Title
Acquired Abstract knowledge in causal induction: a hierarchical Bayesian approach

Author
Lucas, Christopher

Publication Date
2010

Peer reviewed|Thesis/dissertation

Acquired Abstract knowledge in causal induction: a hierarchical Bayesian approach

by

Christopher Guy Lucas

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Psychology

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:
Thomas L. Griffiths, Co-Chair
Alison Gopnik, Co-Chair
John Campbell
Tania Lombrozo

Fall 2010

Acquired Abstract knowledge in causal induction: a hierarchical Bayesian approach

Abstract

Acquired Abstract knowledge in causal induction: a hierarchical Bayesian approach

by

Christopher Guy Lucas
Doctor of Philosophy in Psychology

University of California, Berkeley

Thomas L. Griffiths, Co-Chair

Alison Gopnik, Co-Chair

The human ability to learn quickly about causal relationships requires abstract knowledge that provides constraints and biases regarding what relationships are possible and likely. There has been a long and vigorous debate about the extent to which these biases are learned. Psychologists and philosophers since Plato's time have tried to answer this question with a wide range of techniques, including arguments from intuition, proofs giving bounds on what is learnable, and experimentation.

Hierarchical Bayesian models are a relatively new tool that allows us to address the question of causal induction with a directness that was impossible in the past. Bayesian models solve problems by combining a priori expectations with evidence to learn about unobservable variables like category membership or status as a cause. Using a hierarchical organization allows those expectations to be learned, shaped by yet more abstract inductive biases. Using these models, we can develop hypotheses about the origins of abstract knowledge and make precise predictions which can then be tested experimentally. In some cases, these models reveal new ways to learn abstract properties of the world, and more specifically, causal relationships.

Here I will discuss two kinds of abstract knowledge: knowledge about the forms of causal relationships, and knowledge about the nature of the preferences that cause people to make the choices they do. First, I will show that people can learn that a causal relationship takes a particular form and use that knowledge in their later inferences, consistent with the predictions of a hierarchical Bayesian model. In addition, I will describe experiments with adults and children that indicate that adults' strong inductive biases about the forms of causal relationships may be the result of long experience, rather than innate constraints. Second, I will use a model from economics to explain a range of developmental findings in preference learning, including a shift in which children come to treat other people as having distinct preferences. In both cases, the hierarchical Bayesian models explain developmental differences, offer new predictions about causal learning, and offer a broader view of causal induction.

To my parents, Bill and Barbara,
and Claudia.

Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank the people whose support and help made this dissertation possible. First, I am grateful to Tom Griffiths for the ideas, help and advice he gave me during my time at Berkeley. The methodological and scientific insights he shared with me were invaluable, as were the lessons he taught by example. I hope to show the same poise and generosity toward colleagues, critics, and living creatures in general – from unknown strangers to classroom spiders – that Tom does. I am also lucky to have had Alison Gopnik as an advisor. She introduced me to developmental research and the exciting questions it can answer, made me a better experimentalist, and taught me the value of neoteny to humans in general and scientists in particular. I would like to thank John Campbell, who introduced me to new kinds of questions and new perspectives on the ones I was already asking, and Tania Lombrozo, from whom I learned about concepts, categories, and explanations, and who provided me with useful advice about my dissertation and, separately, chocolate. Chapter 4 would not have been possible without the help of collaborators Fei Xu, Christine Fawcett, Tamar Kushnir, and Lori Markson, who shared valuable comments and the experimental data I used in my analyses. I would also like to thank the members of Alison Gopnik's and Tom Griffith's labs, especially Anna Waismeyer, Brian Waismeyer, and Elizabeth Seiver, who provided me with valuable advice and read drafts of the work in this dissertation, and Kevin Canini, with whom I had useful technical discussions. Lastly, I am grateful for support from the James S. McDonnell Foundations Causal Learning Collaborative to Alison Gopnik, and the Air Force Office of Scientific Research grant number FA9550-07-1-0351 to Tom Griffiths.

Chapter 1

Introduction

Our ability to learn about the causal structure of our environment is vital to our survival. We must constantly use causal knowledge to predict new events, to make plans and achieve goals, and to communicate with the people around us. It is unsurprising, then, that we are efficient at causal induction and generalization, and able to make far-reaching inferences having seen only one or two brief events. Nonetheless, ours is an impressive talent – several factors conspire to make causal learning a difficult problem. One variable can confound or mask the effects of another, hidden variables often influence causal relationships between observed variables, and causal relationships can take many forms. Any one of these factors can make causal learning an underdetermined problem, leaving a large number of causal explanations for many of the events people see in daily life. Given that humans solve this difficult problem frequently and efficiently, we must answer the question of what enables us to learn so well and make the inferences that we do. My goal will be to give a partial answer to this question, starting by laying out my basic explanatory strategy.

Following Marr (1982), we can divide explanations dealing with cognitive phenomena into three kinds. First, there are those that focus on the physical machinery behind learning: if we can unlock the chemical and biological secrets of the neural machinery that supports causal induction, we have an answer to the question. For a committed reductionist, this may be the only answer that is important, but it has some drawbacks. Practical limitations on technology that can measure and manipulate the physical state of the brain still present major difficulties, and if we find it possible to predict human induction with a realistic model of the underlying physical phenomena, that model may defy human comprehension with its complexity; a physical approach (or stance; Dennett, 1989) may not be the most parsimonious way to understand mental phenomena when other strategies are available.

The second kind of explanation deals with the algorithms and processes behind learning, and constitutes a common approach to modeling and understanding phenomena in higher level cognition. Prototype and exemplar theories of concept learning, many connectionist models, and heuristic approaches to judgment and decision-making all fall into this group. If we could map causal induction onto a set of algorithms, we would have another solution to the problem of causal induction, one that would probably be more succinct than a physical-level account.

This work offers a third kind of explanation, at Marr's *computational level*. It identifies the problem being solved as well as the environment to which people have adapted, and derives optimal solutions to that problem. The environment influences the learner by providing information about

what events and regularities are common – a set of inductive biases – which can be represented using probabilities. This framing makes it possible to find optimal solutions in a systematic and straightforward manner, using Bayesian inference[1].

Reflecting the flexibility of the computational-level approach and Bayesian methods, the work presented here covers causal problems in disparate domains. The first is a task in which learners must identify the causes of an effect in a physical system, a classic causal learning problem, while the goal of the second is to identify people's preferences, that is, the latent variables that cause them to make the choices they do. While these problems are different in their structure and the background knowledge that applies to them, their treatment here has a common focus: the abstract knowledge that facilitates causal learning, the acquisition of that knowledge, and the use of hierarchical Bayesian models to understand people's inductive biases. Before introducing the specific empirical and theoretical contributions contained in this dissertation, I will briefly discuss these common themes, specifically the importance and possible origins of inductive biases in causal learning, and the relevance of hierarchical Bayesian methods to studying them.

1.1 The importance of inductive biases

Learning, with the exception of rote memorization of past observations, requires learners to make assumptions about what's being learned – the structure of the environment and the nature of any regularities that might be present. This fact has fueled philosophical debates (Hume, 1748; N. Goodman, 1955) and provided caveats for those who aim to build general-purpose classification systems (Duda, Hart, & Stork, 2000). Many inductive problems may have a similar structure, and systems with few apparent, a priori assumptions can generalize in interesting ways, but it is important not to confuse widely-applicable assumptions with the absence of assumptions. As an example of the cost of knowing too little about a problem, consider the case of predicting whether an effect e will occur, given causes[2] c_1 through c_k. If the relationship between the causes and the effect is deterministic and there are no hidden variables or time dependencies, there are 2^{2^k} possible relationships, and all possible combinations of causes must be observed to learn the true relationship. If, on the other hand, the learner could assume that causes independently generate or prevent the effect, then observing k (or fewer) distinct values for the causes would suffice to learn the relationship.

Inductive biases can vary in their specificity. They may be quite concrete, leading a learner to reject some events as physically impossible, as with Spelke's principles of object perception (Spelke, 1994), where, for instance, infants appear to understand that two objects cannot occupy the same physical location. More abstract inductive biases, which may apply to relations between concrete observations rather observations themselves, are also possible. As an example, the specific choice of similarity function in categorization models, e.g., using Minkowsky distance with $p = 1$ versus $p = 2$, leads to different category boundaries. Generalization by artificial neural networks, too, is due to inductive biases, albeit relatively opaque ones – using sinusoidal activation functions rather than radial or sigmoid ones in a feed-forward network will result in dramatically different generalization behavior. Similarly, the number of hidden layers and the connectivity in an artifi-

[1] Rather than spending significant time on defending the optimality of Bayesian solutions to inductive problems, I refer the reader to previous treatments of the subject (Oaksford & Chater, 2007; Jaynes, 2003).

[2] Throughout this document, the term "cause" will be used in the sense given by Pearl (2000).

cial neural network leads to different generalizations, with some choices precluding learning some regularities altogether.

This is all to emphasize that inductive biases are essential in any system that generalizes beyond the data already observed. It is certainly possible to obscure those biases, and to choose biases that enable narrower or farther-reaching inferences, but they are implicitly or explicitly present in any model that explains inductive inferences, at any level of explanation.

1.2 The origins of inductive biases

Debates about the role and origins of inductive biases are frequently marked by two opposing positions, empiricism and nativism. The empiricist position discounts the role of a priori assumptions and emphasizes flexible, general-purpose learning. Connectionist models (e.g., Colunga & Smith, 2005) are frequently identified with an empiricist point of view, because they often invoke simple learning rules and structural assumptions while being applied to a wide range of phenomena. The nativist position appeals to early, domain-specific knowledge (Spelke & Kinzler, 2007) and "poverty of the stimulus" arguments (Chomsky, 1965), which address the human ability to make farther-reaching inferences than appear to be licensed by the available data. The idea of a universal grammar and specialized, innate mechanisms for learning about others' minds (Leslie, 1994) are well-known examples of nativist arguments. Note that domain-specificity can to an extent be decoupled from the amount of structure built into a set of inductive biases: one can specify domain-specific biases that have weak and abstract implications for inferences, as well as highly-structured, domain-general biases.

It is widely accepted that the nativism-empiricism debate is a matter of degree, where the challenge is in determining how much must be built-in to explain the inferences we observe in children and adults. The tools used to answer this question, in addition to experimental research, have included learnability theorems (e.g., Gold's theorem, Gold, 1967) that place limits on what can be learned without a priori knowledge, and connectionist proofs of concept that, for instance, demonstrate that certain linguistic relationships can be learned using architectures that are given little a priori problem-specific knowledge (Elman et al., 1996; Colunga & Smith, 2005).

Another approach is to view learning problems in a probabilistic light, allowing different kinds inductive biases to be tested directly. Hierarchical Bayesian methods, in particular, give us the ability to investigate the origins and development of abstract knowledge. By abstract knowledge, I mean knowledge that allows us to understand and predict not just concrete events, but broad regularities in our environment, and applies to a wide range of events rather than a handful of circumscribed contexts. For example, a hypothesis about a specific set of causal relations is not abstract, but an understanding that informs the formation of such hypotheses is. I will use the term "hierarchical" to refer to any Bayesian model where the learner is not merely estimating the probability of hypotheses given data, but is acquiring or updating theory-like knowledge (Tenenbaum & Niyogi, 2003), higher-order hypotheses, or "overhypotheses" (Kemp, Perfors, & Tenenbaum, 2007; N. Goodman, 1955).

1.3 Précis

The remainder of this document is divided into four chapters. The first addresses the acquisition and use of abstract knowledge in a standard causal learning scenario, revealing certain inductive biases and showing that adults can acquire abstract knowledge about the forms of causal relationships, facilitating later learning. The second chapter presents developmental results showing that children learn efficiently about the abstract forms of causal relationships, with judgments that differ from adults' in ways that shed light on the origins and learnability of our inductive biases in causal learning. The third chapter considers the problems faced by toddlers and preschoolers when they must learn about others' preferences as well as the abstract structure of preferences in general. It offers a model that makes inferences to hidden causes in the form of subjective utilities, and explains two recent sets of empirical results. A hierarchical extension provides a rational explanation for a developmental shift in preference understanding. The final chapter concludes with a discussion of implications and future directions for these lines of research. Chapters 2 through 4 are empirical papers written in collaboration with others who study these issues; see the acknowledgments for details.

Chapter 2

Learning the forms of causal relationships

2.1 Introduction

Causal inference – learning what causal relationships are present in the world by observing events – is often taken to rely primarily on universal cues such as spatiotemporal contingency or reliable covariation between effects and their prospective causes (Hume, 1748). While recognizing covariation may be central to causal learning, there is both experimental and intuitive support for the idea that people also use domain-specific knowledge. For instance, children come to conclusions that they would not be able to reach if they were ignorant about the mechanisms behind causal relationships, as shown by experiments where they pick mechanistically plausible causes over spatially and temporally contingent ones (Shultz, 1982), or prefer mechanical causes for mechanical effects except when physical constraints make such a relationship unrealistic (Bullock, Gelman, & Baillargeon, 1982). There is also evidence that this kind of knowledge is learned. For example, children use mechanism knowledge differently as they age, with older children showing an understanding of a broader range of mechanisms (Shultz, 1982; Bullock et al., 1982).

Despite these indications that prior knowledge plays a role in causal learning, popular accounts of causal learning frame prior knowledge as orthogonal to (e.g., Cheng, 1997) or incompatible with (e.g., Shultz, 1982; Ahn, Kalish, Medin, & Gelman, 1995) statistical information provided by covariation between causes and effects. Even accounts that accept that knowledge about mechanisms, domains, and categories needs to be combined with covariation information tend not to provide for how we acquire such knowledge (e.g., Waldmann, 1996). The topic has not been ignored, however: in recent years some studies have explicitly considered the interplay of abstract knowledge and statistical information, examining how knowledge about categories and the nature of causal relationships affect the conclusions people draw from covariational evidence (Lien & Cheng, 2000; Kemp, Goodman, & Tenenbaum, 2007; Waldmann, 2007).

In this paper, we present a framework that reconciles covariation-based inference with acquired knowledge about kinds of causal relationships. Our goal is to explain how children and adults extract abstract information about causal relationships from experience, and use this information to guide causal learning. We focus here on knowledge about the *functional form* of causal relationships – the nature of the relationship between causes and effects – and specifically on how

causes interact in producing their effects. For example, imagine a light and two switches. Each switch is connected to the light, and affects whether the light is on or off. However, there are a variety of ways in which the light and the switches could be connected: the light might only turn on when both switches are pressed, or turn on when either is pressed, or turn on only some of the time when either switch is pressed but more often when both are pressed. Similar possibilities hold for other causal systems, with causes either acting independently or in conjunction to bring about or prevent effects. If learners know how causes influence effects in a particular kind of causal system, they can use this information to inform their reasoning about the existence of causal relationships.

Our perspective on how people acquire and use knowledge about the functional form of causal relationships follows in the spirit of rational analysis (Anderson, 1990) and previous accounts of causal inference (Cheng, 1997): we are concerned with clearly specifying the underlying computational problem and comparing human inferences to the optimal solution to this problem. We take the basic challenge of causal induction to be acquiring rich, useful representations of cause and effect that can be represented using *causal graphical models* (Pearl, 2000; Spirtes, Glymour, & Schienes, 1993), a formal language for representing causal relationships which we describe in more detail below. This formal language allows us to clearly characterize the role that functional form plays in causal learning, and to develop a mathematical framework in which we can analyze how knowledge of functional form is acquired and used.

We make two contributions towards understanding how people combine prior knowledge and covariational evidence in causal induction. Our first contribution is showing that the problem of learning the functional form of a causal relationship can be formalized using *hierarchical Bayesian models*, in which information about causal relationships is maintained at multiple levels of abstraction, reflecting both hypotheses about which causal relationships exist among specific sets of variables and more general theories about how causes relate to their effects. This general approach is compatible with other recent work on the acquisition of causal knowledge (Tenenbaum & Niyogi, 2003; Griffiths & Tenenbaum, 2007) and can be applied to phenomena beyond those we consider explicitly. Our second contribution is a series of experiments that test the qualitative predictions made by this approach, as well as the quantitative predictions made by a specific hierarchical Bayesian model. Our experiments focus on learning about a specific kind of causal system, related to the "blicket detector" used in previous work on causal learning (Gopnik & Sobel, 2000; Sobel, Tenenbaum, & Gopnik, 2004). We use this system to explore acquisition and use of abstract knowledge about functional form when the causal structure is specified, as well as when it is learned from contingency data. We also show that the scope of this knowledge seems to be restricted to the domain in which it is learned.

The plan of the paper is as follows. The next section reviews previous experiments that have explored the learning and use of knowledge of functional form. We then summarize standard accounts of how people infer causal relationships from statistical evidence, and consider the role they provide for abstract knowledge. We go on to outline the hierarchical Bayesian approach to analyzing the role of knowledge in causal induction and indicate how this approach applies to functional form. This is followed by a series of experiments that test the predictions of our model, revealing some sources of information that people use and kinds of knowledge they acquire. We conclude by discussing related models and considering new questions we hope to answer with this line of work.

2.2 Using knowledge of functional form in causal induction

Most previous work on the effects of prior knowledge on causal induction has focused on factors such as the plausibility of a causal relationship (Koslowski, 1996; Alloy & Tabachnik, 1984) and the types of entities in a domain (Lien & Cheng, 2000). However, three studies have explicitly looked at how people learn and use information about the form of causal relationships.

Zelazo and Shultz (1989) tested the ability of adults to predict the magnitude of an effect as a function of two causal variables, and found that their inferences depended on the form of the relationship indicated by the physical system. Specifically, when asked to learn from two training events and predict how far a counter-weighted balance would tilt or an obstructed set of blocks would slide, people tended to make inferences consistent with the dynamics of the specific physical system involved. In contrast, Zelazo and Shultz found that 9-year-olds' inferences reflected an understanding that the magnitude of the effect increased with the size of one block and decreased with the size of the other, but did not capture the differences between specific forms of the two relationships. One interpretation of this developmental difference is that adults had more experience with the two causal systems, and were thus able to make more precise inferences from the training events.

Waldmann (2007) also found that knowledge of the form of causal relationships strongly influences evidence-based inference, using a subtler manipulation: he presented adults with the task of determining the effect of consuming colored liquids on the heart rate of an animal, varying whether the stated mechanism by which the liquids influenced heart rate was their taste or their strength as a drug. In the taste case, judgments made by the participants indicated a tendency to believe that the effect of combining both liquids would be the average of their individual effects, whereas judgments in the strength case were consistent with believing the combined effect would be the sum of the individual effects.

Finally, Beckers, De Houwer, Pineno, and Miller (2005) found evidence that the inferences people draw from a set of evidence are shaped by having seen earlier "pre-training" events suggesting that the magnitude of an effect will be an additive or sub-additive function of its combined causes. Importantly, the different pre-training events did not suggest different stories: Beckers et al. only manipulated the strength of the effect in the presence of two causes, so that in the additive condition it was the sum of the strength of the individual causes and in the sub-additive condition it was the maximum. Participants then saw data from a standard "blocking" design, in which the effect occurred in the presence of one cause, A, alone, as well as in the presence of the two causes A and B. Those participants who had received additive pre-training showed a much stronger blocking effect, believing that B alone was unlikely to cause the effect since it did not seem to increase the magnitude of the effect when paired with A. This result is consistent with the assumptions of standard associative learning models (e.g., Rescorla & Wagner, 1972), which assume additive combination of causes.

These three studies illustrate that when people make inferences about the presence or strength of a causal relationship, they are sensitive to the form that any such relationship is likely to take, and that they can learn the form of a relationship from data. Inspired by these examples, the remainder of the paper explores the question of how we might use a computational model to explain how people learn about the functional form of causal relationships and apply that knowledge. We do this by laying out a general formal framework for modeling such learning, and then testing the predictions of this approach within a specific causal system. We start by summarizing the key ideas

behind existing models of causal learning.

2.3 Models of causal learning

Causal learning has been a topic of extensive study in cognitive psychology, resulting in a large number of formal models of human behavior (for a review, see Perales & Shanks, 2007). Our emphasis here will be on *rational* models of causal learning: models that explain human behavior as an optimal solution to a problem posed by the environment (Marr, 1982; Anderson, 1990). In the case of causal learning, the problem amounts to estimating the strength of a causal relationship, or inferring that such a relationship exists (Cheng, 1997; Griffiths & Tenenbaum, 2005). In this section, we will briefly describe some of the most prominent rational models of causal learning, dividing these models into those that focus on learning causal strength and those that focus on learning causal structure. In each case we will summarize the assumptions that these approaches make about functional form – strength-based approaches tend to assume a single fixed functional form, while structure-based approaches make weaker assumptions. First, however, we will describe the formal language of causal graphical models, which we will use to characterize the computational problem of causal induction.

2.3.1 Causal graphical models

Causal graphical models, or causal Bayes nets, are a formalism for representing and reasoning about causal relationships (Pearl, 2000; Spirtes et al., 1993). In a causal graphical model, all relevant variables are represented with nodes, and all direct causal relationships are represented using directed links, or edges, between nodes. In addition to carrying information about statistical relationships between variables, the link structure provides information about the effects of interventions and other exogenous influences on a causal system: acting to change a variable V may only influence its descendants, i.e., variables associated with nodes reachable by following the links directed away from V.

A causal graphical model depicting the example scenario from the introduction is shown in Figure 2.1, in which one might, by intervening on the states of two switches, influence the activation of a light. Here, the states of the switches and the activation of the light are the variables being represented, so each is assigned a node. The links between the switch nodes and the light node indicate that the switches are direct causes of the light.

Specifying a probability distribution for each variable conditioned on its parents in the graph (i.e. those variables that are its direct causes) defines a joint distribution on all of the variables which can be used to reason about the probability of observing particular events and the consequences of intervening on the system. However, an edge from one variable to another in a causal graphical model does not imply a specific functional form for the relationship between cause and effect. Rather, direct causal (and thus statistical) dependence is all it indicates.

The full specification of any causal graphical model includes its *parameterization*, which defines the probability that any given variable will take a particular value, conditional on its direct causes. In our example, one might suppose that the probability that the light activates will be close to zero if either of the switches is not flipped, and close to one otherwise. The use of probabilities permits reasoning under incomplete information, which applies here as the reasoner may not know

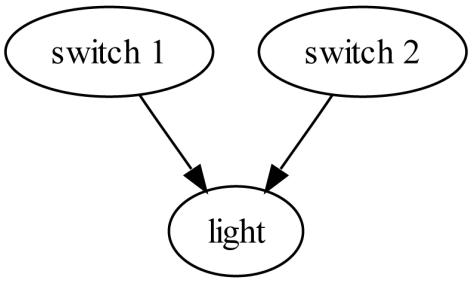


Figure 2.1: Example of a causal graphical model describing the causal relationships behind the operation of a light, in which two light switches are both causes of its activation.

about the internal state of the light – there may be a loose wire or the filament might be broken. The functional form of a causal relationship is captured by the parameterization, allowing, for example, the complete causal graphical model to distinguish a situation where both switches independently turn on the light from one where a single switch turns it off, or where the light activates unreliably.

2.3.2 Learning causal strength

One approach to evaluating a prospective cause c is to take the difference between the probability of the target effect e in its presence, $P(e|c)$, and the probability of the effect in its absence, $P(e|\overline{c})$. This quantity, $P(e|c) - P(e|\overline{c})$, is known as ΔP, where the probabilities $P(e|c)$ and $P(e|\overline{c})$ are computed directly from contingency data. Proponents of ΔP argue that it is a general-purpose measure of the strength of a causal relationship (Jenkins & Ward, 1965) as might result from an associative learning process (Shanks, 1995), but it does make assumptions about the nature of the causal relationship. This can be seen by viewing ΔP from the perspective of learning the structure and parameterization of a causal graphical model (Tenenbaum & Griffiths, 2001; Griffiths & Tenenbaum, 2005). First, ΔP assumes that a particular "focal set" of causes be identified, which is to say that the set of causes or equivalently the structure of the causal graphical model is known in advance. Second, it assumes a parameterization under which the probability of the effect given multiple causes is a linear combination of its probability under separate causes: if the probability of the effect in the presence of a single cause C_i is w_i, then the probability of the effect given the values of its causes $C_1, \ldots, C_n$ is

$$P(e|c_1, \ldots, c_n) = \min(1, \sum_{i=1}^{N} c_i w_i) \tag{2.1}$$

where c_i takes the value one when the i^{th} cause is present, zero otherwise. Under these assumptions, the value computed by the ΔP rule for a particular cause C_i is the w_i that maximizes the probability of the events observed by the learner, i.e., the maximum-likelihood estimate.

The limitations of ΔP as a model of human judgments motivated the development of the Power PC theory (Cheng, 1997), which takes causal learning to be a problem of inferring the "causal power" of prospective causes. The power of a generative causal relationship is defined to

be $\frac{P(e|c)-P(e|\bar{c})}{1-P(e|\bar{c})}$, which is simply ΔP divided by the probability of the effect in the absence of the prospective cause. As with ΔP, the assumptions of causal power can be made explicit by casting it as inference about causal graphical models. As before, the structure is assumed *a priori* via the choice of focal set, but here the assumption about parameterization is that the probability of an effect given its causes follows a "noisy-OR" function (Cheng, 1997; Pearl, 1988), in which each cause has an independent chance to produce its effect (Griffiths & Tenenbaum, 2005; Glymour, 1998), with

$$P(e|c_1, \ldots, c_n) = 1 - \prod_{i=1}^{n}(1 - w_i)^{c_i} \tag{2.2}$$

where c_i is defined as in Equation 2.1. As with ΔP, causal power computes the value of w_i that maximizes the probability of the observed events.

Lu, Yuille, Liljeholm, Cheng, and Holyoak (2007, 2008) recently proposed an extension of the causal power model in which Bayesian inference is used to identify the strength of a causal relationship. In this model, a prior distribution is defined on the strength of the relationship w_i, either being uniform or favoring stronger relationships, and this information is combined with contingency data to obtain a posterior distribution. A single estimate of the strength can be derived from this posterior distribution in several ways, such as taking the most probable or the average value. The basic assumptions behind this model are the same as those of the earlier causal power model (Cheng, 1997), taking the noisy-OR to be the appropriate functional form for the causal relationship, but Lu et al. (2007, 2008) showed that using an appropriate prior can improve the predictions that the model makes.

Finally, Novick and Cheng (2004) explored a way of extending the causal power model to accommodate relationships between multiple causes that go beyond the noisy-OR. In this extension, interactions between causes are handled by introducing new variables that represent combinations of causes. By considering possible generative and inhibitory effects of both the simple causes and their combinations, this model is able to express a richer repertoire of functional forms. Yuille and Lu (2007) have shown that this approach can be used to capture any pattern of dependencies between multiple causes and an effect, with any conditional distribution being expressible as a combination of noisy logic gates.

2.3.3 Learning causal structure

While ΔP and causal power are rational measures of the strength of a causal relationship given certain assumptions about the nature of those relationships, other recent work has explored an alternative view of the problem of causal induction, focusing on the *structural* decision as to whether a causal relationship exists rather than its strength (Griffiths & Tenenbaum, 2005). There are two general approaches to learning what causal structure underlies a set of observations.

The first approach to structure learning has been called the "constraint-based" approach and involves using statistical tests of independence (such as χ^2) to determine which variables are related to one another, and then reasoning deductively from patterns of dependencies to causal structures (e.g., Pearl, 2000; Spirtes et al., 1993). Constraint-based causal learning typically makes no assumptions about the nature of causal relationships, using statistical tests to search for violations of statistical independence and exploiting the fact that the structure of a graphical model implies certain independence relations, to identify the causal structure underlying the available evidence. This

makes it possible to recover causal structures regardless of the functional form of the underlying relationships. However, this flexibility comes at the cost of efficiency. People can learn causal relationships from just a few observations of cause and effect (Shultz, 1982; Gopnik & Sobel, 2000), but constraint-based algorithms require relatively large numbers of data to determine causal structure (enough to produce statistical significance in a test of independence).[1]

The second approach to structure learning is to frame causal induction as a problem of Bayesian inference. In this approach, a learner must determine which hypothetical causal graphical model h is likely to be the true one, given observed data d and a set of beliefs about the plausibility of different models encoded in the *prior* probability distribution $P(h)$. Answering this question requires computing the *posterior* probability $P(h|d)$, which can be found by applying Bayes' rule

$$P(h|d) = \frac{P(d|h)P(h)}{\sum_{h'} P(d|h')P(h')} \tag{2.3}$$

where $P(d|h)$ indicates the probability of observing d assuming that h is true, and is known as the *likelihood*. The likelihood is computed using the probability distribution associated with the causal graphical model h, and thus reflects the expectations of the learner about the functional form of causal relationships. However, in applications of Bayesian structure learning in machine learning (e.g., Cooper & Herskovits, 1992), minimal assumptions are made about the functional form of causal relationships – typically just that the probability of the effect differs in the presence and absence of a cause.

2.4 Modeling the effects of knowledge of functional form

The models of causal induction outlined in the previous section provide a basic framework in which to explore how prior knowledge influences the inferences that people make about causal relationships. However, none of these models directly addresses the problem of learning the functional form of a causal relationship and using that knowledge to inform future causal learning. Existing models either assume that causal relationships have a specific functional form, or make no assumptions about the functional form of a causal relationship. Causal power and ΔP do not allow for the possibility that people might assume different functional forms in different contexts, and constraint-based algorithms and standard Bayesian structure learning invoke minimal knowledge at the cost of efficiency. However, the Bayesian approach to causal learning provides us with the tools we need in order to explore how knowledge of functional form affects causal induction and how this knowledge is acquired. We will focus on learning causal structure, although a similar approach could be applied for learning causal strength.

2.4.1 Using knowledge of functional form

Bayesian inference uses two pieces of abstract knowledge. The first is some prior beliefs about which hypotheses are more likely than others, encoded in the prior distribution $P(h)$. The second is a set of expectations about what effects one should observe given that certain causes are

[1] Nothing prevents constraint-based approaches from including assumptions about functional form to facilitate rapid learning – in such cases one would test specific classes of relationships reflecting the assumptions rather than general statistical independence – but such assumptions lead to the same restrictions that face other fixed-form models.

present, encoded in the likelihood function $P(d|h)$. While most Bayesian structure learning algorithms make relatively generic assumptions about the prior and likelihood, we can make stronger assumptions in order to reflect the knowledge that learners possess. In particular, we can capture knowledge about the functional form of a causal relationship through our choice of likelihood.

We assume that the causal structures under consideration contain variables in two classes (prospective causes and effects), the class to which each variable belongs is known, and that the only relationships that could potentially exist are those between causes and effects. The data, d, are events observed by the learner, consisting of causes being present or absent and effects either occurring or not. We also assume that the events in d are independent once the underlying causal structure is known, so that $P(d|h) = \prod_k P(d_k|h)$ where d_k is a single event. Finally, we assume that the probability of a cause being present in a given event does not depend on the causal structure h, with the causal structure merely determining the probability that the effect occurs. Defining the likelihood then reduces to specifying the probability with which the effect occurs, given the number of its causes that are present.

Returning to the example of the light, different mechanisms translate into different probabilities for certain events, and consequently different likelihood functions. For instance, one might expect that a deterministic conjunctive function applies, giving the effect probability 1 when all causes are present, and 0 otherwise.[2] Alternately, a deterministic disjunctive function might apply, giving the effect probability 1 when at least one cause is present, and 0 otherwise. If there is reason to believe a noisy-OR relationship is at work, the probability of the effect is given by Equation 2.2, as discussed previously.

Assuming that a causal system follows a specific functional form such as one of the above provides constraints that aid causal learning. For example, under a disjunctive function without background causes, just one observation of the presence of a cause and the occurrence of the effect is sufficient to indicate that a causal relationship exists. By exploiting this kind of information, Bayesian models incorporating knowledge about the functional form of causal relationships are able to identify causal structure from limited data in the same way as human learners (Tenenbaum & Griffiths, 2003).

This approach to characterizing how knowledge of functional form might be used is consistent with previous work in causal learning. The idea that the functional form is expressed through the likelihood, defining the probability of the effect given the cause, is standard in statistical interpretations of causal strength estimation (Cheng, 1997; Glymour, 1998; Griffiths & Tenenbaum, 2005; Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2007, 2008). These models allow for a variety of functional forms, including noisy-OR relationships and their generalization to incorporate interactions and other kinds of noisy-logical circuits (Novick & Cheng, 2004; Yuille & Lu, 2007). By considering functional forms that allow linear combinations of causes (including averaging), the results of Waldmann (2007) and Beckers et al. (2005) could also be accommodated within this framework. However, a more significant challenge is explaining how people acquire knowledge of functional form that is appropriate to a given domain.

[2] A note on terminology: we have chosen to use the terms *disjunctive* and *conjunctive* instead of, e.g., OR and Noisy-AND, in the interest of accuracy. A *disjunctive* (generative) causal relationship is one in which the probability of the effect increases at most linearly with the number of causes present, and a *conjunctive* relationship is one in which the probability effect increases more sharply once some number of causes ($n > 1$) is exceeded. Noisy-OR and noisy-AND functions are special cases of these, and are used where appropriate.

2.4.2 Learning causal theories

Bayesian inference provides a simple way to make use of abstract knowledge about the functional form of a causal relationship. More generally, the abstract knowledge needed to perform Bayesian inference can be expressed as a "theory" about a domain, identifying types of objects, the plausibility of causal relationships, and the form of those relationships (Griffiths, 2005; Griffiths & Tenenbaum, 2007). The Bayesian approach also allows us to analyze how these theories themselves might be learned. The notion that we learn theories – complex, abstract, and consistent representations that like scientific theories reflect and facilitate inference about the structure of the real world – has a long history in cognitive development (Carey, 1991; Gopnik & Wellman, 1992). Recent work in computational modeling of causal learning has begun to extend the Bayesian approach to inferring the structure of causal graphical models to the level of theories, using hierarchical Bayesian models (Tenenbaum & Griffiths, 2003; Griffiths & Tenenbaum, 2007).

The basic idea behind a hierarchical Bayesian model is to perform probabilistic inference at multiple levels of abstraction. In the case of causal learning, these levels are the hypothetical causal graphical models under consideration – the hypotheses h we have been discussing so far – and the abstract theories that generalize over these hypotheses. In the resulting probabilistic model, we assume that each theory t defines a probability distribution over hypotheses $P(h|t)$, just as each hypothesis defines a probability distribution over data, $P(d|h)$. To return to the example of the light and switches, we might thus characterize our knowledge of how electrical systems tend to work in terms of a probability distribution over a set of causal graphical models that differ in their parameterization, reflecting conjunctive, disjunctive, or other possible kinds of relationships. Each kind of parameterization would correspond to a different theory, t, with each theory defining a prior distribution over causal graphical models featuring that parameterization.

Like structure learning, theory learning can be reduced to a problem of Bayesian inference. If our goal is to infer a theory t from data d, we can do this by applying Bayes' rule, with

$$P(t|d) = \frac{P(d|t)P(t)}{\sum_{t'} P(d|t')P(t')} \tag{2.4}$$

where $P(t)$ is a prior distribution on theories. The likelihood $P(d|t)$ is obtained by summing over all hypothetical causal structures, with $P(d|t) = \sum_h P(d|h,t)P(h|t)$ where $P(d|h,t)$ is the probability of the data given the structural hypothesis h under the theory t (reflecting the assumptions of the theory about the functional form of a causal relationship) and $P(h|t)$ is the probability of that hypothesis given the theory (reflecting assumptions about the plausibility of particular relationships).

Equation 2.4 applies Bayes' rule in the same way as it was applied for hypothetical causal structures in Equation 2.3. However, theories are defined at a higher level of abstraction than hypotheses about the causal structure relating a specific set of variables. As a consequence, this knowledge supports generalization. For example, upon learning that two switches need to be pressed in order to turn on a light, a learner might believe that a conjunctive relationship is likely to apply in similar settings in the future. The idea that unknown variables can be shared across datasets is at the heart of hierarchical Bayesian models (for details, see Gelman, Carlin, Stern, & Rubin, 1995) and makes it possible for information from one dataset to guide inferences from a second. More formally, learners who encounter a dataset $d^{(1)}$ can update their posterior distributions over theories,

computing $P(t|d^{(1)})$ as in Equation 2.4. Upon encountering more data, $d^{(2)}$ in a similar setting, this posterior distribution over theories can be used to guide inferences about causal structure. Specifically, it takes the role of the prior over theories for interpreting the new data. The joint distribution on hypotheses and theories is given by

$$P(h, t|d^{(1)}, d^{(2)}) \propto P(d^{(2)}|h, t)P(h|t)P(t|d^{(1)}) \tag{2.5}$$

where the constant of proportionality is obtained by normalizing the distribution to sum to one over all h and t. The probability of a given causal structure is then obtained by summing over all theories, with

$$P(h|d^{(1)}, d^{(2)}) = \sum_t P(h, t|d^{(1)}, d^{(2)}) \tag{2.6}$$

allowing the abstract knowledge about causal relationships gleaned from $d^{(1)}$ to influence the conclusions drawn from $d^{(2)}$. Hierarchical Bayesian models thus allow learners to identify the abstract principles that organize a domain, updating their expectations as more data are observed.

This account of how knowledge of functional form can be acquired is the main novel contribution of our approach. It provides a way to understand how learners might make inferences about functional form from observing a causal system, and then use this knowledge later when learning about causal relationships. It differs from the previous approaches to modeling causal learning discussed above in allowing learners to flexibly identify a specific functional form for causal relationships in a given setting, instead of assuming a fixed functional form for all causal relationships, or making weak and generic assumptions about functional form. A similar approach was recently used by Lu, Rojas, Beckers, and Yuille (2008) to explain the results of Beckers et al. (2005) (see the General Discussion for details), and could be used to explain how children and adults form generalizations about the relationship between physical variables in the experiments of Zelazo and Schultz (1989) , with different kinds of relationships corresponding to different causal theories. However, we chose to test the predictions of this account by considering a novel causal system in which we can manipulate a variety of factors related to functional form.

2.5 Testing the predictions of the hierarchical Bayesian approach

The approach outlined in the previous section can be applied to the problem of acquiring and using knowledge of the functional form of causal relationships: abstract theories can express different assumptions about functional form, allowing learners to infer what kind of functional form is most appropriate for a given domain. There are two qualitative predictions that distinguish this hierarchical Bayesian approach from other models of causal learning: (1) that people can make inferences appropriate to causal relationships with more than one kind of functional form, and (2) that people can use evidence from one data set to inform their inferences from another involving different variables. In the remainder of the paper we present a test of these predictions, using a specific causal system to explore whether people can form generalizations about the functional form underlying a causal relationship and what factors influence this process. To do so, we use a causal inference problem in which knowledge of functional form is important, not known in advance, and which permits us to generate quantitative predictions with a specific hierarchical Bayesian model.

Suppose a learner is faced with the problem of identifying which of a set of objects are

Table 2.1: Evidence presented to participants in Experiment 1.

Block	Evidence	Blicket
Conjunctive training	A-B-C-AB-AC+BC-	A,C
Noisy disjunctive training	A+B-C-AB-AC+BC-	A
Deterministic disjunctive training	A+B-C-AB+AC+BC-	A
Test	D- D- D- E- DF+ DF+	

"blickets" using a "blicketosity meter" knowing only that blickets possess something called blicketosity and that the meter sometimes lights up and plays music ("activates").[3] The hypotheses under consideration by the learner are partitions of objects into blicket and non-blicket classes. Since activating the meter is the result of a causal relationship between the object and the detector, these hypotheses can be expressed as causal graphical models, where objects are prospective causes, activation is the effect, and a causal relationship exists between cause and effect if and only if the corresponding object is a blicket.

Crucially, two objects can be placed on the blicketosity meter simultaneously, making it possible to explore different functional forms for the underlying causal relationship. Different functional forms have strong implications for how the learner should interpret different events. For instance, if the learner believes that two blickets' worth of blicketosity are necessary to activate the meter, then seeing a single object fail to activate the meter is uninformative. Under such a belief, two objects that together activate the meter are both blickets, whereas that event under a disjunctive relationship suggests only that one or both of the objects is probably a blicket.

If people assume that the functional form of a causal relationship is fixed (e.g., linear or noisy-OR), or they make minimal assumptions about the functional form of causal relationships, then they will make the same inferences about casual structure regardless of evidence about functional form. Consequently, testing predictions in cases where structural inferences are guided by knowledge of functional form is a way to evaluate the central claims of a hierarchical Bayesian approach: that people entertain abstract theories about the functional form of causal relationships and update their beliefs about these theories in light of evidence. This logic motivated the experiments that we present in this paper. In describing these experiments, we will also define a specific hierarchical Bayesian model, constructed by choosing likelihoods and priors that are simple, flexible, and appropriate to the cover story given to participants, and compare its numerical predictions to those generated by other models of causal inference.

In our experiments, participants were first presented with one of three sets of events involving three objects (A,B,C).[4] Participants saw events that had high probability under either a deterministic disjunctive, conjunctive, or noisy disjunctive theory (see Table 2.1). Next, all groups saw a set of events with three new objects (D,E,F) that were compatible with any of the three theories, D- D- D- E- DF+ DF+, i.e., object D failing to activate the meter three times, E fail-

[3] Previous work using a similar device has referred to it as a "blicket machine" or "blicket detector" (Gopnik & Sobel, 2000; Sobel et al., 2004). We chose to call it a "blicketosity meter" as this gave minimal cues to functional form, while "blicket detector" seems more consistent with an OR function.

[4] We will represent events as a set of present or active prospective causes and the presence or absence of an effect, e.g., if possible causes A and B are present and the effect is observed, the event can be written down as $(\{a, b\}, e)$, or more concisely, $AB+$.

ing once, and a combination of D and F succeeding twice. If participants acquired knowledge about functional form using the first block of evidence, then they would be be expected to come to different conclusions about which objects in the second were blickets.

Using the terms introduced in the previous section, the events involving A, B, and C form an initial dataset $d^{(1)}$, and the events involving D, E, and F form a second dataset $d^{(2)}$. After seeing $d^{(1)}$, learners can compute a posterior distribution over theories, $P(t|d^{(1)})$ by applying Equation 2.4. This posterior distribution informs inferences about the hypothetical causal structures that could explain $d^{(2)}$, as outlined in Equations 2.5 and 2.6. The key prediction is that varying $d^{(1)}$ should affect inferences from $d^{(2)}$. This prediction is not made by any non-hierarchical model, because the events in $d^{(1)}$ do not involve any of the prospective causes in $d^{(2)}$. Without abstract acquired knowledge that spans multiple contexts, there is no continuity between the two sets of evidence.

While the qualitative predictions tested in our experiments are made by any hierarchical Bayesian model, we can also obtain quantitative predictions by considering a specific model that gives values to $P(d|h,t)$, $P(h|t)$, and $P(t)$. The appropriate model will depend on the context, as the hypotheses and theories that are relevant may vary. Our goal was thus not to define a general-purpose model that could explain all instances of learning of functional form, but a simple model appropriate for characterizing the inferences that people make about functional form for the blicketosity meter. The main purpose of developing this model was to illustrate how people's judgments should be affected by our experimental manipulations, assuming a reasonable but relatively restricted set of possible functional forms.

Our hierarchical Bayesian model is defined in terms of the distributions $P(d|h,t)$, $P(h|t)$, and $P(t)$. We will discuss $P(h|t)$ first, then turn to $P(d|h,t)$ and $P(t)$. This is partly motivated by the fact that $P(h|t)$ is the simplest part of the model: we take all hypotheses regarding causal structure to be *a priori* equally likely, yielding $P(h|t) = P(h)$ where $P(h)$ is identical for all structures. If the learner learns what the causal structure is, e.g., he or she is told what causal structure is present in the training block, then $P(h)$ is updated to reflect that knowledge, so that the probability of the known structure becomes 1.

As for $P(d|h,t)$, we have already provided examples of how different functional forms might translate into likelihood functions, and chose for our model a space of theories that can approximate noisy-OR, AND, and other relationships while being broadly compatible with the cover story provided to participants and requiring a small number of parameters. Specifically, we selected a sigmoid likelihood function:

$$P(e|N_{\text{blickets}} = n) = \frac{1}{1 + \exp\{-g(n-b)\}} \tag{2.7}$$

where N_{blickets} is the number of blickets among the present objects, b is the bias, i.e., the number of blickets necessary to make the meter equally likely to activate as not, and g gives the gain of the function. To give a sense of the generality of this function, when $b = 0.5$ and $g >> 1$ one obtains a deterministic-OR function, $b = 1.5$, $g >> 1$ gives a deterministic conjunctive function, and $b = 0.81$, $g = 7.37$ closely approximates a noisy-OR function with $w_i = 0.8$ for all i (see Figure 2.2). Under this specification, the theory held by a learner amounts to his or her beliefs about the probable values of b and g.

Finally, we must provide a prior over theories, which under our model amounts to a prob-

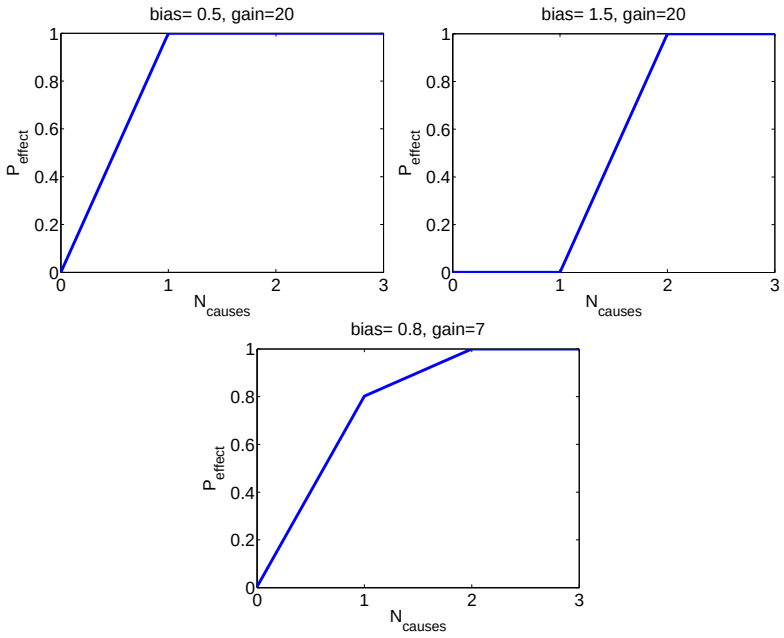


Figure 2.2: Examples of different kinds of relationship forms expressed as sigmoid functions.

ability density for b and g. For the sake of simplicity, we chose exponential priors for both b and g, each with a single hyperparameter (λ_b and λ_g) setting how rapidly the probabilities of values of b and g decrease. The probabilities of b and g were thus proportional to $\exp\{-\lambda_b b\}$ and $\exp\{-\lambda_g g\}$ respectively. If the mean of the prior distribution for the bias parameter is less than one (ie. $\lambda_b > 1$), the prior beliefs favor disjunctive relationships while a large value for the hyperparameter for the gain, λ_g, favors deterministic rather than noisy functions.[5] While we believe that our specification for the space of functional forms is generally appropriate for the cover stories in our experiments, we do not assert that it encompasses all theories that people can entertain. For example, Shanks and Darby (1998) discuss an experiment in which people learn that causes can independently bring about an effect but jointly do not, a functional form that cannot be specified in terms of the logistic function. We return to the issue of more general models of functional form learning in the General Discussion.

With the training data we selected, our model predicts a disordinal interaction in which the rank-order of the ratings for objects D and E reverses: in the conjunctive conditions object D was expected to be judged more likely to be a blicket than object E, and vice versa in the disjunctive conditions. This interaction, which emerges with a wide range of plausible values for λ_b and λ_g, is a consequence of the fact that the D- events should be taken as evidence against D being a blicket under a disjunctive theory, while under a conjunctive theory the D- events are uninformative and the DF+ events indicate that D is a blicket. To make the quantitative predictions that we test in our

[5] Based on the suspicion that people would strongly favor deterministic functions, we considered using an inverse-exponential prior for the steepness – which assigns very low probability to values near zero – but abandoned it in favor of the exponential which can also strongly favor deterministic theories with the right parameter and is compatible with a wider range of beliefs, such as that the meter is acting randomly.

experiments, a single value for each of λ_b and λ_g was chosen to minimize sum squared error when compared to participants' ratings over all experiments, resulting in $\lambda_b = 4.329$ and $\lambda_g = 0.299$. The predictions were fairly insensitive to these values, a point that we explore in detail in the General Discussion.

In the remainder of the paper we present a series of experiments testing the predictions that discriminate our account from others. Experiment 1 examines learning about the functional form of causal relationships when the causal structure is known. Experiment 2 addresses the problem of learning about causal structure and functional form simultaneously, and Experiment 3 provides control conditions to confirm our interpretations of the results of Experiment 2. Experiment 4 tests additional predictions about the consequences of acquiring knowledge about functional forms, and Experiment 5 deals with the domain-specificity of this knowledge.

2.6 Experiment 1: Known causal structure

Experiment 1 was designed to be a direct test of the predictions that distinguish a hierarchical Bayesian account from others, namely that events involving one set of variables will influence later inferences drawn about a different set of variables. Here, we simplified the task by providing participants with the causal structure behind the first event set, leaving them only the problem of learning about functional form.

2.6.1 Methods

Participants

Participants were 57 undergraduates from the University of California, Berkeley who received course credit for participation. These participants were divided into *deterministic disjunctive* ($n = 20$), *noisy-disjunctive* ($n = 20$), and *conjunctive* ($n = 17$) conditions.

Materials and Procedure

The stimuli were six identical beige 2"×2"×2" cubes, labeled A, B, C, D, E, and F, and a green 5"×7"×3" box with a translucent orange panel on top, called a "blicketosity meter". A hidden foot-controlled switch allowed the experimenter to toggle the meter between a state in which it activated when objects were placed on it, and a state in which it did not. When activated, the box played a short melody and the orange panel was illuminated by a light inside.

The experiment consisted of two trials. In the first trial, participants were told that some blocks are blickets, some are not, and that blickets possess blicketosity while non-blickets possess no blicketosity. They were also told that the blicketosity meter had a binary response: it could either activate or not. One or two objects were then identified as blickets (see Table 2.1, third column). They were given no more information about the nature of blickets or the blicketosity meter. The experimenter provided participants with evidence by placing the first three objects (A, B and C), singly or in groups, on the box. By surreptitiously toggling the hidden foot switch, the experimenter was able to control which combinations of objects appeared to activate the machine. There were three different training evidence conditions, labeled by the form of causal relationship

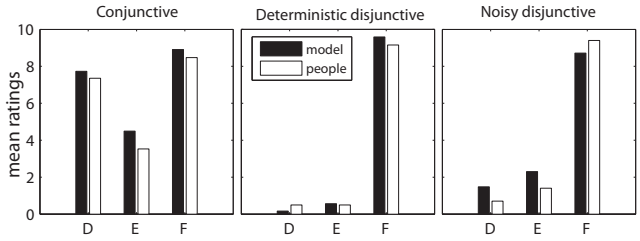


Figure 2.3: Results of Experiment 1, showing the model predictions and human ratings of the probability that test-condition objects are blickets.

they suggested: *deterministic-disjunctive*, *noisy-disjunctive*, and *conjunctive*. Table 2.1 gives the specific events presented in each condition.

After the training block was presented, the objects were set aside, and three new objects labeled D, E, and F were introduced. Participants then saw a block of test evidence that was the same across all conditions, D- D- D- E- DF+ DF+. These specific events were chosen to lead to different beliefs about which objects were blickets, depending on what kind of relationship participants believed applied.

After participants saw the evidence in the test block, they recorded the probability they assigned to each of the test objects D, E, and F being blickets on a 0-10 scale, having been told that a ten indicated they were absolutely certain the object was a blicket, a zero indicated absolute certainty it was not, and a 5 indicated that it was equally likely to be a blicket as not.

Finally, after all participants had recorded their ratings, they were prompted to record in plain English their theories of how the meter interacted with blickets and non-blickets, and how the meter operated.

2.6.2 Results and Discussion

The mean ratings are shown in Figure 2.3. The prediction that causal knowledge derived from one set of objects constrains subsequent inferences was supported: one-way ANOVAs found an effect of the Trial 1 data on judgments in Trial 2 for both object D ($F(2, 54) = 58.1$, $p < 0.001$, $\eta^2 = 0.68$) and object E ($F(2, 54) = 11.353$, $p < 0.001$, $\eta^2 = 0.2$). The specific effects we expected were that object D would be given higher ratings in the *conjunctive* condition than in the *deterministic-disjunctive* condition, and that object E would have a higher rating in the *noisy-disjunctive* condition than in the *deterministic-disjunctive* condition. We found support for the first of these effects ($t(35) = 8.759$, $p < 0.001$, $d = 1.64$) and a trend consistent with the second ($t(38) = -1.603$, $p = 0.117$, $d = -0.50$). The data also supported the hypothesis that many participants in the *conjunctive* test block were inferring that a conjunctive relationship applied to the events: the mean rating of D in the *conjunctive* condition was significantly higher than 5 ($t(16) = 3.15$, $p < 0.01$, $d = 0.76$), indicating that D was considered more likely than not to be a blicket, something that is inconsistent with all linear and noisy disjunctive interpretations.[6]

[6]Hypothetically, a noisy disjunctive relationship with a high failure rate coupled with a belief that almost all objects are blickets could lead to such an inference, but such an explanation is incompatible with ratings participants gave for E.

The numerical predictions of our model closely matched participants' behavior: the within-condition rank orders of ratings were in perfect agreement, mean squared error was 0.38, and the linear correlation between the ratings and participants' judgments was 0.99. We can evaluate the performance of the model by comparing it to alternative models that also make quantitative predictions. As mentioned previously, ΔP and causal power do not predict that responses will vary between any of our conditions. Under ΔP, the predicted response for D is $P(e|D) - P(e|\overline{D})$, or $\frac{2}{5}$, for a rating of 4. Causal power normalizes that quantity by $1 - P(e|\overline{D})$, also giving 4 for D. The predicted ratings for E and F under both ΔP and causal power are 0 and 10, respectively. These models' predictions are accurate in the disjunctive conditions where their assumptions regarding functional form are appropriate, but in the *conjunctive* condition their predictions diverge from the data, leading to an overall mean squared error per rating of 5.83 and a correlation of 0.82 with human ratings. We will evaluate some extensions of these models in the General Discussion.

We can also compare the performance of our model to that of the best possible model that is blind to the training evidence. This model would make optimal predictions for the ratings of D, E, and F, under the constraint that the ratings must be the same across all three conditions. The best predictions for D, E, and F – in terms of minimizing error and maximizing correlation – are equal to their observed means across all three conditions. Such predictions yield a mean-squared error of 3.97 and a linear correlation of 0.83. The best possible model that ignores condition thus performs much worse than our Bayesian model, since this model fails to capture the change in ratings produced by providing different information about the functional form of the causal relationship.

In order to establish that participants' ratings reflected their beliefs about the form of the causal relationships and were likely to generalize appropriately to evidence beyond that provided in our experiment, we analyzed the plain-English theories participants expressed after the second trial. Two hypothesis- and condition-blind research assistants coded participants' theories about how the meter worked, resolving any differences through discussion. The features recorded for each theory included (1) whether the theory was interpretable, (2) whether blickets and non-blickets increased, decreased, or had no influence on the probability of the effect, (3) whether the effect was perfectly, imperfectly, or not predictable given its observable causes, and (4) whether or not the relationship was conjunctive. The overall proportion of theories that was interpretable was 0.74, and did not differ significantly between conditions (Fisher's exact text, $p > 0.5$).

Differences in participants' explicit theories were consistent with learning about the form of the causal relationship: a greater proportion of interpretable theories in the *conjunctive* condition were consistent with a conjunctive relationship (12 of 13) than in the *deterministic-* and *noisy-disjunctive* conditions (3 of 16 and 4 of 13, respectively) (Fisher's exact test, $p < 0.01$). Participants in the *noisy-disjunctive* condition expressed theories that involved noise or imperfect predictive ability more frequently (5 of 13 interpretable theories, versus 1 of 16 for the *deterministic-disjunctive* condition and 0 of 13 for the *conjunctive* condition (Fisher's exact test, $p < 0.05$).

The results of the experiment support our account of causal learning: people developed different theories about the functional form of the causal relationship, and used these theories when reasoning about the existence of individual causal relationships. However, since people were provided with information about the relationships that existed among the objects presented in the training block, this remains a modest test of their ability to learn the functional form of causal relationships. As a more ambitious test, we conducted Experiment 2, in which participants were forced to learn about causal structure and functional form simultaneously.

2.7 Experiment 2: Unknown causal structure

Having found the predicted pattern of judgments when participants knew which training objects were blickets, we conducted Experiment 2 to test the prediction that people can concurrently learn about the causal structure and functional form of causal relationships. We did this using a procedure identical to Experiment 1, save that we withheld the identities of the blickets in the training block, effectively hiding the underlying causal structure. In addition, we took the opportunity to address an alternative interpretation of the results of Experiment 1, under which participants took there to be the same number of blickets in both blocks of evidence and mapped objects in the test block to objects in the training block. We ran a new condition to test this possibility, in which participants saw evidence that was likely given a deterministic-OR relationship and two blickets. If the alternative explanation were true, then participants would be expected to pick out two of the D, E, and F objects as blickets. If people are using information about the functional form of causal relationships to make inferences in the test condition, then their judgments should be similar to those in the one-blicket deterministic disjunctive condition.

2.7.1 Methods

Participants

Participants were 102 undergraduates from the University of California, Berkeley who received course credit for participation, again divided into *deterministic-disjunctive* ($n = 26$), *noisy-disjunctive* ($n = 26$), and *conjunctive* ($n = 24$) conditions, with an additional deterministic disjunctive *base-rate control* ($n = 26$) condition.

Materials and Procedure

The first three conditions were identical to those in Experiment 1, but participants were told nothing about which objects were blickets in the training block, and instead were asked to provide probabilities as in the test block. These will again be referred to as the *deterministic-disjunctive*, *noisy-disjunctive*, and *conjunctive* conditions. The *base-rate control* condition was an additional control to establish that participants were not merely using base-rate information to infer that a specific number of blickets were present in the test block. The evidence participants saw was intended to be compatible with a deterministic disjunctive relationship, but with two blickets present rather than one. The procedure was the same as in the previous three conditions, but the specific training training evidence participants saw was A+ B+ C- AB+ AC+ BC+.

2.7.2 Results and Discussion

Our first analyses focused on the three conditions from Experiment 1, the *deterministic-disjunctive*, *noisy-disjunctive*, and *conjunctive* conditions. As in Experiment 1, there was an effect of the training block on judgments in Trial 2 for D ($F(2, 73) = 6.026$, $p < 0.01$, $\eta^2 = 0.14$). More specifically, object D was given higher ratings in the *conjunctive* condition than in the *deterministic-disjunctive* condition ($t(48) = 3.472$, $p < 0.01$, $d = 0.89$). Contrary to our predictions, the mean rating given to object E was not higher in the *noisy-disjunctive* condition than in the *conjunctive* condition, which we discuss below. With the exception of this reversal, the ordinal match between

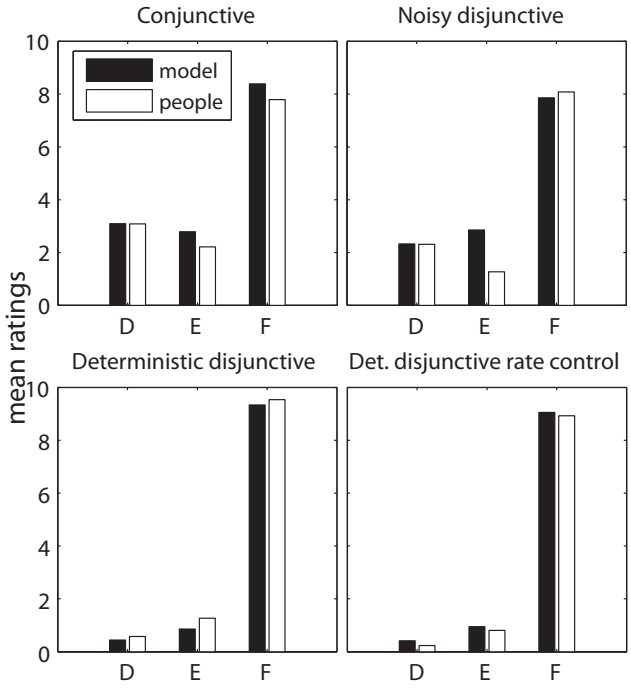


Figure 2.4: Results of Experiment 2, showing the model predictions and human ratings of the probability that test-condition objects are blickets.

numerical predictions of our model and participants' ratings was exact. Mean squared error was 0.29, and the linear correlation between the predictions and participants' ratings was 0.99. The mean ratings and model predictions are shown in Figure 2.4. For comparison, the predictions of both ΔP and causal power yielded an MSE of 4.03 and a correlation of 0.90. Because the ratings varied less between conditions, the best-possible training-blind predictions were better than in Experiment 1, giving an MSE of 0.63 and a correlation of 0.97. As in Experiment 1, participants' explicit theories mentioned conjunctive relationships more often in the *conjunctive* condition (13 of 17) than the *noisy-disjunctive* (6 of 23) and *deterministic-disjunctive* (5 of 23) conditions (Fisher's exact test, $p < 0.001$)

In retrospect, the higher-than-expected rating for D in the *noisy-disjunctive* condition is unsurprising given that the training events were also compatible with complex deterministic theories, and previous work suggests that people tend to believe that complex deterministic causal relationships are more likely than simple stochastic relationships (L. Schulz & Sommerville, 2006). The theories participants expressed were compatible with this interpretation: only 13 (3 of 23) percent of the interpretable theories mentioned noise or an imperfectly predictable effect, versus 38 (5 of 13) percent in the *noisy-disjunctive* condition of Experiment 1, although this difference was not significant (Fisher's exact test, $p = 0.11$). If people are inferring that deterministic relationships outside our space of theories apply, then making it clear that the generative relationship is subject to occasional failure will bring participants' ratings back in line with our predictions. We explore this possibility in the next experiment.

Finally, the results of our additional control condition were consistent with our account. In the *base-rate control* condition the mean ratings for objects D and E were lower than in the original *deterministic-disjunctive* condition, giving a larger effect when comparing *conjunctive* condition ratings for D ($t(48) = 4.18$, $p < 0.001$, $d = 1.02$), as predicted by our model and contrary to what an alternative explanation based on the frequency with which blickets appear in the training data would predict.

2.8 Experiment 3: The effects of noise

In Experiment 2 we found that participants' judgments in the *noisy-disjunctive* condition deviated from the predictions of our model, and speculated that participants were inferring that complex deterministic relationships produced the events. To test whether this was the case, we ran two new conditions in which we provided explicit evidence that the meter was subject to unpredictable failures to activate. In the first, we gave participants training events that were incompatible with a deterministic explanation by prepending two failure events (A- A-) to the *noisy-disjunctive* data in Experiment 2. In the second we told participants that the meter was failure-prone. If participants' unexpected judgments in Experiment 2 were the result of their rejecting the possibility that the relationship was noisy, then both interventions should lead to ratings for D being lower than those for E.

2.8.1 Methods

Participants

Participants were 41 undergraduates from the University of California, Berkeley who received course credit for participation, divided into *event-based noise* ($n = 13$) and *description-based noise* ($n = 28$) conditions.

Materials and Procedure

The two conditions used a procedure based on that in the *noisy-disjunctive* condition of Experiment 2. In the *description-based noise* condition participants were told "Sometimes, the blicketosity meter randomly fails to go off when it should." The *event-based noise* condition added two events in which object A failed to activate the meter to the *noisy-disjunctive* training block, so that participants saw the events A+ A- A- B- C- AB- AC+ BC- in the training block. The two conditions were otherwise identical. As in Experiments 1 and 2, the test evidence was D- D- D- E- DF+ DF+

2.8.2 Results and Discussion

Under both manipulations, the mean participant ratings for D were lower than E, a result that was significant after aggregating data from the two manipulations ($t(40) = -2.03$, $p = 0.049$, $d = 0.32$). Comparing the effect of these two manipulations on the difference between the D and E ratings to the $D - E$ difference in the *noisy-disjunctive* condition of Experiment 2, yielded a

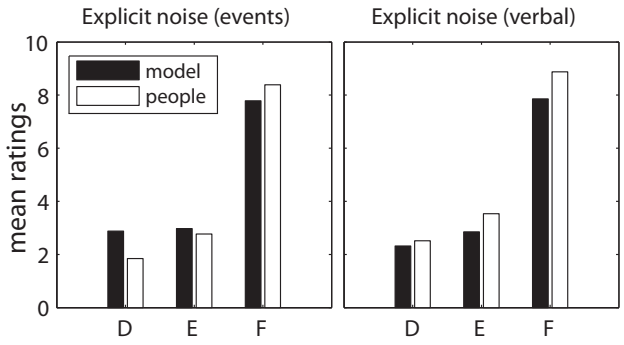


Figure 2.5: Results of Experiment 3, showing the model predictions and human ratings of the probability that test-condition objects are blickets.

significant difference in the *description-based noise* condition ($t(52) = 2.41$, $p = 0.019$, $d = 0.63$), and a trend in the *event-based noise* condition ($t(37) = 1.92$, $p = 0.062$, $d = 0.63$).

The correlation between the ratings of participants and the predictions of the model was 0.98, and the mean squared error was 0.55. Mean ratings and model predictions are given in Figure 2.5. With this indication that people make the inferences one would expect when they have evidence for a failure-prone system, we can test another prediction of our model: that people can use covariation evidence to learn how noisy a class of causal relationships is and use that information to make more accurate inferences about causal structures involving novel objects.

2.9 Experiment 4: Manipulating causal strength

If people are transferring knowledge about the strength of the relationship between the presence of blickets and activation of the meter, then one would expect to see different ratings for the probability of D, E, and F being blickets as the training set is manipulated to suggest higher or lower failure rates for the meter. Specifically, our intuitions and model predict that the probability that D is a blicket given the test data is higher under a high-failure-rate noisy disjunctive relationship than one that fails infrequently; under a nearly failure-free relationship, the three failures to activate under D constitute strong evidence against D being a blicket, while the evidence against E – a single failure – is weaker. At the opposite extreme, when the meter rarely activates for a blicket, the three failures constitute weak evidence that D is not a blicket, while the activation under D and F together is now positive evidence for D being a blicket. We tested this prediction with another experiment.

2.9.1 Methods

Participants

Participants were 41 undergraduates from the University of California, Berkeley, who received course credit for participation, divided into *high-noise* ($n = 20$), and *low-noise* ($n = 21$)

conditions. One participant was excluded for explaining that the blicketosity meter detected non-blickets.

Materials and Procedure

The procedure was similar to that used in the *event-based noise* condition in Experiment 3, but the evidence was varied between two conditions to indicate different failure rates. The evidence in the *low-noise* condition was A+ A+ A+ A- A+ A+ B- C- AB- AC+ BC-, where the meter activated five out of six times given object A alone. The evidence in the *high-noise* condition was A- A- A- A+ A- A- B- C- AB- AC+ BC-, where the meter activated one out of six times given object A alone.

2.9.2 Results and Discussion

The mean rating for D was higher in the *high-noise* condition (3.8, SD=2.4), than in the *low-noise* condition (1.8, SD=2.4), ($t(39) = 2.57$, $p = 0.014$, $d = 0.753$), consistent with the predictions of the model. As in the previous experiments, the quantitative predictions of the model were accurate, with a correlation with ratings of 0.98 and a mean squared error of 0.52.

At this point we have shown that people used covariational evidence to learn about the functional form of causal relationships – including whether the basic structure of the relationship was conjunctive or disjunctive, and the strength of disjunctive relationships – and used that knowledge to guide their later inferences. However, it might be argued that what we have observed was not the acquisition of abstract knowledge, but rather a sort of domain-general priming effect, in which participants' inferences after the training block might broadly favor kinds of relationships consistent with the evidence they had seen, and confabulated when recording their theories. One might also argue that our results were peculiar to our particular cover story or manner of presenting evidence to participants. We designed Experiment 5 to address these possibilities.

2.10 Experiment 5: Effects of domain

This experiment had three goals: (1) To gather support for the idea that the learning demonstrated in previous experiments is a matter of acquiring domain-specific knowledge rather than priming kinds of causal relationships in a domain-independent way, (2) To establish that the effects observed do not depend on the use of a live demonstration, and (3) To show that the transfer-learning effect is not restricted to the "blicketosity meter" cover story. Accordingly, we used a survey procedure in which we described a causal learning scenario and manipulated the domains in which the training and test stimuli were presented, as well as whether those domains matched or differed from one another. The critical prediction is that we should see knowledge of functional form acquired through training having a far greater effect on inferences at test when the domains of the training and test scenarios match.

2.10.1 Methods

Participants

Participants were 60 undergraduates from the University of California, Berkeley, split equally over four conditions corresponding to two test domains crossed with whether the training domain matched or differed from the test domain.

Materials and Procedure

Each participant completed one of four surveys, which varied according to two factors. The first was whether the test block of evidence had a cover story identical to that used in Experiments 1-4, or a novel one which replaced the activation of the meter with a fearful response by a cat and blickets with "Daxes" – rodents of a particular kind. The second factor was the use of a matched or different cover story for the training block of evidence, which had the same structure as the *conjunctive* condition in Experiment 1.

The first page of each survey contained the training block cover story, evidence, and ratings questions, which were identical to those used in the previous experiments. The second page contained "answers" identifying which prospective causes in the training block were blickets/daxes, in order to maximize the effect of training as predicted by the model and observed in previous experiments. The third page contained the cover story and evidence and ratings questions for the transfer block, and the fourth page contained a question about participants' beliefs about the mechanism behind the blicket-meter or dax-cat causal relationship. In this experiment the items corresponding to D, E, and F were identified as X, Y, and Z but will be referred to as D, E, and F here for the sake of clarity.

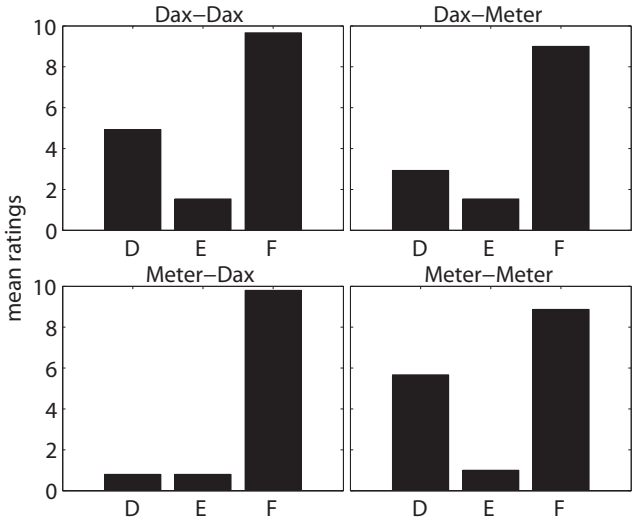
2.10.2 Results and Discussion

Mean ratings for all test objects in all four conditions are shown in Figure 2.6. The variable of interest was the rating capturing participants' beliefs that D was a cause. A two-way ANOVA (cross-domain by transfer domain) revealed a main effect of changing domains between the training and test blocks ($F(1, 56) = 12.8$, $p < 0.001$, $\eta^2 = 0.18$), a non-significant effect of domain ($F(1, 56) = 2.231$, $p = 0.141$, $\eta^2 = 0.03$), and no interaction $F(1, 56) = 0.532$, $p = 0.532$. The main effect of crossing domain, absent any interaction, indicates that participants did not blindly map the prospective causes in the second block to those in the first, and that the transfer learning effect was not merely a consequence of learning an abstract function without attaching it to a context or domain.

2.11 General Discussion

Previous work has shown that people can use and acquire knowledge of the functional form of causal relationships. We have outlined a general formal framework providing a rational analysis of these processes in terms of hierarchical Bayesian inference, and presented a series of experiments that test the predictions of this account. Experiment 1 showed that people can learn and generalize the functional form of a causal relationship when they are provided with explicit

Figure 2.6: Mean ratings of the probability that the test-block items X,Y and Z are blickets or daxes, by condition. The first label term gives the training cover story and the second term gives the test cover story.



information about causal structure. Experiment 2 showed that such inferences can be made directly from covariational evidence. Experiments 3 and 4 showed that people's inferences about functional form are appropriately sensitive to manipulations of noise and the strength of causal relationships. Experiment 5 showed that transfer of knowledge of functional form from one causal learning scenario to another was greater when those scenarios used the same causal system than when they came from quite different domains. The inferences that people made in all of these experiments were consistent with the predictions of our model, both qualitatively and quantitatively.

In this section, we turn to several important issues that have arisen at various points in the paper and deserve further discussion. First, we consider robustness of the model predictions to variation in parameter values. Second, we discuss individual differences in our data, and how these individual differences line up with model predictions. Finally, we provide a more detailed discussion of how our hierarchical Bayesian approach relates to other models of causal learning.

2.11.1 Robustness and interpretation of parameters

Our model has only two parameters – the hyperparameters λ_b and λ_g that determine the prior on functional form – and we used a single pair of values for these parameters to predict the results from all of the experiments. However, understanding the consequences of manipulating these parameters is an important part of evaluating our model. As with any model with fitted parameters, it is possible that the specific values of λ_b and λ_g we selected were crucial for the model to make accurate predictions of human judgments. Given that only two parameters were used to predict 18 distinct ratings across the experiments we presented, the concern is not with over-fitting the data, but that we need to understand what range of parameter values defines an appropriate space of theories.

To address this issue, we examined the sensitivity of the model to parameter choices by

evaluating the model's performance given all combinations of $\frac{1}{\lambda_b}$ values ranging from 0.15 to 0.5 in increments of 0.025 and $\frac{1}{\lambda_g}$ values ranging from 2.5 to 5 in increments of 0.25. We used the reciprocal of λ_b and λ_g because those quantities correspond to the mean gain and bias sampled from the resultant prior distribution. We excluded the *description-based noise* condition from this analysis because the only way to express the additional verbal information would have been to introduce a new condition-specific parameter altering the prior distribution over gain and bias. The results of this investigation are displayed in Figure 2.7, which shows the mean squared error of the model over all experiments as a function of $\frac{1}{\lambda_b}$ and $\frac{1}{\lambda_g}$. This analysis shows that it is important that the mean of the prior on the bias be low ($\frac{1}{\lambda_b} > 0.35$) but the mean of the gain distribution is not especially important.

These results are interesting not just in terms of understanding the robustness of the model predictions, but in what they reveal about the inductive biases of human learners. The range of parameter values that produce a low MSE are those that are consistent with a prior over theories that favors disjunctive relationships over conjunctive relationships. Such a prior seems appropriate for the blicketosity meter, and also fits well with the large body of work suggesting that people assume a noisy-OR relationship for many kinds of causal systems (Cheng, 1997; Griffiths & Tenenbaum, 2005).

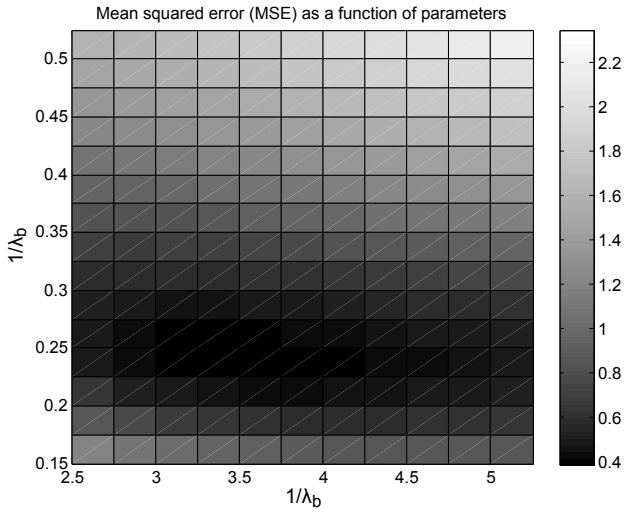


Figure 2.7: Mean squared error of the model for all conditions except verbal explicit-noise as a function of parameter values. For comparison, the lowest possible MSE for a model that does not take advantage of training-block information is 1.92.

2.11.2 Individual differences

When evaluating our model, we compared its predictions to the mean responses of our participants. While this is a common practice, it leaves open the question of whether the observed fits are artifacts of averaging different modes in the responses, each of which is poorly fit by the model (Estes, 1956; Myung, Kim, & Pitt, 2000; Navarro, Griffiths, Steyvers, & Lee, 2006). Explor-

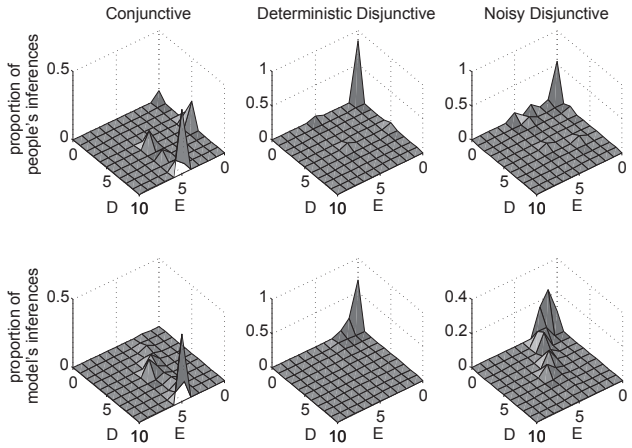


Figure 2.8: Frequencies of specific pairs of ratings for objects D and E, organized by condition. The upper row contains the model predictions and the lower row contains participants' ratings. The two distributions are generally in close concordance.

ing individual differences also gives us the opportunity to conduct a more fine-grained analysis of how well our model predicts the responses of the participants in our experiments, including information about the variability in responses as well as the mean.

To examine the correspondence between the model predictions and the data produced by individual participants, we used our model to generate predictions about the distribution of ratings in Experiment 1 and compared these predictions to the observed distribution of participants' ratings. The predictive distribution was generated in accordance with previous research indicating that individuals *probability-match*: rather than averaging over all hypotheses to produce a single judgment, they select a single hypothesis randomly with probability equal to its subjective probability of being true (Vul & Pashler, 2008). Specifically, we sampled 5000 hypotheses per condition of Experiment 1, each with probability equal to the hypothesis' probability of being true, conditioned on the training block for that condition and the common test block. Each hypothesis gives a probability that D, E, and F are blickets. We mapped these to the participants' scale by multiplying by ten and rounding to the nearest whole number and compared their distribution to participants responses for objects D and E given the same evidence, leaving out ratings for object F because they were not essential to our earlier analysis, they varied less between conditions, and they would have make visualizing the distribution more difficult. The two sets of distributions are shown in Figure 2.8, which indicates that there are no major disparities. Moreover, many of the differences can be ascribed to participants preferentially selecting values of 0, 5 and 10, which were mentioned explicitly in the task instructions.

2.11.3 Related work

In the body of the paper we discussed some well-known models of causal inference, and showed that these models are unable to explain our experimental data. The main problem with these models is that they are not designed to predict the effects of learning about functional form.

However, some more recent work bears on the kind of problem we have considered. We will briefly summarize the most salient examples and discuss the novelty of our contributions in light of them.

Alternative priors on functional forms

Much previous work has proceeded on the assumption that a single functional form is appropriate for describing all contexts in which causal learning takes place (e.g., Cheng, 1997; Novick & Cheng, 2004). If one makes this assumption, it becomes natural to ask whether there exist generic priors over kinds of function that apply across a wide range of domains. Lu, Yuille, Liljeholm, Cheng, and Holyoak (2006) make such a suggestion, albeit without appealing to hierarchically-structured knowledge or cross-context generalization, and we believe that identifying generic priors is an exciting direction for future work. We suspect, however, that such priors must be more abstract and flexible than any specific proposals to date to account for human causal inference, and note that a hierarchical model incorporating the "necessary and sufficient" priors described by Lu et al. do not appear to explain participants' capacity to make inferences consistent with expecting a conjunctive causal relationship.

To test this suspicion, we implemented a hierarchical model like our own but with a space of theories that reflected "necessary and sufficient" priors: all causal relationships were taken to have a noisy-OR form, where blickets had the same weight w and there was a latent background cause with weight w_n. The prior probability of a pair of weights was the same as given in Lu et al.: $\frac{1}{Z}[\exp^{-\alpha(w_n - w)} + \exp^{-\alpha(w - w_n)}]$, where Z is a normalizing constant. Once we optimized α, giving $\alpha = 2.40$, this model performed much better than those that assume a fixed relationship, with an MSE of 0.71 (approximately twice the MSE resulting from using our sigmoid theory space). Nonetheless, this model failed to make the key prediction that people would infer that two blickets were necessary to activate the meter in the Experiment 2 conditions, leading us to conclude that any psychologically real generic prior must be more flexible than these "necessary and sufficient" priors.

2.11.4 Causal coherence

Lien and Cheng (2000) presented a theory explaining how people might identify the level of generality at which causal inferences should be made and use that knowledge to aid later inference. For example, in learning about the effects of a set of chemicals, one might either infer a relationship between a specific chemical and an outcome, or form generalizations about the effects of particular types of chemicals. There are some similarities between the basic structure of their experimental design and our own that raise the question of how well the notion of "causal coherence" they articulated might explain our results.

The basic argument in Lien and Cheng (2000) is that people learn what level of generality is best for representing particular causal variables by selecting the best level from a set identified *a priori*, where the best level maximizes $P(e|c) - P(e|\overline{c})$, with c denoting a cause identified at that level and e the effect. Lien and Cheng illustrate this idea with an example using cigarettes and lung cancer: given enough data, one could infer that lung cancer is caused by (1) smoking particular brands of cigarettes, (2) smoking cigarettes, or (3) inhaling any kind of fumes. Option (2) is preferable to (1) under the contrast criterion because $P(e|\mathrm{brand}) \approx P(e|\mathrm{cigarettes})$ and $P(e|\overline{\mathrm{brand}}) >$

$P(e|\overline{\text{cigarettes}})$, and option (2) is preferable to (3) under the assumption that "fumes" includes substances such that $P(e|\text{fumes}) - P(e|\overline{\text{fumes}}) < P(e|\text{cigarettes}) - P(e|\overline{\text{cigarettes}})$.

While it does predict certain kinds of transfer of knowledge, such a theory cannot explain our data: the only levels of generality our participants could have identified were specific objects, "blickets" and "all objects", and inference to any of these as causes leads to the problems faced by any non-hierarchical model. A more general variation on the idea of identifying the appropriate level of abstraction leads to the argument that a learner could use pre-existing domain knowledge to identify arbitrary events as prospective causes and use the same contrast criterion to select amongst those. For instance, the "levels of abstraction" could include "one blicket", "two blickets", "one blicket and one non-blicket" and so on as possible causes. This might explain the results of Experiment 1, but more machinery is necessary to explain the fact that participants concurrently learned causal structure and functional form in Experiment 2. Moreover, this variation leads to so many possible causes that a graded measure of plausibility is necessary along with a description of how that would interact with contrast. We believe that filling these holes in a natural, parsimonious way would lead to a model indistinguishable from a hierarchical Bayesian approach.

2.11.5 Other hierarchical Bayesian models

Other hierarchical Bayesian models of causal inference have been presented recently, including one concerned with learning "causal schemata" (Kemp, Goodman, & Tenenbaum, 2007) and an account of sequential causal learning (Lu, Rojas, Beckers, & Yuille, 2008). We would like to make clear that our approach is not in competition with these. Rather, these perspectives are complementary. Kemp et al.'s contribution extends the *infinite relational model* (Kemp, Tenenbaum, Griffiths, Yamada, & Ueda, 2006) to account for learning concurrently about types and the causal relationships between them, but it makes no provision for flexibly learning the form of causal relationships. Integrating a hierarchical representation of functional form with causal schemata would provide a natural answer to the question of how people learn about functional form across diverse contexts.

Lu et al.'s (2008) sequential learning research touches on inferring the form of causal relationships, but it commits to using two explicit forms – additive (linear sum) and sub-additive (noisy-MAX) functions for continuous variables – being largely concerned with explaining the effects of presentation order. While their model cannot explain the range of phenomena we have discussed, it nicely complements our focus on functional forms for binary variables. Taken together with the results we have presented in this paper, this work suggests that a hierarchical Bayesian approach has the potential to provide a unifying framework for explaining how people learn the functional form of causal relationships.

2.11.6 Towards a more general model of functional form learning

The specific hierarchical Bayesian model we used to generate quantitative predictions, in which the set of possible functional forms is constrained to those that are consistent with the logistic function, was motivated by its simplicity, flexibility, and consistency with the cover stories that framed the evidence that participants saw in our experiments. While this model was consistent with the judgments that participants made in these experiments, it was not intended as a general account of how people learn the functional form of causal relationships. In particular, it is inconsistent with

functional forms that have previously been explored in the causal learning literature, such as the experiment by Shanks and Darby (1998) in which people learned that two causes produced an effect independently, but not when they occurred together.

The hierarchical Bayesian framework we have presented provides the basis for a more general model of functional form learning, but would need to be supplemented with a richer set of theories concerning possible functional forms. The challenge in doing so lies in defining a systematic way to specify these theories. One possibility is suggested by the recent work of Yuille and Lu (2007) , who showed that noisy-logical circuits consisting of combinations of variables interacting through noisy-OR, noisy-AND, and negation operations could approximate any discrete probability distribution. This result suggests that noisy-logical circuits might provide a reasonable foundation for characterizing a richer hypothesis space of functional forms, with the prior probability of a particular functional form depending on the complexity of its expression in terms of these operations (for examples of this kind of approach in categorization, see Feldman, 2000; N. D. Goodman, Tenenbaum, Feldman, & Griffiths, 2008).

One important step towards understanding how people learn the functional form of causal relationships more generally is identifying the prior probability assigned to different kinds of relationships. The simple model we used in this paper made it possible to draw inferences about this prior directly from people's judgments, through the parameter λ_b and λ_g. Identifying priors over richer sets of functions poses more of a challenge, and will require experiments investigating the difficulty that people encounter in learning functions of different forms. We are currently conducting experiments looking at a wider range of functional forms, and exploring the possibility of using a probabilistic logic to capture the human ability to make inferences consistent with a wide range of functional forms while still exploiting prior knowledge, as in the cases we consider here.

2.12 Conclusion

The results of our experiments show that people efficiently learn about the functional forms of causal relationships using covariation data, category information, and verbal cues, making judgments that are accurately predicted by a hierarchical Bayesian model. These results are compatible with earlier experimental results suggesting that people are sensitive to causal mechanisms and with developmental theories about domain knowledge and framework theories, but they are not predicted by most existing models of covariation-based causal inference. The Bayesian approach we have taken in this paper is able to explain not just how knowledge of causal mechanisms should influence causal inference, but how that knowledge could itself be acquired.

If human causal inference is tightly coupled to abstract knowledge, then some questions remain to be answered. How flexible is this knowledge? Do we possess general inductive mechanisms that permit us to learn a broader set of kinds of causal relationships than are common in the world given sufficient evidence, or do we operate under tight constraints? Where does abstract knowledge about categories and properties intersect with causal inference? Work along these lines is in progress, and we ultimately hope to begin to chart both the structure of adults' causal theories and the developmental trajectory of the acquisition of this knowledge.

Chapter 3

Developmental differences in causal learning

Recent work suggests that children are skilled at inferring specific causal relationships from patterns of data (Gopnik et al., 2004; Sobel et al., 2004). For example, they can infer which blocks will activate a machine based on the contingencies between the blocks and the machine's activation. But an additional question is whether children can infer more abstract causal principles from patterns in data, and use those principles to shape their subsequent predictions. For example, can a child infer that a particular type of machine activates reliably, or requires only a single cause to activate? Will those abstract discoveries bias the child's interpretations of new data?

Developmental data suggest that children do have broad inductive biases. For example, in language learning the shape bias and the mutual exclusivity principle influence more specific inferences about word meaning (Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002; Markman & Wachtel, 1988). However there is debate about whether these biases are the result of innate constraints or are themselves the product of learning (Elman et al., 1996; Leslie, 1994). Recent formal work on hierarchical Bayesian models suggests that, at least in principle, the shape bias may itself be learned as a result of normative inferences from patterns of data (Kemp, Perfors, & Tenenbaum, 2007). Similar high-level biases apply to causal learning, and we know that children can learn about causal types (L. E. Schulz, Goodman, Tenenbaum, & Jenkins, 2008), and the plausibility of cross-domain relationships (L. E. Schulz, Bonawitz, & Griffiths, 2007). In this paper, we explore whether children can learn abstract principles about the forms of causal relationships themselves.

The hierarchical Bayesian approach suggests that the nature of inductive biases may change as more evidence becomes available. Absent evidence, a learner without strong built-in biases should assign similar probabilities to a wide range of hypotheses. As data accumulate, the abstract hypotheses consistent with those data become more probable, and the learner discounts any hypotheses that fit the current data but are less compatible with past experience. If this is correct, then we might expect to see different patterns of inductive bias in adults and children. In particular, children might rely less on past experience and more on present evidence than adults. This is a possibility that has not previously been explored in the causal learning literature, and one that we examine through head-to-head comparison of children and adults in a causal learning task that requires making an abstract generalization about the nature of causal relationships.

We test the high-level generalizations made by children and adults by contrasting two

abstract "overhypotheses" (N. Goodman, 1955; Kemp, Perfors, & Tenenbaum, 2007) about how a causal system works. One is a noisy-OR model, in which each object has a certain independent probability of bringing about an effect. This model is pervasive in the literature on adult causal inference (e.g., Cheng, 1997; Griffiths & Tenenbaum, 2005), and adults appear to assume that causal relationships take a noisy-OR form unless they see evidence to the contrary (Cheng, 1997; Lucas & Griffiths, 2009). The other is an AND model in which individual causes are unable to produce an effect, but multiple causes in conjunction can produce an effect. We provided children and adults with evidence for either an AND or OR relationship and then examined how this evidence biased their judgment of a novel, ambiguous pattern of evidence. Would seeing several instances of a machine activated by a conjunction of causes lead them to assume that this would be the case for a new set of blocks? By comparing how children and adults respond to data that support these different overhypotheses, we can examine first whether children are capable of forming appropriate abstract generalizations, and second whether they are more willing to make these generalizations than adults.

The plan of the paper is as follows. First, we consider how an ideal Bayesian learner can gather evidence for overhypotheses relevant to causal induction. We then discuss the specific overhypotheses about the functional form of causal relationships that we contrast in this paper, together with a method that can be used to diagnose whether learners infer these overhypotheses from data. We go on to use this method to compare the abstract generalizations of children and adults in a causal learning task, finding support for the hypothesis that children are more willing to adopt a novel overhypothesis than adults. We close by discussing the implications of these results.

3.1 Causal overhypotheses

Children can identify causes using only a handful of observations (Gopnik et al., 2004), but the extent to which they learn about the abstract properties of causal relationships remains largely unexplored. From a Bayesian standpoint, learning about causal structure requires having *a priori* beliefs – expressed in a prior probability distribution over hypotheses – about what items are plausible causes, and expectations about how a given causal structure leads to different observable events. These expectations can be expressed formally using a likelihood function, which specifies the probability of observing a particular set of events based on the underlying causal structure.

Most work on probabilistic models of causal learning has assumed a specific kind of likelihood function. This likelihood function is based on causes and effects interacting in a "noisy-OR" manner, each having an independent opportunity to produce the effect (Cheng, 1997; Griffiths & Tenenbaum, 2005; Glymour, 1998). More precisely, a noisy-OR relationship implies that the probability that an effect E occurs given the presence of a set of causes $C_1, \ldots, C_N$ is

$$P(E|C_1, ..., C_N) = 1 - \prod_{i=1}^{N}(1 - w_i) \tag{3.1}$$

where w_i is the probability that C_i generates the effect in the absence of other causes.

Despite the popularity of the noisy-OR in models of causal learning, other kinds of causal relationships are clearly possible. For instance, a noisy-OR model cannot describe an AND relationship, where an effect only occurs when multiple causes are present. This might be the case in

an electrical circuit where multiple switches are wired in series, and a light only turns on when all of the switches are flipped. It is important, then, for models of causal inference to accommodate flexible beliefs about the forms relationships can take. Formalizing inferences about the form of a relationship is straightforward, using an expanded likelihood function, $P(E|C_1, ..., C_N, F)$, where F captures information about the form of the causal relationship. For example, F could indicate that the relationship has a noisy-OR form, but another value of F might indicate that a causal relationship has an AND form.

Learning the form of a causal relationship and generalizing that discovery when reasoning about other causal relationships requires inference at multiple levels of abstraction. This kind of inference, in which lessons from one context can be carried forward for future learning, is easily captured by using a hierarchical Bayesian model (Tenenbaum, Griffiths, & Kemp, 2006; Kemp, Perfors, & Tenenbaum, 2007). A learner's abstract beliefs, or overhypotheses, determine the probabilities of more-concrete hypotheses, each encoding specific causal structures and the form a relationship takes. These hypotheses, in turn, determine the likelihood of different patterns of events.

Formally, we can imagine an inference involving variables at three levels: the observed data D, hypotheses about the causal structure underlying those data H, and overhypotheses (or a "theory", as in Griffiths & Tenenbaum, 2009) T representing generalizations relevant to evaluating those hypotheses (see Figure 3.1). Bayes' rule then specifies how the events a learner sees (D) should change the learner's beliefs, both about the casual system at hand (H), and about the higher-level properties of that kind of system (T). Formally, we have

$$p(T|D) = \frac{p(D|T)p(T)}{p(D)} \tag{3.2}$$

where $p(T)$ is the prior probability of the overhypothesis T, $p(T|D)$ is the posterior probability, and $p(D)$ is obtained by summing the numerator over all overhypotheses T. The probability of the data given an overhypothesis is obtained by summing over all hypotheses consistent with that overhypothesis,

$$p(D|T) = \int p(D|H)p(H|T)\, dH, \tag{3.3}$$

and can be interpreted as an average of the probability of the observed data under those hypotheses weighted by the extent to which each hypothesis is consistent with the overhypothesis.

Intuitively, this hierarchical Bayesian approach provides a way to explain how learners can form and use abstract generalizations about causal systems. For example, if a child sees events that are likely under an AND relationship, such as a machine activating only when pairs of causal objects are placed on it, then the probability of an overhypothesis predicting future AND relationships increases. This is because the best hypotheses for explaining the observed events are those that are most likely under this overhypothesis, so Equation 3.3 yields a high value. Incorporating this value into Equation 2, the posterior probability for that overhypothesis will increase.

As the evidence supporting a particular overhypothesis increases, it will be easier to learn about the structure and form of causal systems that are consistent with that overhypothesis. This comes with a cost: if a causal system has strange or rare abstract properties, such as an unlikely functional form, much more evidence will be necessary to learn about it. The implication is that adults, who have seen a great deal of evidence, should find it very easy to learn about the structure and form of causal relationships that have typical properties. Conversely, children, with their lim-

T : Theory/overhypothesis

H : Hypothesis

D : Data

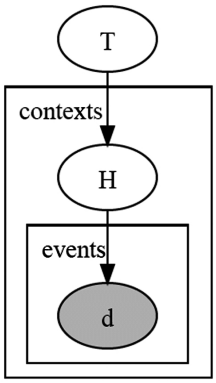


Figure 3.1: The structure of a hierarchical Bayesian model. Theories or overhypotheses (T) include abstract knowledge that may apply to a wide range of circumstances, and they determine what context-specific hypotheses (H), such as beliefs about what variables are the causes of an observed effect, are possible and likely. These hypotheses are in turn used to explain and generalize from a specific set of data (D), including the events the learner observes.

ited experience, should be more sensitive to evidence when learning about relationships that have unusual properties. In the following section, we discuss an experimental design for testing this idea.

3.2 The functional form of causal relationships

If children update their abstract beliefs about causal systems in a manner consistent with Bayesian inference, then the events they see should influence their judgments about different sets of events involving novel causes. To test this hypothesis, we used an experiment with two phases, each with a distinct set of objects. In the first phase, children saw a set of events designed to be likely under one of two abstract overhypotheses about the forms of causal relationships. In the second phase, they saw events where different beliefs about the form of the causal relationship should lead them to make different judgments about which objects are causes. The specific task was to identify the "blickets" within two sets of objects, knowing only that blickets have "blicketness". Prospective blickets could be placed on a "blicketness machine", causing it to either activate by lighting up and playing music or do nothing. Children might entertain a variety of expectations about the relationship between the blickets and the machine, determining how they interpret different events. For example, if they think that two blickets are necessary to activate the machine, then seeing a single object fail to activate it provides no information. At the same time, their expectations about the form of the relationship between blickets and the blicketness machine can be shaped by the events they observe. For instance, seeing two objects fail to activate the machine separately but succeed together suggests that two blickets are necessary for activation.

Participants saw one of two different sets of training events, each designed to lead them to believe that a particular relationship held between blickets' presence and the machine's activation. In the *AND* condition, participants saw a training block of events where three distinct objects, A, B, and C (e.g., a triangle, a circle, and a cube) were placed sequentially on the machine, which failed to activate in all cases. Next, all pairs of the objects were placed on the machine sequentially, with

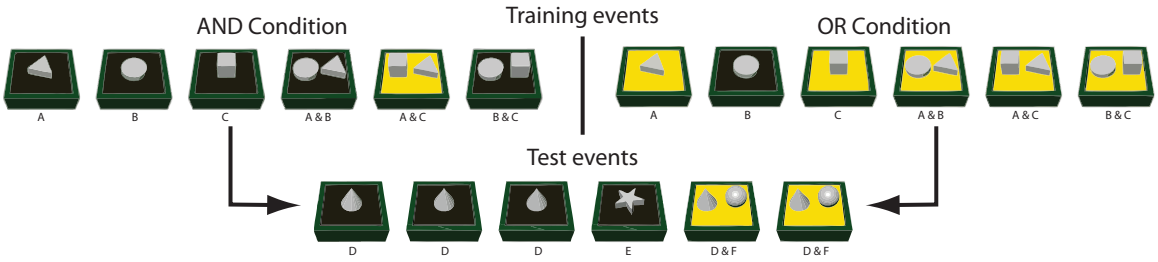


Figure 3.2: Evidence presented to participants in the two training phases, as well as the subsequent test phase which all participants saw. Events are given as a set of prospective causes and the presence or absence of an effect. The bright-paneled machines represent events in which the effect occurs and the dark-paneled machines represent events in which the effect does not occur.

only one pair causing activation. See Figure 2 for a summary of the events in the training and test blocks. Participants were then asked to say which of the objects were blickets.

After making these judgments, participants saw three new objects, D, E, and F (e.g., a cone, star, and a ball) which they had never seen before, and a series of test events intended to be ambiguous, leading to different judgments about which of the new objects were blickets, depending on participants' expectations about the form of the relationship. If children expect that a single blicket suffices to activate the machine, they should believe that F is likely to be a blicket, while D and E are not. If, in contrast, children exploit the information provided by the training block so they conclude that two blickets are necessary to activate the machine, then they should think that D and F are blickets, and be uncertain about E.

Lucas and Griffiths (2009) used a similar design with adults, showing that their inferences about causal structure are driven by their beliefs about the probable forms of causal relationships, which are in turn influenced by events they have seen in the past. The specific pattern of judgments they found was consistent with the predictions of a hierarchical Bayesian model given priors reflecting a strong bias in favor of disjunctive (OR) and deterministic relationships. Such priors are also consistent with adults' performance in other experiments (Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2006). This prior could be chiefly due to adults' experiences revealing that OR relationships are more common, or an innate bias. By comparing the judgments of 4-year-old children to those of adults, we aim to answer that question and better understand the origins of the abstract knowledge that drives efficient causal inference.

3.3 Comparing the prior beliefs of children and adults

Our aim was to answer two questions about the use of causal overhypotheses by children and adults. The first question was whether children, like adults, can use events to update their knowledge about the likely forms of causal relationships, and apply that knowledge to learn the causal structure behind new and ambiguous sets of events. The second question was whether children are more or less sensitive to evidence supporting such high-level generalizations, as opposed to their prior beliefs.

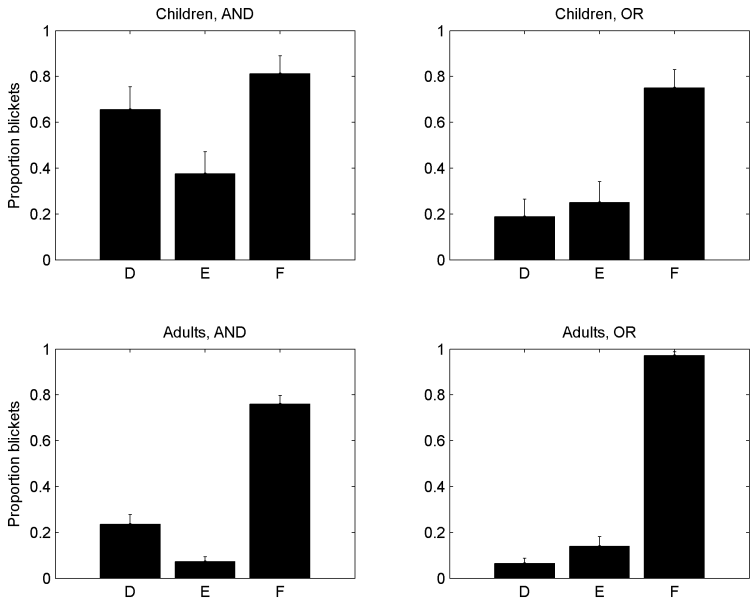


Figure 3.3: Proportions of objects that were judged to be blickets for children (top row) and adults (bottom row) for the *AND* (left column) and *OR* (right column) conditions. Error bars represent standard error of the mean.

If children are more likely than adults to call objects D and E blickets in the *AND* condition, we can conclude that the strong adults bias adults show – toward expecting deterministic relationships where individual causes are sufficient – is due in large part to learning during and after childhood, including, for instance, experience with machines to which OR relationships apply. If children's judgments are indistinguishable from those of adults, we have evidence that learning about the forms of causal relationships occurs early, or plays a minor role in driving our expectations. Finally, if there is no effect of training evidence on children's test-block judgments, we should question the applicability of the model used by Lucas and Griffiths (2009) to causal inference in children, and infer that the ability to form certain overhypotheses is itself a consequence of late-childhood learning or development.

3.3.1 Participants

Children

Thirty-two children were recruited from university-affiliated preschools, divided evenly between the *AND* and *OR* conditions. Children in the *AND* and *OR* conditions had mean ages of 4.46 (SD=0.27) and 4.61 (SD=0.31) years, respectively.

Adults

University of California, Berkeley undergraduates received course credit for participating during lectures of an introductory psychology course. There were 88 participants in the *AND* condi-

tion and 55 in the *OR* condition. Five participants in the *AND* condition were excluded for declining to answer one or more questions.

3.3.2 Methods

Children

Each child sat at a table facing the experimenter, who brought out three gray ceramic objects, each with a different shape, as well as a green box with a translucent panel on top, describing the box as "my blicketness machine".

At the beginning of the experiment, children were prompted to help the experimenter name the objects using their shapes, e.g., "triangle". They were then told that the goal of the game was to figure out which of the objects were blickets, that blickets have blicketness inside them, and blickets cannot be distinguished from non-blickets by their appearance. No other information was provided about the relationship between blickets and the activation of the machine.

The children then observed a set of training events in which the experimenter placed objects alone or in pairs on the machine, which activated in some cases by lighting up and playing music. These events corresponded to either the *OR* condition or *AND* condition training given in Figure 2. After the children saw these events, they were asked whether each object was a blicket or not. Next, the experimenter brought out three objects that the children had not seen before. After the children named the new objects, the experimenter demonstrated the test events listed in Figure 2 and asked whether each of these new objects was a blicket or not. The experiment was repeated a second time for each child, using the same patterns of evidence, but with a distinct set of objects that varied color and had a uniform shape. The identities of the individual objects were counterbalanced, as was whether the children saw the different-shaped or different-colored objects first.

Adults

The adults were tested in groups, and saw demonstrations that were almost identical to what the children saw in the corresponding conditions. Unlike the children, the adults were not asked to name the objects, and they recorded their judgments on sheets of paper rather than responding verbally.

3.3.3 Results

Children

The critical prediction was that children would be more likely to judge object D to be a blicket in the *AND* condition than in the *OR* condition, indicating that they were (1) learning about the form of the relationship between blickets and the machine's activation, and (2) transferring that abstract knowledge to make better inferences about novel objects and otherwise ambiguous events.

Children were more likely to judge object D to be a blicket in the *AND* condition than in the *OR* condition ($p < 0.005$, two-tailed permutation test). There was also a change in the predicted direction for object E, albeit non-significant.

Adults

Adults were also more likely to judge object D to be a blicket in the *AND* condition than in the *OR* condition ($p < 0.005$, two-tailed permutation test), consistent with the results in Lucas and Griffiths (2009). See Figure 3, bottom row, for a summary of their judgments for the test objects.

Differences

In the *AND* condition, the adults judged object D to be a blicket less frequently than children ($p < 0.005$, Fisher's exact test). See Figure 3.3 for a summary of ratings in the three conditions. Children's ratings were also higher for object E ($p < 0.001$, two-sided permutation test), which is consistent with their being quicker to learn that an AND relationship applies: under an AND relationship, the event where E fails to activate the machine is uninformative, so judgments of E being a blicket should reflect the base rate of blickets occurring. The high frequency of other objects being blickets under an AND relationship (4 of 5), plus a belief that blickets are not rare, should lead a learner to expect that a novel object is somewhat likely to be a blicket.

We also applied the hierarchical Bayesian model developed in Lucas and Griffiths (2009) to predict the children's judgments, finding that it yielded accurate predictions when given priors that reflect weak biases about the probably forms of causal relationships. For the specifics of the model, see the previous chapter.

Model fits

We converted children's judgments about blickets to probabilities in order to examine them using the previously-mentioned hierarchical Bayesian model and sigmoid space of hypotheses. We treated is-a-blicket judgments as assertions that objects were definitely blickets, and not-a-blicket judgments as assertions that objects were definitely not blickets. Lucas and Griffiths (2009) found that priors favoring disjunctive, deterministic relationships – predicting a mean gain of 3.34 and a mean bias of 0.23 – fit adults' judgments closely, with a mean squared error of 0.29 per judgment on a zero to ten scale. We found that similar priors best captured adults' judgments in our experiment, giving a mean squared error of 0.80 with a mean gain of 5.30 and bias of 0.11.

These same priors were wildly inconsistent with children's inferences, giving a mean squared error of 6.12. In contrast, priors giving a mean *a priori* gain and bias of 1 – favoring neither AND nor OR relationships – were much more accurate, with a mean squared error of 0.58. The priors that best fit the children's judgments gave a mean gain and bias of 1.45 and 0.85, respectively, with mean squared error of 0.15.

3.4 Discussion

Our experiment was designed to explore two questions: whether children could make high-level generalizations about the form of causal relationships, and whether they were more willing to do so than adults. Our results show that children are capable of making such inferences, and that their judgments were more strongly influenced by the available evidence than adults, whose inferences reflected a bias toward OR relationships. Our results thus support the view that when learning about cause and effect, children are flexible learners whose inexperience may sometimes

let them learn better from sparse evidence, especially in novel situations. These results are also consistent with treating the acquisition and application of causal knowledge as a matter of hierarchical Bayesian inference, where a learner has beliefs expressed at multiple levels of abstraction, with abstract theories driving specific hypotheses which, in turn, enable prediction and categorization.

Before closing, we will address two alternative explanations for our results. The first is that children are more likely than adults to judge any object to be a blicket. This is less consistent with the data than our interpretation, given that adults were more likely than children to call object F a blicket in the *OR* condition, and nearly as likely in the *AND* condition (75 percent of the objects versus 81 percent). A second alternative is that the children were confused by the training data in the *AND* condition, and responded to the novel objects by guessing randomly. This explanation can be ruled out by noting that children judged objects D and F to be blickets more often than chance would predict ($t(15) = 3.529$, $p < 0.005$).

The results of our experiment have implications for understanding causal learning, and for understanding cognitive development more generally. In terms of causal learning, these results suggest that the fundamental biases that lie beneath causal inference are more subtle and abstract than *a priori* preferences for specific kinds of causal relationships. We believe that trying to understand these biases is fertile ground for future research. For cognitive development, the idea that children are more flexible in their commitments about the way that causal systems tend to work seems like not just a necessary consequence of a hierarchical Bayesian approach, but an important insight for understanding how it is that children see the world differently from adults. The plasticity of beliefs that this implies helps to explain the bold exploration and breathtaking innovation that characterizes children's interactions with the world.

Chapter 4

Mental causation and abstract knowledge: learning about preferences

Economists and computer scientists are often concerned with inferring people's preferences from their choices, developing econometric methods (e.g., McFadden & Train, 2000; Boyd & Mellman, 1980) and collaborative filtering algorithms (e.g., Goldberg, Nichols, Oki, & Terry, 1992; Konstan et al., 1997; Breese, Heckerman, & Kadie, 1998) that allow them to assess the subjective value of an item or determine which other items a person might like. Identifying the preferences of others is also a key part of social cognitive development, allowing children to understand how people act and what they want. Young children are thus often in the same position as economists and computer scientists, trying to infer the nebulous preferences of the people around them from the choices they make. In this paper, we explore whether the inferences that children draw about preferences can be explained within the same kind of formalism as that used in economics and computer science, testing the hypothesis that children are making rational inferences from the limited data available.

Before about 18 months of age, children seem to assume that everyone likes the same things that they do, having difficulty understanding the subjective nature of preferences (Repacholi & Gopnik, 1997). However, shortly after coming to recognize that different agents can maintain different preferences, children demonstrate a remarkably sophisticated ability to draw conclusions about the preferences of others from their behavior. For example, two-year-olds seem to be capable of using shared preferences between an agent and themselves as the basis for generalization of other preferences (Fawcett & Markson, 2010), while three- and four-year-olds can use statistical information to reason about preferences, inferring a preference for an object when an agent chooses the object more often than expected by chance (Kushnir, Xu, & Wellman, in press).

This literature in developmental psychology is paralleled by work in econometrics on statistical models for inferring preferences from choices. In this paper, we focus on an approach that grew out of the Nobel prize-winning work of McFadden (see McFadden & Train, 2000 for a review), exploring a class of models known as mixed multinomial logit models (Boyd & Mellman, 1980). These models assume that agents assign some utility to every option in a choice, and choose in a way that is stochastically related to these utilities. By observing the choices people make, we can recover their utilities by applying statistical inference, providing a simple rational standard against which the inferences of children can be compared.

Economists have also explored models for the choices of multiple agents, using hierarchical Bayesian statistics (Train, McFadden, & Ben-Akiva, 1987). These models combine information across agents to make inferences about the properties or values of different options. Similarly, research on preferences in computer science has tended to focus on predicting which options people will like based not just on their own previous choice patterns but also drawing on the choices of other people – a problem known as *collaborative filtering* (Goldberg et al., 1992). This work has led to the development of the now-ubiquitous recommendation systems that suggest which items one might like to purchase based on previous purchases, and has reached notoriety through the recent Netflix challenge (Greene, 2006).

Our contribution in this paper is to bring together these different threads of research to develop rational models of children's inferences about preferences. The following section summarizes developmental work on children's inferences about preferences. We then outline the basic idea behind rational choice models, drawing on previous work in economics and computer science. We go on to show how these models can be used to explain developmental data, first being concerned with inferring preferences from choices, next focusing on using preference information to learn about the properties of hidden objects, and finally addressing the developmental course of preference understanding. We conclude by discussing the implications of our results.

4.1 Children's inferences about preferences

The ability to learn about others' preferences is vitally important: knowing others' preferences enables us to reason about their desires and goals, explain and predict their actions, and cooperate and compete efficiently in social settings. Despite the importance of preference learning, we know little about how children begin to learn about what other people like and dislike, and less about how children come to understand that people have distinct preferences in the first place. This section will describe three past studies. The first reveals the developmental course of preference understanding, the second explores the kinds of statistical evidence that children use in inferring preferences, and the third examines how children make generalizations from consistent patterns of choices.

4.1.1 Recognizing that people have distinct preferences

In order to learn about people's preferences, children must first understand that others have preferences at all, which occurs when they are about 18 months old. Before that age, children do not distinguish the preferences of others from their own. Specifically, Repacholi and Gopnik (1997) showed that 14-month-old children tend to offer other people the items that they themselves prefer rather than the items that those people have previously chosen, while 18-month-old children tend to make offers that reflect the past choices of the offer's recipient.

Repacholi and Gopnik's (1997) study included two experimental conditions. In the *unmatched* condition, each child saw an actor express pleasure after tasting raw broccoli (which the children tended to dislike) and disgust after eating goldfish crackers (which the children tended to like). In the *matched* condition, the actor's pattern of reactions was reversed, matching the child's own. After presenting these reactions, the actor prompted the child to offer a food item by asking "Can you give me some?" and holding out a hand.

Table 4.1: Summary of results from Repacholi and Gopnik (1997), showing number of children in each condition (matched/unmatched) who offered broccoli, cracker, or neither.

	Unmatched				Matched		
Age	**Cracker**	**Broccoli**	**Neither**	**Age**	**Cracker**	**Broccoli**	**Neither**
14-mo	7	1	34	14-mo	13	5	21
18-mo	8	18	15	18-mo	22	7	8

The results of Repacholi and Gopnik (1997) are summarized in Table 4.1. In the *unmatched* condition, almost none (12.5 percent) of the younger children's offers matched the actor's previous choice of broccoli, while 69 percent of the older children's offers were broccoli. In the *matched* condition where the actor chose the cracker, roughly equal proportions of offers by younger and older matched the actor's choice (72 percent and 75.9 percent, respectively). These results suggest that an important transition takes place between 14 and 18 months, with children coming to recognize that other people may not share their preferences.

4.1.2 Learning preferences from statistical evidence

While 18-month-olds are able to infer preferences from affective responses, we often need to make inferences from more limited data, such as the patterns of choices that people make when faced with various options. Recent work by Kushnir and colleagues (Kushnir et al., in press) provided the first evidence that preschoolers can use statistical sampling information as the basis for inferring an agent's preference for toys.[1] Three groups of children were tested in a simple task. Each child was shown a big box of toys. For the first group, the box was filled with just one type of toy (e.g., red discs). For the second group, the box was filled with two types of toys in equal proportions (e.g., 50% red discs and 50% blue plastic flowers). For the third group, the box was filled with two types of toys in different proportions (e.g., 18% red discs and 82% blue plastic flowers). A puppet named Squirrel came in to play a game with the child. Squirrel looked into the box and picked out five toys. The sample consisted of five red discs for all three conditions. Then the child was given three toys – a red disc (the target), a blue plastic flower (the alternative), and a yellow cylinder (the distractor) – and was asked to give Squirrel the one he liked. Each child received two trials with different objects. The results of the experiment showed that the children chose the target (the red disc) 0.96, 1.29, and 1.67 times (out of 2) in the 100%, 50%, and 18% conditions, respectively, suggesting that children used the non-random sampling behavior of Squirrel as the basis for inferring his preferences.

4.1.3 Generalizing from shared preferences

Recognizing that preferences can vary from one agent to another also establishes an opportunity to discover that those preferences can differ in the degree to which they are related to one's own. Fawcett and Markson (2010) asked under what conditions children would use shared

[1] Kushnir et al. also showed that 20-month-old infants make similar inferences, but we will focus on their data from preschoolers, which involved a wider range of conditions and a simpler outcome measure.

preferences between themselves and another agent as the basis for generalization, using a task similar to the "collaborative filtering" problem explored in computer science. Their experiments began with four blocks of training involving two actors. In each block the actors introduced two objects from a common category, including toys, television shows and foods. The actors displayed opposite preferences from each other, with each actor expressing liking for one object and disliking for the other object. One actor had preferences that were matched to the child's in all blocks – her objects had features chosen to be more interesting to the child. After each actor reacted to the objects, the child was given an opportunity to play with the objects, and his or her preference for one object over the other was judged by independent coders, based on relative interest in and play with each object.

After the training blocks, the first test block began. Each actor brought out a new object that was described as being in the same category as the training objects, but was hidden from the child by an opaque container. Each actor then reacted to her novel object in a manner that varied by condition. In the *like* condition, the actor's reaction was to view the object and describe it as her favorite object of the category. In the *dislike* condition, she viewed the object and expressed dislike. In the *indifferent* condition the actor did not see the toy, and professed ignorance about it. The child was then given an opportunity to choose one hidden object to play with. Finally, a second test block began, identical to the first except that the hidden objects were members of a different category from those seen in training. In Experiment 1, members of the new category could be taken to share features with members of the training category, e.g., toys versus books, while in Experiment 2 the new category was chosen to minimize such overlap, e.g., food versus television shows. Children consistently chose the test items that were favored by the agent who shared their own preferences during training, for both toys and the similar category, books. In contrast, when a highly distant category was used during test, children did not show any systematic generalization behaviors. These results suggest that children use shared preferences as the basis for generalization, but they also take into account whether the categories are related or not.

4.2 A rational model connecting choice and preference

In the tradition of rational analysis (Anderson, 1990), we now consider the problem of how a child might *optimally* infer people's preferences from their choices. A first step is positing a specific relationship between people's preferences and their choices. We can then determine how an agent would make optimal inferences from others' behavior given knowledge of this relationship. Fortunately, the relationship between preferences and choices has been the subject of extensive research in economics and psychology.

One of the most basic models of choice behavior is the exponentiated Luce-Shepard choice rule (Luce, 1959; Shepard, 1957), which asserts that when presented with a set of J options with utilities $\mathbf{u} = (u_1, \ldots, u_J)$, people will choose option i with probability proportional to e^{u_i}, with

$$P(c = i|\mathbf{u}) = \frac{\exp(u_i)}{\sum_j \exp(u_j)}, \tag{4.1}$$

where j ranges over the agent's options. Given this choice rule, learning about an agent's preferences is a matter of applying Bayes' rule. Specifically, given an observed sequence of choices

$\mathbf{c} = (c_1, \ldots, c_N)$, the posterior distribution over the utilities of the options is:

$$p(\mathbf{u}|\mathbf{c}) = \frac{P(\mathbf{c}|\mathbf{u})p(\mathbf{u})}{\int P(\mathbf{c}|\mathbf{u})p(\mathbf{u})\, d\mathbf{u}}, \tag{4.2}$$

where $p(\mathbf{u})$ expresses the prior probability of a vector of utilities $\mathbf{u}$. The likelihood $P(\mathbf{c}|\mathbf{u})$ is obtained by assuming that the choices are independent given $\mathbf{u}$, being the product of the probabilities of the individual choices given by Equation 4.1.

While this model can capture preferences among a specific set of objects, it is often important to be able to predict the choices that agents will make about novel objects. This can be done by assuming that options have features that determine their utility, with the utility of option i being the sum of the utilities of its features. If we let $\mathbf{x}_i$ be a binary vector indicating whether an option possesses each of a set of features, and $\boldsymbol{\beta}_a$ be the utility that agent a assigns to those features,[2] we can express the utility of option i for agent a as the inner product of these vectors. The probability of agent a choosing option i is then

$$P(c = i|\mathbf{X}, \boldsymbol{\beta}_a) = \frac{\exp(\boldsymbol{\beta}_a^T \mathbf{x}_i)}{\sum_j \exp(\boldsymbol{\beta}_a^T \mathbf{x}_j)}, \tag{4.3}$$

where $\mathbf{X}$ represents $\mathbf{x}_1, ..., \mathbf{x}_J$, the features of all of the options . We can also integrate out $\boldsymbol{\beta}_a$ to obtain the choice probabilities given just the features of the options, with

$$P(c = j|\mathbf{X}) = \int \frac{\exp(\boldsymbol{\beta}_a^T \mathbf{x}_i)}{\sum_j \exp(\boldsymbol{\beta}_a^T \mathbf{x}_j)} p(\boldsymbol{\beta_a})\, d\boldsymbol{\beta}_a. \tag{4.4}$$

The usual choice of prior for feature utilities $p(\boldsymbol{\beta_a})$ is a normal distribution with a mean of zero, consistent with the view that, in the absence of information, a feature is unlikely to be very desirable or undesirable. This combination of prior and likelihood function corresponds to the mixed multinomial logit model (MML; Boyd & Mellman, 1980), which has been used for several decades in econometrics to model discrete-choice preferences in populations of consumers. A more detailed description of the MML is provided in the Appendix.

The model outlined in this section provides a way to optimally answer the question of how to infer the preferences of an agent from his choices. In the remainder of the paper, we explore how well this simple rational model accounts for the inferences that children make about preferences, applying the model to the key developmental phenomena introduced in the previous section. We will address Kushnir et al.'s result concerning statistical evidence, followed by the preference generalization results found by Fawcett and Markson, and then return to the developmental difference discovered by Repacholi and Gopnik.

4.3 Using statistical information to infer preferences

The experiment conducted by Kushnir and colleagues (Kushnir et al., in press), discussed above, provides evidence that children are sensitive to statistical information when inferring the

[2] We will refer to the utilities of features as "preferences" to distinguish them from the utilities of options.

preferences of agents. In this section, we examine whether this inference is consistent with the predictions of the rational model outlined above.

4.3.1 Applying the MML model

We will briefly describe the intuitions behind the model's predictions for Kushnir et al.'s experiment, providing further technical details in the appendix. Recall that in Kushnir et al.'s experiment, children were asked to pick out the toy that Squirrel prefers, having observed Squirrel choose to play five times with a target object such as a red circle, from a pool of objects that included instances of the target object and an alternate object such as blue flowers. We can decompose this task into learning about Squirrel's preferences, and using that knowledge to offer an object. Squirrel's choices reveal his preferences via their likelihoods: if his choices are much more likely given a strong preference for the target object, then a strong preference is more probable, via Bayes' rule

$$p(\boldsymbol{\beta}_a|\mathbf{c}, \mathbf{X}) \propto P(\mathbf{c}|\boldsymbol{\beta}_a, \mathbf{X})p(\boldsymbol{\beta}_a). \tag{4.5}$$

In the 100% condition, the target object constitutes all of Squirrel's options, so Squirrel's preferences do not determine the likelihood of the choices and no conclusions can be drawn from the choices the child sees. In the 50% condition, the pattern of choices is more likely given a preference for the target object, and strong preferences are especially likely in the 18% condition.

Having learned about Squirrel's preferences, the child must now select an object to give Squirrel, from a set consisting of one target object, one alternative object that was among Squirrel's options in the 50% and 18% conditions, and one novel distractor object. If we suppose each child is choosing as Squirrel would, we can use the Luce-Shepard choice rule (Equation 4.3) to predict the rates at which children should choose the different objects for a particular set of preference values, and average over preference values to predict how often they should choose each item.

4.3.2 Results

Figure 4.1 (a) compares the predictions of the model to the children's choice probabilities. The sum squared error (SSE) of the predictions was 0.0758 when compared with the observed probabilities of selecting the target object, with the correlation of the model predictions and observed data being $r = 0.93$. Figure 4.1 (b) shows that the goodness of fit is generally insensitive to the assumed variance of other people's preferences, σ^2, provided $\sigma^2 > 1$. This is essentially the only free parameter of the model, indicating that there is a close correspondence between the predictions of the rational model and the inferences of the children under a variety of reasonable assumptions about the distribution of preferences.

The only conspicuous difference between the model's predictions and the children's choices was the tendency of children to choose the target object more frequently than alternatives even in the 100% condition. This can be explained by observing that Squirrel was freely choosing to select objects from the box, implicitly indicating that he was choosing these objects over other unobserved options. As a simple test of this explanation, we generated new predictions under the assumption that each choice included one other unobserved option, with features orthogonal to the choices in the box. This improved the fit of the model, resulting in an SSE of 0.05 and a correlation of 0.95.

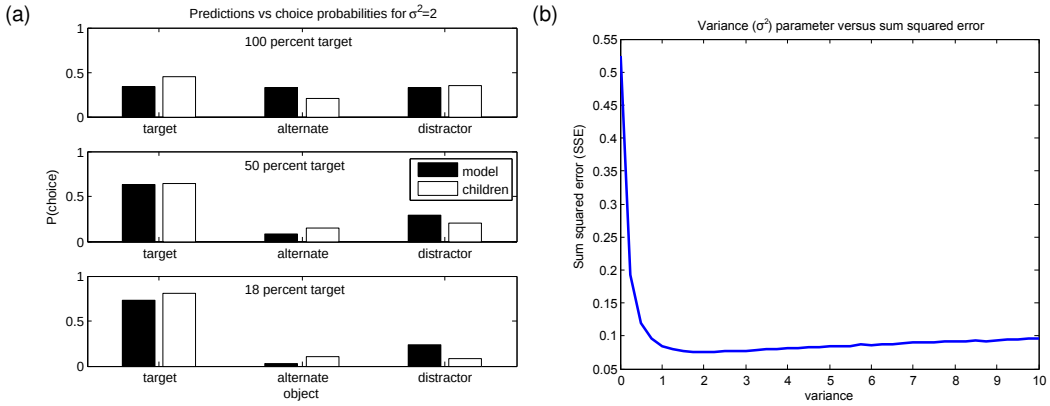


Figure 4.1: Model predictions for data in Kushnir, Xu, and Wellman (in press). (a) Predicted probability that objects will be selected, plotted against observed proportions. (b) Sensitivity of model to setting of variance parameter.

4.4 Generalizing preferences to novel objects

The study of Fawcett and Markson (2010) goes beyond simple estimation of preferences from choices, exploring how children solve the "collaborative filtering" problem of generalizing preferences to novel objects. In this section we will outline how this can be captured using the MML model.

4.4.1 Applying the MML model

Generalization in this task requires two kinds of inferences. The first inference the child must make – learning the actors' preferences by computing $p(\boldsymbol{\beta}_a|\mathbf{X}, \mathbf{c})$ for $a = 1$ (for Actor 1, whose preferences match the child's) and $a = 2$ (for Actor 2, whose preferences differ) – is the same as that necessary for the first set of experiments discussed above. The second inference is estimating the two hidden objects' features via those preferences. Because the actors reacted to the novel objects rather than making choices involving them, we need to modify the model slightly to allow us to assign likelihoods to the positive and negative reactions the actors presented.

Let the features of Actor a's preferred object in round n be $\mathbf{x}_{na}$, the features of the same-category novel object be $\mathbf{x}_{sa}$, and the features of the different-category novel object be $\mathbf{x}_{da}$. When the category is irrelevant, we will use $\mathbf{x}_{*a} \in \{\mathbf{x}_{sa}, \mathbf{x}_{da}\}$ to indicate the features of the novel object. The child's goal is to infer what features $\mathbf{x}_{*a}$ a novel object is likely to have. That knowledge permits her to estimate how much she will like each of the two novel objects, and pick the better one. The available information includes the observed affective response of agent a, the features of the objects from the previous rounds $\mathbf{X}$, and the choices of the agent on the previous rounds $\mathbf{c}$. Under our model, the distribution of a novel object's features given this information is

$$P(\mathbf{x}_{*a}|\mathbf{X}, \mathbf{c}, r_a) = \int P(\mathbf{x}_{*a}|\boldsymbol{\beta}_a, r_a) p(\boldsymbol{\beta}_a|\mathbf{X}, \mathbf{c})\, d\beta_a, \tag{4.6}$$

where $P(\mathbf{x}_{*a}|\boldsymbol{\beta}_a, r_a)$ is the posterior distribution over the features of the novel object given the preferences and affective response of the agent. Computing this distribution requires defining a likelihood $P(r_a|\mathbf{x}_{*a}, \boldsymbol{\beta}_a)$ and a prior on features $P(\mathbf{x}_{*a})$. We deal with these problems in turn.

The likelihood $P(r_a|\mathbf{x}_{*a}, \boldsymbol{\beta}_a)$ reflects the probability of the agent producing a particular affective response given the properties of the object and the agent's preferences. In the experiment, the affective responses produced by the actors were of two types. In the *like* condition, the actor declared that the object was her favorite of the type. If one takes the actor's statement at face value and supposes that the actor has encountered an arbitrarily large number of such objects, then this favorite object has the most desirable possible combination of features, given its category, possessing all category-appropriate features with positive utility, and no features with negative utility:

$$P(\mathbf{x}_{*a}|\boldsymbol{\beta}_a, r_a = \text{“like”}) = 1 \text{ for } \mathbf{x}_{*a} = \arg\max_{\mathbf{x}_{*a}} \boldsymbol{\beta}_a^T \mathbf{x}_{*a}, \text{else } 0. \tag{4.7}$$

In the *dislike* condition, the action – saying “there's a toy in here, but I don't like it”– communicates negative utility, or at least utility below some threshold we fix at zero, so the object could be any member of the category with negative utility:

$$P(\mathbf{x}_{*a}|\boldsymbol{\beta}_a, r_a = \text{“dislike”}) = 1 \text{ for } \boldsymbol{\beta}_a^T \mathbf{x}_{*a} < 0, \text{else } 0. \tag{4.8}$$

In defining the prior distributions from which the features of both the observed objects and the novel objects are sampled, it is important to represent differences between categories. Our feature vectors were concatenations of category-1 features, category-2 features, and multiple-category features, where each feature was present with probability 0.5 if its category could possess the feature, otherwise zero. We arbitrarily chose four features per category, for a total of twelve.

Having computed a posterior distribution over $\mathbf{x}_{*a}$ using this prior and likelihood, the child must combine this information with his or her own preferences to select an object. Unfortunately, we do not have direct access to the utilities of the children in this experiment, so we must estimate them from the children's choice data. We did this using the same procedure as for the adults' utilities. That done, we apply the choice rule a final time and obtain choice probabilities for the novel objects selected by the children.

4.4.2 Results

Figure 4.2 shows the rates at which children chose Actor 1's object and the model's predictions. With the variance parameter set to $\sigma^2 = 2$ and twelve features, the correlation between the predictions and the data was 0.88. When examining only the 28-month-old children ($N = 68$) in the study, the correlation rose to 0.94. The number of features has little influence on predictive accuracy: with 30 features, the correlation was 0.85. The model predicts less-extreme probabilities than were observed in the choices of the children, in particular in the cases where children chose to play with one of Actor 2's objects. This may be attributed to Actor 1's objects having features one would expect people to like a priori, which could be accommodated by using a non-zero mean for the preference prior.

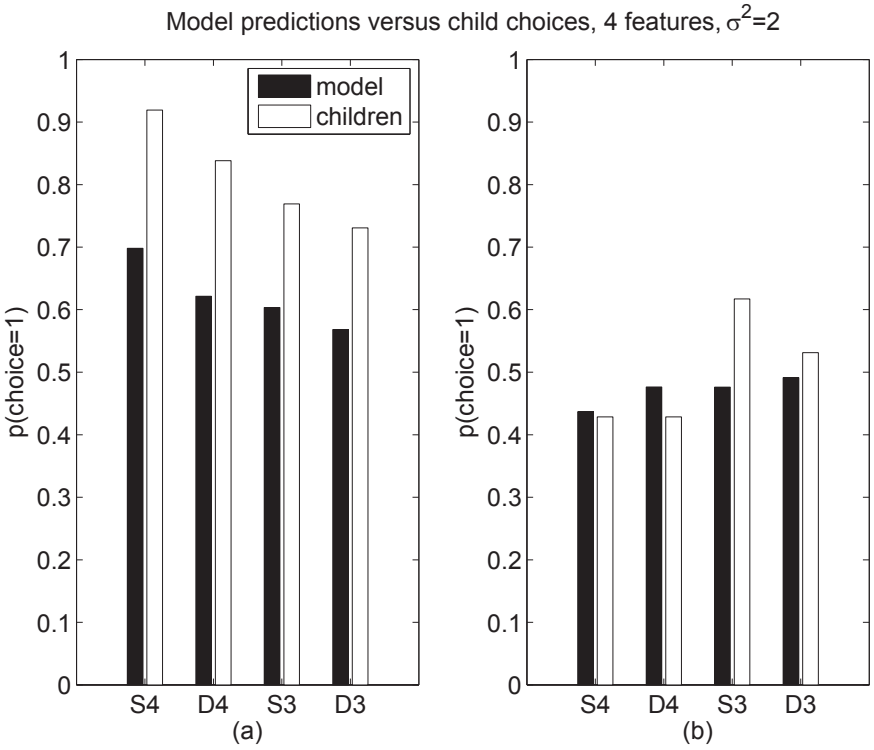


Figure 4.2: Model predictions for data in Experiment 1 of Fawcett and Markson (2010), excluding cases where children had fewer than 4 chances to play with training objects. Graph (a) shows conditions where the actors reacted positively to the hidden object, while graph (b) shows conditions where the actors reacted negatively to the hidden object. The first character for each pair of bars denotes whether the target object was in the same category (S) or a different category (D) from those seen in training. The second character denotes the number of times out of 4 that the child chose Actor 1's objects, reflecting the strength of the child's preferences. Finally, $P(\text{choice} = 1)$ is the probability of selecting Actor 1's novel object.

4.5 The developmental course of preference understanding

The last phenomenon we will consider is the developmental difference found by Repacholi and Gopnik (1997), which provides a challenge for our model and sheds light on the broader developmental course of preference learning. In this section, we will show how our model can explain their effect as the result of children making rational inferences.

4.5.1 Developmental transition as a rational shift in beliefs

Given sparse data and some weak expectations about the strengths of preferences, it may make sense to infer that everyone's preferences are the same[3], even when more numerous observa-

[3]Equivalently, that "preferences" are merely recognition of the intrinsic goodness of the available options.

tions with the same pattern support the belief that people have different preferences. Shifting from one model to another in this way is a consequence of the fact that simpler models tend to be more probable than more complex models with similar accuracy. Complex models, with larger numbers of parameters and the flexibility to explain a wide range of possibilities, assign probability to events not supported by observed data. Until enough events that are improbable under the simpler model are observed, the more complex one should be discounted. As an example, suppose that some postal mail you expect sometimes fails to arrive. If the proportion of letters lost is small, it might be reasonable to believe that the local postal service is reliable, and the letters were lost because of rare events like incorrect addresses – unreliable delivery might explain the losses, but it is unnecessarily flexible, being able to explain any pattern of missing mail, including total stoppage. As more time passes, and more rare events must be invoked to explain the loss of mail, the unreliable delivery hypothesis becomes more likely. In the context of Bayesian model selection, this effect is called the *Bayesian Occam's razor* (Jeffreys & Berger, 1992), and has already been used to explain a developmental shift: Goodman et al. (2006) showed that the classic result in which 3- to 4-year old children begin to pass the standard false belief task (Wimmer & Perner, 1983) can be explained in terms of moving to a more-complex model that includes a new parameter representing others' knowledge of the state of the world.

In the case of preferences, the more flexible model is the one we have been using: each person has a distinct set of preferences, which are drawn from a normal distribution with mean zero. The restricted model assumes that all people have the same preferences, and is otherwise identical to the first. If a learner sees choices made by people with distinct but similar preferences, the simpler model can explain a small number of choices well, as there is insufficient evidence to distinguish between noise and individual differences. As the number of observed choices grows, however, the simpler model will fail to account for the subtle but increasingly reliable differences between individuals, making it more and more likely that the flexible model is correct. We believe that most young children find themselves a situation like this, because their preferences are broadly similar to those of their caregivers and siblings, and it may take quite some time to observe enough evidence to reveal individual differences.

4.5.2 Simulations

In order to establish that our model can predict a developmental shift, it is first necessary to establish sets of choice events corresponding to the child's observations at different ages, which depend on the agents present, their preferences, and the features of their options. We assume that the child observes her own choices and those of her parent and a sibling. The preferences underlying those choices are given in Table 4.2, as are the features we chose for the different food options,[4] four of which are available at any choice event. While we assume that the child lacks direct access to her own preferences, we account for the fact that she observes many more of her own choices than those of others by supposing she sees ten times as many of her own choices as choices by the other agents.[5]

Using these data, we can determine how a rational learner's predictions about a new

[4] We chose the options and features heuristically, with the aim that they be consistent with the preferences about foods exhibited by adults and children in Skinner, Carruth, Bounds and Ziegler (2002).

[5] We found similar results when we provided the model with a large number of self-choices in advance, simulating the child having full access to her own preferences.

Table 4.2: Preferences and features for objects used to establish a developmental transition in theories of preference in simulation of Repacholi and Gopnik (1997).

	Features					
Preferences	fruit/vegetable	soft	sweet	liquid	milky	starchy
self	0	1	1.4	0.6	1.2	1.5
sibling	0.25	2	1.2	0.55	1.3	1.25
parent	0.5	3	1	0.5	1.4	1.0
Options						
raw vegetable	1	0	0	0	0	0
cooked vegetable	1	1	0	0	0	0
fruit	1	0	1	0	0	0
juice	1	0	1	1	0	0
dairy, not dessert	0	0	0	1	1	0
dairy, dessert	0	1	1	0	1	0
bread	0	1	0	0	0	1
cereal	0	0	1	0	0	1
desserts	0	1	1	0	0	1
soft drinks	0	0	1	1	0	0
broccoli	1	0	0	0	0	0
goldfish	0	0	0	0	0	1

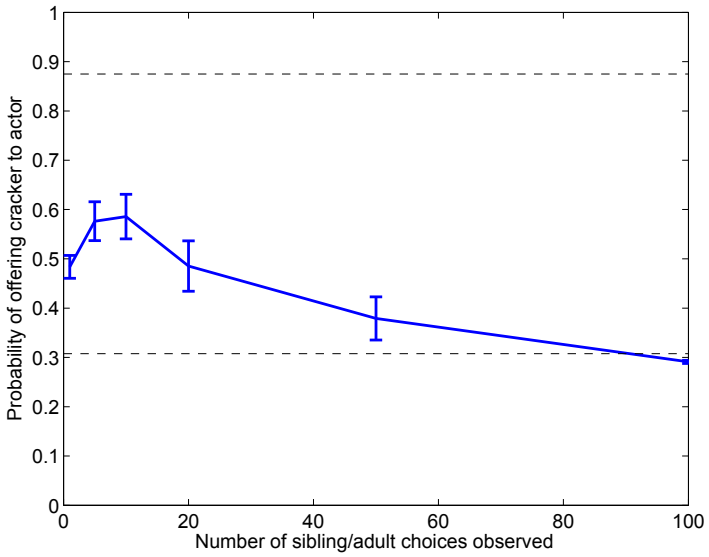
Note: The broccoli and goldfish items were only used in the experimental choice events in which the child selected goldfish and the actor selected broccoli, given both options.

broccoli-choosing actor's preferences should change over time: as the learner observes more choices, her beliefs about whether a simpler model (Model 1 or m_1) or a more-flexible model (Model 2 or m_2) is correct will change, as will her beliefs about what preferences agents should have under each model, leading to different predictions about the probability that the new agent should pick broccoli over goldfish, or vice versa. Formally, if $m \in \{m_1, m_2\}$ denotes the model, d denotes the available data – choices observed along with agent identities and features of options – and c denotes the actor's next choice, then:

$$p(c = \text{broccoli}|d) = \sum_{m \in \{m_1, m_2\}} P(c = \text{broccoli}|m, d)p(m|d) \tag{4.9}$$

where $P(c = \text{broccoli}|m = m_2, d)$ ignores every choice except the new actor's previous broccoli selection: under Model 2, all agents have independent preferences. In contrast, $P(C = \text{broccoli}|m = m_1, d)$ considers every choice event as if it had come from a single agent, so the probability of the actor choosing broccoli again will be dominated by the child's own preferences, which are responsible for most of the observed events. The posterior probability of model m, $p(m|d)$, is proportional to $P(d|m)P(m)$ (see the Appendix for details).

(a)



(b)

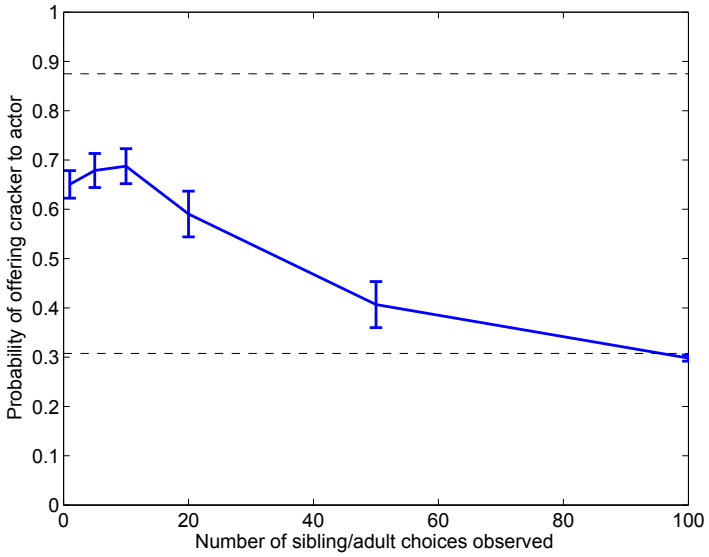


Figure 4.3: Results of simulations of the unmatched condition from Repacholi and Gopnik (1997). Each line shows the mean across 15 simulations, with standard errors. In both plots, the upper dashed line marks the proportion of 14-month-olds who offered the actor goldfish over broccoli (7 of 8), while the lower dashed line marks the proportion of 18-month-olds who did so (8 of 26). Plot (a) assumes equal prior belief in each model, while (b) assumes that the simpler model has a prior probability of 0.9.

4.5.3 Results

As predicted, sparse data favor Model 1, leading to the prediction that the actor would be likely to pick a cracker next. As data accumulate there is a shift toward Model 2, leading to a higher probability that broccoli will be offered to the new actor, because Model 2 treats the actor's choice of broccoli over goldfish crackers as the only event that is diagnostic of her preferences. The specific probabilities are given in Figure 4.3a, assuming that both models are equally likely, *a priori*.

One difference between our predictions and Repacholi and Gopnik's results is that for very small numbers of choice events, the events observed in the experiment compose a significant proportion of the total evidence, leading the more complex Model 2 to be favored. A possible explanation for this is that in the absence of evidence, children believe that Model 1 is more likely. Figure 4.3b shows the inferences that our simulations predict in the case where Model 1 is believed to be correct with a prior probability of 0.9. The resulting predictions are closer to the proportions seen in children's choices.

One remaining question is why so many of the younger children in Repacholi and Gopnik's experiment refused to make any offer at all. Intuitively, you might think that children suspected something in the broccoli choice that did not fit with their preferred model and that this made them less confident in their choices. We can capture this intuition more formally by saying that the children's confidence in a hypothesis is the probability that they assign to it, and that the 14-month-olds assign enough probability to the more complex model that they are less confident in their choices. This might well interact with a general reluctance to make offers to strangers (only 44 percent of the children made offers in a warm-up condition).

4.6 General Discussion

Our goal has been to understand how children reason about the preferences of other people, and to explain their ability to learn from statistical evidence, generalize within and across categories, and discover that other people have their own distinct preferences. To that end, we used a model with roots in econometrics to see what inferences a Bayesian learner might make in these circumstances, making some simple assumptions about how preferences relate to choices. This model's predictions are consistent with children's judgments across a wide range of experimental conditions in Kushnir et al. (in press) and Fawcett and Markson (2010), predicting children's sensitivity to others' options in addition to their choices, their inferences from others' affective displays, and their generalizations across categories. It also shows how conceptual change in preference understanding might be the product of Bayesian inference, adding to a growing body of literature demonstrating that Bayesian methods provide elegant explanations for conceptual change (e.g., N. D. Goodman et al., 2006).

Finally, this work speaks to the development of theory of mind more generally. Most work using probabilistic models has focused on children's understanding of physical causality, such as the action of blocks on machines. This study, along with those of Goodman et al. (2006) and Seiver et al. (2007) , suggests that this kind of modeling can be equally effective in helping us understand children's developing knowledge of psychological causality. In particular, inferring preferences from choices underlies a wide range of more sophisticated understandings of the mind such as the inference of personality traits or intuitive judgments about the decisions of others. We know

that even infants understand that human action is directed towards particular goals (Woodward, 1998). If children assumed or learned an implicit version of the MML model early in development, such assumptions could bootstrap a variety of sophisticated abilities to learn about the minds and decisions of others. Moreover, although much of the focus in the theory of mind literature has been on states of belief, it is arguably even more important, from an evolutionary point of view, for children to be able to infer the desires and preferences of others.

Our analysis of children's inferences about preferences has been organized around a simple model from econometrics – the MML model. The MML model makes several predictions in addition to those we have explored here, which we hope to test in future work. The first is that children should expect that preferences are transitive: an agent who prefers option B to A, and option C to B, will be expected to prefer C to A – a form of transitive inference that follows directly from representing the utility of different outcomes. Another prediction, discussed above, is that children should be able to learn that one object is strongly preferred to another, allowing them to make more predictions than would be possible if they merely recognized the presence of a preference. A third prediction is that children will be able to generalize preferences on the basis of specific features in addition to category membership: if an agent chooses diverse objects that are all red, then children should infer that red objects are desirable to that agent. Conducting new experiments to directly test these predictions will complement the work we have presented in this paper, showing that previous results concerning children's inferences about preferences can be captured by this simple rational model.

Chapter 5

Conclusions

Before discussing the broader implications of this work and the questions it raises, I will recapitulate the basic results of the last three chapters.

Conclusion 1: People acquire and exploit abstract causal knowledge using diverse sources of information

The experiments reported in Chapter 1 show that people learn quickly about the forms of causal relationships using sources of information that include covariation data, category information, and verbal cues. In addition to acquiring inductive biases in a manner broadly consistent with a hierarchical Bayesian perspective, adults' and children's specific judgments are accurately predicted by a hierarchical Bayesian model. These results are compatible both with earlier experimental results suggesting that people are sensitive to causal mechanisms and with developmental theories about domain knowledge and framework theories, but most existing models of covariation-based causal inference fail to predict the transfer learning we have observed. The hierarchical Bayesian approach explains not just how mechanism knowledge should influence causal inference, but how that knowledge can itself be acquired. Returning to the empiricism-nativism debate discussed in Chapter 1, one might view this result as a point in favor of the empiricist perspective, but it is important to remember that this learning depends on pre-existing inductive biases, albeit at a more abstract level than we might have previously assumed. Conversely, the model's use of informative, domain-appropriate priors could be interpreted as support for a nativist stance, but we should note that these priors could themselves be learned, as suggested in Chapter 3.

Conclusion 2: Children can learn abstract principles more quickly than adults

The experiment reported in Chapter 2 shows that children are capable of the same abstract learning that adults are, and that their judgments were more strongly influenced by the available evidence than adults, whose inferences reflected a bias toward OR relationships. These results support the view that when learning about cause and effect, children are flexible learners whose inexperience may let them learn better from sparse evidence, especially in novel situations. This is not to say that children are broadly better at learning causal relationships, or that adults' experience puts them at a disadvantage overall – rather, adults' biases appear to reflect the prevalence of causal relationships in the environment, letting them quickly identify common kinds of relationships at the

expense of being slower to recognize rare kinds. Like the results from Chapter 1, these findings are explainable in terms of hierarchical Bayesian inference, but another level of abstraction might be necessary to make precise predictions about the role of extensive experience and biases that apply across domains.

These results show that the fundamental biases that lie beneath causal inference are more subtle and abstract than a priori preferences for specific kinds of causal relationships. This finding informs the debate about cognitive development in general – the idea that children are more flexible in their commitments about the way that causal systems work is not just a consequence of a hierarchical Bayesian approach, but also an insight for understanding how children see the world differently from adults.

Conclusion 3: A rational model explains children understanding of preferences

Chapter 3 showed that a single model with origins in economics – the mixed multinomial logit – explains how children use statistical information to make inferences about other people's preferences, and their ability to make generalizations that are sensitive to category membership. It also explains an instance of conceptual change, in which 14-month-old children treat preferences as either shared between individuals or merely a reflection of options' intrinsic goodness, while 18-month-olds understand that preferences are idiosyncratic.

5.1 Remaining questions and future directions

Taken together, this work gives evidence that people can acquire and apply abstract knowledge in different domains, that hierarchical Bayesian models are useful tools for understanding how we learn about higher-order causal regularities in our environment, and about what biases precede and follow experience. While this approach answers questions about abstract knowledge and its origins, it raises several questions as well.

First, how might one extend the model in Chapters 2 and 3 to capture the full breadth of human beliefs about the forms of causal relationships? The model in those chapters uses a hypothesis space that was designed to be simple and appropriate for the cover story given to participants, and it fails to include some kinds of causal relationships that should be easy to learn, such as prevention. Given that people seem to be able to learn a wide variety of relationships while still possessing strong biases (e.g., toward disjunctive, generative relationships), one natural answer is to use a compositional model (as in Kemp, Goodman, & Tenenbaum, 2008) with only a minimal set of primitive relationships, but the capacity to represent arbitrary relationships using those elements. A project is planned to enumerate a set of primitives and compositional rules, and to develop experiments to test this proposal.

Second, how do people identify causally relevant variables? Most research into causal inference supposes that the variables have been identified before causal learning takes place, and the work presented here is no exception. In chapters 2 and 3, the model represents the prospective blickets and blicket detector but not other conceivably relevant factors like the objects' geometries and specific positions. Chapter 4 considers a specific example of variable identification in explaining children's shift toward understanding preferences as specific to individuals, but it also assumes that there exists a finite set of observable features and offers no way to discover new ones. Attempting

to incorporate general-purpose variable identification into these models would obscure their contributions, but we cannot ignore the fact that variable choice shapes causal inference. For example, choosing one level of granularity over another licenses dramatically different kinds of inferences, as with understanding organisms in terms of populations, organs, cells, or molecules. Even at a fixed level of granularity, the ways in which people carve the world into causal entities can lead to different predictions, as with representing oxygen as a causal variable rather than phlogiston, the substance that chemists once believed to be responsible for combustion. Future work may address the relationship between functional form and inference to new variables: impossible or unlikely relationships between observed variables can be explained by positing the presence of new causes or mediators, so people's knowledge about the forms of causal relationships could aid in identifying new variables.

Third, what are the new, testable predictions of the model in Chapter 4? In addition to the behaviors it explains, the model makes commitments about how children deal with others' choices and preference. First, it predicts that children should understand that people's preferences can have different extents or strengths, distinct from the weight of evidence that a preference exists at all. This means that if children are given evidence that a person strongly prefers item A to item C, and definitely but weakly prefers item B to item C, children will offer item A to that person rather than item B. The model also predicts that preferences apply to features of options, rather than options themselves, so children should be able to exploit evidence that a particular feature drives a person's like or dislike for an object and to generalize to new items with that feature. Third, the model predicts that certain kinds of experiences should influence the age at which children come to understand that people have distinct preferences. Specifically, children who observe more disagreement about what is desirable, including children with more siblings, should understand at a younger age that not all people share the same preferences. Fourth, the model predicts that children should believe that preferences are transitive: if a person chooses item A over item B, and B over item C, then she should choose item A over item C. There is already some evidence that this is true (Mou, Province, & Luo, 2010), but additional studies with tighter controls are necessary. Given that people's real preferences are not always transitive (Tversky, 1969) and that transitivity violation can lead to suboptimal choices, we have two more questions: are there systematic differences between how people make decisions and their folk theories of decision making? If so, do we treat other people as being less error-prone than they actually are?

In addition to the questions that are specific to the models and phenomena presented here, there are some that apply to all computational-level models of human induction. Anderson's program of rational analysis (Anderson, 1990) devotes a step to the formulation of constraints under which learners operate, but this step is almost universally ignored. Constraints we might consider include limits on time, space (i.e., memory), and the metabolic cost of neural activity, and after accounting for them there are likely cases where Bayesian solutions are not optimal given the costs of producing those solutions. At one level, this is a positive point for researchers interested in computational-level explanations, because it leaves a path for explaining human behaviors that defy simple rational explanations, like judgments that do not maximize outcomes (West & Stanovich, 2003; Vul & Pashler, 2008) and order effects (Lu, Rojas, et al., 2008).

Unfortunately, relatively little is known about the specifics of these constraints, and even if they could be articulated clearly, it is unclear how one would systematically incorporate them into rational models of induction. This is not to say that it is impossible to make progress in un-

derstanding the role of constraints. One response to the challenge this task presents has been the development of "rational process" models which aim to produce approximately Bayesian solutions while appealing to time- and memory-efficient inference and connection to existing process-level explanations (Sanborn, Griffiths, & Navarro, 2006; Shi, Griffiths, Feldman, & Sanborn, n.d.). The study of the physical architecture of the brain may also provide valuable clues about what kinds of inference are possible and what constraints should be considered (Beck et al., 2008).

Hierarchical Bayesian models provide us with a valuable tool for understanding how people acquire the inductive biases that enable us to efficiently recover the causal structure of the world around us. These models give us a direct and flexible way to express hypotheses about how abstract knowledge might result from experience. Nonetheless, they do not obviate the need for process models – indeed, in reifying some inductive problems and their solutions, they highlight just how difficult these problems may be to solve, and thus the need for solutions that consider our limitations.

Bibliography

Ahn, W., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition*, *54*, 299-352.

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: Erlbaum.

Beck, J., Ma, W., Kiani, R., Hanks, T., Churchland, A., Roitman, J., et al. (2008). Probabilistic population codes for Bayesian decision making. *Neuron*, *60*(6), 1142–1152.

Beckers, T., De Houwer, J., Pineno, O., & Miller, R. R. (2005). Outcome additivity and outcome maximality influence cue competition in human causal learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 238-249.

Boyd, J. H., & Mellman, R. E. (1980). Effect of fuel economy standards on the u.s. automotive market: An hedonic demand analysis. *Transportation Research A*, *14*, 367-378.

Breese, J., Heckerman, D., & Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the fourteenth annual conference on uncertainty in artificial intelligence (uai 98).* San Francisco, CA: Morgan Kaufmann.

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.

Carey, S. (1991). Knowledge acquisition: Enrichment or conceptual change? In S. Carey & R. Gelman (Eds.), *The epigenesis of mind: Essays on biology and cognition* (p. 257-292). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

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

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Colunga, E., & Smith, L. (2005). From the lexicon to expectations about kinds: A role for associative learning. *Psychological Review*, *112*, 347–382.

Cooper, G., & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, *9*, 308-347.

Dennett, D. (1989). *The intentional stance*. The MIT Press.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification*. New York: Wiley.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective*. Cambridge, MA: MIT Press.

Estes, W. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, *53*(2), 134-140.

Fawcett, C. A., & Markson, L. (2010). Children reason about shared preferences. *Developmental psychology*, *46*(2), 299.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, *407*, 630-633.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. New York: Chapman & Hall.

Glymour, C. (1998). Learning causes: Psychological explanations of causal explanation. *Minds and Machines*, *8*, 39-60.

Gold, E. M. (1967). Language identification in the limit. *Information and Control*, *10*, 447-474.

Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, *35*(12), 61-70.

Goodman, N. (1955). *Fact, fiction, and forecast*. Cambridge: Harvard University Press.

Goodman, N. D., Baker, C. L., Baraff Bonawitz, E., Mansinghka, V. K., Gopnik, A., Wellman, H., et al. (2006). Intuitive theories of mind: A rational approach to false belief. In *28th annual conference of the cognitive science society.*

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*, 108-154.

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., & 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., & Wellman, H. (1992). Why the child's theory of mind really is a theory. *Mind and Language*, *7*, 145-171.

Greene, K. (2006). The $1 million netflix challenge. *Technology Review*, *6*.

Griffiths, T. L. (2005). *Causes, coincidences, and theories*. Unpublished doctoral dissertation, Stanford University.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 354-384.

Griffiths, T. L., & Tenenbaum, J. B. (2007). Two proposals for causal grammars. In A. Gopnik & L. E. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation.* Oxford: Oxford University Press.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-Based Causal Induction. *Psychological Review*, *116*(4), 661–716.

Hume, D. (1748). *An enquiry concerning human understanding*. Indianapolis, IN: Hackett.

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge: Cambridge University Press.

Jeffreys, W. H., & Berger, J. O. (1992). Ockham's razor and Bayesian analysis. *American Scientist*, *80*(1), 64-72.

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs*, *79*.

Kemp, C., Goodman, N., & Tenenbaum, J. (2008). Theory acquisition and the language of thought. In *Proceedings of thirtieth annual meeting of the cognitive science society.*

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2007). Learning causal schemata. In *the 29th annual conference of the cognitive science society.* Nashville, TN.

Kemp, C., Perfors, A., & Tenenbaum, J. (2007). Learning overhypotheses with hierarchical bayesian models. *Developmental Science*, *10*(3), 307–321.

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., & Ueda, N. (2006). Learning systems of concepts with an infinite relational model. In *21st national conference on artificial intelligence.*

Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R., & Riedl, J. (1997). Grouplens: applying collaborative filtering to usenet news. *Communications of the ACM*, *40*, 77–87.

Koslowski, B. (1996). *Theory and evidence: The development of scientific reasoning*. Cambridge, MA: MIT Press.

Kushnir, T., Xu, F., & Wellman, H. (in press). Young children use statistical sampling to infer the preferences of others. *Psychological Science*.

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.

Lien, Y., & Cheng, P. W. (2000). Distinguishing genuine from spurious causes: A coherence hypothesis. *Cognitive Psychology*, *40*, 87-137.

Lu, H., Rojas, R. R., Beckers, T., & Yuille, A. (2008). Sequential causal learning in humans and rats. In *Twenty-ninth annual conference of the cognitive science society.*

Lu, H., Yuille, A., Liljeholm, M., Cheng, P., & Holyoak, K. (2008). Bayesian generic priors for causal learning. *Psychological review*, *115*(4), 955–984.

Lu, H., Yuille, A., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2006). Modeling causal leanring using bayesian generic priors on generative and preventive powers. In R. Sun & N. Miyake (Eds.), *Twenty-eighth conference of the cognitive science society* (p. 519-524). Erlbaum.

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

Lucas, C., & Griffiths, T. (2009). Learning the Form of Causal Relationships Using Hierarchical Bayesian Models. *Cognitive Science*, *34*(1), 113–147.

Luce, R. D. (1959). *Individual choice behavior*. New York: John Wiley.

Markman, E., & Wachtel, G. (1988). Childrens use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*(2), 121–157.

Marr, D. (1982). *Vision*. San Francisco, CA: W. H. Freeman.

McFadden, D., & Train, K. E. (2000). Mixed MNL models of discrete response. *Journal of Applied Econometrics*, *15*, 447-470.

Mou, Y., Province, J., & Luo, Y. (2010). Can 16-month-old infants make transitive inferences? *Poster session presented at the 2010 International Conference on Infant Studies*.

Myung, I. J., Kim, C., & Pitt, M. A. (2000). Toward an explanation of the power law artifact: Insights from response surface analysis. *Memory & Cognition*, *28*(5), 832-840.

Navarro, D. J., Griffiths, T. L., Steyvers, M., & Lee, M. D. (2006). Modeling individual differences using dirichlet processes. *Journal of Mathematical Psychology*, *50*, 101-122.

Neal, R. M. (1993). *Probabilistic inference using Markov chain Monte Carlo methods* (Tech. Rep. No. CRG-TR-93-1). University of Toronto.

Novick, L. R., & Cheng, P. W. (2004). Assessing interactive causal inference. *Psychological Review*, *111*, 455-485.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press, USA.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Francisco, CA: Morgan Kaufmann.

Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge, UK: Cambridge University Press.

Perales, J. C., & Shanks, D. R. (2007). Models of covariation-based causal judgment: a review and synthesis. *Psychonomic Bulletin and Review*, *14*(4), 577-596.

Repacholi, B. M., & Gopnik, A. (1997). Early reasoning about desires: Evidence from 14- and 18-month-olds. *Developmental Psychology*, *33*(1), 12-21.

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.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2006). A more rational model of categorization. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society.* Mahwah, NJ: Erlbaum.

Schulz, L., & Sommerville, J. (2006). God does not play dice: Causal determinism and preschoolers causal inferences. *Child Development*, *77*(2), 427-442.

Schulz, L. E., Bonawitz, E. B., & Griffiths, T. L. (2007). Can being scared make your tummy ache? naive theories, ambiguous evidence, and preschoolers' causal inferences. *Developmental Psychology*.

Schulz, L. E., Goodman, N., Tenenbaum, J., & Jenkins, A. (2008). Going beyond the evidence: Abstract laws and preschoolers' responses to anomalous data. *Cognition*, *109*(2), 211–223.

Seiver, A., E. Gopnik, & Lucas, C. (2007). Social causal models in children: understanding the interaction of person and situation. *Poster session presented at the Biennial Metting of the Society for Research in Child Development, Boston, MA*.

Shanks, D. R. (1995). *The psychology of associative learning*. Cambridge: Cambridge University Press.

Shanks, D. R., & Darby, R. J. (1998). Feature- and rule-based generalization in human associative learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *24*(4), 405-415.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*, 325-345.

Shi, L., Griffiths, T., Feldman, N., & Sanborn, A. (n.d.).

Shultz, T. R. (1982). Rules of causal attribution. *Monographs of the Society for Research in Child Development*, *47*(Serial no. 194).

Skinner, J. D., Carruth, B. R., Bounds, W., & Ziegler, P. J. (2002). Children's food preferences: A longitudinal analysis. *Journal of the American Dietetic Association*, *102*(11), 1638-1647.

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science*, *13*(1), 13–19.

Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect

evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science*, *28*, 303-333.

Spelke, E. (1994). Initial knowledge: six suggestions. *Cognition*, *50*(1-3), 431–445.

Spelke, E., & Kinzler, K. (2007). Core knowledge. *Developmental Science*, *10*(1), 89–96.

Spirtes, P., Glymour, C., & Schienes, R. (1993). *Causation prediction and search*. New York: Springer-Verlag.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Structure learning in human causal induction. In T. Leen, T. Dietterich, & V. Tresp (Eds.), *Advances in Neural Information Processing Systems 13* (p. 59-65). Cambridge, MA: MIT Press.

Tenenbaum, J. B., & Griffiths, T. L. (2003). Theory-based causal induction. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in Neural Information Processing Systems 15* (p. 35-42). Cambridge, MA: MIT Press.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Science*, *10*, 309-318.

Tenenbaum, J. B., & Niyogi, S. (2003). Learning causal laws. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the 25th annual meeting of the cognitive science society.* Hillsdale, NJ: Erlbaum.

Train, K. E., McFadden, D., & Ben-Akiva, M. (1987). The demand for local telephone service: A fully discrete model of residential calling patterns and service choices. *The RAND Journal of Economics*, *18*(1), 109-123.

Tversky, A. (1969). Intransitivity of preferences. *Psychological review*, *76*(1), 31–48.

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, *19*(7), 645-647.

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

West, R., & Stanovich, K. (2003). Is probability matching smart? Associations between probabilistic choices and cognitive ability. *Memory & Cognition*, *31*(2), 243.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young childrens understanding of deception. *Cognition*, *13*(1), 103–128.

Woodward, A. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, *69*(1), 1–34.

Yuille, A. L., & Lu, H. (2007). The noisy-logical distribution and its application to causal inference. In *Advances in neural information processing systems 20.* Cambridge, MA: MIT Press.

Zelazo, P. R., & Shultz, T. R. (1989). Concepts of potency and resistence in causal prediction. *Child Development*, *60*, 1307-1315.

Appendix A

Materials for Chapter 1, Experiment 5

The cover story for the dax blocks:

You will read about some rodents. Your goal is to figure out which of them are Daxes – some are and some are not. People can't tell Daxes from non-Daxes, but cats can smell Daxes and are afraid of them. There are three rodents, called A, B and C. The list below describes what happened when a cat was exposed to different rodents or groups of rodents.

The training block evidence in the dax condition:

One time, the cat was exposed to A. The cat did not run away.
Another time, the cat was exposed to B. The cat did not run away.
Another time, the cat was exposed to C. The cat did not run away.
Another time, the cat was exposed to A and B. The cat did not run away.
Another time, the cat was exposed to A and C. The cat ran away.
Another time, the cat was exposed to B and C. The cat did not run away.

The questions people were asked in the dax-condition training block:

Write down for rodents A, B, and C the probability that each is a Dax, using a number from 0 to 10 where 0 means you're absolutely certain it is not a Dax, 10 means you're absolutely certain it is a Dax, and 5 means it is equally likely to be a Dax as not.

Test block materials differed only in the specific evidence given.

Appendix B

The mixed multinomial logit model

B.1 Background

The MML model can approximate the distribution of choices for essentially any heterogeneous population of utility-maximizing agents given appropriate choice of $p(\boldsymbol{\beta}_a)$ (McFadden & Train, 2000). One common variant supposes that $\boldsymbol{\beta}_a$ follows a Gaussian distribution around a population mean which in turn has a Gaussian prior. We assume a single-parameter prior in which different agents' preferences are independent and preferences for individual features are uncorrelated with a Gaussian distribution with mean zero and variance σ^2: $\boldsymbol{\beta}_a \sim \mathcal{N}(0, \sigma^2\mathbf{I})$.

B.2 Inferring preferences from statistical information

In order to make predictions for Kushnir et al.'s experiment, we first obtained a distribution over preferences given Squirrel's choices and options, and used those preferences to predict what object Squirrel would have chosen next.

Let $\boldsymbol{\beta}_a$ be Squirrel's preferences, $\mathbf{c} = (c_1 \dots c_N)$ the sequence of N choices Squirrel makes, and $\mathbf{X}_n = [\mathbf{x}_{n1} \dots \mathbf{x}_{nJ_n}]^T$ the observed features of Squirrel's J_n options at choice event n. The set $\{\mathbf{X}_1, \dots, \mathbf{X}_N\}$ will be denoted with $\mathbf{X}$. Estimating $\boldsymbol{\beta}_a$ entails computing $p(\boldsymbol{\beta}_a|\mathbf{c}, \mathbf{X}) \propto P(\mathbf{c}|\boldsymbol{\beta}_a, \mathbf{X})p(\boldsymbol{\beta}_a)$, analogous to the inference of $\mathbf{u}$ in Equation 4.2. The probability of Squirrel's choices is $P(\mathbf{c}|\mathbf{X}, \boldsymbol{\beta}_a) = \prod_{n=1}^{N} P(c_n|\mathbf{X}_n, \boldsymbol{\beta}_a)$, where $P(c_n = j|\mathbf{X}_n, \boldsymbol{\beta}_a)$ is given by Equation 4.3.

We represented the objects using minimal and orthogonal feature vectors, so red discs (Squirrel's target toy) had features $[1\ 0\ 0]^T$, blue flowers (the alternative option in his choices) had features $[0\ 1\ 0]^T$, and yellow cylinders (the distractor) had features $[0\ 0\ 1]^T$, respectively. Each of the 38 objects was a separate option, where 38, 19, and 7 of the objects had features of the target object in the 100%, 50%, and 18% conditions, respectively, with the remainder having the features of the alternative object. The $N = 5$ choices made by Squirrel thus provide the data $\mathbf{c}$ from which its preferences can be inferred. We constructed an approximation to the posterior distribution over $\boldsymbol{\beta}_a$ given $\mathbf{c}$ using importance sampling (see, e.g., Neal, 1993, for details), drawing 10^6 samples of $\boldsymbol{\beta}_a$ from the prior distribution $p(\boldsymbol{\beta}_a)$ and giving each sample $\boldsymbol{\beta}_a^{(i)}$ weight w_i proportional to the corresponding likelihood $P(\mathbf{c}|\mathbf{X}, \boldsymbol{\beta}_a^{(i)})$.

The probability that the child will offer a particular item can be computed by assuming

that the child matches the predicted probability that Squirrel would choose an object, integrating out $\boldsymbol{\beta_a}$:

$$P(c_{\text{child}} = j|\mathbf{X}, \mathbf{c}) = \int P(c_{\text{child}} = j|\mathbf{X}, \boldsymbol{\beta}_a) p(\boldsymbol{\beta}_a|\mathbf{X}, \mathbf{c})\, d\boldsymbol{\beta}_a, \tag{B.1}$$

which is approximately equal to

$$\frac{1}{\sum_i w_i} \sum_i P(c_{\text{child}} = j|\mathbf{X}, \boldsymbol{\beta}_a^{(i)}) w_i. \tag{B.2}$$

B.3 Developmental differences

To model the developmental shift described in Repacholi and Gopnik (1997), we generated a series of choices by three agents: k choices by a parent, k choices by a sibling, and $10k$ choices by the child, with $k \in \{1, 5, 10, 20, 50\}$. Each choice was made from four options selected randomly without replacement from the set in the middle set of rows in Table 4.2. Two additional choices were appended to these data: one in which the child chose a goldfish cracker over broccoli, and one in which a new agent chose broccoli over the goldfish cracker.

The probability of offering the goldfish cracker to the actor was equal to the probability that the actor would choose the goldfish cracker herself, given by Equation 4.9. The probability of model m_1 (same preferences) versus model m_2 (distinct preferences) is given by

$$P(m|d) = \frac{P(m)P(d|m)}{\sum_{m'} P(m')P(d|m')}, \tag{B.3}$$

where d includes all of the available data, including choices $\mathbf{c}$, options and their features $\mathbf{X}$, and the identity of the agent making the choice, $\mathbf{a} = (a_1, a_2, ..., a_N)$.

Assuming that the identities of the agents and the available options are given,

$$P(d|m) = \int \prod_{i=1}^{N} P(c_i = j|\mathbf{X}, \boldsymbol{\beta}_a) p(\boldsymbol{\beta}_a|m, a_i)\, d\boldsymbol{\beta}_a. \tag{B.4}$$

If $m = m_2$, then a_i determines which agent's preferences are applicable at choice event i, while under m_1, the same preferences are used for all choice events, independent of a_i.

We generated predictions over fifteen separate runs, each with different, randomly selected options at each choice event. We used importance sampling to estimate the model and choice probabilities, drawing 250,000 samples per run, per point in Figure 4.3. For each sampling run, the high-dimensionality of the space of preferences made it necessary to obtain proposal distributions concentrated around the posterior, which we achieved by estimating preferences using 150,000 MCMC samples and generating proposals from a t-distribution fitted to those samples.