$\mathtt{h}_i$`,T)` for the `HiddenCause` $\mathtt{h}_i$. The plates indicate that these causal relationships hold for all trials `T`.

Table 5.5: Graph Structures and Probabilities of Events for Two-Ball Machine

	Event					
Graph	a^+b^+	a^+b^-	a^-b^+	a^-b^-	$a^+\|\text{do}(b^+)$	$b^+\|\text{do}(a^+)$
0	0	0	0	1	0	0
1	0	0	0	1	0	ω
2	0	0	0	1	ω	0
3	0	0	0	1	ω	ω
4	0	$\alpha\omega$	0	$1-\alpha\omega$	$\alpha\omega$	0
5	$\alpha\omega^2$	$\alpha\omega(1-\omega)$	0	$1-\alpha\omega$	$\alpha\omega$	ω
6	0	$\alpha\omega$	0	$1-\alpha\omega$	ω	0
7	$\alpha\omega^2$	$\alpha\omega(1-\omega)$	0	$1-\alpha\omega$	ω	ω
8	0	0	$\alpha\omega$	$1-\alpha\omega$	0	$\alpha\omega$
9	$\alpha\omega^2$	0	$\alpha\omega(1-\omega)$	$1-\alpha\omega$	ω	$\alpha\omega$
10	0	0	$\alpha\omega$	$1-\alpha\omega$	0	ω
11	$\alpha\omega^2$	0	$\alpha\omega(1-\omega)$	$1-\alpha\omega$	ω	ω
12	$\alpha\omega^2$	$\alpha\omega(1-\omega)$	$\alpha\omega(1-\omega)$	$1-2\alpha\omega+\alpha\omega^2$	$\alpha\omega$	$\alpha\omega$
13	$\alpha\omega^2(2-\omega)$	$\alpha\omega(1-\omega)^2$	$\alpha\omega(1-\omega)$	$1-2\alpha\omega+\alpha\omega^2$	$\alpha\omega$	$\omega+\alpha\omega-\alpha\omega^2$
14	$\alpha\omega^2(2-\omega)$	$\alpha\omega(1-\omega)$	$\alpha\omega(1-\omega)^2$	$1-2\alpha\omega+\alpha\omega^2$	$\omega+\alpha\omega-\alpha\omega^2$	$\alpha\omega$
15	$\alpha\omega^2(3-2\omega)$	$\alpha\omega(1-\omega)^2$	$\alpha\omega(1-\omega)^2$	$1-2\alpha\omega+\alpha\omega^2$	$\omega+\alpha\omega-\alpha\omega^2$	$\omega+\alpha\omega-\alpha\omega^2$
16	$(\alpha\omega)^2$	$\alpha\omega(1-\alpha\omega)$	$\alpha\omega(1-\alpha\omega)$	$(1-\alpha\omega)^2$	$\alpha\omega$	$\alpha\omega$
17	$\alpha\omega^2(1+\alpha-\alpha\omega)$	$\alpha\omega(1-\omega)(1-\alpha\omega)$	$\alpha\omega(1-\alpha\omega)$	$(1-\alpha\omega)^2$	$\alpha\omega$	$\omega+\alpha\omega-\alpha\omega^2$
18	$\alpha\omega^2(1+\alpha-\alpha\omega)$	$\alpha\omega(1-\alpha\omega)$	$\alpha\omega(1-\omega)(1-\alpha\omega)$	$(1-\alpha\omega)^2$	$\omega+\alpha\omega-\alpha\omega^2$	$\alpha\omega$
19	$\alpha\omega(2+\alpha-2\alpha\omega)$	$\alpha\omega(1-\omega)(1-\alpha\omega)$	$\alpha\omega(1-\omega)(1-\alpha\omega)$	$(1-\alpha\omega)^2$	$\omega+\alpha\omega-\alpha\omega^2$	$\omega+\alpha\omega-\alpha\omega^2$

example, any dataset in which either ball moves without being intervened upon provides evidence against the causal structure shown in Graph 0. The probabilities of all events involving two balls under all causal structures for a two-ball machine are shown in Table 5.5.

Using this theory, we can compute a posterior distribution over Graphs 0 through 19 for any data $\mathcal{D}$, applying Equation 3.1. This posterior distribution can be connected to the results discussed by Gopnik et al. (2004) and Kushnir et al. (2003) by defining an appropriate mapping between the causal structures generated by the theory and the responses possible in the experiment. For the experiments described by Gopnik et al. (2004), possible responses were that `a` was special, `b` was special, and (in Experiment 3) that a hidden cause was involved. I computed the probability that `a` was special by summing over all graphs in which there is a link from A to B, did likewise for `b` being special, and equated the probability of a hidden cause with the probability of Graph 12. For the experiments described by Kushnir et al. (2003), there were four responses, corresponding to the structures shown in Figure 5.2. The probability of "`a` causes `b`" was evaluated by summing over Graphs 1, 5, 9, and 17, and likewise with the complementary structures for "`b` causes `a`." The "common cause" and "separate causes" structures were equated with Graphs 12 and 16.

The theory shown in Figure 5.8 has five parameters: p, q, s, α, and ω. ω was set empirically, via a small experiment. Ten participants were shown a computer simulation of the stick-ball machine, and reproduced the familiarization trials used by Kushnir et al. (2003): participants were told that when `a` causes `b` to move, it makes it move "almost

always," and were shown that a moved b on four of six trials. They were then asked how often they expected a would move b. The mean and median response was that a would move b on 75% of trials, so $\omega = 0.75$ was used. p was also fixed at 0.01, since any small value should be sufficient, and the remaining parameters were optimized to provide the best fit to the results of Kushnir et al., 2003). Table 5.4 shows the predictions of the model with $q = 0.035$, $s = 4$, and $\alpha = 0.36$, although a range of parameter values show the same qualitative trends. The model captures the major trends in the data, predicting the majority response in each condition, and gives a correlation of $r = 0.96$. The values of the parameters indicate that it is more likely that a ball will have the power to move on its own than that it will be connected to another ball ($q > p$), that balls are relatively unlikely to move (α is low), and that balls are quite likely to have independent causes ($s > 1$). With these parameter settings, the model also predicted the pattern of responses shown in Table 5.3 for the stimuli used by Gopnik et al. (2004).

5.2.2 Alternative accounts

Most computational accounts of causal induction that have been developed by psychologists do not make the distinction between observations and interventions, or consider learning about hidden causes. Consequently, the main alternative accounts come from computer science and statistics. Gopnik et al. (2004) and Kushnir et al. (2003) suggest that their results can be explained by constraint-based algorithms such as that proposed by Spirtes et al. (1993). These algorithms certainly do better than existing psychological models, as they are capable of using information derived from both observations and interventions, as well as identifying hidden causes. However, they cannot explain the data discussed in this section.

Explaining the inferences that people make about stick-balls in terms of these algorithms faces the same objections as arose with blicket detectors: small samples, and graded degrees of belief. The experiments described above illustrate that children and adults can identify the causal structure that holds among a set of variables based on only a handful of observations, far fewer than might be required to obtain statistically significant results from standard statistical tests of dependency. The data that people use to make causal inferences are not sufficient to infer that two variables are dependent. For example, in the *association* condition of Gopnik et al. (2004) and the *one observed cause* condition of Kushnir et al. (2003), all of the stimuli involve intervention on B, and suggest that A will occur

with high probability under such circumstances. These data are insufficient to justify the inference that A and B are dependent: it might just be that A occurs with high probability in general. The inference that B causes A requires an expectation that A is unlikely to occur on its own, and that if B causes A, then A is likely to occur when B does.

The data of Kushnir et al. (2003) also illustrate that people can maintain graded degrees of belief about causal structures. For example, the stimuli in the *independent unobserved causes* and *pointing control* conditions both seem to suggest that separate causes are responsible, but this impression is much stronger for the former than the latter, something that is reflected in people's judgments. Since constraint-based algorithms simply identify causal structures that are consistent with patterns of dependency, they cannot capture the subtle variation in degrees of belief that are exhibited by human subjects.

5.3 Summary

The theory-based approach explains how it is possible for people to infer complex causal relationships from small amounts of data: a causal theory that provides strong constraints on causal structures and on the functional form of causal relationship significantly reduces the amount of data that is required to identify such relationships. This principle explains how Halley was able to infer the existence of a comet with an elliptical orbit from just three observations, and how children can rapidly identify blickets and infer the properties of stickballs. The models presented in this chapter provide an account of human judgments in terms of domain-general statistical inference informed by domain-specific causal knowledge. Such inferences do not result from an additive combination of prior knowledge and covariational evidence, but a complex interaction between the two: intuitive theories determine what counts as data, how that data is interpreted, and which other hypotheses are entertained. Without these assumptions, the models fail: neither conventional statistics nor standard algorithms for causal induction can explain human judgments.

Chapter 6

Continuous physical systems

In the preceding chapters, I presented analyses of how people infer causal structures of increasing complexity, from a single causal relationship to hidden common causes. However, all of these analyses concern settings in which events can be described as discrete trials. While systems that operate in discrete time intervals are mathematically tractable, producing data similar to the contingency tables that are supplied to standard algorithms for learning causal structure, most of our learning about the physical world concerns systems that operate in continuous time. Even the blicket detector and the stick-ball machine are perhaps better characterized as involving events that occur in continuous time, rather than on discrete trials.

Events that occur in continuous time present a number of important challenges to theories of causal induction. First, contingency tables can no longer be used to characterize the relationship between cause and effect, since they indicate the frequency with which cause and effect co-occur over a discrete set of trials. In the absence of discrete trials, different kinds of data need to be used, such as the rate at which events occur in the presence and absence of a cause. Consequently, none of the standard models of causal induction discussed in Chapter 4 can be applied to events that occur in continuous time – a serious shortcoming. Second, events that occur in continuous time can exhibit complex dynamics, with causal relationships having characteristic temporal signatures. Different kinds of causal relationships manifest on different timescales: a collision between two billiard balls has instantaneous consequences, while one might develop symptoms days after contracting a disease. This added dimension of temporal dependence makes causal inference more complicated.

While these complexities might seem to make causal induction with events that occur in continuous time more difficult, there is also an important way in which it is easier: each observation can provide much more information as to the underlying causal structure. Different causal structures will produce expectations not just about the probabilities of different events, but the times at which they occur. The temporal pattern exhibited by a set of observations can rule out certain causal structures, providing overwhelming evidence as to the process that produced them. I will exploit this principle later in the chapter, showing that people can infer a hidden common cause from a single observation, given appropriate expectations about the relationship between time and causality.

Providing a satisfying account of people's inferences about physical systems requires extending the framework discussed in the previous chapters to address events that occur in continuous time. In such cases, causal relationships can manifest themselves in a variety of ways, through the rate at which events occur or their timing. The first task undertaken in this chapter is extending the causal graphical model formalism to allow for continuous time. I will then use the results of this analysis to examine human inferences about two physical systems: particle emissions and explosions. The first system, particle emissions, provides the opportunity to examine causal induction from rates, the continuous analogue of causal induction from contingency data, and introduces some of the innovations required to model events in continuous time. The second system, explosions, provides a simple setting in which to explore the role of dynamics in causal induction (c.f. Michotte, 1963).

6.1 Causal graphical models for continuous time

The causal graphical model formalism introduced in Chapter 2 was based upon the assumption that variables indicate events which have a finite, discrete number of opportunities to occur. In many of the preceding analyses, these discrete opportunities were trials, and causal relationships were quantified over all such trials. Under these assumptions, an arrow between two nodes in a graph indicates that the probability of a particular event occurring on any trial depends upon the occurrence of another event. There are two hurdles to overcome in extending this formalism to events that take place in continuous time – one conceptual, and one technical. The conceptual hurdle is conceiving of causal graphical models as representing causal relationships quantified over an unbounded number of instants in time. An arrow between two nodes in such a model thus indicates that whether an event

occurs at a particular instant in time depends upon the occurrence of another event. The technical hurdle is making this intuition precise, and developing a means of parameterizing causal graphical models of this kind. Overcoming this technical challenge will be the major task of this section.

6.1.1 From Bernoulli to Poisson

If we imagine each unit of time being broken up into a series of N_T trials, each picking out an interval of size ΔT, a causal graphical model can be used to specify the probability that a particular event occurs on any given trial. Extending causal graphical models to continuous time requires considering the limit of this system as $N_T \to \infty$ (and $\Delta T \to 0$). The key intuition behind what will happen as we take this limit (holding the mean number of events that occur in a single unit of time constant) can be obtained by studying Figure 6.1. The graphs in the figure are organized into two columns. The left column shows that the *probability* that an event occurs on a particular trial goes to zero as time is ever more finely divided and N_T increases towards ∞. The right column shows that the *rate* at which events occur, defined as the probability that an event occurs on a particular trial divided by the duration of that trial, ΔT, remains constant. The rate can be used to evaluate how many events will occur in an interval, and to compare different probabilistic models defined over events in continuous time.

The relationship between the rate at which events occur and the probability of the number of events in a time interval can be illustrated by first considering a situation in which the probability of an event is constant over all trials. If we define the probability that an event occurs to be α, then the rate at which events occur as $\lambda = \frac{\alpha}{\Delta T} = \alpha N_T$. Parameterized in terms of λ, the probability that N events occur in one unit of time (i.e. N_T trials) follows a binomial distribution,

$$P(N) = \frac{N_T!}{N!(N_T - N)!} \left(\frac{\lambda}{N_T}\right)^N \left(1 - \frac{\lambda}{N_T}\right)^{N_T - N}.$$

Taking $N_T \to \infty$, we obtain

$$P(N) = e^{-\lambda} \frac{(\lambda)^N}{N!}$$

which is a Poisson(λ) distribution. Applying a similar argument to the distribution of the waiting time between events, it can be shown that the limit of a series of Bernoulli events,

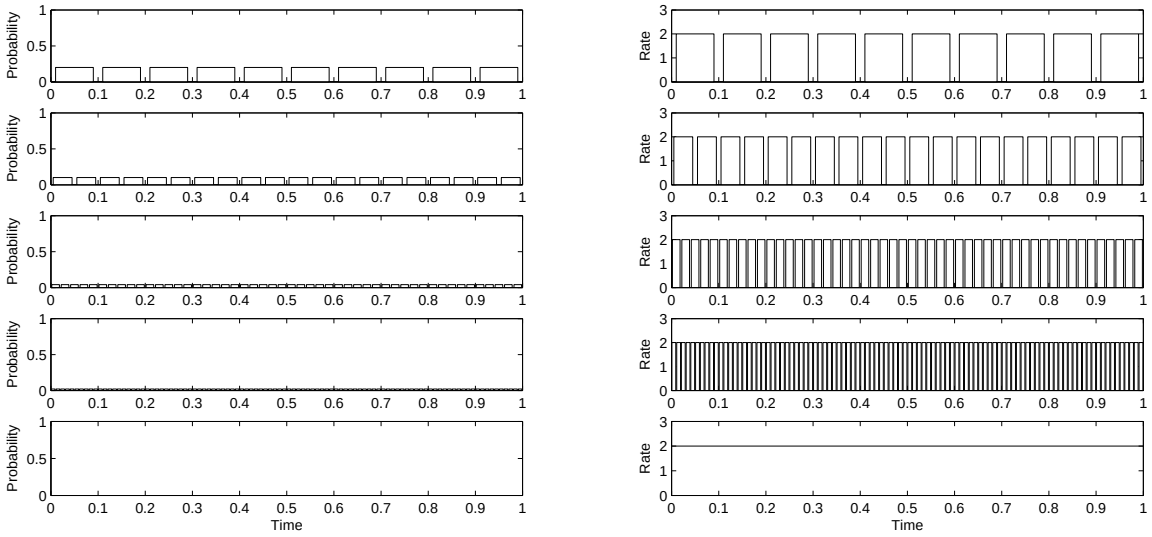

Figure 6.1: The probability of an event on a given trial (left column), and the rate at which events occur (right column) as a unit of time is partitioned into ever-finer intervals. The rows indicate an increase in N_T, the number of intervals per unit time, with $N_T = 10, 20, 50, 100, \infty$.

in which an event occurs with a fixed probability on each of N_T discrete trials, is a *Poisson process* with a constant rate λ for all times `T`. Poisson processes have been extensively studied in statistics, and allow us to compute probability distributions for a variety of quantities of interest, such as the number of events that will occur in a particular interval or the amount of time that will pass between successive events. The properties of Poisson processes that will be used in this chapter are shown in Figure 6.2.

The first of the properties shown in Figure 6.2 provides the basic result that will allow us to define causal graphical models for continuous time. When we move from discrete trials to continuous time, Bernoulli events become Poisson processes, and the probabilities of those events become the rate function for the Poisson process. Parameterizing causal graphical models in a way that is applicable to continuous time will thus require defining a function that gives the rate for each variable at any infinitesimal moment in time `t`. The simplest way to do this is to define a *homogeneous* Poisson process, having a constant rate λ for all times `T`. This is equivalent to an event having a constant probability over all trials. Such an assumption may be appropriate for variables that do not participate in causal relationships, but clearly cannot capture relationships between variables.

In discrete time, C causing E manifests as a change in the probability of E occurring on

1. A Poisson process with rate λ is the limit of a series of N_T Bernoulli trials in which an event occurs on each trial with probability $\frac{\lambda}{N_T}$ as $N_T \to \infty$.
2. The time between events in a Poisson process with rate λ follows an Exponential(λ) distribution.
3. The sum of two Poisson processes with rates λ_1 and λ_2 is a Poisson process with rate $\lambda_1 + \lambda_2$.
4. If a Poisson process is the sum of two Poisson processes with rates λ_1 and λ_2, then the probability that a particular event generated by that process originated from the first of its component processes is $\frac{\lambda_1}{\lambda_1+\lambda_2}$.

Figure 6.2: Important properties of Poisson processes.

any trial when C is present. In continuous time, such a relationship manifests as a change in the rate at which E occurs when C is present. Arrows between variables in a causal graphical model for continuous time thus indicate that the rate of one variable is a function of the other variable. The rate of the resulting Poisson process will not be constant, making it *non-homogeneous*: the rate $\lambda(\mathtt{T})$ depends upon the state of other variables and changes with $\mathtt{T}$. While this dependency could take any form, it is convenient to use a form that is based upon the same assumptions as the noisy-OR parameterization for discrete events.

6.1.2 A continuous equivalent of the noisy-OR

One way to derive the noisy-OR is as a deterministic OR function of a set of events that each occur with constant probability. In the continuous limit, this becomes a deterministic OR function of a set of Poisson processes. Since two Poisson processes never produce events in the same infinitesimal interval, the OR of a set of Poisson processes is simply the sum of those processes. By the third property of Poisson processes shown in Figure 6.2, this is a Poisson process with rate equal to the sum of its component processes.

This result suggests that we might define the continuous equivalent of the noisy-OR for a variable E with parents $X_1, \ldots, X_n$ to be a Poisson process with rate

$$\lambda(\mathtt{T}) = \lambda_0 + \sum_{i=1}^{n} \lambda_i x_i \tag{6.1}$$

where $x_i = 1$ if the sentence represented by X_i is true at time $\mathtt{T}$, and $x_i = 0$ otherwise. This is a simple and intuitive extension of the discrete case. However, it is inaccurate if any of the X_i are true for a finite set of instants in time rather than a set of continuous intervals.

A more correct statement of the parameterization would be

$$\lambda(\mathtt{T}) = \lambda_0 + \sum_i \lambda_i \int_{\mathtt{T}' \in \mathcal{T}_i} \delta(\mathtt{T}, \mathtt{T}')\ d\mathtt{T}' \tag{6.2}$$

where $\delta(\cdot, \cdot)$ is the Dirac delta function, contributing a spike of infinite height when its arguments agree (e.g., Boas, 1983), and $\mathcal{T}_i$ is the set of times at which the sentence represented by X_i is true. The integral is a convolution of the delta function with each point in $\mathcal{T}_i$. If $\mathcal{T}_i$ consists of a finite set of points, the X_i contributes a set of delta functions, each multiplied by λ_i, to $\lambda(\mathtt{T})$ – a discontinuous spike at each moment when that cause is present. If the points form a set of continuous intervals, then the contribution resembles that shown in Equation 6.1, adding λ_i to $\lambda(\mathtt{T})$ for each interval where the sentence represented by X_i is true.[1]

The material presented in this section provide us with the tools we need to examine causal induction in dynamic systems. We can define causal graphical models that apply to continuous time by specifying how the rate of Poisson processes depends upon other variables, and do so by a simple method that is equivalent to the noisy-OR. In the remainder of the chapter, I will use these tools to examine how people learn causal relationships from the rates and times at which events occur, looking at inferences about particle emissions and explosions.

6.2 Particle emissions

Several studies of causal induction have examined how people learn about causal relationships from rate data (Anderson & Sheu, 1995; Wasserman, 1990). Rates are closely related to contingencies, being the number of times the effect occurs in a continuous interval rather than the number of times the effect occurs on a set of discrete trials. Despite this close relationship, the models of contingency data described in Chapter 4, such as ΔP and causal power, cannot be applied to rate data, leading researchers to investigate other models.

Anderson and Sheu (1995, Experiment 2) conducted an experiment in which participants learned whether clicking on a flute icon caused a change in the rate of musical notes produced

[1] The appearance of the delta function in Equation 6.2 produces the equivalence to the noisy-OR, but it may be useful to explore other convolution kernels when specifying rate functions, replacing $\delta(\cdot)$ with some other function of $\mathtt{T}$ and $\mathtt{T}'$. In particular, $\phi_\sigma(\mathtt{T}, \mathtt{T}')$, the Gaussian probability density function with mean $\mathtt{T}'$ and standard deviation σ, is easy to work with and approaches $\delta(\mathtt{T}, \mathtt{T}')$ as $\sigma \to 0$.

by the flute. They found that their results were poorly predicted by the difference in rates, defined as

$$\Delta R = N(c^+) - N(c^-) \tag{6.3}$$

where $N(c^+)$ is the number of times the effect occurred in the interval when the cause, in this case clicking on the flute, was present, and $N(c^-)$ is the number of times the effect occurred when the cause was absent. Anderson and Sheu (1995) found that performance could be better predicted by "grating contrast," which they defined as

$$\text{contrast} = \frac{N(c^+) - N(c^-)}{N(c^+) + N(c^-)} \tag{6.4}$$

and justified by its use as a measure of contrast in psychophysical research. They gave no theoretical motivation for using this measure.

In this section, I will develop a theory-based account of causal induction from rate data, and describe alternative accounts that clarify the relationship of ΔR and grating contrast to models of causal induction from contingency data. These accounts will all be evaluated against human data. Unfortunately, the data from experiments exploring causal induction from rates (Anderson & Sheu, 1995; Wasserman, 1990) cannot be modeled without detailed information about the performance of individual participants. These experiments used a procedure in which participants interacted with objects, and then observed whether there was an alteration in the rate of the effect after their interaction. The models described in this section cannot be applied to this data without a record of the number of periods of interaction and non-interaction. To address this issue I conducted a simple experiment in which participants were provided with information about the rate at which particles were emitted by a chemical compound in the presence and absence of an electric field. The stimuli for this experiment were chosen to test the predictions of the ΔR model.

6.2.1 Experiment 6.1: Causal induction from rates

Method

Participants. 82 Stanford University undergraduates took part in the study.

Stimuli. A questionnaire presented a summary of nine experiments involving different chemical compounds and electrical fields, giving the number of particle emissions

inside and outside the electrical field. The number of particle emissions in each example was selected to give three critical sets of rates (expressed as $\{N(c^+), N(c^-)\}$ pairs): $\{52, 2\}, \{60, 10\}, \{100, 50\}$, for which $\Delta R = 50$, $\{12, 2\}, \{20, 10\}, \{60, 50\}$ for which $\Delta R = 10$, and $\{4, 2\}, \{12, 10\}, \{52, 50\}$, for which $\Delta R = 2$.

Procedure. The instructions outlined a hypothetical laboratory scenario:

> Imagine that you are working in a laboratory and you want to find out whether electrical fields influence the radioactive decay of certain chemical compounds. Below, you can see laboratory records for a number of studies. In each study, a sample of some particular compound was placed inside a particular kind of electrical field for one minute, and the rate of radioactive decay was measured (in number of particles emitted per minute). Each study investigated the effects of a **different** kind of field on a **different** kind of chemical compound, so the results from different studies bear no relation to each other.
>
> Of course, the chemical compounds can emit particles even when not in an electrical field, and they do so at different rates. Some compounds naturally decay at a fast rate, while others naturally decay at a slow rate. Thus, the decay rate of each compound was also measured for one minute in the absence of any electrical field. For each study below, you can see how many particles were emitted during one minute inside the electrical field, and during one minute outside of the electrical field. What you must decide is whether the electrical field increases the rate of particle emissions for each chemical compound.

Participants were instructed to provide ratings in response to a question like that of Experiment 2. Ratings were made on a scale from 0 (the field definitely does not cause the compound to decay) to 100 (the field definitely does cause the compound to decay). Each participant completed the survey as part of a booklet of unrelated experiments.

Results and Discussion

The results are shown in Figure 6.3, together with the model predictions. There was a statistically significant effect of $N(c^-)$ at $\Delta R = 50$ ($F(2, 162) = 12.17$, $MSE = 257.27$, $p < 0.001$), $\Delta R = 10$ ($F(2, 162) = 42.07$, $MSE = 468.50$, $p < 0.001$), and $\Delta R = 2$ ($F(2, 162) = 29.87$, $MSE = 321.76$, $p < 0.001$). Grating contrast fits the data slightly better than ΔR, giving $r = 0.924$, $\gamma = 0.43$, while ΔR gives, $r = 0.899$, $\gamma = 0.05$. Since ΔR predicts that

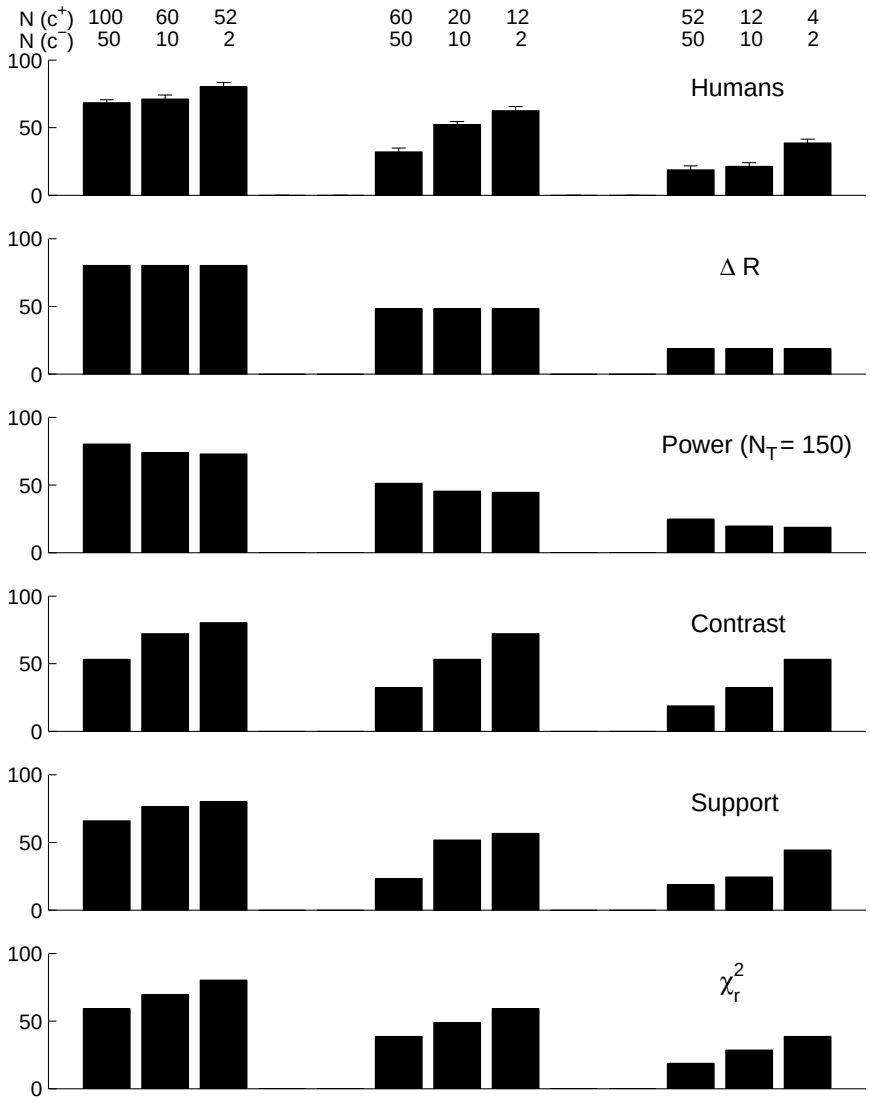

Figure 6.3: Predictions of rational models compared with results of Experiment 6.1. Numbers along the top of the figure show stimulus rates, error bars indicate one standard error.

Ontology:

Types	Number
`Compound`	$N_C \sim P_C$
`Field`	$N_F \sim P_G$
`Time`	$N_T = \mathbb{R}^+$

Predicates	Values	
`Charged(Field,Time)`	`Boolean:`	$\{\texttt{T}, \texttt{F}\}$
`Emission(Compound,Time)`	`Boolean:`	$\{\texttt{T}, \texttt{F}\}$

Plausible relations:

$\texttt{Charged}(\texttt{F}, \texttt{T}) \rightarrow \texttt{Emission}(\texttt{C}, \texttt{T})$
True for all `T` with probability p for each `F`, `C` pair

Functional form:

$\texttt{Charged}(\texttt{F}, \texttt{T}) \sim \text{PoissonProcess}(\cdot)$
$\texttt{Emission}(\texttt{C}, \texttt{T}) \sim \text{PoissonProcess}(\lambda(\texttt{T}))$ for $\lambda(\texttt{T})$ from a continuous noisy-OR:

Cause	Strength	Times
(Background)	$\lambda_0 \sim \text{Power}(1)$	
`Charged(F,T)`	$\lambda_i \lvert \lambda_0 \sim \text{Gamma}(1, \lambda_0)$	$\{\texttt{T} \lvert \texttt{Charged}(\texttt{F}, \texttt{T})\}$

Figure 6.4: Theory for causal induction from particle emissions.

responses within each of the critical sets should be constant, the statistically significant effect of $N(c^-)$ is inconsistent with this model.

6.2.2 Theory-based causal induction

A theory of particle emissions is shown in Figure 6.4. The theory identifies `Compound` and `Field` types, where a field can be `Charged` and a compound can produce an `Emission`. `Time` replaces `Trial` as the dimension along which these events occur, and takes values in the positive real numbers, $\mathbb{R}^+$. The rate of emissions $\lambda(\texttt{T})$ is specified by a continuous noisy-OR, as given in Equation 6.2, with parameters λ_0 and λ_i for each cause i (in this case, for each different `Field`). The prior on λ_0 is a power-law prior, being $P(\lambda_0) \propto \frac{1}{\lambda_0}$. This is an "uninformative" prior (Jeffreys, ?), intended to avoid setting a natural scale on the rate of particle emissions. This lack of a natural scale is carried over into the prior on λ_i through its definition in terms of λ_0. The theory also specifies $\mathcal{T}_i$, the set of times at which cause i affects $\lambda(\texttt{T})$.

With one `Field`, `f`, and one `Compound`, `c`, the hypothesis space generated by this theory is shown in Figure 6.5. The hypothesis space consists of two causal graphical models: using C to indicate `Charged(f,t)` and E to indicate `Emission(c,t)`, Graph 0 shows C and E independent, and Graph 1 shows C causing E. Structurally, these are the same hypotheses as those considered in the account of causal induction from contingency data in Chapter

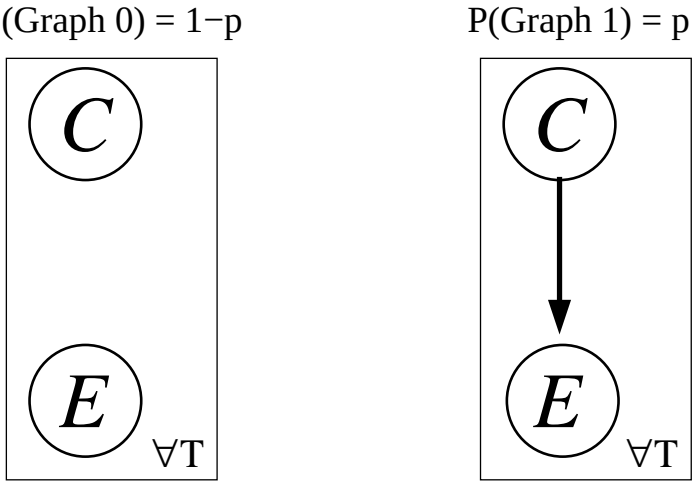

Figure 6.5: Hypothesis space generated by theory of particle emissions with one `Field f` and one `Compound c`. C and E indicate `Charged(f,T)` and `Emission(c,T)` respectively. The plates indicate that these causal relationships hold for all times `T`.

4. The plates shown for these graphical models now quantify over continuous time rather than discrete trials.

In causal induction from rates, our data consist of the number of times the effect occurred in a period of time when the cause was present, and a period of time when the cause was absent. In the experiment, there was a single field and a single compound for each pair of such numbers. The hypotheses to be compared are thus the two causal structures shown in Figure 6.5. From the discussion above, it should be clear that probability of the data under these hypotheses are given by the Poisson distribution, with rate parameter λ_0 for Graph 0 and $\lambda_0 + f\lambda_1$ for Graph 1. We can thus compute "causal support," as in Chapter 4, to evaluate the evidence a data set $\mathcal{D}$ provides in favor of Graph 1 over Graph 0:

$$\text{support} = \log \frac{P(\mathcal{D}|\text{Graph 1})}{P(\mathcal{D}|\text{Graph 0})}.$$

The probabilities $P(\mathcal{D}|\text{Graph 1})$ and $P(\mathcal{D}|\text{Graph 0})$ are computed using the Poisson distribution, integrating over λ_0 and λ_1. These integrals are performed numerically.

Causal support, computed with the Poisson parameterization, gives a good fit to the data shown in Figure 6.3, with $r = 0.978$, $\gamma = 0.35$. As can be seen from the figure, causal support predicts the trend displayed by the stimuli for which $\Delta R = 0$, decreasing as $N(c^-)$ increases. These predictions reflect the fact that the certainty in the value of λ_1 decreases

as $N(c^-)$ increases: if the effect occurs at a high rate in the absence of the cause, it becomes more difficult to determine if an increase in the number of times the effect is observed when the cause is present actually reflects a causal relationship. The effect of $N(c^-)$ is thus a sign that people are attempting to determine the causal structure underlying their observations.

6.2.3 Alternative accounts

The analysis of causal induction from rate data in terms of Poisson distributions can be used to motivate some alternative models, which are the equivalents of the models of causal induction from contingency data discussed in Chapter 4. Considering these models helps to clarify the assumptions behind some previously proposed accounts of causal induction from rate data – ΔR and grating contrast – as well as revealing what features of the theory-based account are responsible for its success.

Parameter estimation and ΔR

In Chapter 4, I showed that the two leading models of causal induction from contingency data – ΔP and causal power – are maximum-likelihood estimates of the strength parameter w_1 under different parameterizations for Graph 1. A similar result can be shown for causal induction from rate data. Under the parameterization specified by Equation 6.2, ΔR is the maximum-likelihood parameter estimate for λ_1 in Graph 1. Consequently, the failure of ΔR to describe the results shown in Figure 6.3 suggests that the construal of human causal induction as a structural inference, rather than parameter estimation, is an important factor in the success of the theory-based account.

It is also possible to identify a direct correspondence between ΔR and ΔP and causal power, rather than relying upon the derivation of the Poisson model given above. To do so, we view the rate information as just the positive events in a contingency table where the total sample size is unknown, so $N(c^+) = N_T P(e^+|c^+)$ and $N(c^-) = NP(e^+|c^-)$ for unknown N_T. If we assume that N_T is fixed across different experiments, we can obtain estimates consistent with the ordering and magnitude implied by ΔP using $\Delta R = N(c^+) - N(c^-) = N_T \Delta P$. If we make the further assumption that N_T is very large, ΔR will also correspond to causal power, since $P(e^-|c^-)$ will tend to 1. The equivalence of ΔP and causal power in this limit is implicit in the linear relationship expressed in Equation 6.2. An intuitive explanation for this result is that the sole difference between the noisy-OR and linear parameterizations

is in their treatment of situations where multiple causes simultaneously influence the effect, and in the continuous limit such events do not occur.

Even though causal power should correspond to ΔR in the limit as $N_T \to \infty$, it is instructive to consider its performance when we assume some finite N_T. The trends predicted by causal power do not vary with the choice of N_T, so the value $N_T = 150$ was chosen to allow these trends to be illustrated. The predictions of this model are shown in Figure 6.3. The predicted trends are clearly at odds with those observed in the data, reflected in the correlation $r = 0.845$, $\gamma = 0.06$.

Structure learning and grating contrast

As in Chapter 4, we can also consider alternative accounts based upon different approaches to learning causal structure. A rate-based equivalent of Pearson's χ^2 test for independence is

$$\chi^2_r = \frac{(N(c^+) - N(c^-))^2}{N(c^-)}. \tag{6.5}$$

which is derived in Appendix B. This statistic bears the same relationship to causal support for rates as the χ^2 test for independence does for contingency data: it is a frequentist independence test that evaluates the hypothesis that two variables are dependent. Like the χ^2 test for independence, χ^2_r does not make assumptions about the nature of the causal relationship – a cause can either increase or decrease the rate at which the effect occurs. Figure 6.3 shows the predictions of this model, which provide a strong correlation with the data, $r = 0.980$, $\gamma = 0.01$. It thus seems that the construal of causal induction from rates as a structural inference is more relevant to explaining these data than any specific assumptions about the form of the relationship.

Computing χ^2_r involves dividing the squared difference between the observed rates by the variance of the rate in the absence of the cause, comparing the magnitude of the effect of introducing the cause to the variation that should arise by chance. This may provide an explanation for the success of the grating contrast measure in predicting human judgments: inspection of Equation 6.4 reveals that grating contrast has many of the same ingredients as χ^2_r.

6.3 Explosions

Developing an account of causal induction from rates provides the opportunity to address events that occur in continuous time. However, since there are no dependencies across time, it remains only a small step from causal induction from contingency data. In many dynamic physical systems, the timing of events provide critical clues as to the underlying causal structure. A classic example of causal inference concerning such systems is Michotte's (1963) extensive investigation of the perception of collisions. In these studies, a mechanical device was used to generate the impression of two objects interacting. Typically, these displays showed one object at rest while another approached it. If the resting object began to move at the moment when the approaching object came into contact with it, people reported the impression that the motion of one object had caused the motion of the other. This impression proved very sensitive to the timing of the motion of the objects, with either temporal or spatial separation between the end of one object's motion and the start of the other reducing the sense of causality. Similar effects can be found with infants, suggesting an early origin for the perception of causality (Leslie, 1982; 1984; Leslie & Keeble, 1987; Oakes & Cohen, 1990; Cohen & Oakes, 1993)). Several studies have examined how the phenomena investigated by Michotte extend to more complex physical systems (Choi & Scholl, in press; White & Milne, 1997; 1999).

Whether people identify a Michottean event as causal depends heavily upon timing: the impression is strongest if the second ball moves at exactly the moment when the first ball strikes it. This sensitivity to timing can be explained under a Bayesian account of causal induction, since it would be a great coincidence to see such an event if there were no underlying causal relationship. However, there are a number of other factors that make people's inferences about such events quite complex, such as the positions, velocities, and overlap of the stimuli (e.g., Choi & Scholl, in press; Scholl & Nakayama, 2002). In the remainder of the chapter, I will focus on a dynamic physical system that removes these complexities, focussing just on the timing of events. This system is based upon the causal properties of explosives.

Explosives provide rich opportunities for examining human causal inferences, as they can explode for a variety of reasons, many of which require no physical contact. The times at which a set of cans explode are thus the result of a complex causal process, and identifying the elements of this process is a significant inductive challenge. Explosions

can occur spontaneously, as the result of instability of the explosive compound, or trigger one another in a chain reaction. Since such chain reactions can propagate by invisible shockwaves, force waves, the timing of a set of explosions provides the only information about the underlying causal structure, and is not accompanied by any other perceptual cues. Explosives thus provide the opportunity to study the role of timing in causal induction, without the other factors involved in Michottean displays.

The first task undertaken in this section is developing a simple theory of explosives. I then consider some of the questions about dynamic physical systems that can be answered using such a theory. Experiment 6.2 examines whether people's responses to these questions correspond to the predictions of this theory-based account. The section ends with a brief consideration of alternative accounts.

6.3.1 Theory-based causal induction

Developing a theory of explosives requires not just specifying the probabilities of events that occur in continuous time, but identifying how these events influence events at other times. I will address the first of these problems using the method introduced above defining a theory that specifies the rate at which different predicates become true. Unlike the theory given in the previous section, this theory will generate causal graphical models that reflect dependencies over time.

Figure 6.6 gives a simple theory of explosives that fulfills this requirement. Each set of explosives is assumed to be stored in a `Can`, which explodes at a particular time. Cans explode spontaneously as the result of some kind of hidden cause becoming active – local temperature, vibration, or cosmic rays. Hidden causes can also influence multiple cans, producing simultaneous explosions. Since I assume cans only explode at one time and occupy only one point in space, the `ExplosionTime` and `Position` of cans are encoded as multi-value predicates. Likewise, I assume that hidden causes only become active once, at some `ActivationTime`.[2] The ontology notes the range of values each of these predicates can take on, and indicates that two of them – the time predicates – refer to the same set of values. An equivalent theory that uses only Boolean predicates is given in Appendix D, making explicit the connection to the results derived earlier in the chapter.

Under this theory, the time at which a hidden cause explodes follows an Exponential(α)

[2] An alternative theory in which hidden causes can become active on multiple occasions is discussed in Griffiths, Baraff, and Tenenbaum (2004), and produces the same qualitative predictions.

Ontology:

Types	Number
`Can`	$N_C \sim P_C$
`HiddenCause`	$N_H = \infty$

Predicates	Values
`Explodes(Can, Time)`	`Time:` $\mathbb{R}^+$
`Active(HiddenCause, Time)`	`Time:` $\mathbb{R}^+$
`Location(Can)`	`Space:` $\mathbb{R}^2$

Plausible relations:

$\texttt{ExplosionTime}(\texttt{C}_1) \rightarrow \texttt{ExplosionTime}(\texttt{C}_2)$
True for with probability 1 for each $\texttt{C}_1 \neq \texttt{C}_2$ pair

$\texttt{Active}(\texttt{H}, \texttt{T}) \rightarrow \texttt{Explodes}(\texttt{C}, \texttt{T})$
Each `C` has an edge from some `H` with probability 1, which holds for all `T`. The particular `H` is chosen based upon the number of existing edges:

$$P(\texttt{ActivationTime}(\texttt{H}) \rightarrow \texttt{ExplosionTime}(\texttt{C})) \propto \begin{cases} M_{\texttt{H},i} & M_{\texttt{H},i} > 0 \\ s & \texttt{H is new} \end{cases}$$

where $M_{\texttt{H},i}$ is the number of edges from `H` when the edges are chosen for the ith can.

Functional form:

$\texttt{ActivationTime}(\texttt{H}) \sim \text{Exponential}(\alpha)$

$\texttt{ExplosionTime}(\texttt{C}_1) \sim \text{Exponential}(\lambda(\texttt{T}))$ for $\lambda(\texttt{T})$ from a continuous noisy-OR:

Cause	Strength	Times
(Background)	$\lambda_0 = 0$	
$\texttt{ActivationTime}(\texttt{H})$	$\lambda_i = \omega$	$\texttt{ActivationTime}(\texttt{H})$
$\texttt{ExplosionTime}(\texttt{C}_2)$	$\lambda_i = \omega$	$\texttt{ExplosionTime}(\texttt{C}_2) + D(\texttt{C}_1, \texttt{C}_2)/\mu$

Figure 6.6: Theory for causal induction with explosives. $D(\texttt{C}_1, \texttt{C}_2)$ is the distance between the locations of cans $\texttt{C}_1$ and $\texttt{C}_2$.

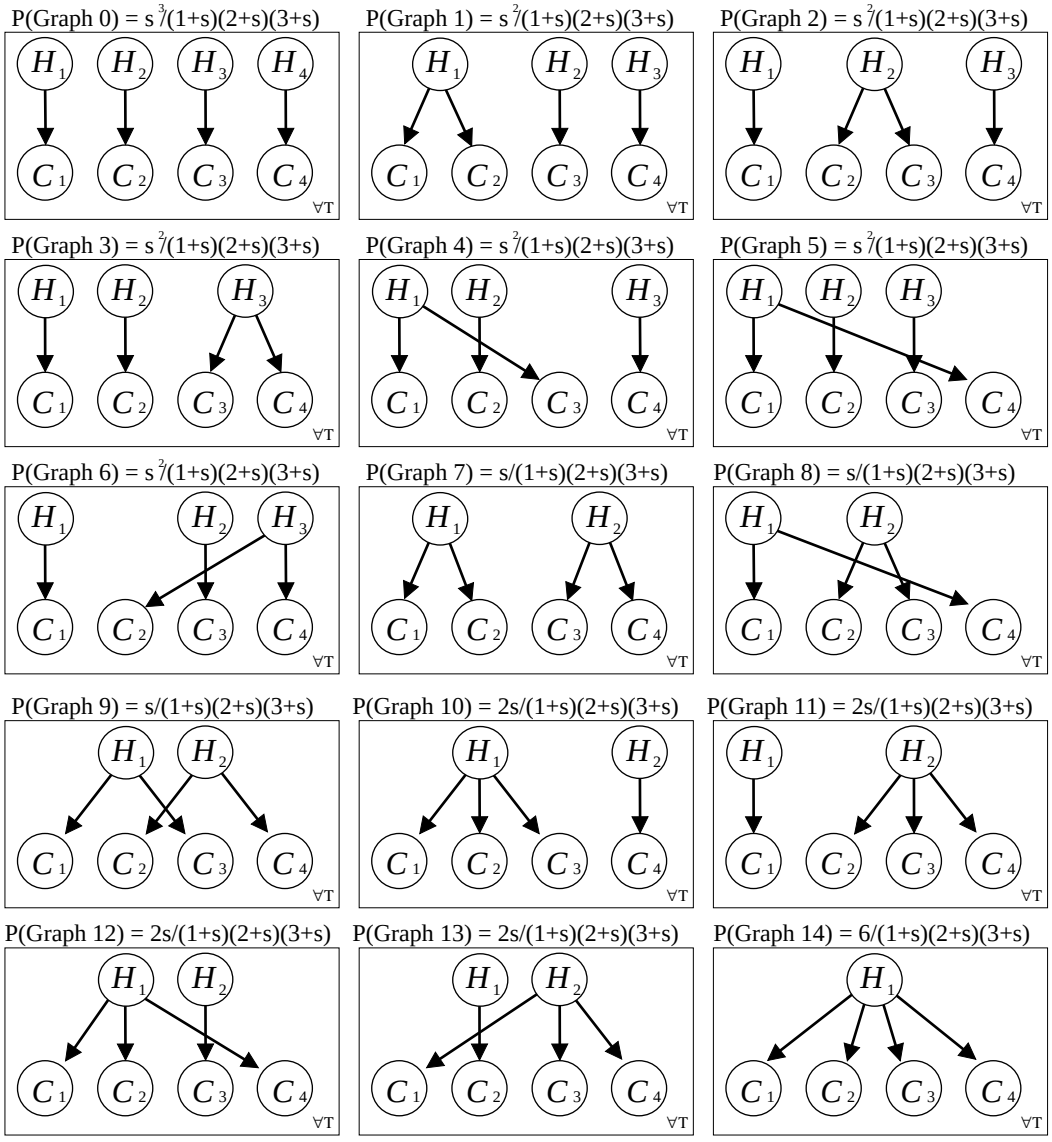

Figure 6.7: Hypothesis space for four cans of Nitro X. C_i indicates `ExplosionTime(c`$_i$`)` for `Can c`$_i$, while H_i indicates `ActivationTime(h`$_i$`)` for `HiddenCause h`$_i$. The dependence of `ExplosionTime(c`$_i$`)` on `ExplosionTime(c`$_j$`)` and `Position(c`$_i$`)` is suppressed.

distribution, a consequence of being the first arrival from a Poisson process (by the second property of Poisson processes given in Figure 6.2). Likewise, the time at which can `c` explodes has an Exponential($\lambda(\mathtt{T})$) distribution. Since the activation of hidden causes and the explosion of other cans are both events that occur at a finite number of points in a continuous interval, the rate function $\lambda(\mathtt{T})$ is a sum of delta functions, each of which contribute a spike multiplied by ω to the rate when their argument is equal to zero. The details of evaluating the probability of events under such a theory are discussed in Appendix D.

Since the causal relationships among cans are constant, the variation among the causal graphical models generated by the theory reduces to the configuration of causal relationships between hidden causes and explosions. Figure 6.7 depicts all configurations generated by the theory for a set of four cans, $N_C = 4$. The 15 graphs shown in the figure correspond to all partitions of four objects into different sets, where the objects are cans and all cans within a set share a hidden cause. The distribution over these partitions is provided by the Chinese restaurant process, introduced in Chapter 5. These causal structures are similar to those generated by the theory of the stick-ball machine described in the previous chapter. The fundamental difference between explosions and stick-balls is in the functional form, with the effects of causal relationships among explosions having a characteristic timecourse which requires attention to dynamics.

6.3.2 Three questions about dynamic systems

The theory developed in the previous section can be used to answer a number of questions about the causal properties of dynamic systems. I will show how the theory can be used to identify the causes of particular events, to learn the parameters that describe a particular kind of explosive, and to infer the causal structure that underlies a set of explosions. Experiment 6.2 will compare human judgments with the answers that the theory provides for these questions.

What caused what?

Working out which events caused others is one of the most basic problems posed by a causal system. This kind of inference – the determination of the "actual cause" – is generally quite complex (e.g., Halpern & Pearl, 2001). However, the theory given in Figure 6.6 can be used to obtain relatively simple answers for the case of explosions. In particular, the theory can

be used to evaluate whether a can exploded as the consequence of a hidden cause becoming active, or as the consequence of the explosion of a previous can. For example, upon seeing a set of explosions like those shown in Figure 6.8, the theory can be used to infer that the first can exploded spontaneously and that each subsequent explosion was the consequence of a previous explosion, forming a causal chain.

The fourth property of Poisson processes shown in Figure 6.2 makes it possible to evaluate which of the components of a sum of Poisson processes was responsible for a particular event. Under the theory of explosions shown in Figure 6.6, explosions can occur for two reasons: because of the activation of a hidden cause, or because of the explosion of another can. The Poisson process for explosions is the sum of a set of component processes reflecting each of these possible causes. Thus, we can ask which of these component processes was likely to have been the source of the explosion event. For simplicity, I will assume that the appropriate structure is Graph 0 from Figure 6.7, which will have highest prior probability if s is large. The probability that the hidden cause $\mathtt{h}_i$ is the actual cause of the explosion of can $\mathtt{c}_i$, which I will denote $H_i \Rightarrow C_i$, is

$$P(H_i \Rightarrow C_i|\mathcal{C}, \text{Graph 0}) = \int_0^\infty P(H_i \Rightarrow C_i|h_i, \mathcal{C}, \text{Graph 0}) P(h_i|\mathcal{C})\ dh_i \tag{6.6}$$

where $\mathcal{C} = \{c_1, \ldots, c_{N_C}\}$ is the set of times at which the cans $\mathtt{c}_1, \ldots, \mathtt{c}_{N_C}$ exploded, h_i is the time at which the hidden cause $\mathtt{h}_i$ became active, and the integral considers all of the possible times at which the hidden cause could have become active. In Appendix D, it is shown that this probability is 1 if no other cans could have been responsible, and less than $\frac{1}{2}$ if another can could have been responsible.

What kind of explosive is this?

Making inferences like that expressed in Equation 6.6 requires knowing about the properties of the particular explosive one is reasoning about. The theory given above has three parameters: α, the rate at which hidden causes become active, ω, the influence of causes on explosions, and μ, the velocity at which explosions propagate. Different settings of these parameters will capture the properties of different explosives and environments. For highly unstable explosives, α and ω will be large. For explosives that are likely to become unstable, but are not sensitive to vibration, α might be large while ω is small. For explosives that are relatively inert, but sensitive to vibration and changes in pressure, α would be small but

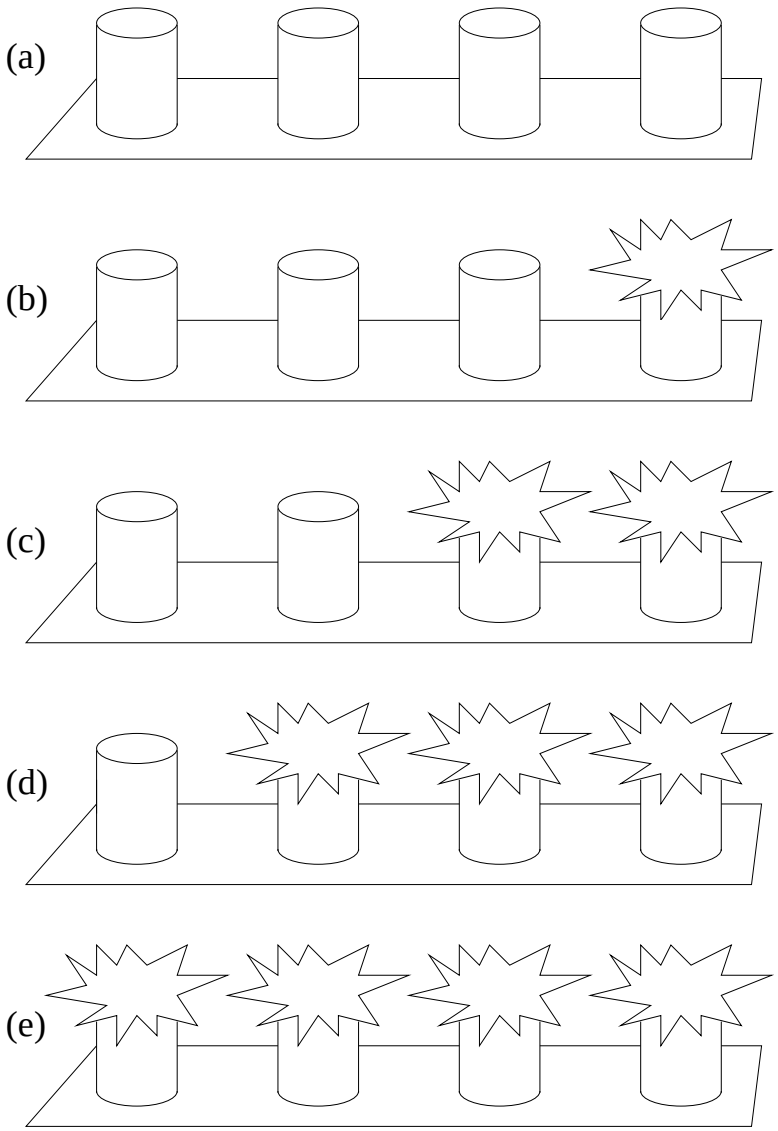

Figure 6.8: (a) Four cans of explosive. (b)-(e) A pattern of explosions consistent with a causal chain.

ω would be large. The theory thus expresses expectations about the general properties of explosives, which can be tuned to capture a the specific properties of a given substance.

The parameters of the theory can be learned through observation of a set of explosions. Each choice of these parameters implies a probability distribution over the times at which explosions occur for each of the different causal structures shown in Figure 6.7. Consequently, we can use Bayesian inference to infer the values of α, ω, and μ for any given dataset. In the following analysis, I will demonstrate how this inference can be performed assuming just one possible causal structure – Graph 0 from Figure 6.7 – but this assumption can be relaxed by integrating over all causal structures in the hypothesis space. Each choice of α, ω, and μ implies a particular distribution $P(\mathcal{C}|\text{Graph } 0, \alpha, \omega, \mu)$. By defining a prior on these parameters, $P(\alpha, \omega, \mu)$, we can use Bayes' rule to obtain a posterior distribution

$$P(\alpha, \omega, \mu|\mathcal{C}, \text{Graph } 0) = \frac{P(\mathcal{C}|\text{Graph } 0, \alpha, \omega, \mu)P(\alpha, \omega, \mu)}{P(\mathcal{C}|\text{Graph } 0)}, \tag{6.7}$$

where the denominator $P(\mathcal{C}|\text{Graph } 0) = \int\int\int P(\mathcal{C}|\text{Graph } 0, \alpha, \omega\mu)P(\alpha, \omega, \mu)\ d\alpha\ d\omega\ d\mu$.

Equation 6.7 can be used to infer α, ω, and μ from patterns of explosions like that shown in Figure 6.8. To make this example more concrete, we might assume that the cans were 2 spatial units apart, the first explosion occurred after 80 temporal units, and the subsequent explosions occurred with a separation of 40 temporal units. Comparing these numbers to the theory shown in Figure 6.6, it can be seen that if this is a genuine chain reaction μ should be 0.05, α should be relatively low, since it took a long time for a spontaneous explosion to occur, and ω should be relatively high, since each can caused another can to explode. Figure 6.9 shows the marginal posterior distributions over α, ω, and μ, obtained via Equation 6.7. These marginal distributions were computed using numerical integration, with $P(\mathcal{C}|\text{Graph } 0, \alpha, \omega, \mu)$ as specified in Appendix D, and independent Exponential(1) priors on each of the parameters. The marginal posteriors indicate that α should be small, ω should be large, and μ has a delta function at 0.05. Thus, from the pattern of explosions exhibited by the cans, Bayesian inference can be used to identify the velocity at which explosions propagate, that spontaneous explosions are relatively infrequent for this kind of substance, and that explosions are quite likely to set one another off.

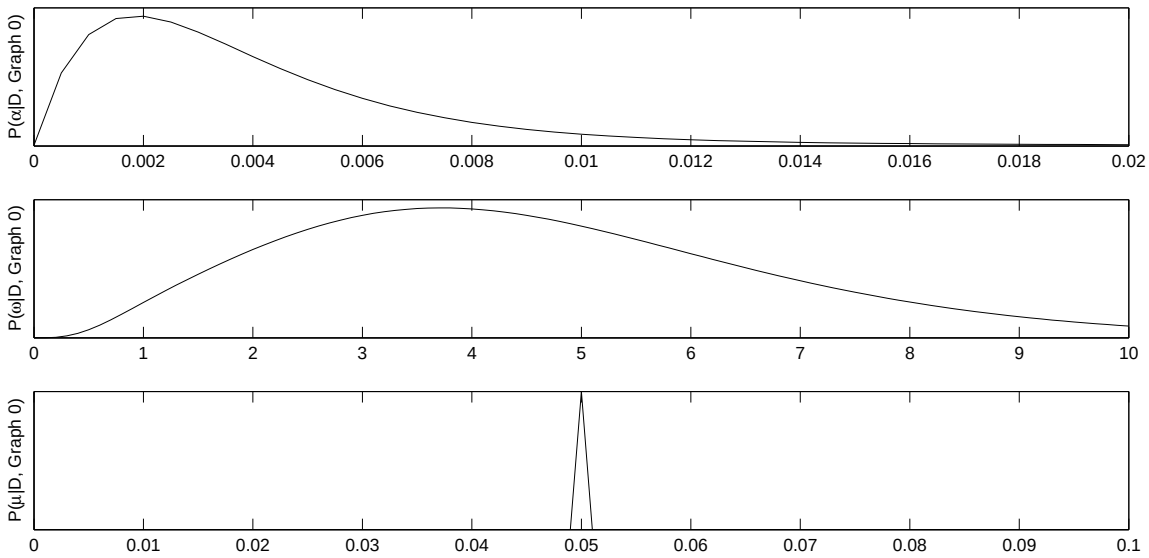

Figure 6.9: Marginal posterior distributions over α, ω, and μ for a set of explosion times $\mathcal{C}$ constituting a chain reaction.

What is the underlying causal structure?

In discussing both inferring the causes of events and learning the parameters of the theory, I made the simplifying assumption that the underlying causal structure was that denoted Graph 0 in Figure 6.7. However, one of the most important things we can learn about a dynamic causal system is the underlying causal structure. When inferring the causes of events and making predictions about what might happen next, it is useful to know what causal relationships exist. In the case of explosions, the variation in causal structure allowed by the theory given in Figure 6.6 concerns the configuration of hidden causes: explosions can occur independently, or be affected by hidden common causes.

The theory-based approach can be used to evaluate the underlying causal structure for any set of observed explosion times $\mathcal{C}$. This involves using Bayes' rule to compute a posterior distribution over the causal structures shown in Figure 6.7, $P(\text{Graph } i|\mathcal{C})$. I will run through this analysis for a particularly interesting case – an array of N_C cans all exploding simultaneously – showing how the posterior probability that the causal structure involves at least one hidden common cause can be computed. This can be done by evaluating the posterior probability of Graph 0 given the observation that the explosion times of all of the cans are equal ($c_i = t$ for all i, denoted $\mathcal{C} = t$).

To apply Bayes' rule, we need two probabilities for each causal structure: the prior probability, $P(\text{Graph } i)$, and the probability of the data under that structure, $P(\lfloor = t|\text{Graph } i)$. The former is supplied directly by the theory, as shown in Figure 6.7. The latter can be computed by integrating over the explosion times of all of the hidden causes. For example, in Graph 14, where there is only a single hidden cause $\mathtt{h}_1$ we have

$$\begin{aligned} P(\mathcal{C} = t|\text{Graph 14}) &= \int_0^\infty P(\mathcal{C} = t|h_1)P(h_1)\ dh_1 \\ &= \int_0^\infty (\omega\delta(t, h_1))^{N_C}\ \alpha \exp\{-\alpha h_1\}\ dh_1 \\ &= \omega^{N_C}\alpha \exp\{-\alpha t\} \end{aligned}$$

where h_1 is the value taken by the explosion time for the hidden cause. The result follows from the properties of the Dirac delta function.

More generally, the probability $P(\mathcal{C} = t|\text{Graph } i)$ depends only upon the number of hidden causes in Graph i, being

$$P(\mathcal{C} = t|\text{Graph } i) = \omega^{N_C}\left(\alpha \exp\{-\alpha t\}\right)^k,$$

where k is the number of hidden causes influencing cans in Graph i. Combining this probability with the prior defined by the theory gives

$$P(\text{Graph } 0|\mathcal{C} = t) = \frac{\zeta^{N_C}}{\prod_{j=0}^{N_C-1}(j+\zeta)} \tag{6.8}$$

where $\zeta = s\alpha \exp\{-\alpha t\}$, and s is a parameter of the theory indicating the relative prevalence of hidden common causes. The probability of the existence of some hidden cause, being $1 - P(\text{Graph } 0|\mathcal{C} = t)$, thus increases as N_C increases for any choice of s and α.

6.3.3 Experiment 6.2: Inferences about Nitro X

The theory-based approach can be used to answer three fundamental questions about explosions: what the cause of a given explosion might be, what properties a particular explosive has, and what the underlying causal structure is for a set of explosions. Experiment 6.2 examined how well the answers the algorithm provides to these questions correspond to human judgments. In the experiment, participants saw several stimuli in which cans of a

novel explosive – Nitro X – exploded in different patterns, and were asked to describe what they saw. The theory-based account makes three predictions, corresponding to each of the three questions explored above:

1. *What caused what?* With the right parameters for this kind of explosive, people should be able to identify the cause of any given explosion.

2. *What kind of explosive is this?* People should be able to infer the parameters of the theory from stimuli like that shown in Figure 6.8.

3. *What is the underlying causal structure?* When shown displays in which all cans explode simultaneously, more people should identify the existence of a hidden common cause as N_C increases.

The experiment was designed to test all three of these predictions.[3]

Method

Participants. Participants were 64 members of the MIT community, recruited through a mailing list for psychology studies. There were 16 participants in each of the four experimental conditions.

Stimuli. The stimuli were pictures of cans sitting on a table, presented on a computer screen. Each stimulus consisted of a new set of cans, all of which exploded at particular times. Explosions were demonstrated with cartoon explosion graphics like those shown in Figure 6.8.

Procedure. The experiment consisted of three familiarization stimuli and five test stimuli. The familiarization stimuli introduced the participants to Nitro X. For the first stimulus, participants were told that Nitro X is very unstable, and this was demonstrated by the experimenter tapping a can (using the mouse) and the can exploding. For the second stimulus, participants saw two cans of Nitro X, the experimenter tapped one can, which exploded, and the can next to it exploded shortly afterwards. Before seeing the third stimulus, participants were again reminded about the instability of Nitro X, and saw a single can explode without any action by the experimenter, after waiting for a few seconds.

The first two test stimuli were identical for all four conditions, and both involved four cans exploding in a causal chain, with a delay between successive explosions. The first

[3]This experiment was partially designed and entirely executed by Liz Baraff, to whom I am very grateful. Preliminary results of the experiment are reported in Griffiths, Baraff, and Tenenbaum (2004).

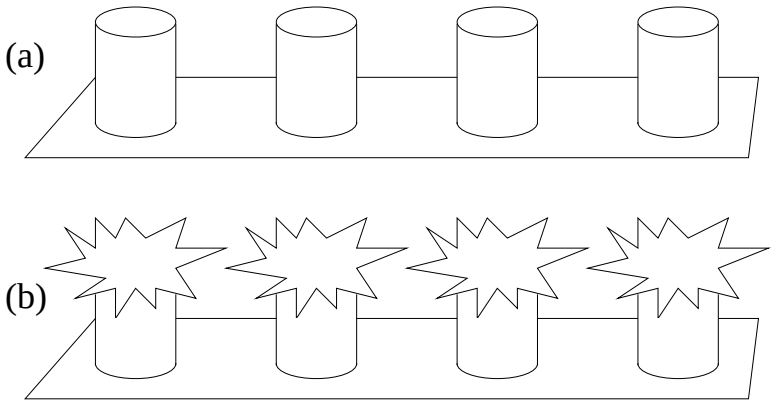

Figure 6.10: The third stimulus used in Experiment 6.2, with $N_C = 4$. (a) Four dormant cans. (b) A simultaneous explosion.

causal chain moved from right to left, as shown in Figure 6.8. The second originated with the second can, with the first and third cans exploding simultaneously, and then the fourth can, each with an appropriate delay. The four conditions differed in the third test stimulus, where the number of cans in the display was varied, being $N_C = 2, 3, 4$ or 6. After a brief delay, all of the cans exploded simultaneously, as shown in Figure 6.10. The last two stimuli allowed the participants to interact with Nitro X by tapping, and will not be discussed further.

After each test stimulus, participants were given a sheet of questions. These sheets gave three options:

1. The first can exploded spontaneously. That explosion caused the other cans to explode, in a chain reaction.

2. Each can exploded spontaneously, all on its own. There was no causal connection between them.

3. Neither of the above is a likely explanation. Please write a plausible alternative here.

The order of the first two options was counterbalanced, but the third option was always last.

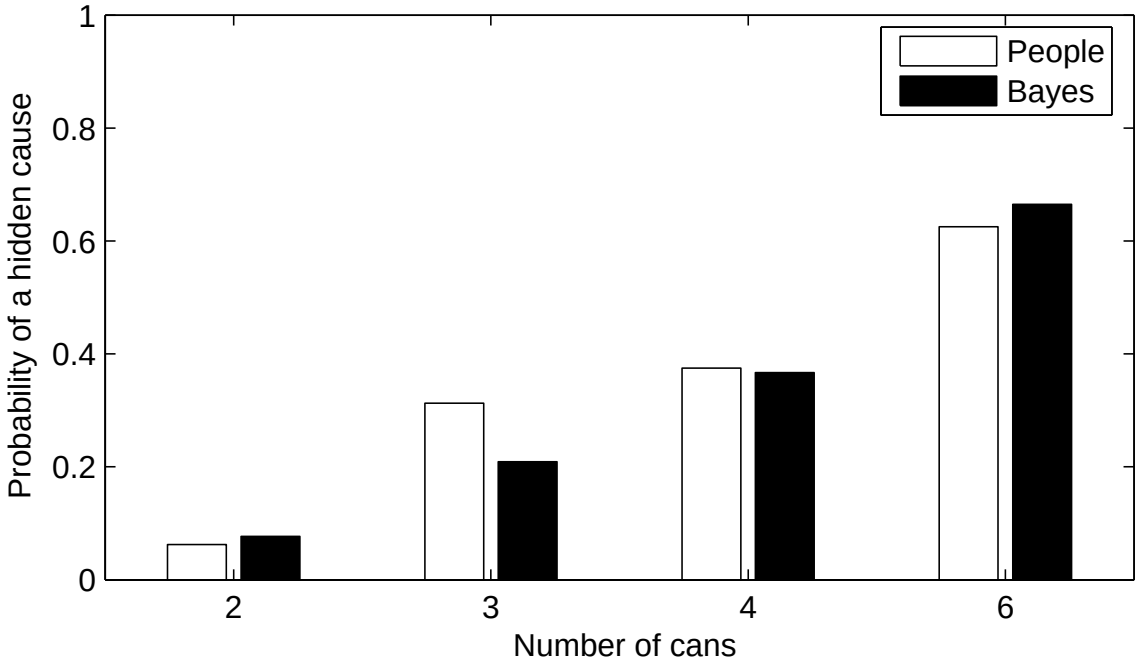

Figure 6.11: Results for the third stimulus in Experiment 6.2, compared with predictions of the theory-based Bayesian account.

Results and Discussion

All participants in all conditions identified the first test stimulus as a causal chain. 16/16, 15/16, 15/16, and 16/16 participants identified the second test stimulus as a causal chain, for $N_C = 2, 3, 4$ and 6 respectively. The participants who indicated otherwise for $N_C = 3$ and 4 both chose the third option, suggesting some combination of spontaneous explosions and a chain reaction. For the third test stimulus, no participants indicated a causal chain – responses were divided between spontaneous explosions and the third option. The responses of people who chose the third option were coded by two raters, who were in 100% agreement in classifying all such responses as indicating a hidden cause. The proportion of participants identifying a hidden cause behind the simultaneous explosion is shown in Figure 6.11. There was a statistically significant effect of N_C, $\chi^2(3) = 11.36$, $p < 0.01$. The number of cans influenced whether people inferred hidden causal stucture, with most people seeing two cans as independent but six as causally related. The figure also shows the probability of the existence of a hidden common cause under the theory-based account, using Equation 6.8 with $\zeta = 12$. These predictions correlate with the data at $r = 0.97$.

The results of the experiment indicate that people's inferences about explosions correspond to the three predictions of the theory-based account. People can infer what caused what: almost every subject correctly identified the first and second test stimuli as being causal chains. This result also indicates that people are able to learn the parameters of a theory, as these events should only be identified as causal chains if people have correctly identified the rate at which explosions propagate, μ. Further evidence for parameter learning comes from the third test stimulus, where no participants appealed to a causal chain, having learned that there is a delay in inter-can causation. Finally, responses to the third test stimulus show that people are capable of inferring hidden causal structure from observations, and the effect of N_C on this inference is entirely consistent with the predictions of the theory-based account.

6.3.4 Alternative accounts

Standard algorithms for learning causal graphical models cannot explain these results. If we imagine that time is broken into discrete intervals, and a can either explodes or does not explode in each interval, then we can construct a contingency table for each pair of cans. Statistical significance tests will identify pairwise dependencies among all cans that explode simultaneously, provided appropriate numbers of non-explosion trials are included. The existence of a hidden common cause is consistent with such a pattern of dependency. However, as a result of reasoning deductively from this pattern, the evidence for such a structure does not increase with N_C: a hidden common cause is merely consistent with the pattern for all $N_C > 2$.

This experiment also illustrates that people are willing to infer hidden causal structure from very small samples – just one datapoint – and from observations alone. Standard constraint-based algorithms cannot solve this problem: while a hidden common cause is consistent with the observed pattern of dependency, causal structures in which the cans influence one another cannot be ruled out without intervention information. People do not consider this possibility because they have learned that the mechanism by which cans influence one another has a time delay.

6.4 Summary

People often make inferences about causal relationships from events that occur in continuous time. Such inferences depend upon the rates at which events occur, or their timing. By extending causal graphical models to events that occur in continuous time, the theory-based approach can be used to explain these inferences. I have shown how such explanations can be formulated for two systems – particle emissions and explosions. These systems illustrate how the theory-based approach overcomes the challenges of causal induction from events in continuous time, showing how causal structure can be inferred from rates, and how dynamics can be introduced into a theory.

The experiment examining how people draw causal inferences from explosions draws much more on perception than the other experiments discussed in this thesis. Traditionally, perceived causality, as typified by the phenomena of Michotte (1963), has been considered fundamentally different from cognitive causal judgment (e.g., Schlottmann & Shanks, ?). My analysis reveals an important sense in which they are fundamentally the same: both involve an inductive inference to a generating process, constrained by a set of assumptions about the nature of causality. Nonetheless, I anticipate that there are important differences between perceptual and cognitive causal inferences, concerning the rigidity of those constraints, their content, and their plasticity. Formulating both kinds of inferences in a common framework provides the opportunity to explore these differences.

Chapter 7

Coincidences

In the last days of August in 1854, the City of London was hit by an unusually violent outbreak of cholera. More than 500 people died over the next fortnight, most of them in a small region in Soho, 250 yards in radius. On September 3, this epidemic caught the attention of John Snow, a medical doctor who had recently begun to argue against the widespread notion that cholera was transmitted by bad air. Snow immediately suspected a water pump on Broad Street as the cause, but could find little evidence of contamination. However, on collecting information about the locations of the cases of cholera, he discovered that they were tightly clustered around the pump. This suspicious coincidence hardened his convictions, and the pump handle was removed. The disease did not spread any further, furthering Snow's (1855) argument that cholera was caused by infected water.

Observing clusters of events in the streets of London does not always result in important discoveries. Towards the end of World War II, London came under bombardment by German V-1 and V-2 flying bombs. It was widespread popular belief that these bombs were landing in clusters, with an unusual number of bombs landing on the poorer parts of the city (Johnson, 1981). This apparent clustering was subsequently dismissed as a mere coincidence. After the war, R. D. Clarke of the Prudential Assurance Company set out to "apply a statistical test to discover whether any support could be found for this allegation" (Clarke, 1946, p. 481). Clarke examined 144 square miles of south London, in which 537 bombs had fallen. He divided this area into small squares and counted the number of bombs falling in each square. If the bombs fell uniformly over this area, then these counts should conform to the Poisson distribution. Clarke found that this was indeed the case, and concluded that his result "lends no support to the clustering hypothesis" (1946, p. 481),

implying that people had been misled by their intuitions.[1]

The suspicious coincidence noticed by John Snow and the mere coincidence that duped the people of London seem to establish a paradox: how is it that coincidences can simultaneously be responsible for important scientific discoveries and widespread false beliefs? The resolution of this paradox has been complicated by the fact that the two approaches that have been taken to the study of coincidences each emphasize only one of these roles. Inspired by examples like that of Snow,[2] one approach has focused on how coincidences relate to causality (Owens, 1992), what constitutes a coincidence (Horwich, 1982; Schlesinger, 1991), and how coincidences are used in scientific argument (Hacking, 1983). The other approach has focused on events like the bombing of London, using them as examples of the shortcomings of human reasoning about chance (Diaconis & Mosteller, 1989; Fisher, 1937; Gilovich, 1993; Plous, 1993).

In addition to their intrinsic interest, coincidences have the potential to help us understand the process of causal discovery. Scientific knowledge is expanded and revised through the discovery of new causal relationships which enrich or invalidate existing theories. In the previous chapters, I have shown how simple intuitive theories can be used to explain how people learn causal relationships. An important problem for this account, and for the more general claim that human knowledge is organized into intuitive theories with a structure analogous to scientific theories (Carey, 1985a; Gopnik & Meltzoff, 1997; Karmiloff-Smith, 1988; Keil, 1989; Murphy & Medin, 1985), is accounting for theory change, in which an existing theory is modified to accommodate new data, or rejected altogether. The connection between coincidences and scientific discovery suggests that they may play a similar role in the development of intuitive theories, helping us to understand how theories change.

The aim of this chapter is to provide a computational account of coincidences, to identify their connections to theory change, and to make quantitative predictions about human

[1] Clarke's investigations were later introduced to a broader audience by the widely-used textbook of Feller (1968), and the Poisson distribution makes several appearances in Pynchon (1973/1995).

[2] Such examples abound. In considering the apparent rotation of stars about the Earth, Aristotle viewed the coincidence between the rate of motion and the distance traversed as evidence for the existence of a single celestial sphere (Franklin, 2001, pp. 133-134). Halley would never have discovered his comet without noticing the surprising regularity in the paths and dates in a table of orbits (Cook, 1998; Hughes, 1990; Yeomans, 1991). Semmelweis might not have developed his theory of contagion without noting the similarity in the symptoms of a doctor injured during an autopsy and those of patients in his ward (Hempel, 1966). Perrin's (1913/1990) argument for the objective reality of molecules was based upon the suspiciously similar estimates of Avogadro's number produced by several quite different methods of measuring molecular magnitudes.

judgments of the strength of coincidences. First, I will argue against a widespread definition of coincidences: the idea that coincidences are unlikely events. I will then present an alternative account, claiming that coincidences are events that provide support for a hypothesis, but not enough support to convince us to accept that hypothesis. I will make this definition precise by expressing it in the Bayesian framework used throughout this thesis, and show how it can be used to make sense of some of the key issues raised by coincidences. After conducting a simple experimental test of this account, I will use it to make quantitative predictions about the strength of coincidences in some of the complex settings where classic examples of coincidences occur: coincidences in space, as in the examples of John Snow and the bombing of London, and coincidences in time, as in the famous "birthday problem". The chapter concludes with an attempt to localize the irrational component of human reasoning about coincidences, and a discussion of the role of coincidences in theory change.

7.1 Coincidences are not just unlikely events

Upon experiencing a coincidence, many people report a response that involves thinking something like "What are the chances of that?" (e.g., Falk, 1981-1982). Subjectively, coincidences are unlikely events: we interpret our surprise at their occurrence as indicating that they have low probability. In fact, it is often assumed that being surprising and having low probability are identical properties: the mathematician Littlewood (1953) suggested that events having a probability of one in a million be considered surprising, and many psychologists make this assumption at least implicitly (e.g., Slovic & Fischhoff, 1977). The notion that coincidences are unlikely events pervades literature addressing the topic, irrespective of its origin. This belief is expressed in books on spirituality ("Regardless of the details of a particular coincidence, we sense that it is too unlikely to have been the result of luck or mere chance," Redfield, 1998, p. 14), popular books on the mathematical basis of everyday life ("It is an event which seems so unlikely that it is worth telling a story about," Eastaway & Wyndham, 1998, p. 48), and even Diaconis and Mosteller (1989) considered the definition "a coincidence is a rare event," but rejected it on the grounds that "this includes too much to permit careful study" (p. 853).

The idea that coincidences are unlikely events often appears as a prelude to the claim that people are irrational in their reactions to coincidences. Such accounts typically assume

that we are justified in attending to events that are sufficiently unlikely to arise by chance, but argue that most coincidences are events that people mistakenly believe to be unlikely. This argument is made explicit in Blackmore and Troscianko's (1985) "chance baseline shift" hypothesis, which suggests that beliefs in the paranormal may be a consequence of underestimating the probability of coincidental events arising by chance. This hypothesis has been explored in a number of studies, with mixed results (Blackmore, 1997; Bressan, 2002; Brugger, Landis & Regard, 1990; Brugger & Taylor, 2003; Musch & Ehrenberg, 2002).

The simplest version of the idea that coincidences are unlikely events refers only to the probability of a single event. Thus, some data, $\mathcal{D}$, might be considered a coincidence if the probability of $\mathcal{D}$ occurring by chance is small. On September 11, 2002, exactly one year after terrorists destroyed the World Trade Center in Manhattan, the New York State Lottery "Pick 3" competition, in which three numbers from 0-9 are chosen at random, produced the results 9-1-1 (Associated Press, September 12, 2002). This seems like a coincidence,[3] and has reasonably low probability: the three digits were uniformly distributed between 0 and 9, so the probability of getting such a combination by chance is $(\frac{1}{10})^3$ or 1 in 1000. If $\mathcal{D}$ is a sequence of eight coinflips that are all heads, which I will denote HHHHHHHH, then its probability under a fair coin is $(\frac{1}{2})^8$ or 1 in 256. If $\mathcal{D}$ is an event in which one goes to a party and meets four people, all of whom are born on August 3, and we assume birthdays are uniformly distributed, then the probability of this event is $(\frac{1}{365})^4$, or 1 in 17,748,900,625. Consistent with the idea that coincidences are unlikely events, these values are all quite small.

The fundamental problem with this account is that while coincidences may be unlikely events, there are many unlikely events that are certainly not coincidences. It is easy to find events that have the same probability, yet differ in whether we consider them a coincidence. Since the probability of an event arising by chance in each of the previous examples depends only on the number of observations, any outcome with the same number of observations has the same probability. Thus, September 11 lottery results of 7-2-3, meeting people with birthdays of May 14, July 8, August 21, and October 4, and obtaining HHTHTTTH all

[3]Indeed, many people sought explanations other than chance: the authorities responsible for the New York lottery were sufficiently suspicious that they initiated an internal investigation, and a variety of sources claimed that mystical forces were involved. The St Petersburg Times quoted one psychologist as saying that "It could be that, collectively, the people in New York caused those lottery numbers to come up 9-1-1... If enough people all are thinking the same thing, at the same time, they can cause events to happen," while a psychic they interviewed suggested that the lottery was used by divine forces to communicate a message to the American people (DeGregory, September 24, 2002).

have the same probability as the coincidental examples given above, even though they are not particularly surprising. This indicates that the definition of a coincidence must refer to something more than just the probability of an event. Teigen and Keren (2003) use several other examples of this kind, supported by experimental results, to illustrate the weak relationship between the surprisingness of events and their probability.

A more sophisticated version of the idea that coincidences are unlikely events is to define coincidences as being events of a "kind" that is unlikely. Hints of this view appear in experiments on coincidences conducted by Falk (1989), who suggested that people are 'sensitive to the extension of the judged event' (p. 489) when evaluating the surprisingness of coincidences. Similarly, Falk (1981-1982) suggested that when one hears a story about a coincidence, "One is probably not encoding the story with all its specific details as told, but rather as a more general event 'of that kind' " (p. 23). Similar ideas have been proposed by psychologists studying figural goodness and subjective randomness (e.g., Garner, 1970; Kubovy & Gilden, 1991). This account was worked out in most extensive detail by the philosopher Schlesinger (1991), who explicitly considered coincidences in birthdays. Under this view, meeting four people all born on August 3 is a bigger coincidence than meeting those born on May 14, July 8, August 21, and October 4 because the former is of the kind *all on the same day* while the latter is of the kind *all on different days*. The probability of observing four birthdays on the same day is $365(\frac{1}{365})^4 = \frac{1}{48627125}$ while the probability of observing four birthdays on different days is $\frac{365\times364\times363\times362}{365\times365\times365\times365} = \frac{47831784}{48627125}$. Coincidences are thus surprising because events of their kind are rare.

The "unlikely kinds" definition faces a number of problems. The first of these is specifying what might count as a kind of event. In some domains, such as coinflips, this might be relatively easy, but even identifying the kinds of events that might be expressed by sets of birthdays becomes quite involved. This problem is compounded when we consider coincidences that go beyond birthdays. While the notion of a "kind" seems reasonably natural for discrete events, many coincidences involve observations that are not naturally assimilated to particular kinds. For example, the bombing of London involved a coincidence based upon bomb locations, which are not easily classified into kinds. To provide a complete account of coincidences, we need to be able to identify the kinds relevant to any contexts, including those involving continuous stimuli.

Even when a meaningful set of kinds can be found, any event is going to be an instance of many different kinds. For example, the set of birthdays {August 3, August 3, August 3,

August 3} is an instance of the *four birthdays on the same day* kind, the *four birthdays on August 3* kind, the *four birthdays in August* kind, and the *four birthdays between June 22 and November 26* kind, not to mention being just *four birthdays*. This raises the question of how to evaluate the probability of this event: which kind should we choose when calculating the probability of an event "of that kind"? In the case of four birthdays on August 3, it seems obvious that we should choose something like *four birthdays on the same day*, although it is difficult to account for why this is the appropriate choice. It is certainly not the most specific kind – *four birthdays on August 3* is a much smaller set. This point can be made even more clearly by considering a set of dates like {January 12, March 22, March 22, July 19, October 1, December 8}, in which the most specific hypothesis would be *six birthdays on exactly these dates*. The best kind for this data might be something relatively complicated, like *six birthdays, two of which are on the same day*, which identifies a pattern that is expressed by only a subset of the observations.

Finally, even when a meaningful set of kinds can be identified and the problem of multiple kinds is avoided by allowing each event to be of only one kind, it is possible to find counterexamples to the "unlikely kinds" definition. For instance, a common way of explaining why a sequence like HHHH is judged less random (and more coincidental) than HHTT is that the former is of the kind *four heads* while the latter is of the kind *two heads, two tails* (c.f. Garner, 1970; Kubovy & Gilden, 1991). Since one is much more likely to obtain a sequence with two heads and two tails than a sequence with four heads when flipping a fair coin four times, the latter seems like a bigger coincidence. The probability of N heads from N_T trials is

$$P_{\text{kind}}(D) = \frac{\binom{N_T}{N}}{2^{N_T}}, \tag{7.1}$$

so the probability of the *four heads* kind is $\frac{\binom{4}{4}}{2^4} = 0.0625$, while the probability of the *two heads, two tails* kind is $\frac{\binom{4}{2}}{2^4} = 0.375$. However, we can easily construct a sequence of a kind that has lower probability than *four heads*: the reasonably random HHHHTHTTHHHTHTH-HTHTTHHH is but one example of the *fifteen heads, eight tails* kind, which has probability $\frac{\binom{23}{15}}{2^{23}} = 0.0584$.

In addition to the technical problems with the definition of coincidences as unlikely events, this account seems to neglect one of the key components of coincidences: their meaningfulness. This aspect of coincidences is what makes them so interesting, and is emphasized in accounts such as that of Jung (1960), and Diaconis and Mosteller (1989),

who consider a coincidence "a surprising concurrence of events, perceived as meaningfully related, with no apparent causal connection" (p. 853). This meaningfulness is tied to the role of coincidences in scientific discoveries. In the remainder of the chapter, I will argue that we notice coincidences not just because they manifest arbitrary low-probability patterns, but because these patterns suggest novel causal explanations.

7.2 Approaching coincidences via causal induction

The characterization of coincidences as unlikely events is associated with a focus on the ways in which the illustrate human irrationality. An alternative approach is to try to explain how it is that coincidences lead to meaningful scientific discoveries. Taking this perspective suggests that we might gain insight into what constitutes a coincidence by considering their role in causal induction. In particular, we can use the framework developed in the preceding chapters to try to explain what makes an event a coincidence.

In this section, I will use the problem of causal induction to develop an account of what makes an event a coincidence, and spell out how this account can explain the paradoxical nature of coincidences. I will then provide a more detailed formal analysis of one simple kind of coincidence – coincidences in coinflips – indicating how this account differs from the idea that coincidences are unlikely events. The section ends by identifying the empirical predictions made by this account, which are tested in the remainder of the chapter.

7.2.1 What makes a coincidence?

Assume that a learner has data $\mathcal{D}$, and two hypotheses as to the nature of the system that produced that data, h_1 and h_0. In the problem of causal induction considered thus far – that of inferring causal structure from data – h_1 and h_0 might refer to causal graphical models. Under other circumstances, h_1 and h_0 could be hypotheses defined at higher levels, such as two causal theories. Given these hypotheses, the posterior odds in favor of h_1 can be computed by applying Bayes' rule:

$$\frac{P(h_1|\mathcal{D})}{P(h_0|\mathcal{D})} = \frac{P(\mathcal{D}|h_1)}{P(\mathcal{D}|h_0)} \frac{P(h_1)}{P(h_0)}. \tag{7.2}$$

This equation identifies the two factors that determine the posterior odds: the *likelihood ratio*, which indicates the support that $\mathcal{D}$ provides in favor of h_1 over h_0, and the *prior*

odds, which express the a priori plausibility of h_1 as compared to h_0.

To make this analysis more concrete, consider the specific example of evaluating whether a new form of genetic engineering influences the sex of rats. The treatment is tested through a series of experiments in which female rats receive a prenatal injection of a chemical, and the sex of their offspring is recorded at birth. In the formal schema above, h_1 is a causal graphical model in which injection of the chemical influences sex, and h_0 is a causal graphical model in which injection and sex are independent, structurally equivalent to Graph 1 and Graph 0 in Figure 4.3. Under Graph 0, the probability that a rat is male should be 0.5, while under Graph 1, rats injected with the chemical have some other probability of being male. Imagine that in the experimental test, the first eight rats were all born male. This would provide relatively strong support for the existence of a causal relationship, such a relationship seems a priori plausible, and as a consequence you might be inclined to conclude that the relationship exists.

Now contrast this with a different case of causal induction. Your friend insists that she possesses the power of psychokinesis. To test her claim, you flip a coin in front of her while she attempts to influence the outcome. Again, the hypotheses h_1 and h_0 are equivalent to Graph 1 and Graph 0. Under Graph 1, she can influence the probability that a coin comes up heads. Under Graph 0, she cannot. The first eight flips are all heads. The likelihood ratio provides just as much support for a causal relationship as in the genetic engineering example, but the existence of such a relationship has lower prior probability. As a consequence, you might conclude that she does not possess psychic powers, and that the surprising outcome of the coinflips was just a coincidence.

This example illustrates the central claim of this chapter: that coincidences are events that provide support for a hypothesis, but not enough support to convince us to accept that hypothesis. This definition can be formalized using the Bayesian machinery introduced above. In most cases of causal induction, such as establishing whether a chemical influences the sex of rats, we learn about causal relationships that seem relatively plausible, and the likelihood ratio and prior odds are roughly in agreement. A coincidence is an event that produces a likelihood ratio in favor of h_1 that is insufficient to overwhelm the prior odds against h_1, resulting in middling posterior odds. The likelihood ratio provides a measure of the *strength* of a coincidence, indicating how much support the event provides for h_1. Under this definition, the strongest coincidences can only be obtained in settings where the prior odds are equally strongly against h_1. Thus, like the test of psychokinesis,

canonical coincidences typically involve data that produce a high likelihood ratio in favor of a hypothesis that has low prior odds.[4]

This account of coincidences allows us to begin to address the apparent paradox of coincidences with which I began the chapter, explaining why coincidences are relevant to scientific discovery, and why they often lead us to false conclusions. The answer to both of these questions is that, under my definition, coincidences provide an opportunity to make a discovery that is inconsistent with our current account of how the world works. The low prior odds in favor of h_1 indicates that this hypothesis is rendered implausible by the remainder of a learner's knowledge, while the high likelihood ratio suggests that h_1 should be taken seriously. In the case of scientific discovery, an accepted scientific theory might endorse h_0, while h_1 gives a more accurate description of the world. Making the discovery that h_1 is the better hypothesis requires collecting data that favor h_1. Thus, by necessity, scientific discoveries will involve coincidences, since such data will have a high likelihood ratio in favor of a hypothesis that has low prior odds.

The fact that many coincidences support false hypotheses is a by-product of modern adults having a relatively accurate intuitive account of how the world works, h_0. If our understanding of the world is accurate, then coincidences can only be false alarms: cases where events that arise by chance provide support for a different hypothesis, h_1. Our susceptibility to being misled by coincidences is thus partly a consequence of our success in causal discovery making one of the major sources of clues redundant. For anybody with a less accurate account of how the world works than a modern adult, such as an early scientist or a young child, coincidences are a rich source of information as to how a theory might be revised, and should be given great attention. This account also explains why many of

[4]This analysis of coincidences has strong parallels with an account developed independently in philosophy of science by Horwich (1982). In reviewing Horwich's book, Good (1984) indicated that he held a similar view of coincidences, stating that "we tend to be surprised ... when the result of an observation has much greater probability on some other, not entirely untenable, hypothesis" (p. 164). In his *Philosophical Essay on Probabilities*, Pierre Simon Laplace, one of the fathers of Bayesian statistics, presented a similar view:

> If we seek a cause wherever we perceive symmetry, it is not that we regard a symmetrical event as less possible than the others, but, since this event ought to be the effect of a regular cause or that of chance, the first of these suppositions is more probable than the second. On a table we see letters arranged in this order, *C o n s t a n t i n o p l e*, and we judge that this arrangement is not the result of chance, not because it is less possible than the others, for if this word were not employed in any language we should not suspect it came from any particular cause, but this word being in use among us, it is incomparably more probable that some person has thus arranged the aforesaid letters than that this arrangement is due to chance. (Laplace, 1795/1951, p. 16)

However, Laplace did not make the connection to coincidences explicit.

the most compelling coincidences, such as the September 11 lottery results, are associated with mysticism. Since h_0 represents the sum of our knowledge of nature, h_1 will have to postulate the existence of a supernatural force.

The relevance of coincidences to scientific discovery provides the basis for giving h_1 and h_0 the vague title of "hypotheses", rather than identifying them with specific causal structures, as in the previous chapters. While each coincidental event may involve an inference about a specific causal structure, the implications of these inferences typically reach much further. For example, evaluating your friend's claims about psychokinesis can be formulated as a comparison of two causal graphical models, one in which a causal relationship exists and one in which it does not, but the discovery that your friend actually has psychokinetic powers would not just influence your beliefs about this specific relationship, but about what kind of forces operate in the world. If your friend has such powers, you might find it more likely that others would have them, and possibly even reconsider some other mystical phenomena you may have dismissed. This reaction is qualitatively different from that produced by learning that a chemical influences the sex of rats, which seems to have few other implications. The difference arises because psychokinesis is inconsistent with your current understanding of how the world works, and its existence suggests that the theory should itself be revised. For many coincidences, h_1 and h_0 do not just concern individual causal relationships, but completely different causal theories. I will return to the question of how coincidences guide theory change later in the chapter.

7.2.2 Coincidences in coinflips

Having laid the conceptual groundwork for this account of coincidences, I will make these claims precise by analyzing one simple setting in which coincidences arise: flipping a coin. This analysis helps to clarify how the Bayesian account given above relates to the idea that coincidences are events of unlikely kinds. I will use the theory-based causal induction framework developed in the preceding chapters to analyze inferences about the efficacy of psychokinesis on the basis of coinflips.

Figure 7.1 show two theories about the efficacy of psychokinesis. One theory, h_0, stipulates that there can be no relationship between thinking about a coin, represented by `Meditating(P,C,T)` for some `Person P`, `Coin C`, and `Trial T`, and whether the coin comes up heads, represented by `Heads(C,T)`. The other theory, h_1, stipulates that anybody can influence the outcome of a coin toss by focussing their mind appropriately, and specifies

h_0:

Ontology:

Types	Number	Predicates	Values
`Person`	$N_P \sim P_P$	`Meditating(Person,Coin,Trial)`	`Boolean` : {H, T}
`Coin`	$N_C \sim P_C$	`Heads(Coin,Trial)`	`Boolean` : {H, T}
`Trial`	$N_T \sim P_T$		

Plausible relations:

Functional form:

$$\begin{array}{lcl} \texttt{Meditating(P,C,T)} & \sim & \text{Bernoulli}(\cdot) \\ \texttt{Heads(C,T)} & \sim & \text{Bernoulli}(0.5) \end{array}$$

h_1:

Ontology:

Types	Number	Predicates	Values
`Person`	$N_P \sim P_P$	`Meditating(Person,Coin,Trial)`	`Boolean` : {H, T}
`Coin`	$N_C \sim P_C$	`Heads(Coin,Trial)`	`Boolean` : {H, T}
`Trial`	$N_T \sim P_T$		

Plausible relations:

`Meditating(P,C,T)` $\rightarrow$ `Heads(C,T)`
True for all `T` for each `P`, `C` pair.

Functional form:

$$\begin{array}{lcl} \texttt{Meditating(P,C,T)} & \sim & \text{Bernoulli}(\cdot) \\ \texttt{Heads(C,T)} & \sim & \begin{cases} \text{Bernoulli}(0.5) & \texttt{Meditating(P,C,T) = F} \\ \text{Bernoulli}(\omega) & \texttt{Meditating(P,C,T) = T} \end{cases} \end{array}$$

Figure 7.1: Theories for coincidences in coinflipping.

probability of the coin coming up heads under such influence using a parameter ω.

Given one `Person` and one `Coin`, each of these theories generates one causal graphical model: h_0 generates a graphical model structurally equivalent to Graph 0 from Figure 4.3, while h_1 generates a graphical model structurally equivalent to Graph 1.[5] Assume that the data, $\mathcal{D}$, consists of N_T trials in the presence of somebody concentrating on a coin, of which N trials produce heads. Since h_0 asserts that these outcomes are all the result of chance,

[5]If the two theories were not quite so extreme in the probabilities that they assigned to the existence of a relationship, they would each generate two graphical models, equivalent to Graph 0 and Graph 1, but assign those models different prior probabilities. The inference comparing these hypotheses could still be performed, but would require summing over the set of causal graphical models generated by each theory. I discuss the evaluation of theories that generate multiple causal graphical models later in the chapter.

$P(\mathcal{D}|h_0)$ is just $\binom{N_T}{N} 0.5^{N_T}$. Evaluating $P(\mathcal{D}|h_1)$ requires making assumptions about the parameter ω. If we define a prior distribution $P(\omega)$, then we can compute

$$\begin{aligned} P(\mathcal{D}|h_1) &= \int_0^1 P(\mathcal{D}|\omega, h_1)P(\omega)\ d\omega \\ &= \int_0^1 \binom{N_T}{N} \omega^N (1-\omega)^{N_T-N} P(\omega)\ d\omega. \end{aligned}$$

Taking $P(\omega)$ to be a uniform distribution over the range $[0, 1]$, we obtain

$$P(\mathcal{D}|h_1) = \frac{1}{N_T + 1}$$

from which it follows that the likelihood ratio in favor of h_1 is

$$\frac{P(\mathcal{D}|h_1)}{P(\mathcal{D}|h_0)} = \frac{2^{N_T}}{\binom{N_T}{N}(N_T + 1)} \tag{7.3}$$

which increasingly favors h_1 as N deviates from $N_T/2$.

Since the term $\binom{N_T}{N}$ appears in both $P(\mathcal{D}|h_1)$ and $P(\mathcal{D}|h_0)$, Equation 7.3 also gives the likelihood ratio in favor of h_1 for $\mathcal{D}$ consisting of any particular sequence of N_T coinflips with N heads. This expression can be rewritten as

$$\frac{P(\mathcal{D}|h_1)}{P(\mathcal{D}|h_0)} = \frac{1}{P_{\text{kind}}(\mathcal{D})(N_T + 1)}$$

where $P_{\text{kind}}(\mathcal{D})$ is defined in Equation 7.1, being the probability of a sequence of the same "kind" as $\mathcal{D}$, where kinds of sequence are differentiated by the number of heads in the sequence. Consequently, the support for h_1, and the strength of the coincidence associated with $\mathcal{D}$, will increase as the probability of a sequence of the same kind as $\mathcal{D}$ decreases. This is consistent with the "unlikely kinds" account of coincidences. This observation reveals why it is possible to construct examples that are broadly consistent with the "unlikely kinds" account of coincidences: it approximates the Bayesian solution to this problem.

Despite this connection, the Bayesian account does not face the same difficulties as the "unlikely kinds" account of coincidences. Firstly, it automatically indicates which kinds are going to be relevant to evaluating coincidences. This is a consequence of formulating the problem as a comparison of two hypotheses. In this example, the kinds are differentiated by the number of heads in a sequence because h_1 and h_0 differ in the probabilities with which

they predict a coin will produce heads. In other settings, the set of kinds will be defined based upon the kind of regularities that can discriminate between the two hypotheses: the kinds considered in a given inference will be those that have implications for the underlying causal structure. Since h_1 and h_0 are defined in terms of probability distributions, the Bayesian account extends naturally to continuous stimuli, as I will demonstrate later in the chapter, unlike the "unlikely kinds" account. The formulation of the comparison of these hypotheses as a Bayesian inference also implicitly solves the problems with multiple kinds, and removes other technical problems. For example, the appearance of the $(N_T + 1)$ term in the denominator of Equation 7.3 corrects for the fact that there are many more kinds of longer sequences when kinds are differentiated by the number of heads. This is the issue that made it possible for a sequence of the kind *fifteen heads, eight tails* to be less likely than a sequence of the kind *four heads*. Under Equation 7.3, the former provides weaker support for h_1 than the latter, as there are 24 kinds of sequence of length 23, and only 5 kinds of sequence of length 4.

7.2.3 Empirical predictions

Having given a precise definition of what constitutes a coincidence, we can evaluate how well this definition accords with human judgments. The Bayesian account presented above makes clear empirical predictions. First and foremost, an event will be considered a coincidence when the likelihood ratio in favor of a hypothesis h_1 is high, but insufficient to overwhelm the prior odds against h_1. If either the likelihood ratio or the prior odds increase, it will simply be considered evidence for that hypothesis. I discuss this prediction in more detail in the next section, and test it in Experiment 7.1 for the case of coinflipping. I then consider how this account can be extended to some of the more complex settings that have featured in arguments about the rationality of the human sense of coincidence, assessing the adequacy of the likelihood ratio in favor of h_1 as a measure of the strength of coincidences in Experiments 7.2 and 7.3.

7.3 The transition from coincidence to evidence

> "Well, Watson, what do you make of this?" asked Holmes, after a long pause.
> "It is an amazing coincidence."
> "A coincidence! Here is one of the three men whom we had named as possible

> actors in this drama, and he meets a violent death during the very hours when we know that that drama was being enacted. The odds are enormous against its being coincidence. No figures could express them. No, my dear Watson, the two events are connected – *must* be connected. It is for us to find the connection."
>
> Sir Arthur Conan Doyle (1986a), *The adventure of the second stain*, p. 909.

What seems like a coincidence to one person can be considered compelling evidence by another.[6] In the analysis given above, whether an event is a coincidence or simply evidence for a hypothesis comes down to whether it ultimately justifies believing in the hypothesis, the result of an interaction between likelihood ratio and prior. Holmes and Watson could thus differ in their construal of a violent death if they differed in the probabilities with which they thought such an event might arise independently or as the result of a connection to their case, or if they differed in the prior probability they assigned to the existence of such a connection.

Under this account, events make a transition from coincidence to evidence as the posterior odds in favor of h_1 increase. If the posterior odds exceed a threshold, an event ceases being a coincidence and simply becomes evidence, as illustrated schematically in Figure 7.2. The figure also identifies an intermediate stage between considering an event a mere coincidence and finding it compelling evidence. This intermediate stage consists of "suspicious coincidences": events that we can neither dismiss nor accept as evidence, but suggest further investigation may be appropriate. The region in which events are considered suspicious coincidences should contain posterior odds of 1, indicating no preference for h_1 or h_0, with the thresholds for mere coincidence and evidence corrsponding to posterior odds slightly above and below this value.

As indicated in Equation 7.2, the posterior odds increase if either the prior odds or the likelihood ratio increases. Such changes can thus result in a transition from coincidence to evidence. An example of the former was provided above: eight male rats in a row seems like evidence in the context of a genetic engineering experiment, but eight heads in a row is mere coincidence in a test of psychokinesis, where the prior odds are smaller. Tests of psychokinesis can also be used to illustrate how a change in the likelihood ratio can produce a transition from mere coincidence, to suspicious coincidence, to evidence: eight heads in a

[6] Coincidences played an important role in the "logical" method of deduction endorsed by Sherlock Holmes, with the notion appearing in 13 of his 60 published cases.

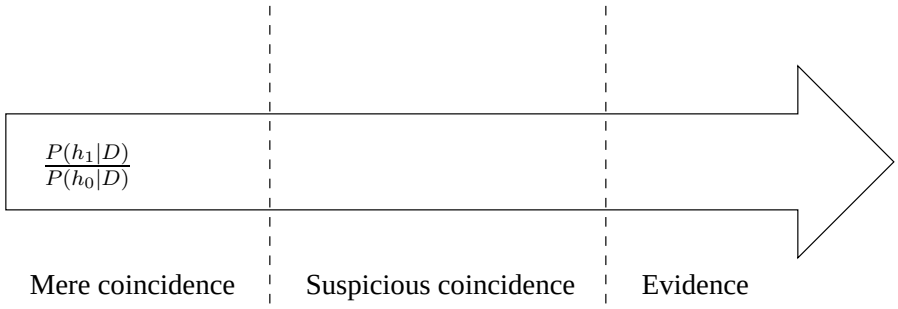

Figure 7.2: Mere and suspicious coincidences both feature a high likelihood ratio and low prior odds in favor of h_1, but in suspicious coincidences the posterior odds exceed a threshold that makes it seem possible that h_1 could actually be true.

row is a mere coincidence, but sixteen might begin to raise suspicions about your friend's powers, or the fairness of the coin. At ninety heads in a row you might, like Guildenstern in Stoppard's (1967) play, begin entertaining the possibility of divine intervention, having relatively unambiguous evidence that something unusual is taking place.

Experiment 7.1 was designed to examine this transition from coincidence to evidence. The experiment uses the two scenarios discussed above – genetic engineering and psychokinesis – to assess whether people's designation of events as mere or suspicious coincidences is affected by changes in the likelihood ratio and prior odds. If an event is judged "just a coincidence" when it provides insufficient support to overcome the prior, we should expect to see events with higher likelihood ratios considered a mere coincidence when people are evaluating claims about psychokinesis. More specifically, if people's assessment of events as coincidences or evidence is based upon the posterior probability of h_1, we should expect to see a negative correlation between this posterior probability and the proportion of people who consider an event a coincidence. Since these predictions rely upon a subtle interaction between likelihood ratio and prior, they are inconsistent with accounts of coincidences that do not incorporate both of these components, such as the definition of coincidences as events of unlikely kinds.

7.3.1 Experiment 7.1: Psychokinesis and genetics

Method

Participants. Participants were 101 undergraduates from Stanford University, participating for course credit. Of these participants, 24 were assigned to the *psychokinesis,posterior* condition, 20 to the *genetics*, *posterior* condition, 28 to the *psychokinesis*, *coincidence* condition, and 29 to the *genetics*, *coincidence* condition.

Stimuli. Two basic cover stories were constructed that would allow the same data to be presented in different contexts. The data consisted of a table of frequencies that showed how many times a heads or tails (males or females) were produced from 100 trials. These data showed 8 trials on which 47, 51, 55, 59, 63, 70, 87, and 99 heads (males) were obtained. Participants receiving the *psychokinesis* cover story saw:

> A group of scientists investigating paranormal phenomena have conducted a series of experiments testing people who claim to possess psychic powers. All of these people say that they have psychokinetic abilities: they believe that they can influence the outcome of a coin toss. The scientists tested this claim by flipping a fair coin 100 times in front of each person as they focus their psychic energies. Under normal circumstances, a fair coin produces heads and tails with equal probability. The results of these experiments are shown below: the identities of the people are concealed with subject numbers, but you are given the number of times the coin came up heads or tails while that person was focusing their psychic energies.

while those receiving the *genetics* cover story saw:

> A group of scientists investigating genetic engineering have conducted a series of experiments testing drugs that influence the development of rat fetuses. All of these drugs are supposed to affect the sex chromosome: they are intended to affect whether rats are born male or female. The scientists tested this claim by producing 100 baby rats from mothers treated with the drugs. Under normal circumstances, male and female rats are equally likely to be born. The results of these experiments are shown below: the identities of the drugs are concealed with numbers, but you are given the number of times male or female rats were produced by mothers treated with each drug.

These cover stories were presented with the data in a short questionnaire, together with further instructions on how to respond to the stimuli.

Procedure. Each participant received a questionnaire listing the eight target data sets in one of two random orders. Orthogonal to the manipulation of the cover story, participants either received the *posterior* or the *coincidence* instructions. The *posterior* instructions for the *psychokinesis* condition were:

> For each of the lines below, please rate HOW LIKELY you think it is that the person has psychic powers, taking into account the results of the experiment. Use a scale from 1 to 10, where 1 indicates NOT AT ALL LIKELY and 10 indicates EXTREMELY LIKELY.

Likewise, the instructions for the *genetics* condition were:

> For each of the lines below, please rate HOW LIKELY you think it is that the drug affects the sex of rats, taking into account the results of the experiment. Use a scale from 1 to 10, where 1 indicates NOT AT ALL LIKELY and 10 indicates EXTREMELY LIKELY.

The eight sets of frequencies were accompanied by lines on which participants could write their responses. The *coincidence* instructions for the *psychokinesis* condition asked people to choose between a mere coincidence and evidence:

> For each of the lines below, please decide whether you think the results for that person are JUST A COINCIDENCE, or COMPELLING EVIDENCE for them having psychic powers, by checking either the COINCIDENCE or the EVIDENCE box.

Similarly, the instructions for the *genetics* condition were:

> For each of the lines below, please decide whether you think the results are JUST A COINCIDENCE, or COMPELLING EVIDENCE for that drug influencing the sex chromosome, by checking either the COINCIDENCE or the EVIDENCE box.

The eight sets of frequencies were listed with checkboxes to allow participants to indicate their responses.

Results and Discussion

One participant in the *genetics* condition and two in the *psychokinesis* condition appeared to reverse the rating scale, and were eliminated from the analysis. The results are shown in Figure 7.3. The *posterior* ratings were subjected to a two-way between-within ANOVA examining the effects of condition (*psychokinesis*, *genetics*) and varying frequency. There was a main effect of condition ($F(1, 39) = 9.30$, $MSE = 13.10$, $p < .01$), a main effect of frequency ($F(7, 273) = 91.60$, $MSE = 3.31$, $p < .0001$), and an interaction between the two ($F(7, 273) = 7.86$, $MSE = 3.31$, $p < .0001$). As can be seen from the figure, the rated probability of the conclusion went up as frequency increased, but did so earlier for the *genetics* than the *psychokinesis* condition. The same analysis was performed for the *coincidence* assessments, showing a main effect of condition ($F(1, 55) = 18.78$, $MSE = 0.18$, $p < .0001$), a main effect of frequency ($F(7, 385) = 99.01$, $MSE = 0.08$, $p < .0001$), and an interaction between the two ($F(7, 385) = 7.39$, $MSE = 0.08$, $p < .0001$). These results are due to a similar pattern of effects: the proportion of cases classified as coincidences decreased as the frequency increased, but earlier for the *genetics* than the *psychokinesis* condition.

As predicted, there was a close correspondence between the proportion of cases classified as a mere coincidence and the mean posterior probability of the regular generating process, with a linear correlation of $r = -0.98$. In fact, points that are equivalent in posterior probability are also equivalent in the proportion of cases that were classified as coincidences. Examining Figure 7.3 closely, it can be seen that 87 heads and 63 males produce the same results in both graphs, as do 63 heads and 59 males, and 99 heads and 70 males. This relationship holds despite the fact that responses were binary in one condition and continuous in the other, and obtained from completely different participants.

The assumption that there is a threshold on the posterior odds that determines whether an event is a coincidence or evidence, as indicated in Figure 7.2, suggests that these judgments might be modeled using a sigmoid (logistic) function of the posterior odds

$$P(\text{“evidence”}|\mathcal{D}) = \frac{1}{1 + \exp\left\{-g \log \frac{P(h_1|\mathcal{D})}{P(h_0|\mathcal{D})} - b\right\}} \tag{7.4}$$

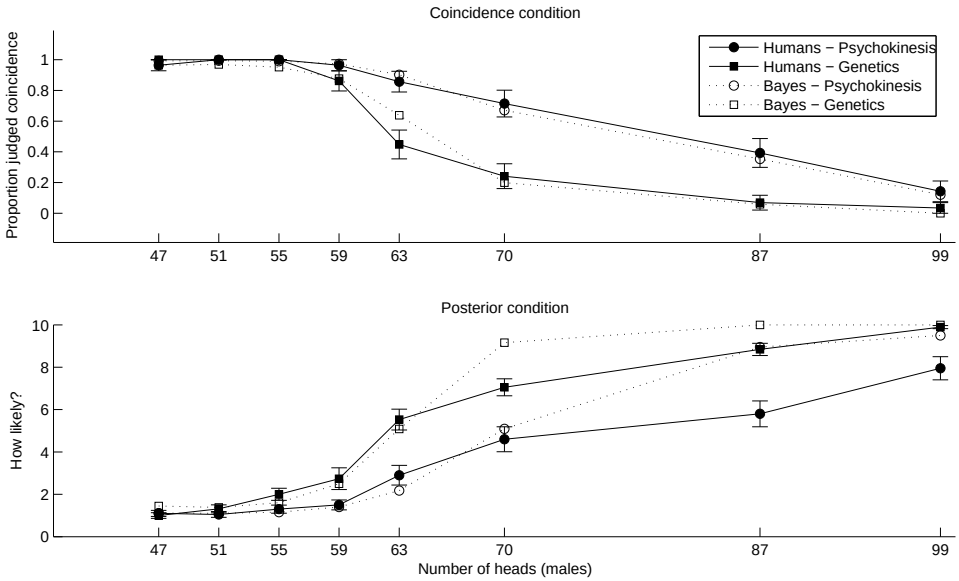

Figure 7.3: Results of Experiment 7.1. The upper panel shows the proportion of cases judged to be coincidences in the *coincidence* condition, and the lower panel shows the mean responses in the *posterior* condition. Dotted lines show model predictions, obtained by estimating prior probabilities for each participant.

where g is the gain of the sigmoid, and b is the bias. As $g \to \infty$, this becomes a step function at the point b. I will assume that $g = 1$ and $b = 0$, meaning that $P(\text{“evidence”}|\mathcal{D})$ is equal to $P(h_1|\mathcal{D})$. Since the likelihood ratio $\frac{P(\mathcal{D}|h_1)}{P(\mathcal{D}|h_0)}$ is given by Equation 7.3, we can estimate the prior odds for each participant by fitting the sigmoid function to their responses, and thus obtain the prior $P(h_1)$.

In the *coincidence* condition, all but one of the participants responded in a fashion consistent with thresholding the posterior odds. It was thus simple to find the value of the prior odds for each participant that maximizes the probability of their responses as predicted by Equation 7.4. This results in a model fit for each participant, and the quality of these fits can be seen from the mean model predictions shown in the upper panel of Figure 7.3. The median values of $P(h_1)$ for the *psychokinesis* and *genetics* conditions were were 0.0004 and 0.23 respectively.

A similar procedure can be used to estimate the prior odds directly from the posterior probabilities provided by the participants in the *posterior* condition. Again fitting a sigmoid function for each participant, this time relative to the squared error, we obtain the fits shown in the lower panel of Figure 7.3. People's more extreme probability judgments can

be seen to be more conservative than those predicted by our Bayesian model, consistent with previous research (e.g., Edwards, 1968). However, this procedure yields similar median values for $P(h_1)$: 0.0011 in the *psychokinesis* condition and 0.20 in the *genetics* condition. Contrary to previous results illustrating deficits in the ability to combine likelihood ratios with prior odds (e.g., Kahneman & Tversky, 1972), people seem quite accurate in assessing the posterior probabilities of causal relationships. This may be a consequence of using priors that are derived from extended experience, rather than base-rates provided in an experimental scenario (c.f. Evans, Handley, Over & Perham, 2002).

The results of this experiment are consistent with the predictions of my Bayesian account of coincidences. Data that provided the same support for h_1 were judged to be coincidences if presented as the results of a test of psychokinesis, and evidence if presented as the results of a test of genetic engineering. The proportion of people who considered an event a coincidence showed a direct correspondence to the posterior probability, with the difference between the two conditions resulting from a difference in the prior probability of a causal relationship. Assuming that people are accurately evaluating the likelihood ratio in favor of h_1 allows us to assess the values of these prior beliefs, which are consistent across experimental procedures and with our intuitions about the efficacy of psychic powers and genetic engineering.

7.4 The strength of coincidences

Experiment 7.1 suggests that the basic constituents of my definition of coincidences are correct: that events are considered a coincidence when they provide support for a hypothesis that is insufficient to convince us of its truth. We can now examine these constituents more carefully. Under this account of coincidences, the likelihood ratio indicates the strength of a coincidence, with higher likelihood ratios indicating more compelling coincidences. In the analysis given in the previous section, I assumed that the likelihood ratio given in Equation 7.3 accurately captured people's assessment of the support that $\mathcal{D}$ gave for h_1 over h_0. Whether people's assessment of the strength of coincidences corresponds to the likelihood ratio in favor of h_1 is an empirical question.

In exploring this question, we have the opportunity to examine people's assessment of coincidences in more realistic settings. The simplicity of coinflipping makes it an effective example with which to explore formal models, but real coincidences, such as the bombing of London, often involve more complex data and more elaborate hypotheses. In these

cases, detecting a coincidence does not just involve recognizing an unusual pattern, but doing so despite the presence of observations that do not express that pattern. These sophisticated inductive inferences have parallels in other aspects of cognition. For example, many problems that arise in cognitive development have exactly this character, requiring a child to notice a regularity that is expressed in only a subset of the data. One such case is word learning: young children are able to learn the relationship between the use of words and the appearance of the objects they identify, despite the fact that only about 70% of the uses of a word by a parent occur when the child is attending to the relevant object (Collins, 1977; Harris, Jones, & Grant, 1983).

I will examine people's judgments about the strength of coincidences from two different kinds of data: spatial data, consisting of the locations of bombs, and temporal data, concerning the dates of birthdays. These two cases were chosen because they have connections to two of the most prominent examples that are used to argue for the irrationality of human reasoning about coincidences: the "birthday problem" and the bombing of London. I will show that people's assessment of the strength of coincidences corresponds extremely well with the rational predictions of the Bayesian account developed above, suggesting that this apparent irrationality must result from a different aspect of people's reasoning about coincidences. I make some suggestions as to the locus of this irrationality towards the end of the chapter.

7.4.1 Coincidences in date

How often have you been surprised to discover that two people share the same birthday? Matching birthdays are a canonical form of coincidence, and are often used to demonstrate errors in human intuitions about chance. The "birthday problem" – evaluating the number of people that need to be in a room to provide a 50% chance of two sharing the same birthday – is a common topic in introductory statistics classes, since students are often surprised to discover that the answer is only 23 people. In general, the number of people required to have a 50% chance of a match on a variable with k alternatives is approximately $\sqrt{k}$, since there are $\binom{N_P}{2} \approx N_P^2$ opportunities for a match between N_P people. Using a set of problems of this form that varied in k, Matthews and Blackmore (1995) found that people expect N_P to increase linearly with k, explaining why such problems produce surprising results. Diaconis and Mosteller (1989) argued that many coincidences are of similar form to the birthday problem, and that people's faulty intuitions about such problems are one

source of errors in reasoning about coincidences.

In this section, I will examine how people evaluate coincidences in date, through a novel "birthday problem": assessing how big a coincidence it would be to meet a group of people with a particular set of birthdays. In contrast with the tasks that have been used to argue that coincidences are an instance of human irrationality, this is not an objective probability judgment. It is a subjective response, asking people to express their intuitions. In many ways, this is a more natural task than assessing the probability of an event. It is also, under my characterization of the nature of coincidences, a more useful one: knowing the probability of an event is generally less useful than knowing how much evidence it provides for a hypothesis. By examining the structure of these subjective responses, we have the opportunity to understand the principles that guide them.

Imagine you went to a party, and met people with a set of birthdays such as {August 3, August 3, August 3, August 3}. Figure 7.4 gives two theories that might be used to explain such an event. One theory, h_0, asserts that the presence of people at the party is independent of their birthday. This theory generates one causal graphical model for any number of people N_P, which is denoted Graph 0 in Figure 7.5. The other theory, h_1, suggests that the presence of some subset of the people at the party was dependent upon their birthdays. As with the theory of bombing presented above, this theory generates 2^{N_P} causal graphical models for N_P people, consisting of all partitions of those people into subsets whose presence either depends or does not depend upon their birthday. Both graphs shown in Figure 7.5 are generated by h_1 with $N_P = 6$. A priori, h_0 seems far more likely than h_1, so a set of birthdays that provides support for h_1 constitutes a coincidence.

The data $\mathcal{D}$ in this setting consists of the birthdays of the people encountered at the party. Since only the people present at the party can be encountered, these are `conditional` data. If B_i indicates `Birthday`($\mathtt{p}_i$) and P_i indicates `Present`($\mathtt{p}_i$) for `Person` $\mathtt{p}_i$, our data are the values of B_i conditioned on $P_i = p_i^+$ for all i. Under h_0, B_i and P_i are independent, so we have

$$P(\mathcal{D}|h_0) = \left(\frac{1}{365}\right)^{N_P^+} \tag{7.5}$$

where N_P^+ is the number of people who are present at the party.

Evaluating $P(\mathcal{D}|h_1)$ is slightly more complicated, due to the possible dependence of B_i on P_i and the functional form of that dependence. Under h_1, if `Birthday`(`P`) influences `Present`(`P`), then `Present`(`P`) is true with probability α only if `Birthday`(`P`) belongs to the

h_0:

Ontology:

Types	Number	Predicates	Values
`Person`	$N_P \sim P_P$	`Present(Person)`	`Boolean` : $\{\mathtt{T}, \mathtt{F}\}$
		`Birthday(Person)`	`Date` : $\{1, \ldots, 365\}$

Plausible relations:

Functional form:

$$\begin{array}{lcl} \mathtt{Birthday(P)} & \sim & \text{Uniform}(\{1, \ldots, 365\}) \\ \mathtt{Present(P)} & \sim & \text{Bernoulli}(\alpha) \end{array}$$

h_1:

Ontology:

Types	Number	Predicates	Values
`Person`	$N_P \sim P_P$	`Present(Person)`	`Boolean` : $\{\mathtt{T}, \mathtt{F}\}$
		`Birthday(Person)`	`Date` : $\{1, \ldots, 365\}$

Plausible relations:

`Birthday(P)`→`Present(P)`
True with probability p for each `P`.

Functional form:

$$\begin{array}{lcl} \mathtt{Birthday(P)} & \sim & \text{Uniform}(\{1, \ldots, 365\}) \\ \mathtt{Present(P)} & \sim & \begin{cases} \text{Bernoulli}(\alpha) & \mathtt{Birthday(P)} \rightarrow \mathtt{Present(P)} \text{ and } \mathtt{Birthday(P)} \in \mathcal{B} \\ \text{Bernoulli}(0) & \mathtt{Birthday(P)} \rightarrow \mathtt{Present(P)} \text{ and } \mathtt{Birthday(P)} \notin \mathcal{B} \\ \text{Bernoulli}(\alpha) & \text{otherwise} \end{cases} \end{array}$$

Figure 7.4: Theories for coincidences in birthdays.

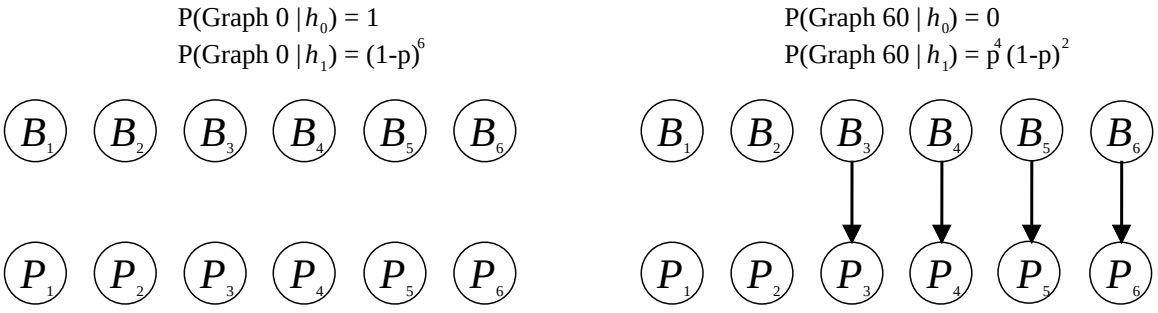

Figure 7.5: Causal graphical models generated by theories of birthdays. B_i indicates `Birthday`($\mathtt{p}_i$), and P_i indicates `Present`($\mathtt{p}_i$)

set $\mathcal{B}$. This set indicates the "filter" that was applied to determine whether people would be invited to the party, identifying some subset of admissable dates. Computing $P(\mathcal{D}|h_1)$ requires making some assumptions about the nature of $\mathcal{B}$.

As a first step towards evaluating $P(\mathcal{D}|h_1)$, we can consider the probability of $\mathcal{D}$ conditioned on a particular $\mathcal{B}$. There are two possibilities for the component of the causal structure that corresponds to each person $\mathtt{p}_i$: with probability $1-p$, B_i and P_i are independent, and with probability p, B_i and P_i are dependent. If B_i and P_i are independent, the probability distribution of B_i conditioned on P_i is just the unconditional distribution of B_i, which is uniform over $\{1,\ldots,365\}$. If B_i and P_i are dependent, the distribution of B_i conditioned on P_i is uniform over the set $\mathcal{B}$, since P_i has constant probability when $B_i \in \mathcal{B}$ and zero probability otherwise. It follows that the probability distribution for each B_i conditioned on $P_i = p_i^+$ is a mixture of two uniform distributions, and

$$P(\mathcal{D}|\mathcal{B}) = \prod_{i=1}^{N_P^*} \left[\frac{1-p}{365} + I(b_i \in \mathcal{B}) \frac{p}{|\mathcal{B}|} \right] \tag{7.6}$$

where $I(\cdot)$ is an indicator function that takes the value 1 when its argument is true and 0 otherwise, and $|\mathcal{B}|$ is the number of dates in $\mathcal{B}$.

We can use Equation 7.6 to compute $P(\mathcal{D}|h_1)$. If we define a prior, $P(\mathcal{B})$ on filter sets $\mathcal{B}$, we have

$$P(\mathcal{D}|h_1) = \sum_{\mathcal{B}} P(\mathcal{D}|\mathcal{B}) P(\mathcal{B}). \tag{7.7}$$

The extent to which a set of birthdays will provide support for h_1 will thus be influenced by the choice of $P(\mathcal{B})$. We want to define a prior that identifies a relatively intuitive set of filters that might be applied to a set of birthdays to determine the presence of people at a party. An enumeration of such regularities might be: falling on the same day, falling on adjacent days, being from the same calendar month, having the same calendar date (e.g., January 17, March 17, September 17, December 17), and being otherwise close in date. With 365 days in the year, these five categories identify a total of 11,358 different sets $\mathcal{B}$: 365 consisting of a single day in the year, 365 consisting of neighboring days, 12 consisting of calendar months, 31 consisting of specific days of the month, and 10,585 having to do with general proximity in date (from 3-31 days). This is not intended to be an exhaustive set of the kinds of regularities one could find in birthdays, but is a simple choice for the values that $\mathcal{B}$ could take on that allows us to test the predictions of the model. Given this

set, I will define a prior, $P(\mathcal{B})$, by taking a uniform distribution over the hypotheses in the first four categories, and giving all 10,585 hypotheses in the fifth category as much weight as a single hypothesis in one of the first four. Equation 7.7 can then be evaluated numerically by explicitly summing over all of these possibilities.

The second term in Equation 7.6 has an important implication: the influence of a filter $\mathcal{B}$ on the assessment of a coincidence decreases as that filter admits more dates. Thus, while the set {August 3, August 3, August 3, August 3} consists of birthdays that all occur in August, the major contribution to the support for h_1 having been responsible for producing this outcome is the fact that all four birthdays fall on the same day. This sensitivity to the size of the set $\mathcal{B}$ is equivalent to the "size principle" that plays a key role in Bayesian models of concept learning and generalization (Tenenbaum, 1999b; 1999a; Tenenbaum & Griffiths, 2001). The filtering procedure by which people come to be present at the party under h_1 is one means of deriving this size principle.

We can use Equations 7.5 and 7.7 to compute the likelihood ratio $\frac{P(\mathcal{D}|h_1)}{P(\mathcal{D}|h_0)}$ for any set of birthdays. Experiment 7.2 compares this likelihood ratio with human ratings of the strength of coincidence for different sets of birthdays.

7.4.2 Experiment 7.2: Birthdays

Method

Participants. Participants were 93 undergraduate students from Stanford University, participating for course credit.

Stimuli. Stimuli were sets of dates, chosen to allow assessment of the degree of coincidence associated with some of the regularities enumerated above. Fourteen potential relationships between birthdays were examined, using two choices of dates. The sets of dates included: 2, 4, 6, and 8 apparently unrelated birthdays for which each date was chosen from a different month, 2 birthdays on the same day, 2 birthdays in 2 days across a month boundary, 4 birthdays on the same day, 4 birthdays in one week across a month boundary, 4 birthdays in the same calendar month, 4 birthdays with the same calendar dates, and 2 same day, 4 same day, and 4 same date with an additional 4 unrelated birthdays, as well as 4 same week with an additional 2 unrelated birthdays. These dates were delivered in a questionnaire. One choice of dates, in the order specified above, was:

February 25, August 10

February 11, April 6, June 24, September 17
January 23, February 2, April 9, July 12, October 17, December 5
February 22, March 6, May 2, June 13, July 27, September 21, October 18, December 11
May 18, May 18
September 30, October 1
August 3, August 3, August 3, August 3
June 27, June 29, July 1, July 2
January 2, January 13, January 21, January 30
January 17, April 17, June 17, November 17
January 12, March 22, March 22,July 19, October 1, December 8
January 29, April 26, May 5, May 5, May 5, May 5, September 14, November 1
February 12, April 6, May 6, June 27, August 6, October 6, November 15, December 22
March 12, April 28, April 30, May 2, May 4, August 18

Procedure. Participants completed the questionnaire as part of a booklet of other short psychology experiments. Each participant saw one choice of dates, with the regularities occurring in one of six random orders. The instructions on the questionnaire read as follows:

> All of us have experienced surprising events that make us think "Wow, what a coincidence". One context in which we sometimes encounter coincidences is in finding out about people's birthdays. Imagine that you are introduced to various groups of people. With each group of people, you discuss your birthdays. Each of the lines below gives the birthdays of one group, listed in calendar order.
>
> Please rate how big a coincidence the birthdays of each group seem to you. Use a scale from 1 to 10, where 1 means 'Very small (or no) coincidence', and 10 means 'Very big coincidence'.

The sets of dates were then given on separate lines, in calendar order within each line, with a space beside each set for a response.

Results and Discussion

The mean responses for the different stimuli are shown in Figure 7.6. The birthdays differed significantly in their judged coincidentalness ($F(13, 1196) = 185.55$, $MSE = 3.35$, $p < .0001$). The figure also shows the predictions of the Bayesian model. The ordinal correlation between the likelihood ratio $\frac{P(\mathcal{D}|h_1)}{P(\mathcal{D}|h_0)}$ and the human judgments was $\rho = 0.921$.

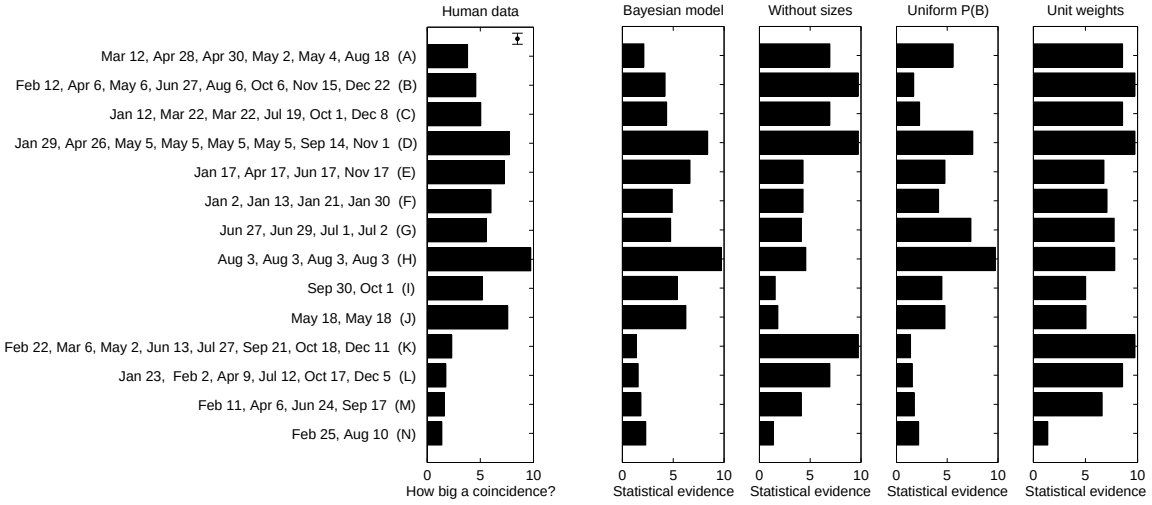

Figure 7.6: The leftmost panel shows the mean judgment of the strength of coincidences from human participants in Experiment 7.3. Error bars indicating one standard error in either direction are shown in the upper right hand corner of the panel. The second panel shows the predictions of the Bayesian model, the third shows the consequences of removing the size principle, and the third shows the consequences of using a uniform prior on filters, $P(\mathcal{B})$. The fifth panel shows the combined effects of these two omissions, illustrating the performance of the model when each filter $\mathcal{B}$ contributes equally to $P(\mathcal{D}|h_1)$.

The values shown in the Figure were obtained by applying a nonlinear transformation, $y = \text{sign}(x)\text{abs}(x)^\gamma$ where x is the log likelihood ratio and $\gamma = 0.60$, which gave a linear correlation of $r = 0.958$.

The predictions of the Bayesian model correspond closely to people's judgments of the strength of coincidences. Each of the parts of this model – the size principle, the set of filters $\mathcal{B}$, and the prior over filters $P(\mathcal{B})$ – contributes to this performance. Figure 7.6 illustrates the contributions of these different components: the panel labelled "Without sizes" shows the effect of removing the size principle; "Uniform $P(\mathcal{B})$" shows the effect of removing $P(\mathcal{B})$; and "Unit weights" shows the effect of removing both of these elements of the model and simply giving equal weight to each filter $\mathcal{B}$ consistent with B_i. I will discuss how each of these modifications reduces the fit of the model to the data, but the basic message is clear: simply specifying a set of regularities is not sufficient to explain people's judgments. The model explains many of the subtleties of people's performance on this task as the result of rational statistical inference.

The "Without sizes" model shown in Figure 7.6 replaces the $\frac{p}{|\mathcal{B}|}$ term in Equation 7.6

with just p, removing the effect of the size principle. The model fit is significantly worse, with a rank-order correlation of $\rho = 0.12$, and $\gamma = 1.00$ giving a linear correlation of $r = -0.079$. The worse fit of this model illustrates the importance of the size of the extension of the judged event in determining the strength of a coincidence, consistent with Falk's (1981-1982; 1989) results. This effect can be seen most clearly by examining the stimuli that consist of four dates: {August 3, August 3, August 3, August 3} is more of a coincidence than {January 17, April 17, June 17, November 17}, which is in turn more of a coincidence than {January 2, January 13, January 21, January 30}. This ordering is consistent with the size of the regularities they express: a set of four birthdays falling on August 3 cover only one date, August 3, while there are 12 dates covered by the hypothesis corresponding to dates falling on the 17th day of the month, and 31 dates covered by the hypothesis corresponding to dates in January.

The size of the extension of the hypothesis is not the only factor influencing the predictions of the Bayesian model. While the size of $\mathcal{B}$ is important in determining $P(\mathcal{D}|h_1)$, the prior $P(\mathcal{B})$ also has a large effect. In the basic model, $P(\mathcal{B})$ gives less weight to the extremely large number of regularities corresponding to intervals of between 3 and 31 days. The importance of this prior over hypotheses is illustrated by the "Uniform $P(\mathcal{B})$" model, which gives equal probability to all of the filters $\mathcal{B}$. This model gives too much weight to the filters that correspond to intervals of dates, resulting in a fit of $\rho = 0.776$, and $r = 0.806$ with $\gamma = 0.80$. The main error made by this model is not predicting the apparent equivalence of {January 17, April 17, June 17, November 17} and {June 27, June 29, July 1, July 2}, despite the fact that the former is of size 12 and the latter of size 7. In the basic model, the effect of the sizes of the regularities is overwhelmed by $P(\mathcal{B})$, corresponding to the fact that dates falling within seven days over a month boundary is not a particularly salient regularity.

The effects of the size principle and $P(\mathcal{B})$ interact in producing the good performance of the basic Bayesian model. These two factors determine which regularities influence the strength of a coincidence. Simply having a sensible set of filters $\mathcal{B}$ provides no guarantee of a good model of coincidence judgments. This can be seen in the "Unit weights" model, in which all filters $\mathcal{B}$ are given unit weight, removing the size principle and using a uniform prior $P(\mathcal{B})$. The model gives a fit of $\rho = 0.099$, and $r = 0.158$ with $\gamma = 0.002$. In this model, the major contributors to the strength of a coincidence are the number of dates and their proximity.

The main discrepancy between the basic Bayesian model and the data is the ordering of the random dates. The model predicts that the longer lists of unrelated dates should be considered less of a coincidence, while people seem to believe the opposite. To explore this curious effect further, we conducted a second survey with a separate group of 73 Stanford undergraduates, showing them a subset of 8 of the 14 stimuli used in the experiment that included the four sets of random dates. The participants were asked to rate the strength of the coincidences, as before, and to state why they gave the rating they did. Of the 73 participants, 49 did not identify any kind of pattern in the random dates, 23 noted a regularity, and one gave a high rating because of a match with her own birthday. The regularity identified by the 23 subjects had to do with the fact that the "random" birthdays were suspiciously evenly spaced throughout the year, not overlapping at all in month or date. This regularity could be accommodated by using a more elaborate set of filters, at the cost of greater complexity of the model.

7.4.3 Coincidences in space

John Snow's inference to the cause of the Broad Street cholera outbreak and the mistaken beliefs of the populace during the bombing of London were both based upon coincidences in space – clusters in the locations of patients and bombs respectively. By studying people's assessment of the strength of coincidences in space, we can gain insight into how people make such inferences. Formulating these coincidences in terms of causal induction also extends the theory-based causal induction framework to another kind of data. In the previous chapters, I considered how causal structure could be inferred from contingency data, rates, and times. Here, the data consist of points in space.

As in the temporal case discussed in Chapter 6, spatial coordinates are continuous. By analogy to discrete trials, we could imagine cutting up space into a discrete set of regions, and defining the probability of a particular event for each region. By the same arguments as those presented in Chapter 6, when we take the limit of the number of regions to infinity we obtain a *spatial* Poisson process, which gives a rate for each point in space. The number of events that occur in a particular region can be computed using this rate, just as the number of events that occur in a particular interval can be computed using the rate of the temporal Poisson process. Consequently, the functional form component of causal theories for spatial events needs to specify how the rate for each variable as a function of its parents.

Figure 7.7 gives two theories of bombing.[7] The theory identified as h_0 assumes that each bomb has its own target, while the theory identified as h_1 allows some bombs to share a common target.[8] The points at which a bomb explodes has a Gaussian distribution around the location of its target, with covariance matrix Σ. Targets are distributed uniformly throughout the region in which bombs fall, $\mathcal{R}$.

The theory h_0 generates only one causal graphical model, denoted Graph 0 in Figure 7.8. In this model, each `Bomb` $\mathtt{b}_i$ has a single `Target` $\mathtt{t}_i$, and the points at which the bombs explode are independent. Using X_i to indicate the point at which $\mathtt{b}_i$ explodes, `ExplosionTime`($\mathtt{b}_i$), and L_i to indicate the location of the target $\mathtt{t}_i$, `Location`($\mathtt{t}_i$`t`), we have

$$\begin{aligned} P(x_i|h_0) &= \int_{\mathcal{R}} P(x_i|\ell_i)P(\ell_i)\ d\ell_i \\ &\propto \int_{\mathcal{R}} \phi_\Sigma(x_i,\ell_i)\ d\ell_i \\ &\approx 1 \end{aligned}$$

where x_i is the value taken by X_i, ℓ_i is the value taken by L_i, and $\phi_\Sigma(x,\ell)$ is the value of the multivariate Gaussian density with mean ℓ and covariance matrix Σ at point x. The approximation in the last line is a consequence of the fact that if ℓ_i is near the boundary of $\mathcal{R}$, some of the mass of $P(x_i|\ell_i)$ will fall outside $\mathcal{R}$, making values of x_i near the boundary slightly less likely. This effect will be negligible if $\mathcal{R}$ is large relative to Σ, so $P(x_i|h_0)$ is well approximated by a uniform distribution over $\mathcal{R}$. With $\mathcal{D}$ consisting of the locations of N_B bombs, we thus have

$$P(\mathcal{D}|h_0) = \left(\frac{1}{|\mathcal{R}|}\right)^{N_B} \tag{7.8}$$

where $|\mathcal{R}|$ is the area of $\mathcal{R}$.

The theory h_1 generates 2^{N_B} causal graphical models, corresponding to each partition of N_B bombs into two sets, one in which each bomb $\mathtt{b}_i$ has a unique target $\mathtt{t}_i$ and one in which

[7] As with the theory of explosions presented in Chapter 6, these theories can be defined entirely in terms of boolean predicates at the loss of some notational efficiency. A version of h_1 in this form appears in Appendix E, and reveals some interesting parallels with the temporal case.

[8] A more elaborate version of h_1 might allow multiple common targets, in line with the theories used to analyze the stick-ball machine and Nitro X. While such a theory might be a more accurate account of people's expectations, evaluating quantities like $P(X|h_1)$ becomes a challenging computational problem. Consequently, in this chapter I restrict myself to cases in which a single common target is sufficient to describe the regularities present in the data.

h_0:

Ontology:

Types	Number	Predicates	Values
`Target`	$N_T = \infty$	`Location(Target)`	`Space` : $\mathcal{R} \subset \mathbb{R}^2$
`Bomb`	$N_B \sim P_B$	`ExplosionPoint(Bomb)`	`Space` : $\mathcal{R} \subset \mathbb{R}^2$

Plausible relations:

`Location(T)`→`ExplosionPoint(B)`
Each `B` has an edge from a unique `T`.

Functional form:

$$\begin{aligned} \texttt{Location(T)} &\sim \text{Uniform}(\mathcal{R}) \\ \texttt{ExplosionPoint(B)} &\sim \text{Gaussian}(\texttt{Location(T)}, \Sigma) \end{aligned}$$

h_1:

Ontology:

Types	Number	Predicates	Values
`Target`	$N_T = \infty$	`Location(Target)`	`Space` : $\mathcal{R} \subset \mathbb{R}^2$
`Bomb`	$N_B \sim P_B$	`ExplosionPoint(Bomb)`	`Space` : $\mathcal{R} \subset \mathbb{R}^2$

Plausible relations:

`Location(T)`→`ExplosionPoint(B)`
Each `B` has an edge from some `T`. With probability p, `T` = $\texttt{t}_\texttt{C}$, a common target, otherwise `T` is unique to `B`.

Functional form:

$$\begin{aligned} \texttt{Location(T)} &\sim \text{Uniform}(\mathcal{R}) \\ \texttt{ExplosionPoint(B)} &\sim \text{Gaussian}(\texttt{Location(T)}, \Sigma) \end{aligned}$$

Figure 7.7: Theories for coincidences in bombing.

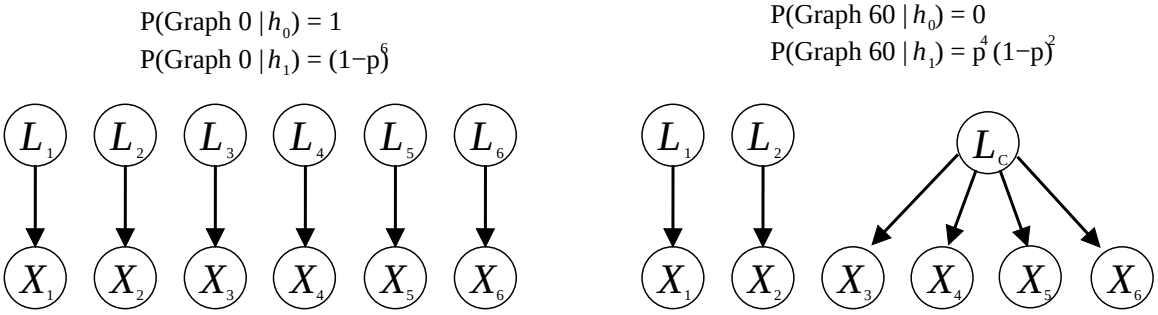

Figure 7.8: Causal graphical models generated by theories of bombing. L_i indicates $\texttt{Location}(\texttt{t}_i)$, X_i indicates $\texttt{ExplosionPoint}(\texttt{b}_i)$, and $\texttt{t}_\texttt{C}$ is the common target.

each bomb shares a common target $\mathtt{t_C}$. Both of the causal graphical models shown in Figure 7.8 are generated by this theory, with $N_B = 6$. Evaluating $P(\mathcal{D}|h_1)$ requires summing over all of these different causal models, a procedure that is discussed in Appendix E. Evaluating this probability is facilitated by the fact that h_1 implies that each X_i is drawn from a *mixture* of a uniform and a Gaussian, giving

$$\begin{aligned} P(\mathcal{D}|\Sigma, p, \ell_{\mathtt{C}}) &= \prod_{i=1}^{N_B} [P(x_i|L_i{\rightarrow}X_i)P(L_i{\rightarrow}X_i|p) + P(X|\Sigma, \ell_{\mathtt{C}}, L_{\mathtt{C}}{\rightarrow}X_i)P(L_{\mathtt{C}}{\rightarrow}X_i|p)] \\ &= \prod_{i=1}^{N_B} \frac{1-p}{|\mathcal{R}|} + p\ \phi_{\Sigma}(x_i, \ell_{\mathtt{C}}) \end{aligned}$$

where $P(L_i{\rightarrow}X_i|p)$ is the probability that $\mathtt{b}_i$ has a unique target, and $P(L_{\mathtt{C}}{\rightarrow}X_i|p)$ is the probability that $\mathtt{b}_i$ shares the common target. Each of these possibilities implies a different distribution for X_i, being uniform and Gaussian respectively, and their probabilities provide the weights with which these distributions are mixed, being $1-p$ and p respectively. Computing $P(\mathcal{D}|h_1)$ thus reduces to the problem of computing the marginal probability of data under a mixture distribution, a problem that has been studied extensively in statistics (e.g., Emond, Raftery, & Steele, 2001).

These results provide the basic constituents of the likelihood ratio, which indicates the support that data $\mathcal{D}$ provide for h_1. Experiment 7.3 was designed to investigate how well this quantity predicts people's assessment of the strength of coincidences in bombing. Participants were informed that h_0 was in fact the correct account of the data, meaning that any support for h_1 would constitute a coincidence.

7.4.4 Experiment 7.3: Bombing

Method

Participants. Participants were 235 undergraduates from Stanford University, participating for course credit.

Stimuli. Stimuli were 12 images containing points at different locations within a 10 by 10 square, ranging from -5 to 5 in two directions. No markers on the axes indicated this scale, but we provide the information to give meaning to the parameters listed below. Nine of these stimuli were generated from a mixture of a uniform and a Gaussian distribution, with parameters selected to span four different dimensions – number of points, proportion of

Table 7.1: Parameters Used in Generating the Stimuli for Experiment 7.3.

Property	Parameters		
Number	$N_B = 20$	$N_B = 50$	$N_B = 200$
Proportion	$p = 0.5$	$p = 0.3$	$p = 0.1$
Location	$\ell_{\mathtt{C}} = \begin{bmatrix} -3 \\ -3 \end{bmatrix}$	$\ell_{\mathtt{C}} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$	$\ell_{\mathtt{C}} = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$
Spread	$\Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$	$\Sigma = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$	$\Sigma = \begin{bmatrix} \frac{1}{5} & 0 \\ 0 & \frac{1}{5} \end{bmatrix}$

points within the cluster, location of the cluster, and spread of the cluster. The basic values of the parameters used in generating the stimuli were $N_B = 50, p = 0.3, \ell_{\mathtt{C}} = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$ and $\Sigma = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$, which were varied systematically to produce the range of stimuli described above. The parameter values used to generate these stimuli are given in Table 1. The other three stimuli were generated by sampling 50 points from the uniform distribution. All 12 images are shown in Figure 7.9, with repetition of the stimulus embodying the basic parameter values accounting for the presence of 15 images in the Figure. The stimuli were delivered in a questionnaire.

Procedure. Participants completed the questionnaire as part of a booklet of other short psychology experiments. Each participant saw all 12 images, in one of six random orders. The instructions on the questionnaire read as follows:

> During World War II, the city of London was hit repeatedly by German bombs. While the bombs were found to be equally likely to fall in any part of London, people in the city believed otherwise.
>
> Each of the images below shows where bombs landed in a particular part of London for a given month, with a single point for each bomb. On the lines at the bottom of the page corresponding to each image, please rate HOW BIG A COINCIDENCE the distribution of bombs seems to you. Use a scale from 1 to 10, where 1 means 'Very small (or no) coincidence', and 10 means 'Very big coincidence'.

The images were labelled with alphabetical letters, and correspondingly labelled lines were provided at the bottom of the questionnaire for responses.

Results and Discussion

The mean responses are shown in Figure 7.9. Planned comparisons were computed for each of the manipulated variables, with statistically significant outcomes for number ($F = 22.89$, $p < .0001$), proportion ($F = 10.18$, $p < .0001$), and spread ($F = 12.03$, $p < .0001$), and a marginally significant effect of location ($F = 2.0$, $p = 0.14$). The differences observed among responses to the three sets of points generated from the uniform distribution were not statistically significant ($F = 0.41$, $p = 0.66$). All planned comparisons had $df = 2, 2574$, and $MSE = 6.21$.

Values of $\frac{P(\mathcal{D}|h_1)}{P(\mathcal{D}|h_0)}$ were computed for each image using the method outlined in Appendix E. The predictions of the Bayesian model are shown in Figure 7.9. The ordinal correlation between the raw statistical evidence and the responses was $\rho = 0.965$. The values shown in the figure are a result of the transformation $y = \text{sign}(x)\text{abs}(x)^{0.32}$ for $x = \log \frac{P(\mathcal{D}|h_1)}{P(\mathcal{D}|h_0)}$, which gave a linear correlation of $r = 0.981$. People's assessment of the strength of coincidences shows a remarkably close correspondence to the predictions of this Bayesian account. The main discrepancy is an overestimate of the effect of strength of coincidence for the stimulus with the least spread. This may have been a consequence of the fact that the dots indicating the bomb locations overlapped in this image, making it difficult for participants to estimate the number of bombs landing in the cluster.

7.5 The locus of human irrationality

> "Singular coincidence, Holmes. Very smart of you to notice it, but rather uncharitable to suggest that it was cause and effect."
>
> Sir Arthur Conan Doyle (1986b), *The adventure of the dying detective*, p. 396.

The results of Experiments 7.2 and 7.3 are consistent with the prediction that the perceived strength of a coincidence should correspond to the support for an improbable hypothesis. In both experiments, the likelihood ratio in favor of h_1 gave a remarkably good fit to human judgments. These results have implications for understanding how coincidences sometimes lead people to false conclusions. In the terminology I have introduced, the irrational errors associated with coincidences involve mistaking mere coincidences for suspicious ones, and erroneously taking suspicious coincidences as compelling evidence for a conclusion. Under my analysis, these errors result from over-estimating the posterior odds

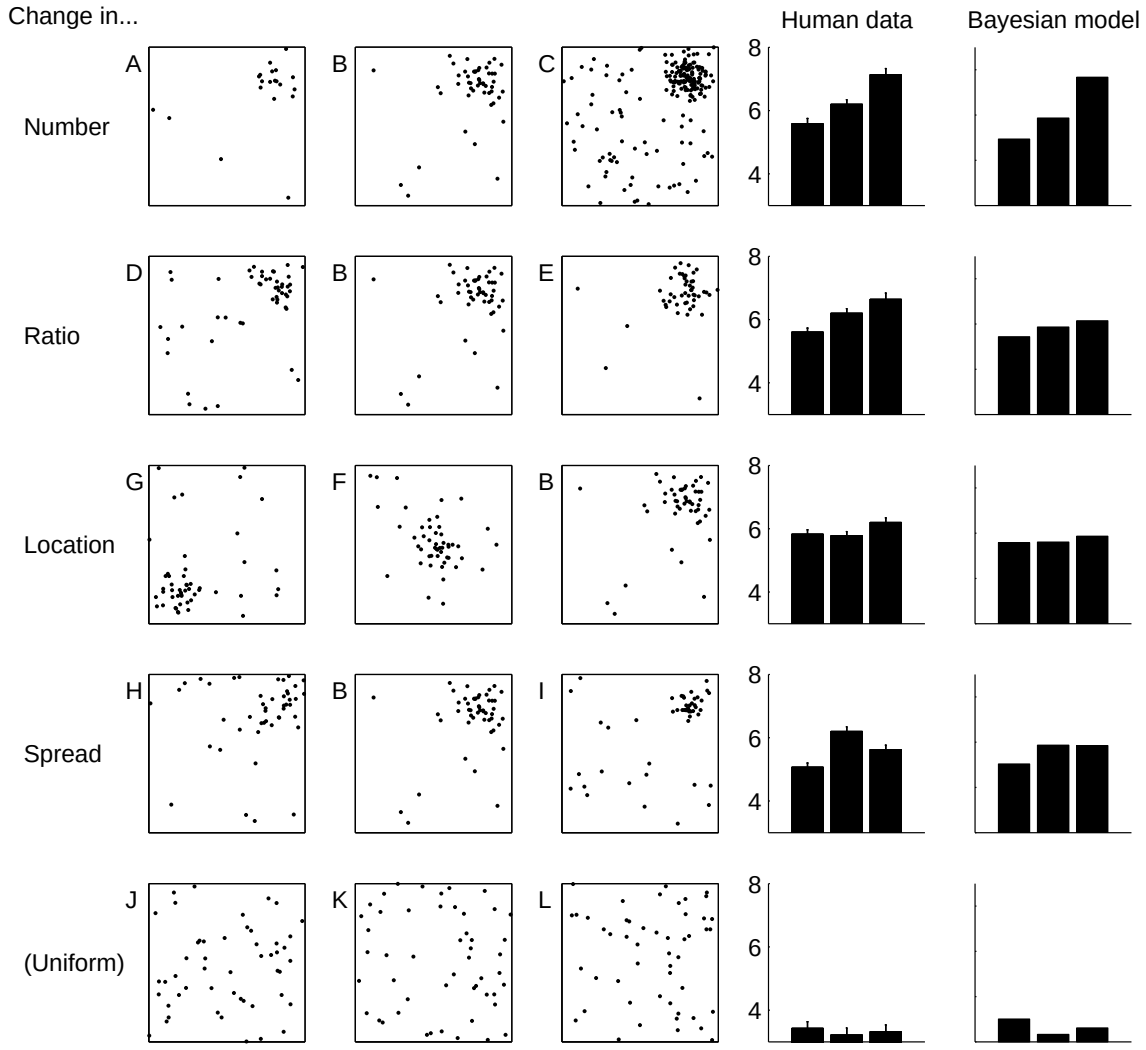

Figure 7.9: Results of Experiment 7.2. Each line shows the three stimuli used to test the effects of manipulating one of the statistical properties of the stimulus, together with the mean judgments of strength of coincidences from human participants and the predictions of the Bayesian model. Error bars show one standard error, and letters label the different stimuli.

in favor of a hypothesis. The results of the preceding experiments suggest that people are accurate in their assessment of the support that data provide for a particular hypothesis, raising the possibility that these errors are due to inaccurate prior odds. Thus, people reach false conclusions as the consequence of a coincidence when they fail to accurately assess the a priori plausibility of the hypothesis that the coincidence suggests.

The suggestion that people can accurately assess the evidence that a set of events provides for a conclusion is consistent with some of the ideas that appear in the literature on judgment and decision making. Tversky and Koehler (1994) argued that many of the irrational aspects of people's probability judgments can be understood by viewing these judgments as reflecting the support that a set of observations provide for a particular hypothesis. In order to use this information, people have to be able to actually compute some measure of support. While various measures have been suggested, a Bayesian measure of support similar to our measure of evidence has been found to provide reasonable results on at least some cognitive tasks (Koehler, White, & Grondin, 2003). This is consistent with the results of Experiments 7.2 and 7.3. However, accurately assessing the support for a hypothesis does not guarantee a valid conclusion about the truth of that hypothesis, just as accurate results from a statistical analysis do not guarantee a valid conclusion. Reaching the right conclusion requires having well-calibrated priors.

The importance of veridical prior knowledge can be illustrated by an analogy to logical deduction. The validity of an argument is a consequence of both the truth of the premises and the soundness of the reasoning, which correspond to the calibration of the prior beliefs and the calculation of the evidence in a Bayesian inductive inference. Invalid conclusions can be reached if a single premise is false, or the deduction mechanism is unsound. The claim that miscalibration of priors contributes to false inferences is analogous to asserting that the deduction mechanism behind people's conclusions is sound, but the premises from which they reason are not always true. Our knowledge about the world is complex and often inconsistent, and the right evidence in the wrong place can lead us astray. If we are promiscuous believers in paranormal events, we can easily take mere coincidences to be suspicious.

A classic example of a case in which the miscalibration of priors may have resulted in an unusual conclusion is provided by the history of synchronicity. Jung (1960) introduced synchronicity as an acausal force connecting events, "equal in rank to causality as a principle of explanation" (p. 435). Jung was led to believe in the existence of synchronicity through a

series of coincidences, including a spontaneous explosion during a conversation with Freud, the mysterious appearance of a scarab beetle while a patient was relating a dream about a scarab, and an alarmingly extended series of encounters with fish. The fact that it may seem possible for many of us to go through similar experiences without believing them more than mere coincidences may illustrate a difference in our prior beliefs about the validity of our theories about the world. As Hardy, Harvie, and Koestler (1973) suggest, "Whether one believes that some highly improbable meaningful coincidences are manifestations of some such unknown principle operating beyond physical causality, or are produced by that immortal monkey at the typewriter, is ultimately a matter of inclination and temperament" (p. 230).

One suggestive hypothesis as to why we might be more willing to believe in the existence of causal relationships than we could be comes from developmental psychology. Gopnik and Meltzoff (1997) argue that the scientific behavior of adults is an extension of the capacity for causal discovery that is essential for the cognitive development of children. It is quite understandable that children might be willing to believe in the hypotheses suggested by coincidences, since they are surrounded by events that really do involve novel causal relationships. Small children are justified in being conspiracy theorists, since their world is run by an inscrutable and all-powerful organization possessing secret communications and mysterious powers. If our scientific capacities really are for solving these childhood mysteries, then our disposition to believe in the existence of unexpected causal relationships might lag behind our current state of knowledge, leading us to see causes where none exist.

Further opportunities for erroneous inferences are provided by cases where suspicious coincidences are not tested through further investigation. If we examine the contexts in which coincidences lead people to false beliefs, we see that many of them involve situations where it is hard to conduct convincing experiments that invalidate a hypothetical causal relationship. Synchronicity, extrasensory perception, and other paranormal forces are all quite slippery subjects of investigation, for which it is challenging to construct compelling experimental tests (e.g., Diaconis, 1978). The bombing of London involved a similarly untestable hypothesis, compounded by the fear and uncertainty associated with being under attack. The cases where coincidences have resulted in rational discoveries, in science and detective stories, are all cases where a coincidence suggests a hypothesis which can be established through further investigation. Halley's hypothesis was spectacularly validated

by the return of his comet in 1758, and Holmes has the chance to test his theories by collecting further clues. Without this kind of detailed investigation, all but the most compelling coincidences should be treated as nothing more than suspicious.

7.6 Coincidences and theory change

Many cognitive scientists have suggested that the growth and organization of knowledge can be understood by examining similar processes in scientific theories (Carey, 1985a; Gopnik & Meltzoff, 1997; Karmiloff-Smith, 1988; Keil, 1989; Murphy & Medin, 1985). One of the major problems that arises in this "theory theory" is understanding the process of theory change. The formal analyses I have presented in this chapter have characterized coincidences as involving data that provide support for a theory that has low a priori probability. Coincidences thus constitute an opportunity to discover that one's current theory of how the world works is false. This characterization of coincidences suggests that they may play an important role in theory change, similar to the role of anomalies in Kuhn's (1970) influential account of scientific discovery.

The theory theory draws extensively upon work in philosophy of science, and in particular upon Kuhn's (1970) analysis of science in terms of a succession of scientific revolutions. One of the major topics of Kuhn's work is the factors contributing to scientific discovery and subsequent theoretical change. Principal among these factors is the growing awareness of "anomalies," with Kuhn (1970) claiming that "discovery commences with the awareness of anomaly, i.e., with the recognition that nature has somehow violated the paradigm-induced expectations that govern normal science" (p. 52). Kuhn (1970) argued that the process of discovery often follows a particular course:

> Initially, only the anticipated and usual are experienced even under circumstances where anomaly is later to be observed. Further acquaintance, however, does result in awareness of something wrong or does relate the effect to something that has gone wrong before. That awareness of anomaly opens a period in which conceptual categories are adjusted until the initially anomalous has become the anticipated. At this point the discovery has been completed. (p. 64)

Anomalies can also be responsible for large-scale theoretical change, inducing a crisis that

is resolved by the development of a new theory. However, Kuhn (1970) noted that "if an anomaly is to evoke crisis, it must usually be more than just an anomaly" (p. 82).

Anomalous scientific results can be of two kinds. The strongest kind of anomaly is an event that is impossible under a particular scientific theory, having zero probability. Such an event contributes infinite evidence against the theory, and suggests that it should be replaced. However, most anomalies are of a different kind: events that are improbable under a theory. Salmon (1990) suggested that a Bayesian approach to comparing theories might be consistent with Kuhn's characterization of theory change. Salmon characterized an anomaly as "a phenomenon that appears to have a small, possibly zero, likelihood given that theory" (1990, p. 193). This assertion is similar to the claim that coincidences are unlikely events, defining anomalies only in terms of their probability under the current theory and not considering alternatives. Just as we can construct cases in which events are equally unlikely but not equally coincidental, we can construct cases in which events are equally unlikely but not equally anomalous. A full account of anomalies needs to compare this likelihood with some alternative, as in my account of coincidences.

The consistency of Salmon's (1990) statistical definition of an anomaly with the accounts that appear in the literature on coincidences suggests that there may be some correspondence between the two notions. Kuhn's characterization of anomalies is very similar to my definition of coincidences: anomalies are patterns of results that suggest a structure not predicted by the current theory, which can come to motivate theoretical change once sufficient evidence mounts. Kuhn's (1970, p. 64) description of the process by which anomalies lead to discoveries bears a remarkable similarity to the process by which mere coincidences become suspicious. Initially, a few surprising coincidences will be dismissed as the result of chance. However, as one comes to consider the possibility of other processes being involved, and as the number of coincidences increases, the evidence provided by this set of events begins to promote suspicions. Further exploration of the source of these events might reveal an unexpected causal relationship. Once one is aware of this relationship, the events that were previously coincidences become anticipated, and merely provide further evidence for a known relationship. Likewise, the statement that crises are provoked by anomalies that are not just anomalies expresses the same sentiment as our notion of suspicious coincidences – in order to result in a change in beliefs, a coincidence must be more than just a coincidence.

7.7 Summary

Coincidences pose an interesting paradox, playing key roles in both significant scientific discoveries and in propagating false beliefs. Resolving this paradox requires going beyond the common idea that coincidences are just unlikely events, and considering their relationship to causality. The theory-based causal induction framework provides the tools that are needed in order to make this relationship clear. Using this framework, it is possible to define coincidences as events that provide support for a hypothesis, but not enough support to convince us to believe in that hypothesis. This account can be used to identify what goes wrong when we are misled by coincidences, and to clarify the role that coincidences play in the revision of intuitive theories.

Chapter 8

Implications for intuitive theories

The variation in the formal theories presented in the preceding chapters illustrates the diversity of the knowledge that guides human causal induction. The quality of the correspondence between computational models and data suggests that these theories give a good characterization of the assumptions that people make in these different settings. Having established the descriptive virtues of theory-based causal induction, I will use this chapter to examine some of the implications of this framework for the role of intuitive theories in causal induction. The formal nature of the framework makes it possible to make precise claims about some of the key issues surrounding intuitive theories introduced in Chapter 1: domain-specificity, the role of mechanism knowledge, and theory acquisition. I will discuss these issues in turn.

8.1 Domain specificity

Physics, biology, and psychology all involve quite different causal principles, such as force, growth, and desire. Even young children are sensitive to this variation, having different expectations about the causal relationships participated in by biological and non-biological (Springer & Keil, 1991) and social and non-social (Gelman & Spelke, 1981; Shultz, 1982a) entities. Consequently, it has been suggested that each of these domains is characterized by a separate domain-specific theory, identifying the principles appropriate to that domain (see Hirschfeld & Gelman, 1994, for a variety of views on this issue). The early manifestation of domain-specific causal inferences, such as knowledge of the causal properties of objects (e.g., Spelke, Breinlinger, Macomber, & Jacobson, 1992) has led to claims that these inferences

are the result of distinct and specialized cognitive modules (e.g., Leslie, 1994).

Claims of domain-specificity are highly controversial. Recently, Schulz and Gopnik (in press) reported a series of studies that seemed to provide evidence against the domain-specificity of causal reasoning, demonstrating that young children made the same inferences in different domains and were willing to identify causal relationships that involved domain-inappropriate causes. Schulz and Gopnik (in press) used these results to argue that children have a domain-general capacity for learning causal relationships.

In the theory-based framework introduced in this thesis, causal induction is viewed as the result of domain-general statistical inference that is informed by a domain-specific (or at least domain-sensitive) theory. The theory provides constraints on the causal structures that are evaluated by statistical inference. In this section, I will use this framework to explain Schulz and Gopnik's (in press) results, demonstrating that they are consistent with domain-specific prior knowledge playing a role in causal induction. I will do so by considering how domain might be expected to influence expectations about the functional form and plausibility of causal relationships.

8.1.1 The effect of domain on functional form

Causal relationships in different domains involve very different causal mechanisms. For example, you would probably use different methods to move a heavy box a yard to the left from those that you would use to move a friend a yard to the left. However, this difference in the mechanism by which effects are brought about need not be reflected in a difference in the assumed functional form of the underlying relationship. If attempting to drag a heavy box and asking your friend to move are both successful about 90% of the time, these two relationships can be described by a similar functional form. The mapping from domain-specific mechanism to functional form is many-to-one, with a variety of different mechanisms reducing to the same set of qualitative assumptions about functional form. Consequently, causal induction involving systems in quite different domains can have much the same character: even if the *content* of theories differs, the *constraints* they imply for causal relationships can be the same.

An experiment conducted by Schulz and Gopnik (in press, Experiment 3) illustrates this point. In this experiment, children learned about causal relationships in two different domains: biology and psychology. In the biology domain, children were asked to infer which flowers caused a toy monkey to sneeze, while in the psychology domain, they learned

Table 8.1: Effect of Domain on Functional Form

Condition	C	All	Other
test (biology)	**0.78 (0.90)**	0.11 (0.00)	0.11 (0.10)
control (biology)	0.05 (0.06)	**0.89 (0.86)**	0.05 (0.08)
test (psychology)	**0.67 (0.90)**	0.28 (0.00)	0.05 (0.10)
control (psychology)	0.05 (0.06)	**0.83 (0.86)**	0.11 (0.08)

Note: Numbers indicate the proportion of children identifying C as the cause, A, B, and C as causes, or producing some other response, from Schulz & Gopnik (in press, Experiment 3). Predictions of Bayesian model are given in parentheses. Boldface indicates majority.

which animals scared a toy rabbit. There were two conditions in each domain. Using A, B, and C to indicate the presence of each of three flowers (or animals) and E to indicate a sneezing monkey or a scared rabbit, the *test* condition consisted of four events: $e^-|a^+b^-c^-$, $e^-|a^-b^+c^-$, $e^+|a^-b^-c^+$, and $e^+|a^+b^+c^+$. The *control* condition featured four different events: $e^+|a^+b^-c^-$, $e^+|a^-b^+c^-$, $e^+|a^-b^-c^+$, and $e^+|a^+b^+c^+$. Children made quite similar inferences across the two domains, as shown in Table 8.1, identifying C as the cause in the *test* condition, and all of A, B, and C as causes in the *control* condition.

Schulz and Gopnik (in press) used the results of this experiment to argue that children's ability to infer causal relationships is domain independent. A different interpretation of these results is that they indicate that the same functional form can be assumed in different domains (and for different underlying causal mechanisms). If the theories characterizing two systems are isomorphic, then causal inferences using those theories will be identical. A pair of isomorphic theories for sneezing and scaring is shown in Figures 8.1 and 8.2. There is a direct correspondence between the types of entities identified by these theories and the predicates applied to those entities, with `Flower` and `Beast`, `Monkey` and `Rabbit`, and `Sneezes` and `Scared` all playing the same roles. The theories are identical in their assumptions about the plausibility of causal relationships and the functional form of those relationships. For the stimuli shown to the children, both theories generate the same hypothesis space of causal graphical models, shown in Figure 8.3.

Under the theories shown in Figures 8.1 and 8.2, the influences of multiple causes on both sneezing and scaring are described by the noisy-OR parameterization. The use of the same functional form across the two theories is a consequence of the applicability of the same set of assumptions about the nature of causal relationships in the domain: that causes influence their effects probabilistically, and that these each of these influences has

Ontology:

Types	Number
`Flower`	$N_F \sim P_B$
`Monkey`	$N_M \sim P_M$
`Trial`	$N_T \sim P_T$

Predicates	Values	
`Present(Flower,Trial)`	`Boolean:`	$\{\texttt{H},\texttt{T}\}$
`Sneezes(Monkey,Trial)`	`Boolean:`	$\{\texttt{H},\texttt{T}\}$

Plausible relations:

$\texttt{Present}(\texttt{F},\texttt{T}) \rightarrow \texttt{Sneezes}(\texttt{M},\texttt{T})$
True for all `T` with probability p for each `F`, `M` pair

Functional form:

$\texttt{Present}(\texttt{F},\texttt{T}) \sim$ Bernoulli$(\cdot)$
$\texttt{Sneezes}(\texttt{M},\texttt{T}) \sim$ Bernoulli(ν) for ν from a noisy-OR:

Cause	Strength
(Background)	$w_0 = \epsilon$
$\texttt{Present}(\texttt{F},\texttt{T})$	$w_i = 1 - \epsilon$

Figure 8.1: Theory for causal induction with "biology" (sneezing monkeys).

Ontology:

Types	Number
`Beast`	$N_B \sim P_B$
`Rabbit`	$N_R \sim P_R$
`Trial`	$N_T \sim P_T$

Predicates	Values	
$\texttt{Present}(\texttt{Beast},\texttt{Trial})$	`Boolean:`	$\{\texttt{T},\texttt{F}\}$
$\texttt{Scared}(\texttt{Rabbit},\texttt{Trial})$	`Boolean:`	$\{\texttt{T},\texttt{F}\}$

Plausible relations:

$\texttt{Present}(\texttt{B},\texttt{T}) \rightarrow \texttt{Scared}(\texttt{R},\texttt{T})$
True for all `T` with probability p for each `B`, `R` pair

Functional form:

$\texttt{Present}(\texttt{B},\texttt{T}) \sim$ Bernoulli$(\cdot)$
$\texttt{Scared}(\texttt{R},\texttt{T}) \sim$ Bernoulli(ν) for ν from a noisy-OR:

Cause	Strength
(Background)	$w_0 = \epsilon$
$\texttt{Present}(\texttt{B},\texttt{T})$	$w_i = 1 - \epsilon$

Figure 8.2: Theory for causal induction with "psychology" (scared rabbits).

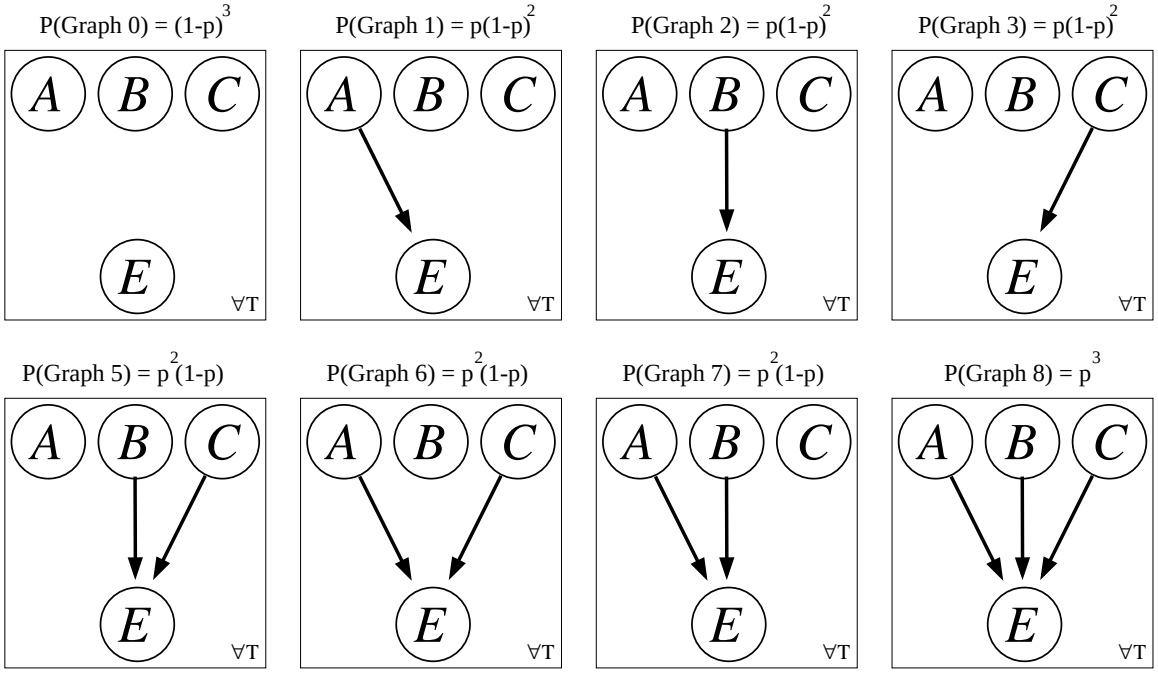

Figure 8.3: Hypothesis space for causal induction in both the "biology" and "psychology" settings. A indicates `Present(A,T)` for either a `Flower` or `Beast a`, and likewise for B and C. E indicates either `Sneezes(M,T)` or `Scared(R,T)`, for a `Monkey m` or `Rabbit r`. The plates indicate that these causal relationships hold for every `Trial T`. The same hypothesis space applies to causal induction across domains, with A and B indicating the presence of `InDomain` causes, and C indicating the presence of an `OutDomain` cause.

an independent opportunity to do so. This shared functional form results in the same predictions for the two conditions, as shown in Table 8.1. These predictions used $p = 0.5$ and $\epsilon = 0.05$ for both biology and psychology.

8.1.2 The effect of domain on plausibility

In many of the examples discussed in the previous chapters, assumptions about the plausibility of a causal relationship played a less important role than assumptions about the underlying ontology and the functional form of causal relationships. In part, this is because the stimuli used in psychology experiments on causal learning tend to involve variables among which causal relationships are quite plausible. Studying problems of causal learning involving variables from different domains provides an opportunity to explore the effect of domain on the plausibility of causal relationships. In particular, one might expect that plausible relationships would be restricted to causes that use forces appropriate to that domain. For example, asking a box to move is far likely to be successful than dragging it.

Shultz (1982b, Experiment 4) demonstrated that young children have strong expectations about the plausibility of different kinds of causal relationships. For example, he found that children know that a lamp is more likely than a fan to produce a spot of light, that a fan is more likely than a tuning fork to blow out a candle, and that a tuning fork is more likely than a lamp to produce resonance in a box. All of these kinds of relationship are in the physical domain, but they involve different mechanisms. Schulz and Gopnik (in press) have recently extended this kind of investigation, examining how children assess the plausibility of causal relationships across domains, and how this assessment interacts with statistical evidence.

Schulz and Gopnik (in press, Experiment 4) introduced children to causal systems in two domains. The *physical* domain involved a machine that made noise, with the candidate causes of the activation of the machine being two magnetic buttons, the in-domain objects `a` and `b`, and speech, the out-domain object `c`. The *psychological* domain involved reasoning about what might make a person giggle. The in-domain objects `a` and `b` were silly faces, and the out-domain object `c` was a switch. In each domain, children were first asked which objects were likely to produce the effect, and unanimously identified the in-domain causes. They then saw a series of trials exactly the same as those used in the *test* condition of Schulz and Gopnik's (in press) Experiment 3, discussed above. As shown in Table 8.2, the majority of the children now identified the out-domain object `c` as the cause, despite its low

Table 8.2: Effect of Domain on Plausibility

Condition	c	a, b	Other
physical	**0.75 (0.79)**	0.00 (0.08)	0.25 (0.13)
psychological	**0.81 (0.79)**	0.00 (0.08)	0.19 (0.13)

Note: Numbers indicate the proportion of children identifying c as a cause, a and/or b as a cause, or producing some other response, from Schulz & Gopnik (in press, Experiment 4). Predictions of Bayesian model are given in parentheses. Boldface indicates majority.

initial plausibility.

A simple theory that characterizes both the physical and the psychological stimuli used by Schulz and Gopnik is shown in Figure 8.4. Under this theory, both in-domain and out-domain objects can influence the effect, but the plausibility of such relationships differs. The probability of an in-domain relationship is set by p, while the probability of an out-domain relationship is set by q. The hypothesis space generated by this model for objects a, b, and c is shown in Figure 8.3, using A to indicate the presence of a, B to indicate the presence of b, C to indicate the presence of c, and E to indicate the activation of the effect.

The initial responses of the children in Schulz and Gopnik's experiment indicates that q is much less than p. The predictions of the model with $p = 0.5$, $q = 0.05$, and $\epsilon = 0.5$ are shown in Table 8.2. The model identifies c as a cause, despite its low plausibility, because of the strong assumptions about functional form. This effect can be best understood by considering the limit as $\epsilon \to 0$. If E never occurs in the absence of a cause, then seeing E occur in the presence of c provides unambiguous evidence that C causes E. Thus, provided q takes on some value greater than zero, the probability that C causes E will be 1.00. Allowing ϵ to take on values greater than zero increases the influence of q on the outcome. In particular, if ϵ is somewhat greater than q, it becomes more likely that A and B are causes of E, and the causal relationship simply failed to manifest on the trials when a and b were present.

Schulz and Gopnik (in press) interpreted this experiment as indicating that children are aware of domain-specific constraints on causal relationship, but that these constraints can be over-ridden by domain-general principles of causal learning. In the analysis above, the probability of an out-domain variable being involved in a causal relationship, q, has little effect on the predictions of the model: the assumptions about the functional form of the causal relationship mean that C will be identified as the cause even if q is very small. The

Ontology:

Types	Number
`Cause`	$N_C \sim P_C$
`InDomain`	·
`OutDomain`	·
`Effect`	$N_E \sim P_O$
`Trial`	$N_T \sim P_T$

Predicates	Values	
`Present(Cause, Trial)`	`Boolean:`	{H, T}
`Active(Effect, Trial)`	`Boolean:`	{H, T}

Plausible relations:

`Present(C, T)` → `Active(E, T)`
True for all `T` with probability p for each `C, E` pair where `C` is an `InDomain` cause, and with probability q for each `C, E` pair where `C` is an `OutDomain` cause

Functional form:

`Present(C, T)` ∼ Bernoulli(·)
`Active(E, T)` ∼ Bernoulli(ν) for ν from a noisy-OR:

Cause	Strength
(Background)	$w_0 = \epsilon$
`Active(C, T)`	$w_i = 1 - \epsilon$

Figure 8.4: Theory for causal induction across domains.

model predicts that the value of q would have a greater effect given ambiguous evidence. For example, seeing $e^+|a^+b^-c^+$ and $e^+|a^-b^+c^+$ would suggest that A and B cause E if q is small, and that C causes E if q is large. Exploring cross-domain inferences in the face of ambiguous evidence might reveal a more subtle, graded interaction between domain-specific constraints and domain-general statistical learning.

Different domains operate by different causal princples, a fact that can be captured in the theory-based framework by using different causal theories and by allowing relationships to differ in their plausibility. However, such differences need not always result in different behavior: as shown by the results of Schulz and Gopnik (in press, Experiment 3), if the theories that describe causal systems in two domains imply the same constraints on causal graphical models, then we should expect causal inferences to have much the same character across those systems. Likewise, while knowledge of the causal principles by which different domains operate can influence the plausibility of causal relationships, strong assumptions about functional form can overwhelm the effects of plausibility, as in Schulz and Gopnik (in press, Experiment 4). These experiments thus leave open the possibility that domain-specific knowledge guides causal induction. Analyzing their results using the theory-based

causal induction framework suggests further experiments which might shed light upon this important issue.

8.2 Theories and mechanisms

As mentioned in Chapter 1, psychological theories about causal induction have traditionally fallen into two camps (Newsome, 2003): covariation-based approaches characterize human causal induction as the consequence of a domain-general statistical sensitivity to covariation between cause and effect (e.g., Cheng & Novick, 1990; 1992; Shanks & Dickinson, 1987), while mechanism-based approaches focus on the role of prior knowledge about the mechanisms by which causal force can be transferred (e.g., Ahn & Kalish, 2000; Shultz, 1982b; White, 1995). The theory-based approach emphasizes the interaction between prior knowledge and statistical learning in causal induction. In this section I will attempt to clarify how the formal theories appealed to by the theory-based approach connect to the notion of causal mechanism.

When researchers refer to "causal mechanism", they typically mean the chain of events mediating between cause and effect, as illustrated in Figure 8.5 (a) (e.g., Bullock, Gelman, & Baillargeon, 1982; Glymour, & Cheng, 1998; Shultz, 1982b; see Shultz & Kestenbaum, 1985, for a discussion of different kinds of mechanism). However, a detailed understanding of the mechanisms mediating between cause and effect is clearly not necessary for causal induction – if one possessed such knowledge there would be nothing to learn. Furthermore, recent studies investigating the limits of people's understanding of causal systems suggests that in fact, our mechanism knowledge may look more like Figure 8.5 (b). For example, Rozenblit and Keil (2002) found that when asked to explain how mechanical systems like crossbows and helicopters work, people radically over-estimated the extent of their mechanism knowledge. It seems that, in general, our causal knowledge identifies the fact that a mechanism exists, but does not necessarily articulate all of the steps that connect cause and effect (Keil, 2003).

Results like those of Rozenblit and Keil (2002) raise an interesting question: if our knowledge of causal mechanisms is as shallow as it appears to be, how is it possible for this knowledge to inform causal induction? The theory-based account provides an answer to this question. Under this account, prior knowledge plays two important roles in causal induction: identifying which relationships are plausible, and characterizing the functional

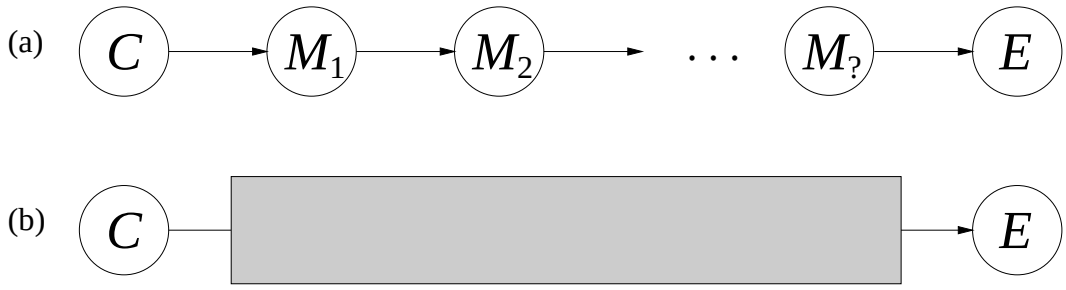

Figure 8.5: Two conceptions of causal mechanism knowledge. (a) The causal mechanism specifies the chain of events mediating between a cause C and its effect E. (b) Often, people know that some mechanism exists, but not the details.

form of those relationships. The shallow mechanism knowledge described by Keil (2003) is sufficient to fulfill these roles. Whether a causal relationship seems plausible is affected by mechanism knowledge, but the key determinant in this decision is not the particular details of the causal mechanism, but whether such a mechanism could exist. Similarly, evaluating the functional form of a causal relationship does not require knowing every step between cause and effect, but knowing what kind of relationship those steps might produce.

Neither assessing plausibility nor specifying functional form requires a detailed account of a chain of events from cause to effect. Several authors have equated the plausibility of a causal relationship with the existence of a potential mechanism by which the cause could influence the effect (e.g., Ahn & Kalish, 2000; Schlottmann, 1999). Koslowski (1996) provides a succinct exposition of this view:

> The mechanism in a causal situation explains the process by which a cause brings about an effect. Without the availability of a plausible process, causation is unlikely to be seen as taking place. Covariation is sometimes seen as indicating cause and is at other times seen as merely artifactual precisely because plausible mechanisms are seen as operating in the former instances and as unlikely to be operating in the latter. (p. 13)

Koslowski and her colleagues (Koslowski, 1996; Koslowski & Okagaki, 1986; Koslowski, Okagaki, Lorenz, & Umbach, 1989) have conducted a series of experiments investigating this claim, finding that people consider causal relationships more plausible when supplied with a potential mechanism, and less plausible when the most likely mechanisms are ruled out.

Lack of complete mechanism knowledge is likewise no impediment to reasoning about the functional form of a causal relationship. This point is illustrated by the discovery of Halley's comet, a causal inference that required the prior knowledge supplied by Newton's theory of physics. Newton's theory was notoriously amechanistic, departing from its forebears by introducing forces unmediated by particles (e.g., Westfall, 1977). Physicists are still engaged in the project of providing a mechanistic account of Newton's ideas, and in particular the force of gravity. While Halley did not know how the mass of stars and planets influenced the orbits of comets, he was still able to use information about the form of this influence to reason about the cause of the events that he observed. Indeed, in introducing his own account of causality, Pearl (1996) reduces mechanisms to "...nothing but ordinary physical laws, cast in the form of deterministic equations" (p. 432), being no more than the specification of the functional form of the relationship between two variables.

The causal theories used in theory-based causal induction can thus be viewed as expressing the consequences of the shallow knowledge people possess of the mechanisms that operate in different domains. However, these theories do not express that knowledge directly: they are just as amechanistic as Newton's theory of physics, characterizing the possible relationships among entities and their form. The theories I have described are the constraints on the functional relationships among variables that can be the *consequence* of mechanism knowledge rather than the knowledge itself. Many mechanisms can imply the same set of constraints, as illustrated in the discussion of domain-specificity in the previous section.

Distinguishing between theories and mechanisms provides an important insight into how causal induction is possible. A major challenge for mechanism-based accounts of causal induction is explaining how new causal mechanisms might be learned: if all causal induction requires mechanism knowledge, one can never discover a relationship that suggests a new mechanism (e.g., Cheng, 1993). If theory and mechanism are distinct, it becomes possible to learn a set of causal relationships without knowing their underlying mechanism. The existence of these relationships can then encourage the search for a mechanism that accounts for them, and the discovery of such a mechanism justifies further inferences about possible causal relationships. Such a pattern is extremely common in science – most of the suspicious coincidences that suggest new causal relationships are followed by a search for a mechanism. For example, Hempel (1966) describes a famous discovery of exactly this kind, made by the Hungarian physician Ignaz Semmelweis in the 19th century. Semmelweis noticed a

suspicious similarity in the symptoms of patients in a particular ward and the symptoms exhibited by a colleague injured during an autopsy. Inferring that a causal relationship might exist, he began to search for the mechanism responsible – a search that brought him remarkably close to the modern theory of germs.

8.3 Theory acquisition

Analyses of causal induction that postulate a common representation for causal relationships and intuitive theories can construe cognitive development and the everyday discovery of new causal relationships as the same process, with theories being enriched and revised through the addition of those new relationships (e.g., Gopnik & Glymour, 2002). I have argued that causal induction actually involves multiple levels of representation, with causal graphical models being adequate to express specific causal relationships, but causal theories defining laws that constrain those relationships. Under this account, cognitive development and everyday learning do not act upon the same representations: cognitive development can involve the revision of theories, while everyday learning uses those theories when identifying new causal relationships. My focus in this thesis has been on the latter process, explaining the role of causal theories in causal induction. Understanding how those theories themselves are learned is an important open problem. Under the analogy between language comprehension and causal induction introduced in Chapters 2 and 3, this is the explanatory problem, the equivalent of accounting for how people learn grammars.

The question of how people learn causal theories can be answered using the methods I have applied to the question of how people learn causal structure. The three levels of representation assumed by this account (Figure 3.1) and the assumption that each level generates the one below define a hierarchical Bayesian model. As discussed in Chapter 5, data can be used to make inferences about theories as well as inferences about causal structure. Given data $\mathcal{D}$, the posterior distribution over theories T is

$$P(T|\mathcal{D}) = \frac{P(\mathcal{D}|T)P(T)}{P(\mathcal{D})}, \tag{8.1}$$

where $P(\mathcal{D}) = \sum_T P(\mathcal{D}|T)P(T)$, and

$$P(\mathcal{D}|T) = \sum_{i=1}^{|\mathcal{H}_T|} P(\mathcal{D}|\text{Graph } i)P(\text{Graph } i|T)$$

which sums over all causal graphical models generated by the theory.

Equation 8.1 can be used to assess the posterior probability of any set of theories. However, most cases of theory change do not involve comparing a completely arbitrary set of theories. Rather, one of the components of an existing theory is modified, distinguishing new types of entities, allowing new causal structures, or specifying a different form for a known relationship. In the preceding chapters, I discussed two examples of this kind of inference. In Chapter 5, I showed how Equation 8.1 could be used to select between two theories that differed in the functional form of the relationship between blickets and blicket detectors. The same approach can be used to evaluate a range of theories that differ in their assumptions about functional form. In Chapter 7, the coincidences I analyzed all involved a decision between two theories that differed in their assumptions about causal structure. The same analysis can be extended to allow more general comparison of theories that differ in the forces they posit.

The question of how people infer that there are different types of entities that engage in different patterns of causal relationships has recently been explored by Tenenbaum and Niyogi (2003). Tenenbaum and Niyogi (2003) found that people could learn about the existence of different types of entities based purely upon the causal relationships in which they participated. Kemp, Griffiths, and Tenenbaum (2004) have developed a computational model that explains how such learning can take place, using Bayesian inference to simultaneously identify the number of types, the types themselves, and the plausibility of causal relationships among entities of those types. This approach performs the computation identified in Equation 8.1 for a slightly more impoverished class of theories than those presented here (Tenenbaum & Griffiths, in prep).

These examples of theory acquisition just begin to scratch the surface of the question of how people learn intuitive theories. More detailed development and testing this account raises a rich set of questions to be explored in future research. One issue is the source of prior probabilities for theories – the distribution $P(T)$ in Equation 8.1. As discussed in the next section, such prior probabilities can be construed as the consequence of higher-level theories. Another source of priors over theories is linguistic interaction. While suspicious coincidences provide some of the most compelling examples of scientific theory change, the development of intuitive theories is intimately related to the development of linguistic abilities (e.g., Gopnik & Meltzoff, 1997). The Bayesian approach discussed in this section has the potential to provide insight into how these factors can be combined, with data

driving theory change through suspicious coincidences and linguistic interaction modifying the prior plausibility of candidate theories.

8.4 Higher-level theories

The causal theories that I presented in the previous chapters have the same constituents as the intuitive theories discussed in accounts of cognition and cognitive development – ontologies and causal laws – but are far more limited in scope. The analogy to Wellman's (1990; Wellman & Gelman, 1992) notion of a framework theory is useful in indicating the way that multiple levels of causal knowledge interact, but misleading in suggesting that the causal theories I describe are at the same level of generality as framework theories. Framework theories are supposed to provide the fundamental principles used in organizing an entire domain, such as intuitive physics, biology, or psychology. The theories I have presented characterize the principles that underlie very simple causal systems, with well delineated boundaries. For example, while the theory of Nitro X makes predictions for any array of arbitrarily many cans of explosive, it hardly provides a complete theory of all physical systems.

The theories I have discussed in this thesis are at a relatively low level of generality, identifying the causal principles involved in specific kinds of system. This is partly a consequence of attempting to provide the simplest theory that provides the constraints necessary to explain hman inferences. Doing justice to the notion of a framework theory requires postulating the existence of higher-level theories, which express principles common to many systems. As indicated in Equation 8.1, applying Bayes' rule at the level of theories requires having a prior distribution over such theories. Just as people have strong expectations about the causal relationships that might hold in a given system, they have strong expectations about the kind of causal relationships that could operate in a domain. For example, theories of the stick-ball machine are constrained by beliefs about how *any* physical system could work. These constraints are expressed in higher-level domain theories. Such higher-level theories can act as hypothesis-space generators for theories at a lower level, defining a hypothesis space of theories, and a prior on those theories. This view is explored in detail in Tenenbaum and Griffiths (in prep).

Identifying the kind of principles that should be included in these higher-level theories requires investigating the assumptions that guide causal induction across a wide range of

systems. By examining inferences involving different systems in physics, biology, and psychology, it should become possible to pick out the domain-specific principles that generate the theories of these systems that people use. The different physical systems discussed in Chapters 5 and 6, such as the stick-ball machine and Nitro X, suggest what some of these principles might be for the domain of physics. The theories that I used to explain people's inferences about these systems had much the same character, using hidden causes ("prime movers") to inject mechanical energy into the system, and having rules about the circumstances under which events occurred that gave a qualitative correspondence to Newton's first two laws of physics – that no object changes state without a cause, and that causes produce changes in state with high probability. These principles may be a part of the higher-level theory that organizes knowledge about physical causality, a hypothesis that can be explored by examining people's inferences about other physical systems.

Providing an account of how people learn causal theories of specific systems that appeals to higher-level causal theories raises a new problem: explaining how these higher-level theories are learned. Changes in the causal theories of domains constitute some of the most interesting phenomena in cognitive development, and in the history of science. However, at this point, concerns about an infinite recursion, providing no ultimate solution to the question of how people learn causal relationships, seem justified. There are three reasons not to be concerned by such a recursion. First, the mechanism by which the inference is performed at each step is the same – regardless of the level of representation, inferences about the level above can be made using Bayesian inference. There is thus no mysterious new force of learning that enters at any point. Second, this inference becomes simpler at each level. While more data might be required, the space of possible theories of physics must be smaller than the space of possible theories of the stick-ball machine, since the latter is generated by the former. Third, the recursion is not infinite. At some point, it grounds out in a set of basic assumptions about the nature of causality, which provide constraints on the most general domain theories.

8.5 Summary

The development of any kind of computational model of cognition provides the opportunity to demonstrate that a particular set of assumptions is sufficient to predict human judgments. The theory-based causal induction framework casts human causal induction as the result of

a domain-general statistical inference, informed by domain-specific prior knowledge. Having a formal computational account makes it possible to explain why effects of domain might not be observed in particular experiments, even if people are making use of domain-specific knowledge. It also helps to clarify the role of mechanism knowledge in causal induction, identifying the key factor as not the specific mechanisms that mediate between cause and effect, but the constraints on causal relationships that are implied by such mechanisms. The domain-general statistical principles that are used to infer causal structure can be extended to the level of causal theories, with the potential to provide insight into how those theories are learned. This perspective suggests a hierarchy of theories, with more general theories providing constraints that are used in learning theories at levels below, and provides the potential to explore the content and acquisition of those theories.

Chapter 9

Conclusion

Edmond Halley, James Currie, John Snow, and Ignaz Semmelweis all made scientific discoveries by evaluating a set of hypotheses constrained by their prior knowledge. Through the interaction of theory and data, they were able to infer the existence of new causal relationships. In this thesis, I have presented a series of examples that suggest that everyday causal inferences can be explained in the same way, as the consequence of a rational statistical inference evaluating hypothetical causal structures generated by an intuitive theory. This theory-based approach explains how people can learn so much from so little: they have expectations about which causal relationships are possible, and the form that those relationships should take. Given such constraints, only a small amount of data is required to infer causal relationships.

The theory-based causal induction framework specifies how shallow mechanism knowledge can be integrated with statistical inference. The formal theories I have presented spell out the aspects of this knowledge that have implications for causal induction: knowing the types of entities in a domain, the causal relationships in which they can participate, and the functional form of those relationships. By viewing a theory as a hypothesis-space generator, it becomes clear how such theories play a role in causal induction: they identify possible causal structures, and inform our beliefs about the prior probability of such structures. Taking this perspective allows the top-down influence of theories and the bottom-up influence of data to be combined via Bayesian inference. It also reveals how domain-general statistical principles can be integrated with domain-specific theories to facilitate the learning of new causal relationships.

While I have used causal graphical models as the basic representation of causal structure

throughout the thesis, I have argued that they cannot represent all of the content of human causal knowledge. Causal graphical models can express causal relationships, but they cannot express causal laws that state which relationships are possible and the form they take. Such laws are a key constituent of both scientific and intuitive theories (e.g., Carey, 1985b; Laudan, 1977; Wellman, 1990; Wellman & Gelman, 1992). In making this point, I drew upon an analogy to formal linguistics: causal graphical models play the same role in causal induction as syntactic structures like parse trees play in language comprehension. The grammar of a language places constraints on possible parse trees, and these constraints cannot be expressed in any single parse tree. Likewise, intuitive theories place constraints on causal graphical models, and these constraints cannot be expressed in a single causal graphical model. These claims also have implications for the adequacy of structure-learning algorithms as accounts of human causal induction. Standard structure learning algorithms do not exploit the kind of constraints on causal graphical models that are implied by an intuitive theory. As a consequence, they require much more data than human learners to identify a causal relationship.

The tension between statistical learning and prior knowledge that has motivated much of the literature concerning human causal induction is also behind many of the classic debates in cognitive science, such as similarity- vs. theory-based accounts of categorization (Murphy & Medin, 1985), and association- vs. knowledge-based language acquisition (Skinner, 1957; Chomsky, 1959). Each of these debates centers around a difficult inductive problem, and such problems can only be solved by using some form of statistical learning to choose among a constrained set of hypotheses. The controversy concerns the relative importance of these two factors, with one camp advocating powerful domain-general statistical learning with minimal constraints on hypotheses, and the other endorsing limited learning with strong domain-specific constraints. Examples of these two positions appear in the earliest days of inquiry into human cognition, but the question that they address – how the mind solves inductive problems – is fundamentally a computational question, and one that can be addressed using the tools of modern statistics. Expressing problems of induction as statistical problems makes it possible to go beyond the classical positions, and to formally evaluate how statistical learning interacts with constraints from prior knowledge.

Causal induction is a particularly promising arena for exploring the interaction between statistical learning and prior knowledge in human cognition, for two reasons. First, work in computer science and statistics has begun to make it possible to articulate the computational

problem underlying causal induction, suggesting methods for representing the structure of causal systems. Second, people learn about causal relationships in diverse settings, from being provided with contingency data, to observing dynamic physical systems. Each of these settings draws on different prior knowledge, and the variation in human judgments across settings provides insight into the role that such knowledge plays in causal learning. By establishing the prior knowledge that is necessary to explain causal inferences across a range of settings, we can begin to understand the principles by which that knowledge is organized, and the mechanisms by which it is used. Whether they concern the orbits of comets or computer-generated explosions, inferences about causal relationships provide insight into how people solve the difficult inductive problems that are at the heart of understanding human cognition.

Appendix A

Causal graphical models considered

While they have been (almost surprisingly) enthusiastically welcomed by psychologists, causal graphical models remain controversial in statistics (see, for example, Pearl, 1995, and the following discussion). Several statisticians and philosophers have strongly criticized the causal graphical model formalism and its proponents (in particular, Humphreys & Freedman, 1996; Freedman & Humphreys, 1998; Freedman, 2004; responses are provided by Korb & Wallace, 1997, and Spirtes, Glymour, & Scheines, 1997). My treatment of causal graphical models in Chapter 2 is relatively uncritical, focussing on the features of the formalism rather than its problems.[1] In this appendix, I will summarize these criticisms, and assess how they apply to the project undertaken in the thesis. I will focus on the criticisms presented by Humphreys and Freedman (1996; Freedman & Humphreys, 1998: Freedman, 2004). I see these criticisms as addressing three major issues: the definition of causality, algorithms for causal induction, and the possibility of inferring causation from observation. I have different reactions to these three groups of criticisms. I accept only some of the criticisms of the definition of causality, agree with the criticisms of algorithms for causal induction and believe that they are consistent with my argument, and agree with the skepticism about inferring causation from observation but believe that it is irrelevant to the project undertaken in this thesis.

[1] I thank Persi Diaconis for pointing this out, and for leading me to consider the issues discussed in this appendix.

A.1 Defining causality

Philosophers and statisticians have offered a variety of definitions of causality, including associative, generative, logical, and probabilistic accounts (e.g., Bunge, 1959; Harre & Madden, 1975; Holland, 1986; Rubin, 1990; Sosa & Tooley, 1993; Suppes, 1970). Spirtes et al. (1993), and more explicitly Pearl (2000), define causality as directed functional dependence: the existence of a functional relationship between two variables, where the direction of that relationship is used in inferring the consequences of intervention. i.e. if X causes Y, then $Y = f(X)$ for some f, and the state of Y can be influenced by intervening on X, but the state of X is not altered by intervening on Y.

Humphreys and Freedman (1996) criticize Spirtes et al. (1993) for failing to provide a non-circular definition of causality. This is a valid criticism, since Spirtes et al. (1993) were not explicit about the notion of causality that their account assumed, and did not explore its implications (Spirtes et al. 1997). However, I think subsequent work has diluted the content of this criticism. Pearl (2000) provided an explicit definition of causality along the lines above, and more recently Woodward (2003) has developed a philosophical account that starts from the notion of intervention rather than causality, defines causality in terms of intervention, and derives from this a framework equivalent to that offered by SGS and Pearl. It is this framework that I take as a starting point: a definition of causality as directed functional dependency, with the corresponding implications for statistical models of observation and intervention.

Whether this framework is either metaphysically valid or capable of capturing all aspects of human inferences remains an open question. Leaving aside the metaphysical issue, the work presented in this thesis suggests that the definition of causality as directed functional dependency might be more general than the intuitive use of the term. For example, the results presented in Chapter 4 show that people tend to interpret causation as a generative relation, with causes increasing the probability of their effects. If this is the case, then directed functional dependency is necessary for X to cause Y, but not sufficient.

Despite these shortcomings, the causal graphical model framework provides the best available formalism for exploring the questions that are central to understanding human causal induction. In particular, the relationship between intervention and causality assumed in causal graphical models both clearly defines the difference between causation and association and motivates why an agent should wish to learn about the former. Stating

causal relationships in terms of functional dependency makes it possible to formalize the kinds of expectations about causality that guide human inferences, and to differentiate expectations that concern causal structure from those that concern functional form.

A.2 Algorithms for causal induction

Humphreys and Freedman (1996) provide extensive criticisms of the algorithms of Spirtes et al. (1993). These algorithms involve conducting χ^2 tests to identify dependencies among variables, and then reasoning deductively from the pattern of dependencies to the set of causal structures that could account for those dependencies. As I discuss in depth in Chapter 5, there are two problems with these algorithms: they attempt to reduce an inductive problem to a deductive problem, and they do not incorporate prior knowledge. Humphreys and Freedman (1996) focus on the same points, as do Korb and Wallace (1997), who are otherwise in favor of causal graphical models. These criticisms are thus consistent with my argument, and the weaknesses of these algorithms do not pose a general problem for using the causal graphical model framework to explain human inferences.

A.3 The possibility of inferring causation

Humphreys and Freedman's (1996; Freedman & Humphreys, 1998; Freedman, 2004) most substantial objections are part of a more general concern about the willingness of social scientists to infer causation from correlational analyses, without checking the assumptions of their models and corroborating their analyses with other sources of data. This concern appears elsewhere in Freedman's work (e.g. Freedman, 1991; 1999). His conclusions are pessimistic: that causal relationships cannot be inferred "... unless there is substantial prior knowledge about the mechanisms that generated the data." (Freedman, 2004, p. 1). I agree with this pessimistic conclusion, but think that it does not obviate the use of this framework for examining human inferences. In fact, it supports the project undertaken in this thesis: people readily infer causal relationships, so what strong prior knowledge is allowing them to make these inferences?

The issues that arise in modeling cognition are very different from those that arise in statistical data analysis. This comes down to the fact that statisticians and psychologists have different unknowns. In any causal inference, there are three factors: (1) the observed

data, (2) the existence of the causal relationship, and (3) the set of assumptions under which the inference is made. The statistical questions about causality concern whether (2) can be inferred from (1) given a particular choice of assumptions (3), and whether these assumptions are justified in a given setting. Cognitive modeling involves observing that people *are* willing to infer (2) from (1) in a given setting, and asking what assumptions (3) guided that inference. These two sets of questions are related, in that both require a definition of causality that identifies what pairs of observations and assumptions justify the inference of causation. However, an answer to the psychological questions will not be an answer to the statistical questions, because the assumptions that people make about causality can be identified even if those assumptions are not justified.

This claim can be clarified by pursuing an analogy. Developing models of cognition that are based upon statistical inference is not about statistics, but meta-statistics. It is analogous to a situation in which one is provided with a dataset and the conclusions of a group of statisticians, and attempts to determine the assumptions that the statisticians were making in drawing their conclusions. Some of these assumptions may not have been valid, and some statisticians may have inferred a causal relationship where none exists as a consequence. But their inferences were hopefully rational given their assumptions, and their actions give clues to those assumptions.

Under this analysis, causal graphical models can be useful for cognitive modeling if they provide a means of identifying which pairs of observations and assumptions justify inferring a particular causal structure. The criticisms considered in this appendix fail to demonstrate that this is not the case. Freedman (2004) quotes a passage of personal communication from Judea Pearl that supports the claim that this is exactly what causal graphical models provide:

> Causal analysis with graphical models does not deal with defending modeling assumptions, in much the same way that differential calculus does not deal with defending the physical validity of a differential equation that a physicist chooses to use. In fact no analysis void of experimental data can possibly defend modeling assumptions. Instead, causal analysis deals with the conclusions that logically follow from the combination of data and a given set of assumptions, just in case one is prepared to accept the latter. (Pearl, personal communication, cited in Freedman, 2004, p. 15)

Appendix B

Contingencies

This appendix summarizes some of the technical results concerning the models discussed in Chapter 4. I show that ΔP and causal power are maximum likelihood parameter estimates, and discuss how causal support can be computed, and how it relates to χ^2.

B.1 Maximum likelihood parameter estimates

Both ΔP and causal power are maximum likelihood estimates of the causal strength parameter for C in Graph 1 of Figure 2.1 (a), but under different parameterizations. For any parameterization of Graph 1, the log likelihood of the data is given by

$$\log P(\mathcal{D}|w_0, w_1) = \sum_{e,c} N(e,c) \log P(e|c) \tag{B.1}$$

where $\mathcal{D}$ is contingency data $N(e,c)$, $P(e|c)$ is the probability distribution implied by the model, suppressing its dependence on w_0, w_1, and $\sum_{e,c}$ denotes a sum over all pairs of e^+, e^- and c^+, c^-. Equation B.1 is maximized whenever w_0 and w_1 can be chosen to make the model probabilities equal to the empirical probabilites:

$$P(e^+|c^+, b^+; w_0, w_1) = P(e^+|c^+), \tag{B.2}$$

$$P(e^+|c^-, b^+; w_0, w_1) = P(e^+|c^-). \tag{B.3}$$

To show that ΔP corresponds to a maximum likelihood estimate of w_1 under a linear parameterization of Graph 1, we identify w_1 in Equation 2.4 with ΔP (Equation 4.1), and

w_0 with $P(e^+|c^-)$. Equation 2.4 then reduces to $P(e^+|c^+)$ for the case $c = c^+$ and to $P(e^+|c^-)$ for the case $c = c^-$, thus satisfying the sufficient conditions in Equations B.2-B.3 for w_0 and w_1 to be maximum likelihood estimates. To show that causal power corresponds to a maximum likelihood estimate of w_1 under a noisy-OR parameterization, we follow the analogous procedure: identify w_1 in Equation 2.1 with causal power (Equation 4.2), and w_0 with $P(e^+|c^-)$. Then Equation 2.1 reduces to $P(e^+|c^+)$ for $c = c^+$ and to $P(e^+|c^-)$ for $c = c^-$, again satisfying the conditions for w_0 and w_1 to be maximum likelihood estimates.

B.2 Evaluating causal support

Causal support is defined as the log likelihood ratio in favor of Graph 1 over Graph 0:

$$\text{support} = \log \frac{P(\mathcal{D}|\text{Graph 1})}{P(\mathcal{D}|\text{Graph 0})}. \tag{B.4}$$

We obtain the likelihoods $P(\mathcal{D}|\text{Graph 1}), P(\mathcal{D}|\text{Graph 0})$ by integrating out the parameters w_0, w_1. This means that each value of the parameters is assigned a prior probability, and this probability is combined with the likelihood of the data given the structure and the parameters to give a joint distribution over data and parameters given the structure. We can then sum over all values that the parameters can take on, to result in the probability of the data given the structure. Thus, if we want to compute the probability of the observed data for the structure depicted by Graph 1, we have

$$P(\mathcal{D}|\text{Graph 1}) = \int_0^1 \int_0^1 P(\mathcal{D}|w_0, w_1, \text{Graph 1})\ P(w_0, w_1|\text{Graph 1})\ dw_0\ dw_1 \tag{B.5}$$

and the equivalent value for Graph 0 is given by

$$P(\mathcal{D}|\text{Graph 0}) = \int_0^1 P_0(\mathcal{D}|w_0, \text{Graph 0})\ P(w_0|\text{Graph 0})\ dw_0. \tag{B.6}$$

where the likelihoods $P(\mathcal{D}|w_0, w_1, \text{Graph 1}), P(\mathcal{D}|w_0, \text{Graph 0})$ are specified by the parameterization of the graph, and the prior probabilities $P(w_0, w_1|\text{Graph 1})$, $P(w_0|\text{Graph 0})$ are set a priori. Integrating over all values of the parameters penalizes structures that require more parameters, simply because the increase in the dimensionality of the space over which the integrals are taken is usually disproportionate to the size of the region for which the likelihood is improved.

For generative causes, $P(\mathcal{D}|\text{Graph 1})$ is computed using the noisy-OR parameterization, and for preventive causes, it is computed using the noisy-AND-NOT. We also need to define prior probabilities $P(w_0, w_1|\text{Graph 1})$ and $P(w_0|\text{Graph 0})$, to which we assign a uniform density. Because causal support depends on the full likelihood functions for both Graph 1 and Graph 0, we may expect causal support to be modulated by causal power, but only in interaction with other factors that determine how much of the posterior probability mass for w_1 in Graph 1 is bounded away from zero (where it is pinned in Graph 0).

B.3 An algorithm for computing causal support

Equation B.6 can be evaluated analytically. If w_0 denotes the probability of the effect occurring regardless of the presence or absence of the cause and we take a uniform prior on this quantity, we have

$$P(\mathcal{D}|\text{Graph 0}) = \int_0^1 w_0^{N(e^+)}(1-w_0)^{N(e^-)}\, dw_0 = B(N(e^+)+1, N(e^-)+1) \tag{B.7}$$

where $B(r,s)$ is the beta function, and $N(e+)$ is the marginal frequency of the effect. For integers r and s, $B(r,s)$ can be expressed as a function of factorials, being $\frac{(r-1)!(s-1)!}{(r+s-1)!}$. In general Equation B.5 cannot be evaluated analytically, but it can be approximated simply and efficiently by Monte Carlo simulation. Since we have uniform priors on w_0 and w_1, we can obtain a good approximation to $P(\mathcal{D}|\text{Graph 1})$ by drawing m samples of w_0 and w_1 from a uniform distribution on $[0,1]$ and computing

$$P(\mathcal{D}|\text{Graph 1}) \approx \frac{1}{m}\sum_{i=1}^{m} P(\mathcal{D}|w_{0i}, w_{1i}, \text{Graph 1}) \tag{B.8}$$

where w_{0i} and w_{1i} are the ith sampled values of w_0 and w_1. We thus need only compute the probability of the observed scores D under this model for each sample, which can be done efficiently using the counts from the contingency table. This probability can be written as

$$P(\mathcal{D}|w_{0i}, w_{1i}, \text{Graph 1}) = \prod_{e,c} P(e|c, b^+; w_{0i}, w_{1i})^{N(e,c)} \tag{B.9}$$

where the product ranges over e^+, e^- and c^+, c^-, and $P(e|c; w_{0i}, w_{1i})$ reflects the chosen parameterization – noisy-OR for generative causes, and noisy-AND-NOT for preventive.

As with all Monte Carlo simulations, the accuracy of the results improves as m becomes large. For the examples presented in this paper, we used $m = 100,000$.

B.4 The χ^2 approximation

For large samples we can approximate the value of causal support with the familiar χ^2 test for independence. There are both intuitive and formal reasons for the validity of the χ^2 approximation. Intuitively, the relationship holds because the χ^2 statistic is used to test for the existence of statistical dependency between two variables, and C and E are dependent in Graph 1 but not in Graph 0. A large value of χ^2 indicates that the null hypothesis of independence should be rejected, and that Graph 1 is favored. However, χ^2 assumes a different parameterization of Graph 1 from causal support, and the two will only be similar for large samples.

The formal demonstration of the relationship between χ^2 and causal support is as follows. When the likelihood $P(\mathcal{D}|w_0, w_1)$ is extremely peaked (e.g., in the limit $N\to\infty$), we may replace the integrals in Equation B.5 with supremums over w_0, w_1. That is, the marginal likelihood essentially becomes the maximum of the likelihood, and causal support reduces to the ratio of likelihood maxima – or equivalently, the difference in loglikelihood maxima – for Graph 1 and Graph 0. Under these circumstances causal support reduces to the frequentist likelihood ratio statistic, equal to half of the G^2 statistic (e.g., Wickens, 1989). Correspondingly, Pearson's χ^2 for independence,

$$\chi^2 = N \sum_{e,c} \frac{(P(e,c) - P(e)P(c))^2}{P(e)P(c)}, \tag{B.10}$$

can be shown to approximate twice causal support by a Taylor series argument: the second order Taylor series of $\sum_i p_i \log \frac{p_i}{q_i}$, expanded around $p = q$, is $\frac{1}{2}\sum_i \frac{(p_i - q_i)^2}{q_i}$ (Cover & Thomas, 1991). The χ^2 approximation only holds when ΔP is small and N is large.

For learning with rates, the likelihood ratio statistic for comparing Graph 0 and Graph 1 under the parameterization given in Chapter 6 is

$$N(c^+)\log N(c^+) + N(c^-)\log N(c^-) - (N(c^+) + N(c^-))\log \frac{N(c^+) + N(c^-)}{2}, \tag{B.11}$$

which, by essentially the same argument as that given above for G^2, will approximate the

value of causal support in the large sample limit. Using the Taylor series argument employed in the contingency case, we obtain the χ^2 approximation given in Equation 6.5, which holds only when the difference in rates is small.

Appendix C

Stick-balls

The hypothesis spaces generated by the theory of stick-ball machines presented in Chapter 5 can include structures in which balls can cause one another to move, such as that shown in Figure C.1 (a). Cyclic relationships are usually prohibited in graphical models (Pearl, 1988), although some kinds of cycles can be dealt with in causal graphical models (e.g., Pearl, 2000). One common approach to dealing with cycles is to impose temporal structure on a set of events, unrolling the cycle into a set of dependencies that hold between two variables in successive time slices. In this appendix, I will outline how this approach can be used to deal with cyclic causal relationships in stick-ball machines.

Given N_B balls, with B_i indicating $\texttt{Moves}(\texttt{b}_i, \texttt{T})$ and N_H hidden causes, with H_j indicating $\texttt{Active}(\texttt{h}_j, \texttt{T})$, the following procedure can be used to generate values of B_i and H_j on any given trial:

1. Determine which hidden causes are active by sampling the values of H_j.

2. Determine which balls are moved by the hidden causes by sampling the values of B_i, conditioned just on H_j. If $H_j \rightarrow B_i$ and $h_j = 1$, then $b_i = 1$ with probability ω.

3. Determine which balls move other balls. Every ball that moves has one opportunity to move the balls to which it has causal connections. If $B_i \rightarrow B_j$, b_j is currently 0, and $\texttt{b}_i$ has not previously attempted to move $\texttt{b}_j$, then $b_j = 1$ with probability ω.

4. Repeat step 3 with the balls that were just moved by other balls. This procedure continues until all balls that have moved have had one opportunity to influence each of the balls to which they are connected.

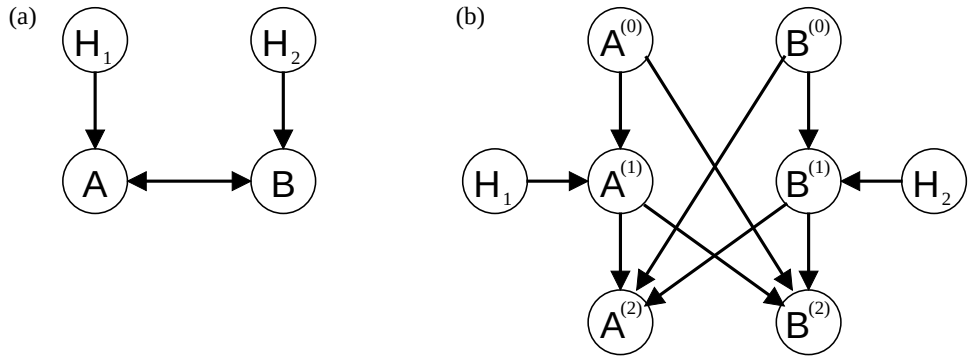

Figure C.1: Dealing with bidirectional causal relationships in stick-ball machines. (a) A causal graphical model generated by the theory of stick-ball machines given in Chapter 5. (b) The same model "unrolled" through time, removing the cyclic causal relationship. Both models show the causal relationships among variables on a single trial, but can be quantified over trials as discussed in Chapter 2.

This procedure implicitly defines a temporal succession of events, with hidden causes becoming active, then moving a subset of the balls, each of which moves some subset of the remaining balls, each of which moves further balls, and so forth. This temporal succession removes the cycles in the underlying causal graphical model, allowing events to unroll through time.

The generative procedure described in the previous paragraph can be expressed as a recipe for constructing an "unrolled" graphical model from the basic graphical model. The unrolled model can be used to compute the probability of events, and is constructed as follows:

1. Create $N_B + 1$ copies of B_i, numbered from 0 to N_B, indicating successive points in time within the trial. I will use $B_i^{(t)}$ to refer to the copy of B_i at time t.

2. $B_i^{(t-1)} \rightarrow B_i^{(t)}$ for all $t \geq 1$. If $H_j \rightarrow B_i$ in the basic model, then $H_j \rightarrow B_i^{(1)}$ in the unrolled model. If $B_i \rightarrow B_j$ in the basic model, then $B_i^{(t-1)} \rightarrow B_j^{(t)}$ and $B_i^{(t-2)} \rightarrow B_j^{(t)}$ for all $t \geq 2$ in the unrolled model.

3. Set $b_i^{(0)} = 0$. Parameterize $B_i^{(1)}$ as a noisy-OR of H_j. For $t \geq 2$, parameterize $B_i^{(t)}$ as

$$P(b_i^{(t)} = 1 | b_\cdot^{(t-1)}, b_\cdot^{(t-2)}) = 1 - (1 - b_i^{(t-1)})(1 - \omega)^{\sum_{j \neq i} b_j^{(t-1)}(1 - b_j^{(t-2)})}.$$

   Under this parameterization $b_i^{(t)} = 1$ if $b_i^{(t-1)} = 1$, and is otherwise a noisy-OR of all

B_j that changed between $t-2$ and $t-1$.

The probability of a set of observed values for B_i in the basic model is the probability of $B_i^{(N_B)}$ taking those values in the unrolled model, summing over all latent variables. Figure C.1 (b) shows the unrolled model for the basic model shown in Figure C.1 (a).

Appendix D

Explosions

This appendix provides some of the technical details behind the analysis of explosions presented in Chapter 6. I will present a theory of explosions that uses only Boolean predicates, and indicate how this reduces to the theory given in the chapter. I will then explain how quantities like $P(\mathcal{D}|\text{Graph 0})$ can be computed under this theory, outline the generative procedure that is described by such probability distributions, and show how these results can be used to answer questions about the causes of particular explosions.

D.1 A Boolean theory

Figure D.1 presents a theory for causal induction with explosives that uses only Boolean predicates. All variables are parameterized as Poisson processes, with rates that depend upon the variables with which they have causal relationships. On the assumption that a can explodes only once and a hidden cause becomes active only once, these events are the first produced by the appropriate Poisson processes. It follows from the second property of Poisson processes given in Figure 6.2 that these times have exponential distributions, resulting in the theory given in Figure 6.6.

A causal graphical model generated by this theory is shown in Figure D.2. This model is similar in spirit to a dynamic Bayesian network (e.g., Murphy, 2003), depicting the relationships that hold among a set of variables over time. There are two important differences from standard depictions of dynamic Bayesian networks. First, this model is defined for continuous time, while dynamic Bayesian networks typically describe discrete trials. Second, the relationships shown here are quantified over all future times, while dynamic Bayesian

Ontology:

Types	Number
Can	$N_C \sim P_C$
HiddenCause	$N_H = \infty$
Time	$\mathbb{R}$
Space	$\mathbb{R}$

Predicates	Values	
Explodes(Can, Time)	Boolean:	{T, F}
Active(HiddenCause, Time)	Boolean:	{T, F}
Located(Can, Space)	Boolean:	{T, F}

Plausible relations:

$\texttt{Explodes}(\texttt{C}_1, \texttt{T}) \rightarrow \texttt{Explodes}(\texttt{C}_2, \texttt{T}')$
True for with probability 1 for all T and all $\texttt{T}' > \texttt{T}$, for each $\texttt{C}_1 \neq \texttt{C}_2$ pair

$\texttt{Active}(\texttt{H}, \texttt{T}) \rightarrow \texttt{Explodes}(\texttt{C}, \texttt{T})$
Each C has an edge from some H with probability 1, which holds for all T. The particular H is chosen based upon the number of existing edges:

$$P(\texttt{Active(H,T)} \rightarrow \texttt{Explodes(C,T)}) \propto \begin{cases} M_{\texttt{H},i} & M_{\texttt{H},i} > 0 \\ s & \texttt{H is new} \end{cases}$$

where $M_{\texttt{H},i}$ is the number of outgoing edges from H when the edges are chosen for the ith can.

Functional form:

$\texttt{Active}(\texttt{H}, \texttt{T}) \sim \text{PoissonProcess}(\alpha)$

$\texttt{Explodes}(\texttt{C}_1, \texttt{T}) \sim \text{PoissonProcess}(\lambda(\texttt{T}))$ for $\lambda(\texttt{T})$ from a continuous noisy-OR:

Cause	Strength	Times
(Background)	$\lambda_0 = 0$	
$\texttt{Active}(\texttt{H}, \texttt{T})$	$\lambda_i = \omega$	$\{\texttt{T} \mid \texttt{Active(H,T)}\}$
$\texttt{Explodes}(\texttt{C}_2, \texttt{T})$	$\lambda_i = \omega$	$\{\texttt{T} \mid \texttt{T} = T_{C_2} + D(\texttt{C}_1, \texttt{C}_2)/\mu\}$

Figure D.1: Theory for causal induction with explosives using Boolean predicates. $T_{\texttt{C}}$ indicates the time at which can C explodes, while $D(\texttt{C}_1, \texttt{C}_2)$ is the distance between the locations of cans $\texttt{C}_1$ and $\texttt{C}_2$.

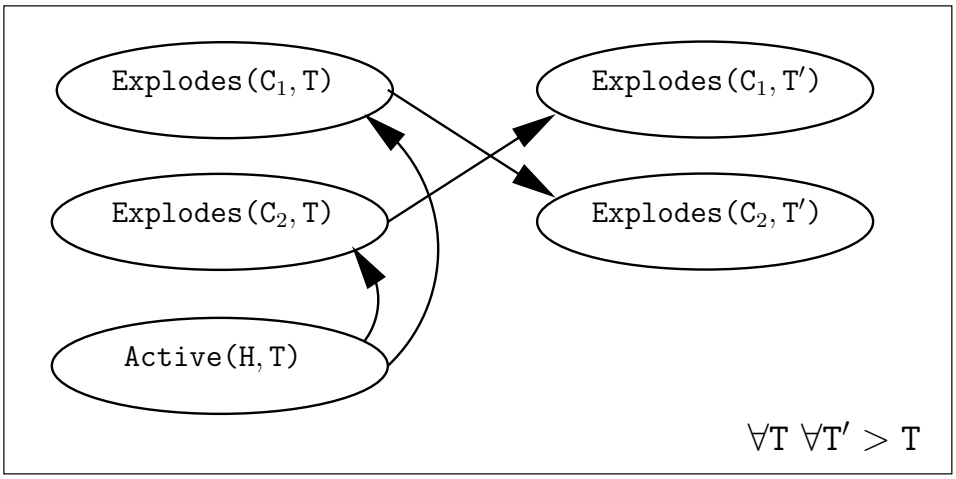

Figure D.2: Causal relationships among variables over time for a system in which $N_C = 2$. The dependence of each variable on its previous state and the dependence of $\texttt{Explodes}(\texttt{c}_i, \texttt{T})$ on $\texttt{Located}(\texttt{c}_i, \texttt{S})$ are not shown in this figure.

networks typically show only the relationships that hold between successive trials. This is fundamentally a constraint of tractability: in general, quantifying over all future times will result in a network in which some nodes have an unbounded number of parents. In this case, the functional form of the causal relationships involved is such that only a small subset of these parents are relevant, and computations can still be performed efficiently.

D.2 Evaluating probabilities

The theory of explosions given in Chapter 6 (and re-expressed above) allows us to compute the probability of any set of explosion times under any causal structure. I will demonstrate how such probabilities can be computed for Graph 0, in which each can has a separate hidden cause. In this causal structure

$$P(\mathcal{C}|\text{Graph } 0, \alpha, \omega, \mu) = \int_0^\infty \cdots \int_0^\infty P(\mathcal{C}|h_1, \ldots, h_{N_C}) \left[\prod_{i=1}^{N_C} P(h_i)\ dh_i \right] \tag{D.1}$$

where h_i is the time at which hidden cause $\texttt{h}_i$ becomes active. Since each h_i follows an Exponential(α) distribution, $P(h_i) = \alpha \exp\{-\alpha h_i\}$. I will derive $P(\mathcal{C}|h_1, \ldots, h_{N_C})$, and use this to compute $P(\mathcal{C}|\text{Graph } 0, \alpha, \omega, \mu)$.

The explosion times of the cans follow Exponential($\lambda(c_i)$) distributions, with an explosion at time c_i having probability $\lambda(c_i)\exp\{-\int_0^{c_i}\lambda(t)\ dt\}$. Let t_{ij}, for $i \neq j$, denote the time at which can $\mathtt{c}_i$ should explode as a result of the explosion of can $\mathtt{c}_j$ at time c_j, being $t_{ij} = c_j + D(\mathtt{c}_i, \mathtt{c}_j)/\mu$. Then, under the continuous noisy-OR parameterization assumed in the theory,

$$P(\mathcal{C}|h_1, \ldots, h_{N_C}) = \prod_{i=1}^{N_C} \omega(\delta(c_i, h_i) + \sum_{j\neq i} \delta(c_i, t_{ij}))\exp\{-\omega(m_i + I(h_i < c_i))\}, \qquad \text{(D.2)}$$

where $m_i = \sum_{j\neq i} I(t_{ij} < c_i)$ is the number of missed opportunities to explode as a result of other explosions, and $I(\cdot)$ is an indicator function taking the value 1 when its argument is true, and 0 otherwise. Substituting this probability into Equation D.1, we obtain

$$\begin{aligned} P(\mathcal{C}|\text{Graph } 0, \alpha, \omega, \mu) &= \omega^{N_C}\exp\{-\omega\sum_{i=1}^{N_C} m_i\}\prod_{i=1}^{N_C}\alpha\exp\{-\alpha c_i\} + \\ &\quad \sum_{j\neq i}\delta(c_i, t_{ij})\left(\exp\{-\alpha c_i\} + (1-\exp\{-\alpha c_i\})\exp\{-\omega\}\right) \qquad \text{(D.3)} \end{aligned}$$

This gives the probability of $\mathcal{C}$ for each choice of α, ω, and μ, and can be used to evaluate the posterior distribution over these parameters as described in Chapter 6.

D.3 A generative procedure

The probability distributions described in the previous section seem relatively complex, but can be derived as the result of a simple generative procedure. This procedure is the continuous analogue of that defined in Appendix C for the stick-ball machine, and proceeds as follows:

1. Sample the time H_j at which each hidden cause $\mathtt{h}_j$ becomes active from an Exponential(α) distribution.

2. Conditioned on H_j, sample $C_i^{(1)}$ for each can. With probability $1-\exp\{-\omega\}$, the can explodes at the time its hidden cause becomes active; with probability $\exp\{-\omega\}$ it does not explode, and we set $C_i^{(1)} = \infty$.

3. Starting with the earliest $c_i^{(1)}$, determine the times t_{ij} at which the shockwave originating from $\mathtt{c}_i$ would encounter each other can $\mathtt{c}_j$. Set $c_i^{(2)} = c_i^{(1)}$. For $j \neq i$, sample

the explosion time $C_j^{(*)}$ that would result from a shockwave originating at $\mathtt{c}_i$: with probability $1 - \exp\{-\omega\}$, $\mathtt{c}_j$ explodes at t_{ij}; with probability $\exp\{-\omega\}$, it does not explode and $c_j^{(*)} = \infty$. Set $c_j^{(2)} = \min(c_j^{(1)}, c_j^*)$.

4. Repeat step 3 with the next earliest $c_i^{(2)}$, and continue this procedure until values of $C_i^{(N_C)}$ are obtained. At this point, shockwaves originating from each can have had the opportunity to influence all other cans.

The probability of a set of explosion times $\mathcal{C}$ is the probability of $C_i^{(N_C)}$ under this generative procedure. This procedure treats each possible cause as a separate process that has an opportunity to produce the explosion. Since the minimum of the times generated from exponential distributions with rates λ_1 and λ_2 has an exponential distribution with rate $\lambda_1 + \lambda_2$, taking the minimum over all of these processes yields the distribution given in Equation D.2.

D.4 What caused what?

In Graph 0, the probability that the hidden cause $\mathtt{h}_i$ is the actual cause of the explosion of can $\mathtt{c}_i$ (denoted $H_i \Rightarrow C_i$), is given by

$$P(H_i \Rightarrow C_i | \mathcal{C}, \text{Graph 0}) = \int_0^\infty P(H_i \Rightarrow C_i | h_i, \mathcal{C}, \text{Graph 0}) P(h_i | \mathcal{C}, \text{Graph 0})\ dh_i. \quad \text{(D.4)}$$

The right hand side of this equation has two parts: the probability that $\mathtt{h}_i$ is the actual cause of the explosion of $\mathtt{c}_i$ conditioned on the time at which $\mathtt{h}_i$ becomes active, h_i, and the explosion times of all of the cans, $\mathcal{C}$, and the posterior distribution on h_i given $\mathcal{C}$. Both parts require elaboration.

Using the fourth property of Poisson processes from Figure 6.2, the probability that a potential cause was responsible for producing an explosion is proportional to the contribution of that cause to the rate of explosions at the time when the explosion occurred. The probability that the hidden cause $\mathtt{h}_i$ produced the explosion of $\mathtt{c}_i$ at time c_i is thus

$$P(H_i \Rightarrow C_i | h_i, \mathcal{C}, \text{Graph 0}) = \frac{\delta(c_i, h_i)}{\delta(c_i, h_i) + \sum_{j \neq i} \delta(c_i, t_{ij})}.$$

The consequences of taking ratios of delta functions can be determined by using the definition of $\delta(\cdot,\cdot)$ as the limit of $\phi_\sigma(\cdot,\cdot)$ as $\sigma\rightarrow 0$. Thus

$$\begin{aligned} P(H_i \Rightarrow C_i | h_i, \mathcal{C}, \text{Graph 0}) &= \lim_{\sigma \longrightarrow 0} \frac{\phi(c_i, h_i)}{\phi(c_i, h_i) + \sum_{j \neq i} \phi(c_i, t_{ij})} \\ &= \lim_{\sigma \longrightarrow 0} \frac{1}{1 + \sum_{j \neq i} \exp\{((c_i - h_i)^2 - (c_i - t_{ij})^2)/2\sigma^2} \\ &= \frac{1}{1 + (\sum_{j \neq i} I(c_i = t_{ij}))/I(c_i = h_i)} \end{aligned} \tag{D.5}$$

where the ratio in the denominator remains well-defined as at least one of its terms must take a non-zero value.

The posterior distribution over all the H_i given $\mathcal{C}$ can be computed by applying Bayes' rule,

$$P(h_1, \ldots, h_{N_C} | \mathcal{C}, \text{Graph 0}) = \frac{P(\mathcal{C} | h_1, \ldots, h_{N_C}, \text{Graph 0}) \prod_{i=1}^{N_C} P(h_i)}{P(\mathcal{C} | \text{Graph 0})},$$

where the likelihood is supplied by Equation D.2, the prior on h_i is Exponential(α), and the denominator is given in Equation D.3. In Graph 0, the H_i are independent conditioned on $\mathcal{C}$, so we can evaluate the posterior for the hidden cause of interest without taking the other hidden causes into account. The result is

$$P(h_i | \mathcal{C}, \text{Graph 0}) = \frac{(\delta(c_i, h_i) + \sum_{j \neq i} \delta(c_i, t_{ij}) \exp\{-\omega I(h_i < c_i)\}) \alpha \exp\{-\alpha h_i\}}{\alpha \exp\{-\alpha c_i\} + \sum_{j \neq i} \delta(c_i, t_{ij}) (\exp\{-\alpha c_i\} + (1 - \exp\{-\alpha c_i)\}) \exp\{-\omega\})}. \tag{D.6}$$

If $c_i \neq t_{ij}$ for all j, then this is a delta function at c_i. If $c_i = t_{ij}$ for some j, then the posterior has support for all values of h_i, but retains a delta spike at c_i.

Equations D.5 and D.6 are the two components we need to compute the probability that $\mathtt{h}_i$ caused $\mathtt{c}_i$ to explode. Equation D.5 shows that $P(H_i \Rightarrow C_i | h_i, \mathcal{C}, \text{Graph 0})$ only takes on non-zero values when $c_i = h_i$. Consequently, Equation 6.6 gives $P(H_i \Rightarrow C_i | \mathcal{C}, \text{Graph 0})$ equal to

$$\frac{1}{1 + (\sum_{j \neq i} I(c_i = t_{ij}))/I(c_i = h_i)} \frac{1}{1 + \frac{1}{\alpha} \sum_{j \neq i} \delta(c_i, t_{ij}) (1 + (\exp\{\alpha c_i\} - 1) \exp\{-\omega\})},$$

where the first term reflects the ratio of the contribution to the rate at $c_i = h_i$, and the second is the mass associated with the delta function at $c_i = h_i$ in the posterior distribution of h_i. This expression has two important qualitative properties: that the probability that $\mathtt{h}_i$ caused $\mathtt{c}_i$ to explode is 1 if no other cans could have been responsible, and that this

probability is less than $\frac{1}{2}$ if another can could have been responsible.

Appendix E

Bombing

This appendix provides some of the technical details behind the analysis of coincidences in bombing presented in Chapter 7. I will present a theory of bombing that uses only Boolean predicates, and indicate how this reduces to the theory given in the chapter. I will then explain how quantities like $P(\mathcal{D}|h_1)$ can be computed by approximating the sum over all possible causal structures.

E.1 A Boolean theory

Figure D.1 presents a theory for causal induction with explosives that uses only Boolean predicates. All variables are parameterized as Poisson processes, with rates that depend upon the variables with which they have causal relationships. Targets arise with constant rate throughout the space. The rate at which bombs fall increases around a target is the result of a continuous noisy-OR. Rather than using the Dirac delta function as a convolution kernel, here I use the probability density function for a multivariate Gaussian with covariance matrix Σ, denoted $\phi_\Sigma(\cdot, \cdot)$.

The Poisson processes defined in the theories given in Figure E.1 become probability density in the theories given in Figure 7.7 through the assumption that $\texttt{Location}(\texttt{T}, \texttt{S})$ and $\texttt{Explodes}(\texttt{B}, \texttt{S})$ are each true at a single point $\texttt{S}$. For example, if L_i denotes the point $\texttt{S}$ such that $\texttt{Location}(\texttt{t}_i, \texttt{S})$ is true, we can derive a probability density function for L_i for each of the theories shown in Figure E.1. It follows from the properties of the Poisson process that the probability density at each point is proportional to the rate at that point, so in this case L_i has a uniform distribution over the space. If X_i denotes the point $\texttt{S}$ such that

h_0:

Ontology:

Types	Number	Predicates	Values	
`Bomb`	$N_B \sim P_B$	`Explodes(Bomb, Space)`	`Boolean:`	{T, F}
`Target`	$N_T = \infty$	`Location(Target, Space)`	`Boolean:`	{T, F}
`Space`	$\mathcal{R} \subset \mathbb{R}^2$			

Plausible relations:

`Location(T, S)`→`Explodes(B, S′)`
Each `B` has an edge from a unique `T` with probability 1, which holds for all `S` and `S′`

Functional form:

`Location(T, S)` ∼ PoissonProcess(α)
`Explodes(B, S)` ∼ PoissonProcess(λ(`S`)) for λ(`S`) from a continuous noisy-OR

Cause	Strength	Points
(Background)	$\lambda_0 = 0$	
`Location(T, S)`	$\lambda_1 = \omega$	{`S`|`Location(T, S)`}

h_1:

Ontology:

Types	Number	Predicates	Values	
`Bomb`	$N_B \sim P_B$	`Explodes(Bomb, Space)`	`Boolean:`	{T, F}
`Target`	$N_T = \infty$	`Location(Target, Space)`	`Boolean:`	{T, F}
`Space`	$\mathcal{R} \subset \mathbb{R}^2$			

Plausible relations:

`Location(T, S)` → `Explodes(B, S′)`
Each `B` has an edge from some `T` with probability 1, which holds for all `S` and `S′`. With probability p, `T` = $\mathtt{t_C}$, a common target, otherwise `T` is unique to `B`.

Functional form:

`Location(T, S)` ∼ PoissonProcess(α)
`Explodes(B, S)` ∼ PoissonProcess(λ(`S`)) for λ(`S`) from a continuous noisy-OR

Cause	Strength	Points
(Background)	$\lambda_0 = 0$	
`Location(T, S)`	$\lambda_1 = \omega$	{`S`|`Location(T, S)`}

Figure E.1: Theories for coincidences in bombing using Boolean predicates.

$\texttt{Explodes}(\texttt{b}_i, \texttt{S})$ is true, then a similar argument applied to the Gaussian convolution kernel used in defining the rate of $\texttt{Explodes}(\texttt{B}, \texttt{S})$ indicates that X_i has a multivariate Gaussian distribution.

E.2 Evaluating probabilities

Computing $P(\mathcal{D}|h_1)$ for $\mathcal{D} = \{x_1, \ldots x_{N_B}\}$ requires evaluating

$$P(\mathcal{D}|h_1) = \sum_{i=0}^{2^{N_B}-1} \left[\int_{\mathcal{R}} \int P(\mathcal{D}|\Sigma, \ell_{\texttt{C}}, \text{Graph } i) P(\Sigma) P(\ell_{\texttt{C}}) \; d\Sigma \; d\ell_{\texttt{C}} \right] \int_0^1 P(\text{Graph } i|p) P(p) \; dp. \quad \text{(E.1)}$$

I will explain how this sum can be computed evaluating the bracketed term for Graph $2^{N_B} - 1$, in which all X_i are drawn from a single Gaussian distribution, summing over the mean $\ell_{\texttt{C}}$ and covariance Σ, and then discussing how the result can be generalized. More details on the kind of computations performed in this section can be found in Minka (2001).

I will assume a uniform prior on $\ell_{\texttt{C}}$, with $P(\ell_{\texttt{C}}) = 1/|\mathcal{R}|$ for $\ell_{\texttt{C}} \in \mathcal{R}$, and an inverse Wishart prior on Σ with parameters $\mathbf{I}, k$, where $\mathbf{I}$ is the d-dimensional identity matrix. Under this prior,

$$P(\Sigma) = \frac{1}{c_{kd}|\Sigma|^{(k+d+1)/2}} \exp\{-\frac{1}{2}\text{tr}(\Sigma^{-1})\}$$

where $c_{kd} = 2^{kd/2}\pi^{d(d-1)/4} \prod_{j=1}^{d} \Gamma((k+1-j)/2)$ and Γ is the standard gamma function (e.g., Boas, 1983). Given Σ and $\ell_{\texttt{C}}$, we have

$$\begin{aligned} P(\mathcal{D}|\Sigma, \ell_{\texttt{C}}, \text{Graph } 2^{N_B}-1) &= \frac{1}{|2\pi\Sigma|^{N_B/2}} \exp\{-\frac{1}{2}\sum_{i=1}^{N_B}(x_i - \ell_{\texttt{C}})^T\Sigma^{-1}(x_i - \ell_{\texttt{C}})\} \\ &= \frac{1}{|2\pi\Sigma|^{N_B/2}} \exp\{-\frac{N_B}{2}(\bar{x} - \ell_{\texttt{C}})^T\Sigma^{-1}(\bar{x} - \ell_{\texttt{C}}) - \frac{1}{2}\text{tr}(\mathbf{S}\Sigma^{-1})\} \end{aligned}$$

where $\bar{x}$ is $\frac{1}{N_B}\sum x_i$ and $\mathbf{S} = \sum_{i=1}^{N_B}(x_i - \bar{x})(x_i - \bar{x})^T$.

Using these definitions, we can express our integral as

$$\begin{aligned} P(\mathcal{D}|\text{Graph } 2^{N_B}-1) &= \int_{\mathcal{R}} \int P(\mathcal{X}|\Sigma, \ell_{\texttt{C}}, \text{Graph } 2^{N_B}-1) P(\Sigma) P(\ell_{\texttt{C}}) \\ &= \frac{1}{|\mathcal{R}| c_{kd} (2\pi)^{dN_B/2}} \int \frac{1}{|\Sigma|^{(N_B+k+d+1)/2}} \exp\{-\frac{1}{2}\text{tr}((\mathbf{S}+\mathbf{I})\Sigma^{-1})\} \\ &\qquad \left[\int_{\mathcal{R}} \exp\{\frac{N_B}{2}(\bar{x} - \ell_{\texttt{C}})^T\Sigma^{-1}(\bar{x} - \ell_{\texttt{C}})\} \; d\ell_{\texttt{C}} \right] d\Sigma. \end{aligned}$$

The bracketed integrand has the form of a Gaussian. The result is upper bounded by $|2\pi\Sigma/N_B|^{1/2}$, with the tightness of the bound increasing with the size of $\mathcal{R}$ relative to Σ/N_B. This reduces the outer integral to

$$\begin{aligned} P(\mathcal{D}|\text{Graph } 2^{N_B}-1) &\approx \frac{1}{|\mathcal{R}|c_{kd}(2\pi)^{d(N_B-1)/2}N_B^{d/2}}\int \frac{1}{|\Sigma|^{(N_B+k+d)/2}}\exp\{-\frac{1}{2}\text{tr}((\mathbf{S}+\mathbf{I})\Sigma^{-1})\}\ d\Sigma \\ &= \frac{1}{|\mathcal{R}|\pi^{d(N_B-1)/2}N_B^{d/2}|\mathbf{S}+\mathbf{I}|^{(N_B+k-1)/2}}\prod_{j=1}^{d}\frac{\Gamma((N_B+k-j)/2)}{\Gamma((k+1-j)/2)} \end{aligned} \tag{E.2}$$

where the result follows from the fact that the integrand has the form of an inverse Wishart distribution. The expression given in Equation E.2 is a measure of the "Gaussianity" of $\mathcal{D}$: it is the probability of $\mathcal{D}$ being produced from some Gaussian distribution, assessed under the priors specified on $\ell_{\mathtt{C}}$ and Σ.

This result can be extended to allow us to evaluate the probability of $\mathcal{D}$ under any other graph. For each graph, the X_i can be partitioned into two sets: those that have their own target and those that share the common target. The X_i that have their own target each have probability $\frac{1}{|\mathcal{R}|}$. The probability of the X_i that share a target can be computed using Equation E.2.

The integral over p is straightforward to evaluate. Taking $P(p)$ to be uniform over $[0,1]$, the probability of Graph i, in which N_B^+ bombs share a common target and N_B^- bombs have their own targets, is

$$\begin{aligned} P(\text{Graph } i) &= \int_0^1 P(\text{Graph } i|p)P(p)\ dp \\ &= \int_0^1 p^{N_B^+}(1-p)^{N_B^-}\ dp \\ &= \frac{\Gamma(N_B^+ + 1)\Gamma(N_B^- + 1)}{\Gamma(N_B+2)} \end{aligned} \tag{E.3}$$

as the integrand takes the form of the beta function (Boas, 1983).

Combining the values of $P(\mathcal{D}|\text{Graph } i)$ obtained via Equation E.2 with with $P(\text{Graph } i)$ from Equation E.3, we need only evaluate the sum over the 2^{N_B} possible graph structures. This can be done by Monte Carlo simulation. The results shown in Figure 7.9 were computed using a form of importance sampling designed for finding the Bayes factors of mixture distributions (Emond, Raftery, & Steele, 2001). The model predictions shown in the figure use 100,000 samples and $k = 4$.

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.

Aldous, D. (1985). Exchangeability and related topics. In *École d'été de probabilités de Saint-Flour, XIII—1983* (pp. 1–198). Berlin: Springer.

Allan, L. G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society*, *15*, 147-149.

Allan, L. G. (1993). Human contingency judgments: Rule based or associative? *Psychological Bulletin*, *114*, 435-448.

Allan, L. G., & Jenkins, H. M. (1983). The effect of representations of binary variables on judgment of influence. *Learning and Motivation*, *14*, 381-405.

Alloy, L. B., & Tabachnik, N. (1984). Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information. *Psychological Review*, *91*, 112-149.

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, J. R., & Sheu, C.-F. (1995). Causal inferences as perceptual judgments. *Memory and Cognition*, *23*, 510-524.

Associated Press. (September 12, 2002). *N.Y. lottery drawing pops up 9-1-1.*

Atran, S. (1995). Classifying nature across cultures. In E. E. Smith & D. N. Osherson (Eds.), *Thinking: An invitation to cognitive science* (Vol. 3, p. 131-174). Cambridge, MA: MIT Press.

Blackmore, S., & Troscianko, T. (1985). Belief in the paranormal: Probability judgements, illusory control, and the "chance baseline shift". *British Journal of Psychology*, *76*, 459-468.

Blackmore, S. J. (1997). Probability misjudgment and belief in the paranormal: A newspaper survey. *British Journal of Psychology*, *88*, 683-689.

Blei, D. M., Griffiths, T. L., Jordan, M. I., & Tenenbaum, J. B. (2004). Hierarchical topic models and the nested Chinese restaurant process. In *Advances in neural information processing systems 16.* Cambridge, MA: MIT Press.

Boas, M. L. (1983). *Mathematical methods in the physical sciences* (2nd ed.). New York: Wiley.

Bressan, P. (2002). The connection between random sequences, everyday coincidences, and belief in the paranormal. *Applied Cognitive Psychology*, *16*, 17-34.

Brugger, P., Landis, T., & Regard, M. (1990). A "sheep-goat effect" in repetition avoidance: Extra-sensory perception as an effect of subjective probability? *British Journal of Psychology*, *81*, 455-468.

Brugger, P., & Taylor, K. I. (2003). Esp: Extrasensory perception or effect of subjective probability? *Journal of Consciousness Studies*, *10*, 221-246.

Buehner, M., & Cheng, P. W. (1997). Causal induction: The Power PC theory versus the Rescorla-Wagner theory. In M. Shafto & P. Langley (Eds.), *Proceedings of the 19th annual conference of the cognitive science society* (p. 55-61). Hillsdale, NJ: Lawrence Erlbaum Associates.

Buehner, M. J., Cheng, P. W., & Clifford, D. (2003). From covariation to causation: A test of the assumption of causal power. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1119-1140.

Bullock, M., Gelman, R., & Baillargeon, R. (1982). The development of causal reasoning. In W. J. Friedman (Ed.), *The developmental psychology of time* (p. 209-254). New York: Academic Press.

Bunge, M. (1959). *Causality and modern science.* Cambridge, MA: Harvard University Press.

Buntine, W. L. (1994). *Learning with graphical models* (Tech. Rep. No. FIA-94-02). NASA Ames Research Center.

Carey, S. (1985a). *Conceptual change in childhood.* Cambridge, MA: MIT Press.

Carey, S. (1985b). Constraints on semantic development. In J. Mehler (Ed.), *Neonate cognition* (p. 381-398). Hillsdale, NJ: Erlbaum.

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

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

Charniak, E. (1993). *Statistical language learning.* Cambridge, MA: MIT Press.

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

Cheng, P. W. (1993). Separating causal laws from causal facts: Pressing the limits of statistical relevance. In *The psychology of learning and motivation* (Vol. 30, p. 215-264). San Diego: Academic Press.

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

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

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

Choi, H., & Scholl, B. J. (in press). Effects of grouping and attention on the perception of causality. *Perception and Psychophysics*.

Chomsky, N. (1956). Threee models for the description of language. *IRE Transactions on Information Theory*, *2*, 113-124.

Chomsky, N. (1959). A review of B.F Skinner's Verbal Behavior. *Language*, *31*, 26-58.

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, MA: MIT Press.

Clarke, R. D. (1946). An application of the Poisson distribution. *Journal of the Institute of Actuaries (London)*, *72*.

Cohen, L. B., & Oakes, L. M. (1993). How infants perceie a simple causal event. *Developmental Psychology*, *29*, 421-433.

Collins, D. J., & Shanks, D. R. (2002). Momentary and integrative response strategies in causal judgment. *Memory and Cognition*, *30*, 1138-1147.

Collins, D. J., & Shanks, D. R. (submitted). Conformity to the Power PC theory of causal induction depends on type of probe question. *Memory and Cognition*.

Collins, G. M. (1977). Visual co-orientation and maternal speech. In H. R. Schaffer (Ed.), *Studies in mother-infant interaction.* London: Academic Press.

Cook, A. (1998). *Edmond Halley: Charting the heavens and the seas.* Clarendon Press.

Cooper, G., & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, *9*, 308-347.

Cover, T., & Thomas, J. (1991). *Elements of information theory.* New York: Wiley.

Currie, J. (1798/1960). Medical reports on the effects of water, cold and warm, as a remedy in fever and other diseases, whether applied to the surface of the body or used internally. including an inquiry into the circumstances that render cold drink, or the cold bath dangerous in health. To which are added observations on the nature of fever and on the effect of opium, alcohol and inanition. In L. Clendening (Ed.), *Source book of medical history* (p. 428-433). New York: Dover.

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

Danks, D., & McKenzie, C. R. M. (under revision). *Learning complex causal structures.*

DeGregory, L. (September 24, 2002). 9-1-1 numerology. *St Petersburg Times.*

Diaconis, P. (1978). Statistical problems in ESP research. *Science*, *201*, 131-136.

Diaconis, P., & Mosteller, F. (1989). Methods for studying coincidences. *Journal of the American Statistical Association*, *84*, 853-861.

Doyle, A. C. (1986a). *Sherlock Holmes: The complete novels and stories* (Vol. 1). New York: Bantam.

Doyle, A. C. (1986b). *Sherlock Holmes: The complete novels and stories* (Vol. 2). New York: Bantam.

Eastaway, R., & Wyndham, J. (1998). *Why do buses come in threes? The hidden mathematics of everyday life.* New York: Wiley.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representation of human judgment.* New York: Wiley.

Emond, M. J., Raftery, A. E., & Steele, R. (2001). *Easy computation of Bayes factors and normalizing constants for mixture models via importance sampling* (Tech. Rep. No. 398). University of Washington.

Enderton, H. B. (1972). *A mathematical introduction to logic.* New York: Academic Press.

Evans, J. S. B. T., Handley, S. J., Over, D. E., & Perham, N. (2002). Background beliefs in Bayesian reasoning. *Memory and Cognition*, *30*, 179-190.

Falk, R. (1981-1982). On coincidences. *Skeptical Inquirer.*

Falk, R. (1989). Judgment of coincidences: Mine versus yours. *American Journal of Psychology*, *102*, 477-493.

Feller, W. (1968). *An introduction to probability theory and its applications.* New York: Wiley.

Fisher, R. A. (1937). *The design of experiments.* London: Oliver & Boyd.

Franklin, J. (2001). *The science of conjecture: Evidence and probability before Pascal.* Baltimore, MD: John Hopkins University Press.

Freedman, D., & Humphreys, P. (1998). *Are there algorithms that discover causal structure?* (Tech. Rep. No. 514). Department of Statistics, University of California at Berkeley.

Freedman, D. A. (1991). Statistical models and shoe leather. *Sociological Methodology*, *21*, 291-313.

Freedman, D. A. (1999). From association to causation: Some remarks on the history of statistics. *Statistical Science*, *14*, 243-258.

Freedman, D. A. (2004). On specifying graphical models for causation, and the identification problem. *Evaluation Review*, *26*, 267-293.

Friedman, N. (1997). Learning belief networks in the presence of missing values and hidden variables. In D. Fisher (Ed.), *Fourteenth international conference on machine learning* (p. 125-133). San Francisco, CA: Morgan Kaufmann.

Friedman, N., Getoor, L., Koller, D., & Pfeffer, A. (1999). Learning probabilistic relational models. In *Proceedings of the 16th internation joint conference on artificial intelligence (IJCAI).* Stockholm, Sweden.

Friedman, N., & Koller, D. (2000). Being Bayesian about network structure. In *Proceedings of the 16th annual conference on uncertainty in ai* (p. 201-210). Stanford, CA.

Garner, W. R. (1970). Good patterns have few alternatives. *American Scientist*, *58*, 34-52.

Gelman, R., & Spelke, E. S. (1981). The development of thoughts about animate and inanimate objects: Implications for research in social cognition. In J. H. Flavell & L. Ross (Eds.), *The development of social cognition in children.* Cambridge: Cambridge University Press.

Gelman, S. A., & Wellman, J. M. (1991). Insides and essence: Early understandings of the non-obvious. *Cognition*, *38*, 213-244.

Gilovich, T. (1993). *How we know what isn't so: The falliblity of reason in everyday life.* New York: Free Press.

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

Glymour, C. (2001). *The mind's arrows: Bayes nets and graphical causal models in psychology.* Cambridge, MA: MIT Press.

Glymour, C., & Cheng, P. W. (1998). Causal mechanism and probability: a normative approach. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition.* Oxford: Oxford University Press.

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

Good, I. J. (1984). A Bayesian approach in the philosophy of inference. *British Journal for the Philosophy of Science*, 161-166.

Gopnik, A. (1996). The scientist as child. *Philosophy of Science*, *63*, 485-514.

Gopnik, A., & Glymour, C. (2002). Causal maps and Bayes nets: A cognitive and computational account of theory-formation. In *The cognitive basis of science.* Cambridge: Cambridge 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*, 1-31.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories.* Cambridge, MA: MIT Press.

Gopnik, A., & Sobel, D. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child Development*, *71*, 1205-1222.

Gopnik, A., Sobel, D. M., Shulz, 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.

Griffiths, T. L., Baraff, E. R., & Tenenbaum, J. B. (2004). Using physical theories to infer hidden causal structure. In *Proceedings of the 26th annual meeting of the cognitive science society.*

Griffiths, T. L., & Ghahramani, Z. (in prep). *Priors for infinite sparse representations.*

Griffiths, T. L., & Tenenbaum, J. B. (in press). Elemental causal induction. *Cognitive Psychology.*

Hacking, I. (1983). *Representing and intervening.* Cambridge: Cambridge University Press.

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

Halley, E. (1752). *Astronomical tables with precepts both in English and in Latin for computing the places of the Sun, Moon &c.* London.

Halpern, J. Y., & Pearl, J. (2001). Causes and explanations: A structural-model approach – Part I: Causes. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence* (p. 194-202). San Francisco, CA: Morgan Kaufmann.

Hamadeh, H. K., Bushel, P. B., Jayadev, S., Martin, K., DiSorbo, O., Sieber, S., Bennett, L., Tennant, R., Stoll, R., Barrett, J. C., Blanchard, K., Paules, R. S., & Afshari, C. A. (2002). Gene expression analysis reveals chemical-specific profiles. *Toxicological Sciences*, *67*, 219-231.

Hardy, A., Harvie, R., & Koestler, A. (1973). *The challenge of chance.* New York: Random House.

Harre, R., & Madden, E. H. (1975). *Causal powers.* Oxford: Blackwell.

Harris, M., Jones, D., & Grant, J. (1983). The nonverbal content of mothers' speech to infants. *First Language*, *4*, 21-31.

Heckerman, D. (1998). A tutorial on learning with Bayesian networks. In M. I. Jordan (Ed.), *Learning in graphical models* (p. 301-354). Cambridge, MA: MIT Press.

Hempel, C. G. (1966). *Philosophy of natural science.* New York: Prentice-Hall.

Hirschfeld, L. A., & Gelman, S. A. (Eds.). (1994). *Mapping the mind: Domain specificity in cognition and culture.* Cambridge: Cambridge University Press.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*, 945-960.

Horwich, P. (1982). *Probability and evidence.* Cambridge: Cambridge University Press.

Hughes, D. W. (1990). Edmond Halley: His interest in comets. In N. J. M. Thrower (Ed.), *Standing on the shouldres of giants: a longer view of Newton and Halley* (p. 324-372). Berkeley and Los Angeles: University of California Press.

Hume, D. (1739/1978). *A treatise of human nature.* Oxford: Oxford University Press.

Hume, D. (1748). *An enquiry concerning human understanding.* Indianapolis, IN: Hackett.

Humphreys, P., & Freedman, D. A. (1996). The grand leap. *British Journal of the Philosophy of Science*, *47*, 113-23.

Inagaki, K., & Hatano, G. (2002). *Young children's thinking about biological world.* New York: Psychology press.

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs*, *79*.

Jenner, E. (1798). *An inquiry into the causes and effects of the variolae vaccinae.*

Johnson, D. (1981). *V-1, V-2: Hitler's vengeance on London.* New York: Stein & Day.

Jordan, M. I. (1998). *Learning in graphical models.* Cambridge, MA: MIT Press.

Jung, C. G. (1960). *The structure and dynamics of the psyche. collected works* (Vol. VIII). London: Routledge & Kegan Paul.

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing.* Upper Saddle River, NJ: Prentice Hall.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*, 430-454.

Karmiloff-Smith, A. (1988). The child is a theoretician, not an inductivist. *Mind and Language*, *3*, 183-195.

Kass, R. E., & Rafferty, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773-795.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development.* Cambridge, MA: MIT Press.

Keil, F. C. (2003). Folkscience: Coarse interpretations of a complex reality. *Trends in Cognitive Science*, *7*, 368-373.

Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist*, *28*, 107-128.

Kemp, C., Griffiths, T. L., & Tenenbaum, J. B. (2004). *Discovering latent classes in relational data* (Tech. Rep. No. 2004-019). MIT AI Memo.

Kersting, K., & Raedt, L. D. (2000). Bayesian logic programs. In J. Cussens & A. Frisch (Eds.), *Proceedings of the work-in-progress track at the 10th international conference on inductive logic programming* (p. 138-155).

Koehler, D. J., White, C. M., & Grondin, R. (2003). An evidential support accumulation model of subjective probability. *Cognitive Psychology*, *46*, 152-197.

Koller, D., & Pfeffer, A. (1997). Object-oriented Bayesian networks. In *Proceedings of the 13th annual conference on uncertainty in artificial intelligence (UAI).* Providence, RI.

Korb, K. B., & Wallace, C. S. (1997). In search of the philosopher's stone: Remarks on Humphreys and Freedman's critique of causal discovery. *British Journal of the Philosophy of Science*, *48*, 543-553.

Koslowski, B. (1996). *Theory and evidence: The development of scientific reasoning.* Cambridge, MA: MIT Press.

Koslowski, B., & Okagaki, L. (1986). Non-Humean indices of causation in problem-solving situations: Causal mechanism, analogous effects, and the status of rival alternative accounts. *Child Development*, *57*, 1100-1108.

Koslowski, B., Okagaki, L., Lorenz, C., & Umbach, D. (1989). When covariation is not enough: The role of causal mechanism, sampling method, and sample size in causal reasoning. *Child Development*, *60*, 1316-1327.

Kubovy, M., & Gilden, D. (1991). Apparent randomness is not always the complement of apparent order. In G. R. Lockhead & J. R. Pomerantz (Eds.), *The perception of structure* (p. 115-127). Washington, DC: American Psychological Association.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press.

Kushnir, T., Gopnik, A., Schulz, L., & Danks, D. (2003). Inferring hidden causes. In *Proceedings of the twenty-fifth annual conference of the cognitive science society.*

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

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

Laplace, P. S. (1795/1951). *A philosophical essay on probabilities* (F. W. Truscott & F. L. Emory, Trans.). New York: Dover.

Laudan, L. (1977). *Progress and its problems.* Berkeley, CA: University of California Press.

Leslie, A. M. (1982). The perception of causality in infants. *Perception*, *11*, 173-186.

Leslie, A. M. (1984). Spatiotemporal continuity and the perception of causality in infants. *Perception*, *13*, 287-305.

Leslie, A. M. (1994). ToMM, ToBY, and agency: Core architecture and domain specificity. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture.* Cambridge: Cambridge University Press.

Leslie, A. M., & Keeble, S. (1987). Do six-month-old infants perceive causality. *Cognition*, *25*, 265-288.

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

Littlewood, J. E. (1953). *A mathematician's miscellany.* London: Methuen.

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

Lopez, F. J., Cobos, P. L., Cano, A., & Shanks, D. R. (1998). The rational analysis of human causal and probability judgment. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (p. 314-352). Oxford: Oxford University Press.

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing.* Cambridge, MA: MIT Press.

Marr, D. (1982). *Vision.* San Francisco, CA: W. H. Freeman.

Matthews, R. A. J., & Blackmore, S. J. (1995). Why are coincidences so impressive? *Perceptual and Motor Skills*, *80*, 1121-1122.

McCloskey, M. (1983). Intuitive physics. *Scientific American*, *284*, 114-123.

Medin, D. L., & Atran, S. (Eds.). (1999). *Folkbiology.* Cambridge, MA: Bradford Books.

Melz, E. R., Cheng, P. W., Holyoak, K. J., & Waldmann, M. R. (1993). Cue competition in human categorization: Contingency or the Rescorla-Wagner rule? comments on shanks (1991). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 1398-1410.

Milch, B., Marthi, B., & Russell, S. (2004). Blog: Relational modeling with unknown objects. In T. Dietterich, L. Getoor, & K. Murphy (Eds.), *Icml 2004 workshop on statistical relational learning and its connections to other fields* (p. 67-73).

Minka, T. (2001). *Inferring a Gaussian distribution.* (http://www.stat.cmu.edu/∼minka/papers/gaussian.html)

Muggleton, S. (1997). Learning from positive data. In *Selected papers from the 6th international workshop on inductive logic programming* (p. 358-376). Springer-Verlag.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*, 289-316.

Murphy, K. (2003). Learning Markov processes. In L. Nadel (Ed.), *The encyclopedia of cognitive science.* London: Nature Publishing Group.

Musch, J., & Ehrenberg, K. (2002). Probability misjudgment, cognitive ability, and belief in the paranormal. *British Journal of Psychology*, *93*, 169-177.

Newsome, G. L. (2003). The debate between current versions of the covariation and mechanism approaches to causal inference. *Philosophical Psychology*, *16*, 87-107.

Newton, I. (1687/1962). *Isaac Newton's mathematical principles of natural philosophy and his system of the world* (A. Motte, Trans.). Berkeley and Los Angeles: University of California Press.

Nichols, S., & Stich, S. (2003). *Mindreading: An integrated account of pretense, self-awareness and understanding other minds.* Oxford: Oxford University Press.

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

Oakes, L. M., & Cohen, L. B. (1990). Infant perception of a causal event. *Cognitive Development*, *5*, 193-207.

Opfer, J. E., & Gelman, S. A. (2001). Children's and adults' models for predicting teleological action: The development of a biology-based model. *Child Development*, *72*, 1367-1381.

Owens, D. (1992). *Causes and coincidences.* Cambridge: Cambridge University Press.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems.* San Francisco, CA: Morgan Kaufmann.

Pearl, J. (1995). Causal diagrams ofr empirical research. *Biometrika*, *82*, 669-688.

Pearl, J. (1996). Structural and probabilistic causality. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *The psychology of learning and motivation: Causal learning* (Vol. 34). San Diego: Academic Press.

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

Pearson, K. (1904/1948). On the theory of contingency and its relation to association and normal correlation. In *Karl Pearson's early statistical papers* (p. 443-475). Cambridge: Cambridge University Press.

Pearson, K. (1911). *The grammar of science* (3rd ed.). London: A. C. Black.

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

The perception of causality. (1963). New York: Basic Books.

Perner, J. (1991). *Understanding the representational mind.* Cambridge, MA: MIT Press.

Perrin, J. (1913/1990). *Atoms* (D. L. Hammick, Trans.). Connecticut: Ox Bow Press.

Piaget, J. (1930). *The child's conception of physical causality.* London: Routledge and Kegan Paul.

Pitman, J. (2002). *Combinatorial stochastic processes.* (Notes for Saint Flour Summer School)

Plous, S. (1993). *The psychology of judgment and decision making.* New York: McGraw-Hill.

Pynchon, T. (1973/1995). *Gravity's rainbow.* London: Vintage.

Redfield, J. (1998). *The Celestine vision: Living the new spiritual awareness.* Sydney, Australia: Bantam.

Rehder, B. (2003). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1141-1159.

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

Robins, J. M. (1986). A new approach to causal inference in mortality studies with sustained exposure periods – applications to control of the healthy worker survivor effect. *Mathematical Modeling*, *7*, 1393-1512.

Robins, J. M. (1987). Addendum to "a new approach to causal inference in mortality studies with sustained exposure periods – applications to control of the healthy worker survivor effect". *Computers and Mathematics, with Applications*, *14*, 923-945.

Rozenblit, L. R., & Keil, F. C. (2002). The missunderstood limits of folk science: an illusion of explanatory depth. *Cognitive Science*, *26*, 521-562.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*, 688-701.

Rubin, D. B. (1990). Formal models of statistical inference for causal effects. *Journal of Statistical Planning and Inference*, *25*, 279-292.

Russell, S. J., & Norvig, P. (2002). *Artificial intelligence: A modern approach* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Salmon, W. C. (1980). *Scientific explanation and the causal structure of the world.* Princeton, NJ: Princeton University Press.

Salmon, W. C. (1990). Rationality and objectivity in science or Tom Kuhn meets Tom Bayes. In C. W. Savage (Ed.), *Scientific theories* (Vol. XIV). Minneapolis, MN: University of Minnesota Press.

Saxe, R., Tenenbaum, J. B., & Carey, S. (in press). Secret agents: Inferences about hidden causes by 10- and 12-month-old infants. *Psychological Science*.

Schlesinger, G. N. (1991). *The sweep of probability.* Notre Dame, IN: University of Notre Dame Press.

Schlottman, A. (1999). Seeing it happen and knowing how it works: How children understand the relation between perceptual causality and underlying mechanism. *Developmental Psychology*, *35*, 303-317.

Scholl, B. J., & Nakayama, K. (2002). Causal capture: Contextual effects on the perception of collision events. *Psychological Science*, *13*, 493-498.

Schulz, L., & Gopnik, A. (in press). Causal learning across domains. *Developmental Psychology*.

Schustack, M. W., & Sternberg, R. J. (1981). Evaluation of evidence in causal inference. *Journal of Experimental Psychology: General*, *110*, 101-120.

Segal, E., Pe'er, D., Regev, A., Koller, D., & Friedman, N. (2003). Learning module networks. In *Proceedings of the nineteenth conference on uncertainty in artificial intelligence (UAI)* (p. 525-534). San Francisco, CA: Morgan Kaufmann.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., & Friedman, N. (2003). Module networks: Identifying regulatory modules and their condition-specific regulartors from gene expression data. *Nature Genetics*, *34*(166-176).

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

Shanks, D. R. (1995b). *The psychology of associative learning.* Cambridge University Press.

Shanks, D. R. (2002). Tests of the Power PC theory of cauasl induction with negative contingencies. *Experimental Psychology*, *49*, 1-8.

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

Shanks, D. R., López, F. J., Darby, R. J., & Dickinson, A. (1996). Distinguishing associative and probabilsitic contrast theories of human contingency judgment. In *The psychology of learning and motivation* (Vol. 34, p. 265-312). San Diego, CA: Academic Press.

Shultz, T. R. (1982a). Causal reasoning in the social and non-social realms. *Canadian Journal of Behavioural Science*, *14*, 307-322.

Shultz, T. R. (1982b). Rules of causal attribution. *Monographs of the Society for Research in Child Development*, *47*(Serial no. 194).

Shultz, T. R., Fisher, G. W., Pratt, C. C., & Rulf, S. (1986). Selection of causal rules. *Child Development*, *57*, 143-152.

Shultz, T. R., & Kestenbaum, N. R. (1985). Causal reasoning in children. *Annals of Child Development*, *2*, 195-249.

Skinner, B. F. (1957). *Verbal behavior.* Englewood Cliffs, NJ: Prentice Hall.

Slovic, P., & Fischoff, B. (1977). On the psychology of experimental surprises. *Journal of Experimental Psychology: Human Perception and Performance*, *3*, 544-551.

Snow, J. (1855). *On the mode of communication of cholera.* London: John Churchill.

Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (in press). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science.*

Sosa, E., & Tooley, M. (Eds.). (1993). *Causation.* Oxford: Oxford University Press.

Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological Review*, *99*, 605-632.

Spelke, E. S., Phillips, A. T., & Woodward, A. L. (1995). Infants' knowledge of object motion and human action. In A. J. Premack, D. Premack, & D. Sperber (Eds.), *Causal cognition: A multidisciplinary debate* (p. 44-77). Oxford: Clarendon Press.

Spirtes, P., Glymour, C., & Scheines, R. (1997). Reply to Humphreys and Freedman's review of *causation, prediction and search. The British Journal of the Philosophy of Science*, *48*, 555-568.

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

Spirtes, P., Glymour, C., & Schienes, R. (2001). *Causation prediction and search* (2nd ed.). Cambridge, MA: MIT Press.

Springer, K., & Keil, F. C. (1991). Early differentiation of causal mechanisms appropriate to biological and non-biological kinds. *Child Development*, *62*, 767-781.

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

Stoppard, T. (1967). *Rosencrantz and Guildenstern are dead.* New York: Grove Press.

Suppes, P. (1970). *A probabilistic theory of causality.* Amsterdam: North Holland.

Teigen, K. H., & Keren, G. (2003). Surprises: low probabilities or high contrasts? *Cognition*, *87*, 55-71.

Tenenbaum, J. B. (1999a). Bayesian modeling of human concept learning. In M. S. Kearns, S. A. Solla, & D. A. Cohn (Eds.), *Advances in neural information processing systems 11* (p. 59-65). Cambridge, MA: MIT Press.

Tenenbaum, J. B. (1999b). *A Bayesian framework for concept learning.* Unpublished doctoral dissertation, Massachussets Institute of Technology, Cambridge, MA.

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

Tenenbaum, J. B., & Griffiths, T. L. (2003). Theory-based causal induction. In *Advances in neural information processing systems 15* (p. 35-42). Cambridge, MA: MIT Press.

Tenenbaum, J. B., & Griffiths, T. L. (in prep). *The place of intuitive theories in rational causal inference.*

Tenenbaum, J. B., & Niyogi, S. (2003). Learning causal laws. In *Proceedings of the 25th annual meeting of the cognitive science society.* Erlbaum.

Tenenbaum, J. B., Sobel, D. M., & Gopnik, A. (submitted). *Bayesian inference in causal learning from ambiguous data: Evidence from adults and children.*

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, *101*, 547-567.

Vallee-Tourangeau, F., Murphy, R. A., Drew, S., & Baker, A. G. (1998). Judging the importance of constant and variable candidate causes: A test of the Power PC theory. *Quarterly Journal of Experimental Psychology*, *51A*, 65-84.

Waldmann, M. R. (1996). Knowledge-based causal induction. In *The psychology of learning and motivation* (Vol. 34, p. 47-88). San Diego: Academic Press.

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

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

Waldnann, M. R., & Martignon, L. (1998). A Bayesian network model of causal learning. In M. A. Gernsbacher & S. J. Derry (Eds.), *Proceedings of the twentieth annual conference of the cognitive science society* (p. 1102-1107). Mahwah, NJ: Erlbaum.

Ward, W. C., & Jenkins, H. M. (1965). The display of information and the judgment of contingency. *Canadian Journal of Psychology*, *19*, 231-241.

Wasserman, E. A. (1990). Detecting response-outcome relations: Toward an understanding of the causal texture of the encironment. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 26, p. 27-82). San Diego, CA: Academic Press.

Wasserman, E. A., Elek, S. M., Chatlosh, D. C., & Baker, A. G. (1993). Rating causal relations: Role of proabbility in judgments of response-outcome contingency. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *19*, 174-188.

Wellman, H. M. (1990). *The child's theory of mind.* Cambridge, MA: MIT Press.

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology*, *43*, 337-375.

Westfall, R. S. (1977). *The construction of modern science: Mechanisms and mechanics.* Cambridge, UK: Cambridge University Press.

White, P. A. (1990). Ideas about causation in philosophy and psychology. *Psychological Bulletin*, *108*, 3-18.

White, P. A. (1995). *The understanding of causation and the production of action: from infancy to adulthood.* Hillsdale, NJ: Lawrence Erlbaum.

White, P. A. (1998). Causal judgement: Use of different types of contingency information as confirmatory and disconfirmatory. *European Journal of Cognitive Psychology*, *10*, 131-170.

White, P. A. (2002). Perceiving a strong causal relation in a weak contingency: Further investigation of the evidential evaluation model of causal judgement. *Quarterly Journal of Experimental Psychology*, *55A*, 97-114.

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

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

White, P. A., & Milne, A. (1997). Phenomenal causality: Impressions of pulling in the visual perception of objects in motion. *American Journal of Psychology*, *110*, 573-602.

White, P. A., & Milne, A. (1999). Impressions of enforced disintegration and bursting in the perception of collision events. *Journal of Experimental Psychology: General*, *128*, 499-516.

Wickens, T. D. (1989). *Multiway contingency tables analysis for the social sciences.* Hillsdale, NJ: Erlbaum.

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

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, *20*, 557-585.

Yeomans, D. K. (1991). *Comets: a chronological history of observation, science, myth and folklore.* New York: Wiley.