COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science (2010) 1–59

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/j.1551-6709.2010.01128.x

Learning to Learn Causal Models

Charles Kemp,[a] Noah D. Goodman,[b] Joshua B. Tenenbaum[b]

[a] *Department of Psychology, Carnegie Mellon University*
[b] *Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology*

Received 6 November 2008; received in revised form 11 June 2010; accepted 14 June 2010

Abstract

Learning to understand a single causal system can be an achievement, but humans must learn about multiple causal systems over the course of a lifetime. We present a hierarchical Bayesian framework that helps to explain how learning about several causal systems can accelerate learning about systems that are subsequently encountered. Given experience with a set of objects, our framework learns a causal model for each object and a *causal schema* that captures commonalities among these causal models. The schema organizes the objects into categories and specifies the causal powers and characteristic features of these categories and the characteristic causal interactions between categories. A schema of this kind allows causal models for subsequent objects to be rapidly learned, and we explore this accelerated learning in four experiments. Our results confirm that humans learn rapidly about the causal powers of novel objects, and we show that our framework accounts better for our data than alternative models of causal learning.

Keywords: Causal learning; Learning to learn; Learning inductive constraints; Transfer learning; Categorization; Hierarchical Bayesian models

1. Learning to learn causal models

Children face a seemingly endless stream of inductive learning tasks over the course of their cognitive development. By the age of 18, the average child will have learned the meanings of 60,000 words, the three-dimensional shapes of thousands of objects, the standards of behavior that are appropriate for a multitude of social settings, and the causal structures underlying numerous physical, biological, and psychological systems. Achievements like

Correspondence should be send to Charles Kemp, Department of Psychology, Carnegie Mellon University, 5000 Forbes Avenue, Baker Hall 340T, Pittsburgh, PA 15213. E-mail: ckemp@cmu.edu

these are made possible by the fact that inductive tasks fall naturally into families of related problems. Children who have faced several inference problems from the same family may discover not only the solution to each individual problem but also something more general that facilitates rapid inferences about subsequent problems from the same family. For example, a child may require extensive time and exposure to learn her first few names for objects, but learning a few dozen object names may allow her to learn subsequent names much more quickly (Bloom, 2000; Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002).

Psychologists and machine learning researchers have both studied settings where learners face multiple inductive problems from the same family, and they have noted that learning can be accelerated by discovering and exploiting common elements across problems. We will refer to this ability as ''learning to learn'' (Harlow, 1949; Yerkes, 1943), although it is also addressed by studies that focus on ''transfer learning,'' ''multitask learning,'' ''lifelong learning,'' and ''learning sets'' (Caruana, 1997; Stevenson, 1972; Thorndike & Woodworth, 1901; Thrun, 1998; Thrun & Pratt, 1998). This paper provides a computational account of learning to learn that focuses on the acquisition and use of inductive constraints. After experiencing several learning problems from a given family, a learner may be able to induce a *schema*, or a set of constraints that captures the structure of all problems in the family. These constraints may then allow the learner to solve subsequent problems given just a handful of relevant observations.

The problem of learning to learn is relevant to many areas of cognition, including word learning, visual learning, and social learning, but we focus here on causal learning and explore how people learn and use inductive constraints that apply to multiple causal systems. A door, for example, is a simple causal system, and experience with several doors may allow a child to rapidly construct causal models for new doors that she encounters. A computer program is a more complicated causal system, and experience with several pieces of software may allow a user to quickly construct causal models for new programs that she encounters. Here we consider settings where a learner is exposed to a family of objects and learns causal models that capture the causal powers of these objects. For example, a learner may implicitly track the effects of eating different foods and may construct a causal model for each food that indicates whether it tends to produce indigestion, allergic reactions, or other kinds of problems. After experience with several foods, a learner may develop a schema (Kelley, 1972) that organizes these foods into categories (e.g., citrus fruits) and specifies the causal powers and characteristic features of each category (e.g., citrus fruits cause indigestion and have crescent-shaped segments). A schema of this kind should allow a learner to rapidly infer the causal powers of novel objects: for example, observing that a novel fruit has crescent-shaped segments might be enough to conclude that it causes indigestion.

There are three primary reasons why causal reasoning provides a natural setting for exploring how people learn and use inductive constraints. First, abstract inductive constraints play a crucial role in causal learning. Some approaches to causal learning focus on bottom-up statistical methods, including methods that track patterns of conditional independence or partial correlations (Glymour, 2001; Pearl, 2000). These approaches, however, offer at best a limited account of human learning. Settings where humans observe correlational data without the benefit of strong background knowledge often lead to weak learning even when large amounts of training data are provided (Lagnado & Sloman, 2004; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003). In contrast, both adults and children can infer causal connections from observing just one or a few events of the right type (Gopnik & Sobel, 2000; Schulz & Gopnik, 2004)—far fewer observations than would be required to compute reliable measures of correlation or independence. Top-down, knowledge-based accounts provide the most compelling accounts of this mode of causal learning (Griffiths & Tenenbaum, 2007).

Second, some causal constraints are almost certainly learned, and constraint learning probably plays a more prominent role in causal reasoning than in other areas of cognition, such as language and vision. Fundamental aspects of language and vision do not change much from one generation to another, let alone over the course of an individual's life. It is therefore possible that the core inductive constraints guiding learning in language and vision are part of the innate cognitive machinery rather than being themselves learned (Bloom, 2000; Spelke, 1994). In contrast, cultural innovation never ceases to present us with new families of causal systems, and the acquisition of abstract causal knowledge continues over the life span. Consider, for example, a 40-year-old who is learning to use a cellular phone for the first time. It may take him a while to master the first phone that he owns, but by the end of this process—and certainly after experience with several different cell phones—he is likely to have acquired abstract knowledge that will allow him to adapt to subsequent phones rapidly and with ease.

The third reason for our focus on causal learning is methodological, and it derives from the fact that learning to learn in a causal setting can be studied in adults and children alike. Even if we are ultimately interested in the origins of abstract knowledge in childhood, studying analogous learning phenomena in adults may provide the greatest leverage for developing computational models, at least at the start of the enterprise. Adult participants in behavioral experiments can provide rich quantitative judgments that can be compared with model predictions in ways that are not possible with standard developmental methods. The empirical section of this paper therefore focuses on adult experiments. We discuss the developmental implications of our approach in some detail, but a full evaluation of our approach as a developmental model is left for future work.

To explain how abstract causal knowledge can both constrain learning of specific causal relations and can itself be learned from data, we work within a hierarchical Bayesian framework (Kemp, 2008; Tenenbaum, Griffiths, & Kemp, 2006). Hierarchical Bayesian models include representations at several levels of abstraction, where the representation at each level captures knowledge that supports learning at the next level down (Griffiths & Tenenbaum, 2007; Kemp, Perfors, & Tenenbaum, 2007; Kemp & Tenenbaum, 2008). Statistical inference over these hierarchies helps to explain how the representations at each level are learned. Our model can be summarized as a three-level framework where the top level specifies a causal schema, the middle level specifies causal models for individual objects, and the bottom level specifies observable data. If the schema at the top level is securely established, then the framework helps to explain how abstract causal knowledge supports the construction of causal models for novel objects. If the schema at the upper level is not yet established, then the framework helps to explain how causal models can be learned primarily from observable data. Note, however, that top-down learning and bottom-up learning are just two of the possibilities that emerge from our hierarchical approach. In the most general case, a learner will be uncertain about the information at all three levels, and will have to simultaneously learn a schema (inference at the top level) and a set of causal models (inference at the middle level) and make predictions about future observations (inference at the bottom level).

Several aspects of our approach draw on previous psychological research. Cognitive psychologists have discussed how abstract causal knowledge (Lien & Cheng, 2000; Shanks & Darby, 1998) might be acquired, and they have studied the bidirectional relationship between categorization and causal reasoning (Lien & Cheng, 2000; Waldmann & Hagmayer, 2006). Previous models of categorization have used Bayesian methods to explain how people organize objects into categories based on their features (Anderson, 1991) or their relationships with other objects (Kemp, Tenenbaum, Griffiths, Yamada, & Ueda, 2006), although not in a causal context. In parallel, Bayesian models of knowledge-based causal learning have often assumed a representation in terms of object categories, but they have not attempted to learn these categories (Griffiths & Tenenbaum, 2007). Here we bring together all of these ideas and explore how causal learning unfolds simultaneously across multiple levels of abstraction. In particular, we show how learners can simultaneously make inferences about causal categories, causal relationships, causal events, and perceptual features.

2. Learning causal schemata

Later sections will describe our framework in full detail, but this section provides an informal introduction to our general approach. As a running example we consider the problem of learning about drugs and their side-effects: for instance, learning whether blood-pressure medications cause headaches. This problem requires inferences about two *domains*—people and drugs—and can be formulated as a *domain-level problem*:

$$\texttt{ingests}(\texttt{person}, \texttt{drug}) \overset{?}{\rightarrow} \texttt{headache}(\texttt{person}) \tag{1}$$

The domain-level problem in Eqn. 1 sets up an *object-level* problem for each combination of a person and a drug. For example,

$$\texttt{ingests}(\texttt{Alice}, \texttt{Doxazosin}) \overset{?}{\rightarrow} \texttt{headache}(\texttt{Alice}) \tag{2}$$

represents the problem of deciding whether there is a causal relationship between Alice taking Doxazosin and Alice developing a headache, and

$$\texttt{ingests}(\texttt{Bob}, \texttt{Prazosin}) \overset{?}{\rightarrow} \texttt{headache}(\texttt{Bob}) \tag{3}$$

represents a second problem concerning the effect of Prazosin on Bob. Our goal is to learn an *object-level causal model* for each object-level problem. In Fig. 1A there are six people

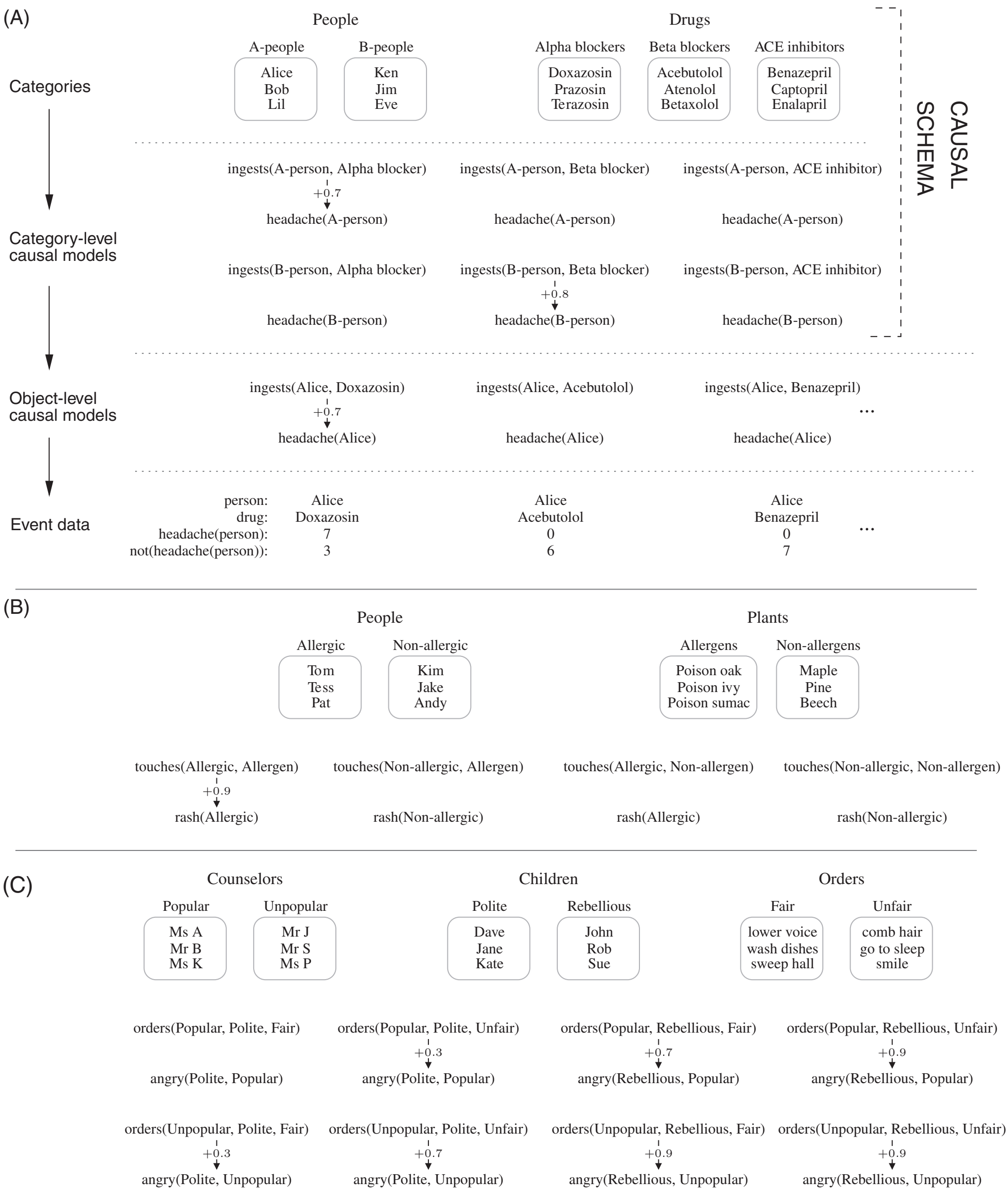


Fig. 1. Three settings where causal schemata can be learned. (A) The drugs and headaches example. The people are organized into two categories and the drugs are organized into three categories. The category-level causal models indicate that alpha blockers cause headaches in A-people and beta blockers cause headaches in B-people. There are 54 object-level causal models in total, one for each combination of a person and a drug, and three of these models are shown. The first indicates that Doxazosin often gives Alice headaches. The event data for learning these causal models are shown at the bottom level: Alice has taken Doxazosin 10 times and experienced a headache on seven of these occasions. (B) The allergy example. The schema organizes plants and people into two categories each, and the object-level models and event data are not shown. (C) The summer camp example. The schema organizes the counselors, the children, and the orders into two categories each.

and nine drugs, which leads to 54 object-level problems and 54 object-level models in total. Fig. 1A shows three of these object-level models, where the first example indicates that ingesting Doxazosin tends to cause Alice to develop headaches. The observations that allow these object-level models to be learned will be called event data or contingency data, and they are shown at the bottom level of Fig. 1A. The first column of event data indicates, for example, that Alice has taken Doxazosin 10 times and has experienced headaches on seven of these occasions.

The 54 object-level models form a natural family, and learning several models from this family should support inferences about subsequent members of the family. For example, learning how Doxazosin affects Alice may help us to rapidly learn how Doxazosin affects Bob, and learning how Alice responds to Doxazosin may help us to rapidly learn how Alice responds to Prazosin. This paper will explore how people learn to learn object-level causal models. In other words, we will explore how learning several of these models can allow subsequent models in the same family to be rapidly learned.

The need to capture relationships between object-level problems like Eqns. 2 and 3 motivates the notion of a causal schema. Each possible schema organizes the people and the drugs into categories and specifies causal relationships between these categories. For example, the schema in Fig. 1A organizes the six people into two categories (A-people and B-people) and the nine drugs into three categories (alpha blockers, beta blockers, and ACE inhibitors). The schema also includes *category-level* causal models that specify relationships between these categories. Because there are two categories of people and three categories of drugs, six category-level models must be specified in total, one for each combination of a person category and drug category. For example, the category-level models in Fig. 1A indicate that alpha blockers tend to produce headaches in A-people, beta blockers tend to produce headaches in B-people, and ACE inhibitors rarely produce headaches in either group. Note that the schema supports inferences about the object-level models in Fig. 1A. For example, because Alice is an A-person and Doxazosin is an alpha-blocker, the schema predicts that ingesting Doxazosin will cause Alice to experience headaches.

To explore how causal schemata are learned and used to guide inferences about object-level models, we work within a hierarchical Bayesian framework. The diagram in Fig. 1A can be transformed into a hierarchical Bayesian model by specifying how the information at each level is generated given the information at the level immediately above. We must therefore specify how the event data are generated given the object-level models, how the object-level models are generated given the category-level models, and how the category-level models are generated given a set of categories.

Although our framework is formalized as a top-down generative process, we will use Bayesian inference to invert this process and carry out bottom-up inference. In particular, we will focus on problems where event data are observed at the bottom level and the learner must simultaneously learn the object-level causal models, the category-level causal models and the categories that occupy the upper levels. After observing event data at the bottom level, our probabilistic model computes a posterior distribution over the representations at the upper levels, and our working assumption is that the categories and causal models learned by people are those assigned maximum posterior probability by our model. We do not discuss psychological mechanisms that might allow humans to identify the representations with maximum posterior probability, but future work can explore how the computations required by our model can be implemented or approximated by psychologically plausible mechanisms.

Although it is natural to say that the categories and causal models are learned from the event data available at the bottom level of Fig. 1A, note that this achievement relies on several kinds of background knowledge. We assume that the learner already knows about the relevant domains (e.g., people and drugs) and events (e.g., ingestion and headache events) and is attempting to solve a problem that is well specified at the domain level (e.g., the problem of deciding whether ingesting drugs can cause headaches). We also assume that the existence of the hierarchy in Fig. 1A is known in advance. In other words, our framework knows from the start that it should search for *some* set of categories and *some* set of causal models at the category and object levels, and learning is a matter of finding the candidates that best account for the data. We return to the question of background knowledge in the General Discussion and consider the extent to which some of this knowledge might be the outcome of prior learning.

We have focused on the drugs and headaches scenario so far, but the same hierarchical approach should be relevant to many different settings. Suppose that we are interested in the relationship between touching a plant and subsequently developing a rash. In this case the domain-level problem can be formulated as

$$\texttt{touches}(\texttt{person}, \texttt{plant}) \overset{?}{\rightarrow} \texttt{rash}(\texttt{person})$$

We may notice that only certain plants produce rashes, and that only certain people are susceptible to rashes. A schema consistent with this idea is shown in Fig. 1B. There are two categories of plants (allergens and nonallergens), and two categories of people (allergic and nonallergic). Allergic people develop rashes after touching allergenic plants, including poison oak, poison ivy, and poison sumac. Allergic people, however, do not develop rashes after touching nonallergenic plants, and nonallergic people never develop rashes after touching plants.

As a third motivating example, suppose that we are interested in social relationships among the children and the counselors at a summer camp. In particular, we would like to predict whether a given child will become angry with a given counselor if that counselor gives her a certain kind of order. The domain-level problem for this setting is:

$$\texttt{orders}(\texttt{counselor}, \texttt{child}, \texttt{order}) \overset{?}{\rightarrow} \texttt{angry}(\texttt{child}, \texttt{counselor})$$

One possible schema for this setting is shown in Fig. 1C. There are two categories of counselors (popular and unpopular), two categories of children (polite and rebellious), and two categories of orders (fair and unfair). Rebellious children may become angry with a counselor if that counselor gives them any kind of order. Polite children accept fair orders

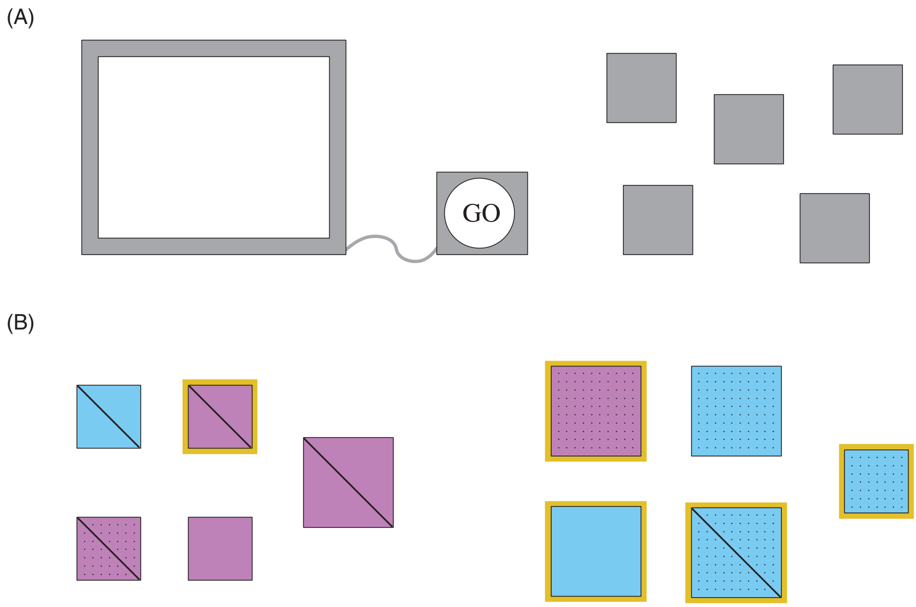


Fig. 2. Stimuli used in our experiments. (A) A machine and some blocks. The blocks can be placed inside the machine and the machine sometimes activates (flashes yellow) when the GO button is pressed. The blocks used for each condition of Experiments 1, 2, and 4 were perceptually indistinguishable. (B) Blocks used for Experiment 3. The blocks are grouped into two family resemblance categories: blocks on the right tend to be large, blue, and spotted, and tend to have a gold boundary but no diagonal stripe. These blocks are based on stimuli created by Sakamoto and Love (2004).

from popular counselors, but may become angry if a popular counselor gives them an unfair order or if an unpopular counselor gives them any kind of order.

Our experiments will make use of a fourth causal setting. Consider the blocks and the machine in Fig. 2A. The machine has a GO button, and it will sometimes activate and flash yellow when the button is pressed. Each block can be placed inside the machine, and whether the machine is likely to activate might depend on which block is inside. The domain-level problem for this setting is:

$$\texttt{inside}(\texttt{block}, \texttt{machine}) \;\&\; \texttt{button\_pressed}(\texttt{machine}) \xrightarrow{?} \texttt{activate}(\texttt{machine})$$

Note that the event on the left-hand side is a compound event which combines a state (a block is inside the machine) and an action (the button is pressed). In general, both the left- and right-hand sides of a domain-level problem may specify compound events that are expressed using multiple predicates.

One schema for this problem might organize the blocks into two categories: *active* blocks tend to activate the machine on most trials, and *inert* blocks seem to have no effect on the machine. Note that the blocks and machine example is somewhat similar to the drugs and headaches example: Blocks and drugs play corresponding roles, machines and people play

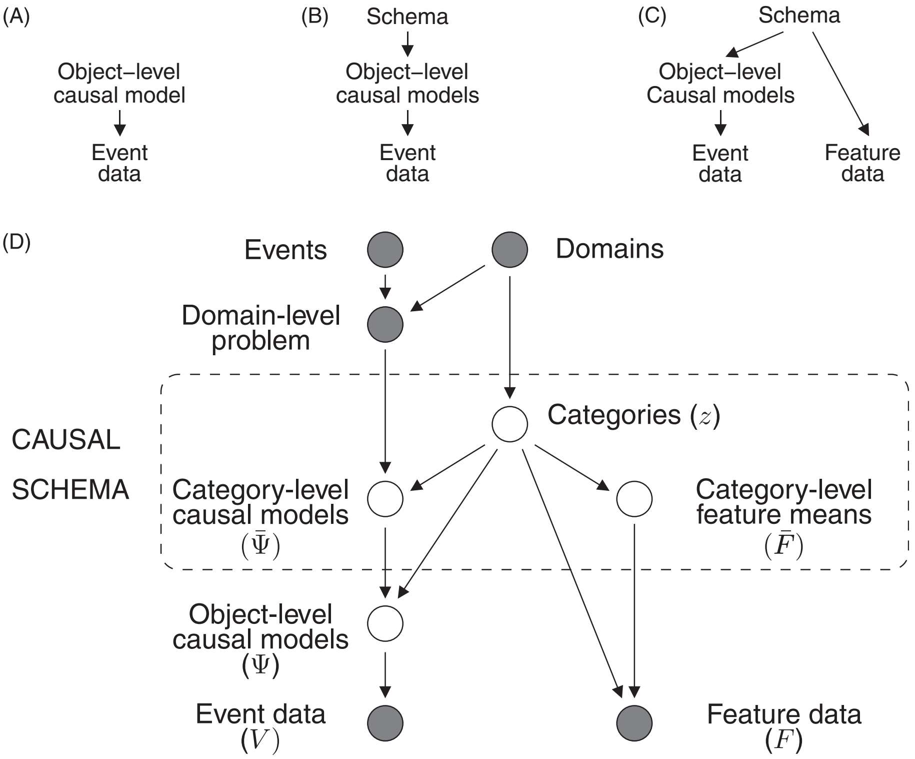

(A)
Object–level causal model
Event data
(B)
Schema
Object–level causal models
Event data
(C)
Schema
Object–level Causal models
Event data
Feature data
(D)
Events
Domains
Domain-level problem
CAUSAL SCHEMA
Categories (z)
Category-level causal models ($\bar{\Psi}$)
Category-level feature means ($\bar{F}$)
Object-level causal models (Ψ)
Event data (V)
Feature data (F)

Fig. 3. A hierarchical Bayesian approach to causal learning. (A) Learning a single object-level causal model. (B) Learning causal models for multiple objects. The schema organizes the objects into categories and specifies the causal powers of each category. (C) A generative framework for learning a schema that includes information about the characteristic features of each category. (D) A generative framework that includes (A)–(C) as special cases. Nodes represent variables or bundles of variables and arrows indicate dependencies between variables. Shaded nodes indicate variables that are observed or known in advance, and unshaded nodes indicate variables that must be inferred. We will collectively refer to the categories, the category-level causal models, and the category-level feature means as a causal schema. Note that the hierarchy in Fig. 1A is a subset of the complete model shown here.

corresponding roles, and the event of a machine activating corresponds to the event of a person developing a headache.

The next sections introduce our approach more formally and we develop our framework in several steps. We begin with the problem of learning a single object-level model—for example, learning whether ingesting Doxazosin causes Alice to develop headaches (Fig. 3A). We then turn to the problem of simultaneously learning multiple object-level models (Fig. 3B) and show how causal schemata can help in this setting. We next extend our framework to handle problems where the objects of interest (e.g., people and drugs) have perceptual features that may be correlated with their categories (Fig. 3C). Our final analysis addresses problems where multiple members of the same domain may interact to produce an effect—for example, two drugs may produce a headache when paired although neither causes headaches in isolation.

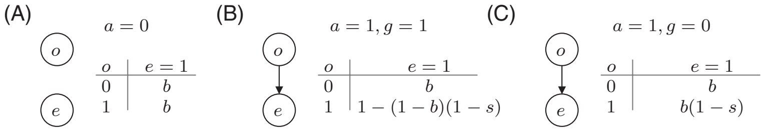


Fig. 4. Causal graphical models that capture three possible relationships between a cause o and an effect e. Variable a indicates whether there is a causal relationship between o and e, variable g indicates whether this relationship is generative or preventive, and variable s indicates the strength of this relationship. A generative background cause of strength b is always present.

Although we develop our framework in stages and consider several increasingly sophisticated models along the way, the result is a single probabilistic framework that addresses all of the problems we discuss. The framework is shown as a graphical model in Fig. 3D. Each node represents a variable or bundle of variables, and some of the nodes have been annotated with variable names that will be used in later sections of the paper. Arrows between nodes indicate dependencies—for example, the top section of the graphical model indicates that a domain-level problem such as

$$\texttt{ingests}(\texttt{person}, \texttt{drug}) \overset{?}{\rightarrow} \texttt{headache}(\texttt{person})$$

is formulated in terms of domains (people and drugs) and events (ingests(·,·) and headache(·)). Shaded nodes indicate variables that are observed (e.g., the event data) or specified in advance (e.g., the domain-level problem), and the unshaded nodes indicate variables that must be learned. Note that the three models in Fig. 3A–C correspond to fragments of the complete model in Fig. 3D, and we will build up the complete model by considering these fragments in sequence.

3. Learning a single object-level causal model

We begin with the problem of elemental causal induction (Griffiths & Tenenbaum, 2005) or the problem of learning a causal model for a single object-level problem. Our running example will be the problem

$$\texttt{ingests}(\texttt{Alice}, \texttt{Doxazosin}) \overset{?}{\rightarrow} \texttt{headache}(\texttt{Alice})$$

where the cause event indicates whether Alice takes Doxazosin and the effect event indicates whether she subsequently develops a headache. Let o refer to the object Doxazosin, and we overload our notation so that o can also refer to the cause event ingests(Alice, Doxazosin). Let e refer to the effect event headache(Alice).

Suppose that we have observed a set of trials where each trial indicates whether or not cause event o occurs, and whether or not the effect e occurs. Data of this kind are often called contingency data, but we refer to them as event data V. We assume that the outcome of each trial is generated from an object-level causal model M that captures the causal relationship between o and e (Fig. 5). Having observed the trials in V, our beliefs about the causal model can be summarized by the posterior distribution $P(M|V)$:

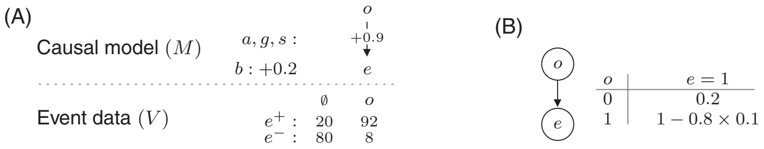


Fig. 5. (A) Learning an object-level causal model *M* from event data *V* (see Fig. 3A). The event data specify the number of times the effect was (e^+) and was not (e^-) observed when *o* was absent (∅) and when *o* was present. The model *M* shown has $a = 1$, $g = 1$, $s = 0.9$, and $b = 0.2$, and it is a compact representation of the graphical model in (B).

$$P(M \mid V) \propto P(V \mid M)P(M). \tag{4}$$

The likelihood term $P(V \mid M)$ indicates how compatible the event data *V* are with model *M*, and the prior $P(M)$ captures prior beliefs about model *M*.

We parameterize the causal model *M* using four causal variables (Figs. 4 and 5). Let *a* indicate whether there is an arrow joining *o* and *e*, and let *g* indicate the polarity of this causal relationship ($g = 1$ if *o* is a generative cause and $g = 0$ if *o* is a preventive cause). Suppose that *s* is the strength of the relationship between *o* and *e*.[1] To capture the possibility that *e* will be present even though *o* is absent, we assume that a generative background cause of strength *b* is always present. We specify the distribution $P(e \mid o)$ by assuming that generative and preventive causes combine according to a network of noisy-OR and noisy-AND-NOT gates (Glymour, 2001).

Now that we have parameterized model *M* in terms of the triple (*a*,*g*,*s*) and the background strength *b*, we can rewrite Eq. 4 as

$$P(a, g, s, b \mid V) \propto P(V \mid a, g, s, b)P(a)P(g)P(s)P(b). \tag{5}$$

To complete the model we must place prior distributions on the four causal variables. We use uniform priors on the two binary variables (*a* and *g*), and we use priors $P(s)$ and $P(b)$ that capture the expectation that *b* will be small and *s* will be large. These priors on *s* and *b* are broadly consistent with the work of Lu, Yuille, Liljeholm, Cheng and Holyoak (2008), who suggest that learners typically expect causes to be necessary (*b* should be low) and sufficient (*s* should be high). Complete specifications of $P(s)$ and $P(b)$ are provided in Appendix A.

To discover the causal model *M* that best accounts for the events in *V*, we can search for the causal variables with maximum posterior probability according to Eq. 5. There are many empirical studies that explore human inferences about a single potential cause and a single effect, and previous researchers (Griffiths & Tenenbaum, 2005; Lu et al., 2008) have shown that a Bayesian approach similar to ours can account for many of these inferences. Here, however, we turn to the less-studied case where people must learn about many objects, each of which may be causally related to the effect of interest.

4. Learning multiple object-level models

Suppose now that we are interested in simultaneously learning multiple object-level causal models. For example, suppose that our patient Alice has prescriptions for many different drugs and we want to learn about the effect of each drug:

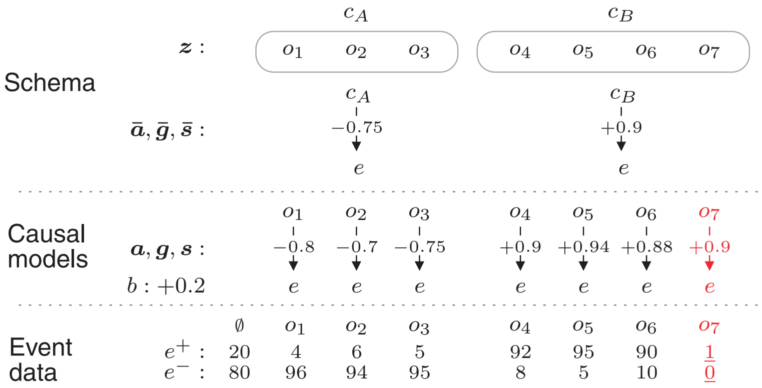
$$\begin{aligned}\texttt{ingests}(\texttt{Alice},\texttt{Doxazosin}) &\overset{?}{\rightarrow} \texttt{headache}(\texttt{Alice})\\ \texttt{ingests}(\texttt{Alice},\texttt{Prazosin}) &\overset{?}{\rightarrow} \texttt{headache}(\texttt{Alice})\\ \texttt{ingests}(\texttt{Alice},\texttt{Terazosin}) &\overset{?}{\rightarrow} \texttt{headache}(\texttt{Alice})\\ &\vdots\end{aligned}$$

For now we assume that Alice takes at most one drug per day, but later we relax this assumption and consider problems where patients take multiple drugs and these drugs may interact. We refer to the ith drug as object o_i, and as before we overload our notation so that o_i can also refer to the cause event `ingests(Alice,` o_i).

Our goal is now to learn a set $\{M_i\}$ of causal models, one for each drug (Figs. 3b and 6). There is a triple (a_i,g_i,s_i) describing the causal model for each drug o_i, and we organize these variables into three vectors, $\boldsymbol{a}$, $\boldsymbol{g}$, and $\boldsymbol{s}$. Let Ψ be the tuple $(\boldsymbol{a}, \boldsymbol{g}, \boldsymbol{s}, b)$ which includes all the parameters of the causal models. As before, we assume that a generative background cause of strength b is always present.

One strategy for learning multiple object-level models is to learn each model separately using the methods described in the previous section. Although simple, this strategy will not succeed in learning to learn because it does not draw on experience with previous objects when learning a causal model for a novel object that is sparsely observed. We will allow information to be shared across causal models for different objects by introducing the notion of a causal schema. A schema specifies a grouping of the objects into categories and includes category-level causal models which specify the causal powers of each category. The schema in Fig. 6 indicates that there are two categories: objects belonging to category c_A tend to prevent the effect and objects belonging to category c_B tend to cause the effect. The strongest possible assumption is that all members of a category must play identical causal roles. For example, if Doxazosin and Prazosin belong to the same category, then the causal models for these two drugs should be identical. We relax this strong assumption and assume instead that members of the same category play similar causal roles. More precisely, we assume that the object-level models corresponding to a given category-level causal model are drawn from a common distribution.

Formally, let z_i indicate the category of o_i, and let $\bar{\boldsymbol{a}}$, $\bar{\boldsymbol{g}}$, $\bar{\boldsymbol{s}}$, and $\bar{b}$ be schema-level analogs of $\boldsymbol{a}$, $\boldsymbol{g}$, $\boldsymbol{s}$, and b. Variable $\bar{a}(c)$ is the probability that any given object belonging to category c will be causally related to the effect, variables $\bar{g}(c)$ and $\bar{s}(c)$ specify the expected polarity and causal strength for objects in category c, and variable $\bar{b}$ specifies the expected strength of the generative background cause. Even though $\boldsymbol{a}$ and $\boldsymbol{g}$ are vectors of probabilities, Fig. 6 simplifies by showing each $\bar{a}(c)$ and $\bar{g}(c)$ as a binary variable. To generate a causal model for each object, we assume that each arrow variable a_i is generated by tossing a coin with



	$\emptyset$	o_1	o_2	o_3	o_4	o_5	o_6	o_7
e^+ :	20	4	6	5	92	95	90	1
e^- :	80	96	94	95	8	5	10	0

Fig. 6. Learning a schema and a set of object-level causal models (see Fig. 3B). z specifies a set of categories, where objects belonging to the same category have similar causal powers, and $\bar{\boldsymbol{a}}$, $\bar{\boldsymbol{g}}$, and $\bar{\boldsymbol{s}}$ specify a set of category-level causal models. Note that the schema supports inferences about an object (o_7, counts underlined in red) that is very sparsely observed.

weight $\bar{a}(z_i)$, that each polarity g_i is generated by tossing a coin with weight $\bar{g}(z_i)$, and that each strength s_i is drawn from a distribution parameterized by $\bar{s}(z_i)$. Let $\bar{\Psi}$ be a tuple $(\bar{\boldsymbol{a}}, \bar{\boldsymbol{g}}, \bar{\boldsymbol{s}}, \bar{b})$ that includes all parameters of the causal schema. A complete description of each parameter is provided in Appendix A.

Now that the generative approach in Fig. 1A has been fully specified we can use it to learn the category assignments z, the category-level models $\bar{\Psi}$, and the object-level models Ψ that are most probable given the events V that have been observed:

$$P(z, \bar{\Psi}, \Psi \mid V) \propto P(V \mid \Psi) P(\Psi \mid \bar{\Psi}, z) P(\bar{\Psi} \mid z) P(z). \tag{6}$$

The distribution $P(V \mid \Psi)$ is defined by assuming that the contingency data for each object-level model are generated in the standard way from that model. The distribution $P(\Psi \mid \bar{\Psi}, z)$ specifies how the model parameters Ψ are generated given the category-level models $\bar{\Psi}$ and the category assignments z. To complete the model, we need prior distributions on z and the category-level models $\bar{\Psi}$. Our prior $P(z)$ assigns some probability mass to all possible partitions but favors partitions that use a small number of categories. Our prior $P(\bar{\Psi} \mid z)$ captures the expectation that generative and preventive causes are equally likely a priori, that causal strengths are likely to be high, and that the strength of the background cause is likely to be low. Full details are provided in Appendix A.

Fig. 6 shows how a schema and a set of object-level causal models (top two levels) can be simultaneously learned from the event data V in the bottom level. All of the variables in the figure have been set to values with high posterior probability according to Eq. 6: for instance, the partition z shown is the partition with maximum posterior probability. Note that learning a schema allows a causal model to be learned for object o_7, which is very sparsely observed (see the underlined entries in the bottom level of Fig. 6). On its own, a single trial might not be very informative about the causal powers of this object, but experience with previous objects allows the model to predict that o_7 will produce the effect about as regularly as the other members of category c_B.

To compute the predictions of our model we used Markov chain Monte Carlo methods to sample from the posterior distribution in Eq. 6. A more detailed description of this inference algorithm is provided in Appendix A, but note that this algorithm is not intended as a model of psychological processing. The primary contribution of this section is the computational theory summarized by Eq. 6, and there will be many ways in which the computations required by this theory can be approximately implemented.

5. Experiments 1 and 2: Testing the basic schema-learning model

Our schema-learning model attempts to satisfy two criteria when learning about the causal powers of a novel object. When information about the new object is sparse, predictions about this object should be based primarily on experience with previous objects. Relying on past experience will allow the model to go beyond the sparse and noisy observations that are available for the novel object. Given many observations of the novel object, however, the model should rely heavily on these observations and should tend to ignore its observations of previous objects. Discounting past experience in this way will allow the model to be flexible if the new object turns out to be different from all previous objects.

Our first two experiments explore this tradeoff between conservatism and flexibility. Both experiments used blocks and machines like the examples in Fig. 2. As mentioned already, the domain-level problem for this setting is:

$$\texttt{inside}(\texttt{block}, \texttt{machine})\ \texttt{\&}\ \texttt{button\_pressed}(\texttt{machine}) \xrightarrow{?} \texttt{activate}(\texttt{machine})$$

In terms of the notation we have been using, each block is an object o_i, each button press corresponds to a trial, and the effect e indicates whether the machine activates on a given trial.

Experiment 1 studied people's ability to learn a range of different causal schemata from observed events and to use these schemata to rapidly learn about the causal powers of new, sparsely observed objects. To highlight the influence of causal schemata, inferences about each new object were made after observing at most one trial where the object was placed inside the machine. Across conditions, we varied the information available during training, and our model predicts that these different sets of observations should lead to the formation of qualitatively different schemata, and hence to qualitatively different patterns of inference about new, sparsely observed objects.

Experiment 2 explores in more detail how observations of a new object interact with a learned causal schema. Instead of observing a single trial for each new object, participants now observed seven trials and judged the causal power of the object after each one. Some sequences of trials were consistent with the schema learned during training, but others were inconsistent. Given these different sequences, our model predicts how and when learners should overrule their schema-based expectations when learning a causal model for a novel object. These predicted learning curves—or ‘‘unlearning curves’’—were tested against the responses provided by participants.

Our first two experiments were also designed in part to evaluate our model relative to other computational accounts. Although we know of no previous model that attempts to capture causal learning at multiple levels of abstraction, we will consider some simple models inspired by standard models in the categorization literature. These models are discussed and analyzed following the presentation of Experiments 1 and 2.

5.1. Experiment 1: One-shot causal learning

Experiment 1 explores whether participants can learn different kinds of schemata and use these schemata to rapidly learn about the causal powers of new objects. In each condition of the experiment, participants initially completed a training phase where they placed each of eight objects into the machine multiple times and observed whether the machine activated on each trial. In different conditions they observed different patterns of activation across these training trials. In each condition, the activations observed were consistent with the existence of one or two categories, and these categories had qualitatively different causal powers across the different conditions. After each training phase, participants completed a ‘‘one-shot learning’’ task, where they made predictions about test blocks after seeing only a single trial involving each block.

5.1.1. Participants

Twenty-four members of the MIT community were paid for participating in this experiment.

5.1.2. Stimuli

The experiment used a custom-built graphical interface that displayed a machine and some blocks (Fig. 2A). Participants could drag the blocks around, and they were able to place up to one block inside the machine at a time. Participants could also interact with the machine by pressing the GO button and observing whether the machine activated. The blocks used for Experiments 1 and 2 were perceptually indistinguishable, and their causal powers could therefore not be predicted on the basis of their physical appearance.

5.1.3. Design

The experiment includes four within-participant conditions and the training data for each condition are summarized in Fig. 7. The first condition ($p = \{0, 0.5\}$) includes two categories of blocks: blocks in the first category never activate the machine, and blocks in the second category activate the machine about half the time. The second condition ($p = \{0.1, 0.9\}$) also includes two categories: blocks in the first category rarely activate the machine, and blocks in the second category usually activate the machine. The remaining conditions each include only one category of blocks: blocks in the third condition ($p = 0$) never activate the machine, and blocks in the fourth condition ($p = 0.1$) activate the machine rarely.

Condition		Training data								
		$\emptyset$	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8
$p = \{0, 0.5\}$	e^+ :	0	0	0	0	0	5	4	6	1
	e^- :	10	10	10	10	1	5	6	4	0
		$\emptyset$	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8
$p = \{0.1, 0.9\}$	e^+ :	0	1	2	1	0	9	8	9	1
	e^- :	10	9	8	9	1	1	2	1	0
		$\emptyset$	o_1	o_2	o_3	o_4	o_5	o_6		
$p = 0$	e^+ :	0	0	0	0	0	0	0		
	e^- :	10	10	10	10	10	10	10		
		$\emptyset$	o_1	o_2	o_3	o_4	o_5	o_6		
$p = 0.1$	e^+ :	0	1	2	1	2	1	2		
	e^- :	10	9	8	9	8	9	8		

Fig. 7. Training data for the four conditions of Experiment 1. In each condition, the first column of each table shows that the empty machine fails to activate on each of the 10 trials. Each remaining column shows the outcome of one or more trials when a single block is placed inside the machine. For example, in the $p = \{0, 0.5\}$ condition block o_1 is placed in the machine 10 times and fails to activate the machine on each trial.

5.1.4. Procedure

At the start of each condition, participants are shown an empty machine and asked to press the GO button 10 times. The machine fails to activate on each occasion. One by one the training blocks are introduced, and participants place each block in the machine and press the GO button one or more times. The outcomes of these trials are summarized in Fig. 7. For example, the $p = \{0, 0.5\}$ condition includes eight training blocks in total, and the block shown as o_1 in the table fails to activate the machine on each of 10 trials. After the final trial for each block, participants are asked to imagine pressing the GO button 100 times when this block is inside the machine. They then provide a rating which indicates how likely it is that the total number of activations will fall between 0 and 20. All ratings are provided on a seven-point scale where one is labeled as ''very unlikely,'' seven is labeled as ''very likely,'' and the other values are left unlabeled. Ratings are also provided for four other intervals: between 20 and 40, between 40 and 60, between 60 and 80, and between 80 and 100. Each block remains on screen after it is introduced, and by the end of the training phase six or eight blocks are therefore visible onscreen. After the training phase two test blocks are introduced, again one at a time. Participants provide ratings for each block before it has been placed in the machine, and after a single trial. One of the test blocks (o^+) activates the machine on this trial, and the other (o^-) does not.

The set of four conditions is designed to test the idea that inductive constraints and inductive flexibility are both important. The first two conditions test whether experience with the training blocks allows people to extract constraints that are useful when learning about the causal powers of the test blocks. Conditions three and four explore cases where these constraints need to be overruled. Note that test block o^+ is surprising in these conditions because the training blocks activate the machine rarely, if at all.

To encourage participants to think about the conditions separately, machines and blocks of different colors were used for each condition. Note, however, that the blocks within each condition were always perceptually identical. The order in which the conditions were presented was counterbalanced according to a Latin square design. The order of the training blocks and the test blocks within each condition was also randomized subject to several constraints. First, the test blocks were always presented after the training blocks. Second, in conditions $p = \{0, 0.5\}$ and $p = \{0.1, 0.9\}$ the first two training blocks in the sequence always belonged to different categories, and the two sparsely observed training blocks (o_4 and o_8) were always the third and fourth blocks in the sequence. Finally, in the $p = 0$ condition test block o^+ was always presented second, because this block is unlike any of the training blocks and may have had a large influence on predictions about any block which followed it.

5.1.5. Model predictions

Fig. 8 shows predictions when the schema-learning model is applied to the data in Fig. 7. Each plot shows the posterior distribution on the activation strength of a test block: the probability $P(e \mid o)$ that the block will activate the machine on a given trial. Because the background rate is zero, this distribution is equivalent to a distribution on the causal power

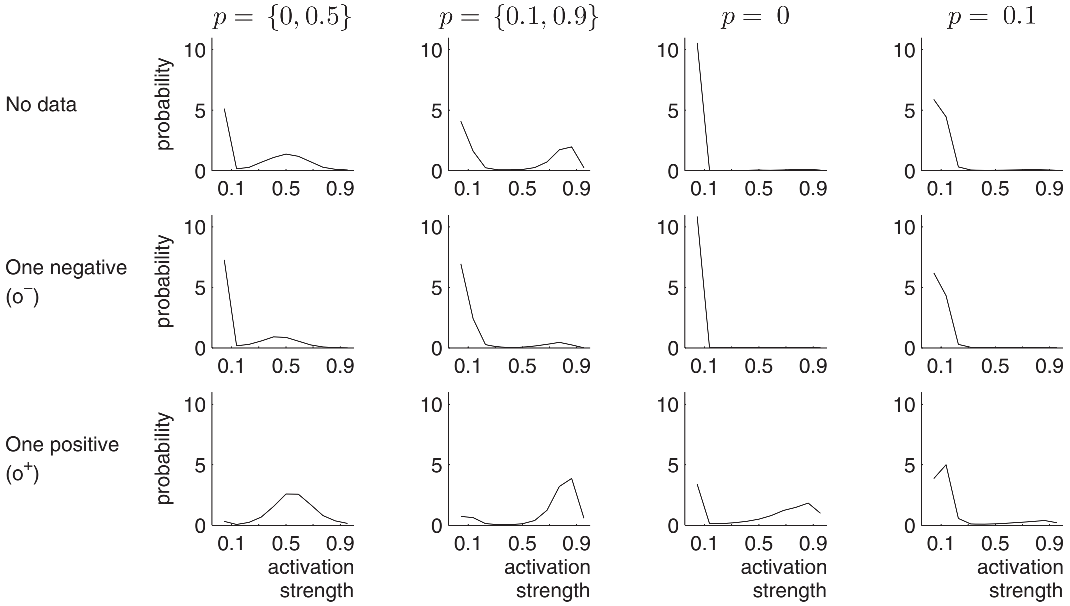


Fig. 8. Predictions of the schema-learning model for Experiment 1. Each subplot shows the posterior distribution on the activation strength of a test block. There are three predictions for each condition: The first row shows inferences about a test block before this block has been placed in the machine, and the remaining rows show inferences after a single negative (o^-) or positive (o^+) trial is observed. Note that the curves represent probability density functions and can therefore attain values greater than 1.

(Cheng, 1997) of the test block. Recall that participants were asked to make predictions about the number of activations expected across 100 trials. If we ask our model to make the same predictions, the distributions on the total number of activations will be discrete distributions with shapes similar to the distributions in Fig. 8.

The plots in the first row show predictions about a test block before it is placed in the machine. The first plot indicates that the model has discovered two causal categories, and it expects that the test block will activate the machine either very rarely or around half of the time. The two peaks in the second plot again indicate that the model has discovered two causal categories, this time with strengths around 0.1 and 0.9. The remaining two plots are unimodal, suggesting that only one causal category is needed to explain the data in each of the $p = 0$ and $p = 0.1$ conditions.

The plots in the second row show predictions about a test block (o^-) that fails to activate the machine on one occasion. All of the plots have peaks near 0 or 0.1. Because each condition includes blocks that activate the machine rarely or not at all, the most likely hypothesis is always that o^- is one of these blocks. Note, however, that the first plot has a small bump near 0.5, indicating that there is some chance that test block o^- will activate the machine about half of the time. The second plot has a small bump near 0.9 for similar reasons.

The plots in the third row show predictions about a test block (o^+) that activates the machine on one occasion. The plot for the first condition peaks near 0.5, which is consistent with the hypothesis that blocks which activate the machine at all tend to activate it around half the time. The plot for the second condition peaks near 0.9, which is consistent with the observation that some training blocks activated the machine nearly always. The plot for the third condition has peaks near 0 and near 0.9. The first peak captures the idea that the test block might be similar to the training blocks, which activated the machine very rarely. Given that none of the training blocks activated the machine, one positive trial is enough to suggest that the test block might be qualitatively different from all previous blocks, and the second peak captures this hypothesis. The curve for the final condition peaks near 0.1, which is the frequency with which the training blocks activated the machine.

5.1.6. Results

The four columns of Fig. 9 show the results for each condition. Each participant provided ratings for five intervals in response to each question, and these ratings can be plotted as a curve. Fig. 9 shows the mean curve for each question. The first row shows predictions before a test block has been placed in the machine (responses for test blocks o^- and o^+ have been combined). The second and third rows show predictions after a single trial for test blocks o^- and o^+.

The first row provides a direct measure of what participants have learned during the training for each condition. Note first that the plots for the four rows are rather different, suggesting that the training observations have shaped people's expectations about novel blocks. A two-factor ANOVA with repeated measures supports this conclusion, and it indicates that there are significant main effects of interval [$F(4,92) = 31.8$, $p < .001$] and condition [$F(3,69) = 15.7$, $p < .001$] but no significant interaction between interval and condition [$F(12,276) = 0.74$, $p > .5$].

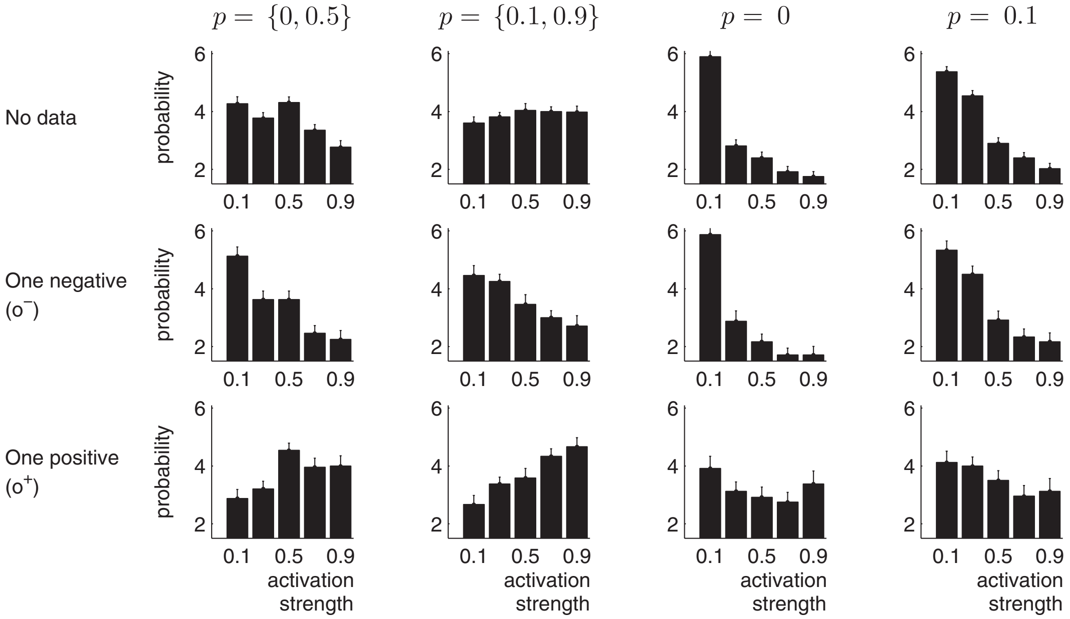


Fig. 9. Results for the four conditions in Experiment 1. Each subplot shows predictions about a new object that will undergo 100 trials, and each bar indicates the probability that the total number of activations will fall within a certain interval. The x-axis shows the activation strengths that correspond to each interval and the y-axis shows probability ratings on a scale from one (very unlikely) to seven (very likely). All plots show mean responses across 24 participants. Error bars for this plot and all remaining plots show the standard error of the mean.

In three of the four conditions, the human responses in the top row of Fig. 9 are consistent with the model predictions in Fig. 8. As expected, the curves for the $p = 0$ and $p = 0.1$ conditions indicate an expectation that the test blocks will probably fail to activate the machine. The curve for the $p = \{0, 0.5\}$ condition peaks in the same places as the model prediction, suggesting that participants expect that each test block will either activate the machine very rarely or about half of the time. The first (0.1) and third (0.5) bars in the plot are both greater than the second (0.3) bar, and paired sample t tests indicate that both differences are statistically significant ($p < .05$, one-tailed). The $p = \{0, 0.5\}$ curve is therefore consistent with the idea that participants have discovered two categories.

The responses for the $p = \{0.1, 0.9\}$ condition provide no evidence that participants have discovered two causal categories. The curve for this condition is flat or unimodal and does not match the bimodal curve predicted by the model. One possible interpretation is that learners cannot discover categories based on probabilistic causal information. As suggested by the $p = \{0, 0.5\}$ condition, learners might distinguish between blocks that never produce the effect and those that sometimes produce the effect, but not between blocks that produce the effects with different strengths. A second possible interpretation is that learners can form categories based on probabilistic information but require more statistical evidence than we provided in Experiment 1. Our third experiment supports this second interpretation and demonstrates that learners can form causal categories on the basis of probabilistic evidence.

Consider now the third row of Fig. 9, which shows predictions about a test block (o^+) that has activated the machine exactly once. As before, the differences between these plots suggest that experience with previous blocks shapes people's inferences about a sparsely observed novel block. A two-factor ANOVA with repeated measures supports this conclusion, and indicates that there is no significant main effect of interval [$F(4,92) = .46$, $p > .5$], but that there is a significant main effect of condition [$F(3,69) = 4.20$, $p < .01$] and a significant interaction between interval and condition [$F(12,276) = 6.90$, $p < .001$]. Note also that all of the plots in the third row peak in the same places as the curves predicted by the model (Fig. 8A). For example, the middle (0.5) bar in the $p = \{0, 0.5\}$ condition is greater than the bars on either side, and paired sample t tests indicate that both differences are statistically significant ($p < .05$, one-tailed). The plot for the $p = 0$ condition provides some support for a second peak near 0.9, although a paired-sample t test indicates that the difference between the fifth (0.9) and fourth (0.7) bars is only marginally significant ($p < .1$, one-tailed). Our second experiment explores this condition in more detail, and it establishes more conclusively that a single positive observation can be enough for a learner to decide that a block is different from all previously observed blocks.

Consider now the second row of Fig. 9, which shows predictions about a test block (o^-) that has failed to activate the machine exactly once. The plots in this row are all decaying curves, because each condition includes blocks that activate the machine rarely or not at all. Again, though, the differences between the curves are interpretable and match the predictions of the model. For instance, the $p = 0$ curve decays more steeply than the others, which makes sense because the training blocks for this condition never activate the machine. In particular, note that the difference between the first (0.1) and second (0.3) bars is greater in the $p = 0$ condition than the $p = 0.1$ condition ($p < .001$, one-tailed).

Although our primary goal in this paper is to account for the mean responses to each question, the responses of individual participants are also worth considering. Kemp (2008) presents a detailed analysis of individual responses and shows that in all cases except one the shape of the mean curve is consistent with the responses of some individuals. The one exception is the o^+ question in the $p = 0$ condition, where no participant generated a U-shaped curve, although some indicated that o^+ is unlikely to activate the machine and others indicated that o^+ is very likely to activate the machine on subsequent trials. This disagreement suggests that the $p = 0$ condition deserves further attention, and our second experiment explores this condition in more detail.

5.2. Experiment 2: Discovering new causal categories

Causal schemata support inferences about new objects that are sparsely observed, but sometimes these inferences are wrong and will have to be overruled when a new object turns out to be qualitatively different from all previous objects. Experiment 1 provided some suggestive evidence that human learners will overrule a schema when necessary. In the $p = 0$ condition, participants observed six blocks that never activated the machine, then saw a single trial where a new block (o^+) activated the machine. The results in Fig. 9 suggest that some participants inferred that the new block might be qualitatively different from the previous blocks. This finding suggests that a single observation of a new object is sometimes enough to overrule expectations based on many previous objects, but several trials may be required before learners are confident that a new object is unlike any of the previous objects. To explore this idea, Experiment 2 considers two cases where participants receive increasing evidence that a new object is different from all previously encountered objects.

5.2.1. Participants

Sixteen members of the MIT community were paid for participating in this experiment.

5.2.2. Design and procedure

The experiment includes two within-participant conditions ($p = 0$ and $p = 0.1$) that correspond to conditions 3 and 4 of Experiment 1. Each condition is very similar to the corresponding condition from Experiment 1 except for two changes. Seven observations are now provided for the two test blocks: for test block o^-, the machine fails to activate on each trial, and for test block o^+ the machine activates on all test trials except the second. Participants rate the causal strength of each test block after each trial and also provide an initial rating before any trials have been observed. As before, participants are asked to imagine placing the test block in the machine 100 times, but instead of providing ratings for five intervals they now simply predict the total number of activations out of 100 that they expect to see.

5.2.3. Model predictions

Fig. 10 shows the results when the schema-learning model is applied to the tasks in Experiment 2. In both conditions, predictions about the test blocks track the observations provided, and the curves rise after each positive trial and fall after each negative trial.

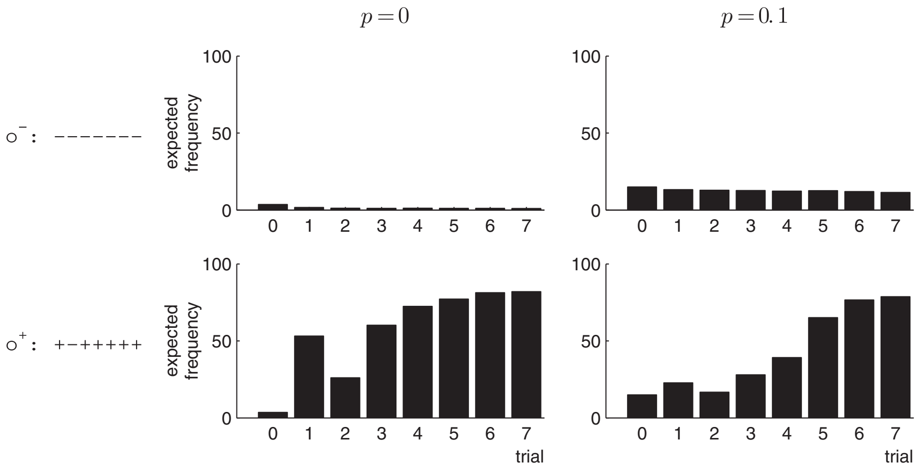


Fig. 10. Predictions of the schema-learning model for Experiment 2. A new block is introduced that is either similar (o^-) or different (o^+) from all previous blocks, and the trials for each block are shown on the left of the figure. Each plot shows how inferences about the causal power of the block change with each successive trial.

The most interesting predictions involve test block o^+, which is qualitatively different from all of the training blocks. The o^+ curves for both conditions attain similar values by the final prediction, but the curve for the $p = 0$ condition rises more steeply than the curve for the $p = 0.1$ condition. Because the training blocks in the $p = 0.1$ condition activate the machine on some occasions, the model needs more evidence in this condition before concluding that block o^+ is different from all of the training blocks.

The predictions about test block o^- also depend on the condition. In the $p = 0$ condition, none of the training blocks activates the machine, and the model predicts that o^- will also fail to activate the machine. In the $p = 0.1$ condition, each training block can be expected to activate the machine about 15 times out of 100. The curve for this condition begins at around 15, then gently decays as o^- repeatedly fails to activate the machine.

5.2.4. Results

Fig. 11 shows average learning curves across 16 participants. The curves are qualitatively similar to the model predictions, and as predicted the o^+ curve for the $p = 0$ condition rises more steeply than the corresponding curve for the $p = 0.1$ condition. Note that a simple associative account might predict the opposite result, because the machine in condition $p = 0.1$ activates more times overall than the machine in condition $p = 0$. To support our qualitative comparison between the o^+ curves in the two conditions, we ran a two-factor ANOVA with repeated measures. Because we expect that the $p = 0$ curve should be higher than the $p = 0.1$ curve from the second judgment onwards, we excluded the first judgment from each condition. There are significant main effects of condition [$F(1,15) = 6.11$, $p < .05$] and judgment number [$F(6,90) = 43.21$, $p < .01$], and a significant interaction between condition and judgment number [$F(6,90) = 2.67$, $p < .05$]. Follow-up paired-sample t tests indicate that judgments two through six are reliably greater in the $p = 0$ condition (in all

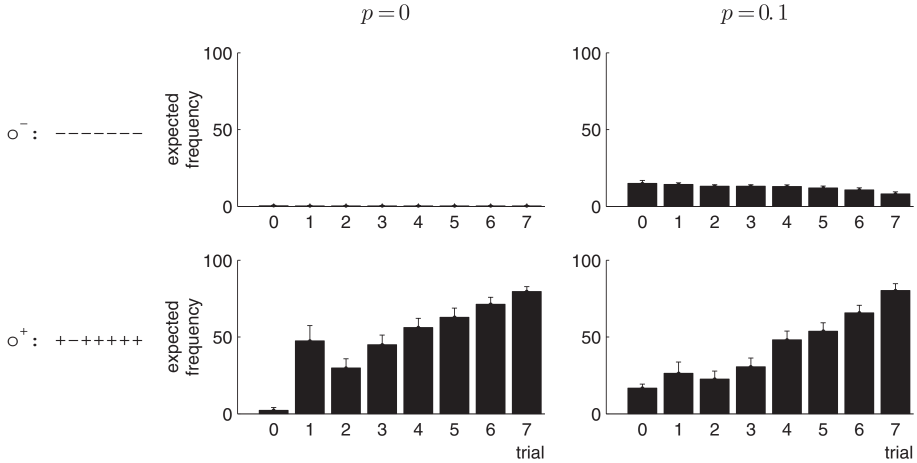


Fig. 11. Mean responses to Experiment 1. The average learning curves closely match the model predictions in Fig. 10.

cases $p < .05$, one-tailed), supporting the prediction that participants are quicker in the $p = 0$ condition to decide that block o^+ is qualitatively different from all previous blocks.

5.3. Alternative models

As mentioned already, our experiments explore the tradeoff between conservatism and flexibility. When a new object is sparsely observed, the schema-learning model assumes that this object is similar to previously encountered objects (Experiment 1). Once more observations become available, the model may decide that the new object is different from all previous objects and should therefore be assigned to its own category (Experiment 2). We can compare the schema-learning model to two alternatives: an *exemplar* model that is overly conservative, and a *bottom-up* model that is overly flexible. The exemplar model assumes that each new object is just like one of the previous objects, and the bottom-up model ignores all of its previous experience when making predictions about a new object.

We implemented the bottom-up model by assuming that the causal power of a test block is identical to its empirical power—the proportion of trials on which it has activated the machine. Predictions of this model are shown in Fig. 12. When applied to Experiment 1, the most obvious failing of the bottom-up model is that it makes identical predictions about all four conditions. Note that the model does not make predictions about the first row of Fig. 8A, because at least one test trial is needed to estimate the empirical power of a new block. When applied to Experiment 2, the model is unable to make predictions before any trials have been observed for a given object, and after a single positive trial the model leaps to the conclusion that test object o^+ will always activate the machine. Neither prediction matches the human data, and the model also fails to predict any difference between the $p = 0$ and $p = 0.1$ conditions.

We implemented the exemplar model by assuming that the causal power of each training block is identical to its empirical power, and that each test block is identical to one of the training blocks. The model, however, does not know which training block the test block will match, and it makes a prediction that considers the empirical powers of all training blocks, weighting each one by its proximity to the empirical power of the test block. Formally, the distribution d_n on the strength of a novel block is defined to be

$$d_n = \frac{\sum_i w_i d_i}{\sum_i w_i} \tag{7}$$

where d_i is the distribution for training block i, and is created by dividing the interval [0,1] into eleven equal intervals, setting $d_i(x) = 11$ for all values x that belong to the same interval as the empirical power of block i, and setting $d_i(x) = 0$ for all remaining values. Each weight w_i is set to $1 - \mid p_n - p_i \mid$, where p_n is the empirical power of the novel block and p_i is the empirical power of training block i. As Eq. 7 suggests, the exemplar model is closely related to exemplar models of categorization (Medin & Schaffer, 1978; Nosofsky, 1986).

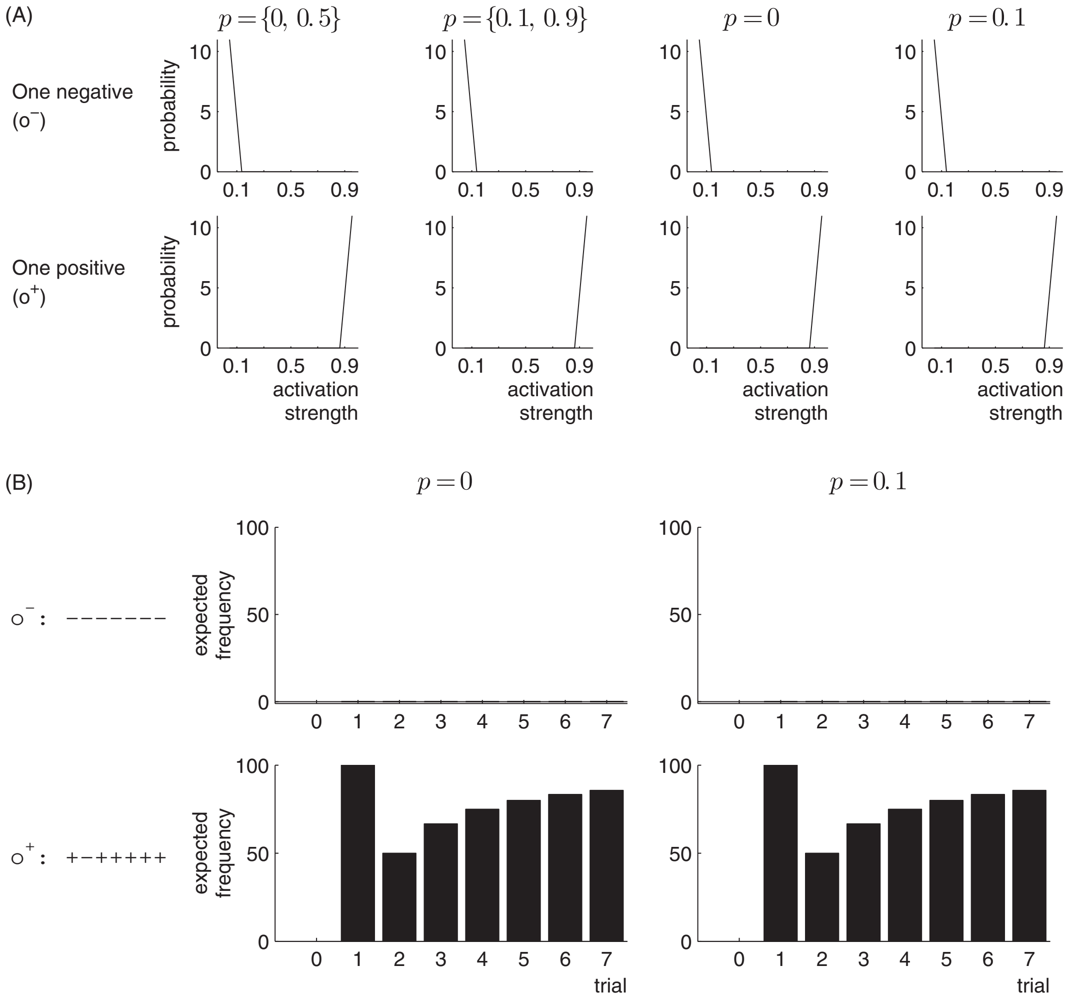


Fig. 12. Predictions of the bottom-up model for (A) Experiment 1 and (B) Experiment 2. In both cases the model fails to account for the differences between conditions.

Predictions of the exemplar model are shown in Fig. 13. The model accounts fairly well for the results of Experiment 1 but is unable to account for Experiment 2. Because the model assumes that test object o^+ is just like one of the training objects, it is unable to adjust when o^+ activates the machine more frequently than any previous object.

Overall, neither baseline model can account for our results. The bottom-up model is too quick to throw away observations of previous objects, and the exemplar model is unable to handle new objects that are qualitatively different from all previous objects. Other baseline models might be considered, but we are aware of no simple alternative that will account for all of our data.

Our first two experiments deliberately focused on a very simple setting where causal schemata are learned and used, but real-world causal learning is often more complex. The

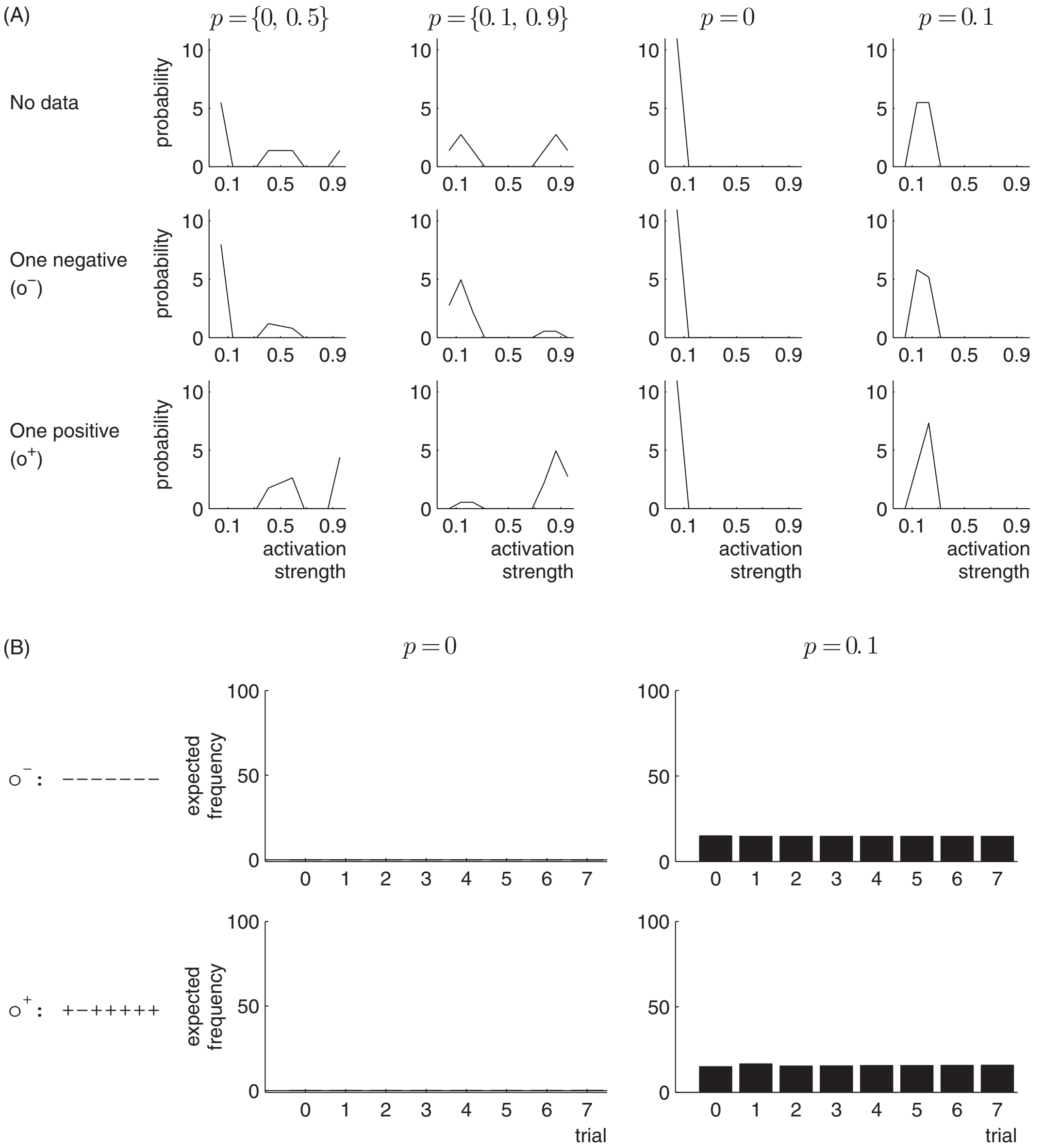


Fig. 13. Predictions of the exemplar model for (A) Experiment 1 and (B) Experiment 2. The model accounts fairly well for Experiment 1 but fails to realize that test block o^+ in Experiment 2 is qualitatively different from all previous blocks.

rest of the paper will address some of these complexities: in particular, we show how our framework can incorporate perceptual features and can handle contexts where causes interact to produce an effect.

6. Learning causal categories given feature data

Imagine that you are allergic to nuts, and that one day you discover a small white sphere in your breakfast cereal—a macadamia nut, although you do not know it. To discover the causal powers of this novel object you could collect some causal data—you could eat it and wait to see what happens. Probably, however, you will observe the features of the object, including its color, shape, and texture, and decide to avoid it because it is similar to other allergy-producing foods that you have encountered.

Our hierarchical Bayesian approach can readily handle the idea that members of a given category tend to have similar features in addition to similar causal powers (Figs. 3C and 14). Suppose that we have a matrix F which captures many features of the objects under consideration, including their sizes, shapes, and colors. We assume that objects belonging to the same category have similar features. For instance, the schema in Fig. 14 specifies that objects of category c_B tend to have features f_1 through f_4, but objects of category c_A tend not to have these features. Formally, let the schema parameters include a matrix $\bar{F}$, where $\bar{f}_j(c)$ specifies the expected value of feature f_j within category c (Fig. 3D). Building on previous models of categorization (Anderson, 1991), we assume that the value of f_j for object o_i is generated by tossing a coin with bias $\bar{f}_j(z_i)$. Our goal is now to use the features F along with the events V to learn a schema and a set of object-level causal models:

$$P(z, \bar{F}, \bar{\Psi}, \Psi \mid F, V) \propto P(F \mid \bar{F}, z)P(\bar{F} \mid z)P(V \mid \Psi)P(\Psi \mid \bar{\Psi}, z)P(\bar{\Psi} \mid z)P(z). \tag{8}$$

There are many previous models for discovering categories of objects with similar features (Anderson, 1991; Love, Medin, & Gureckis, 2004), and feature-based categorization is

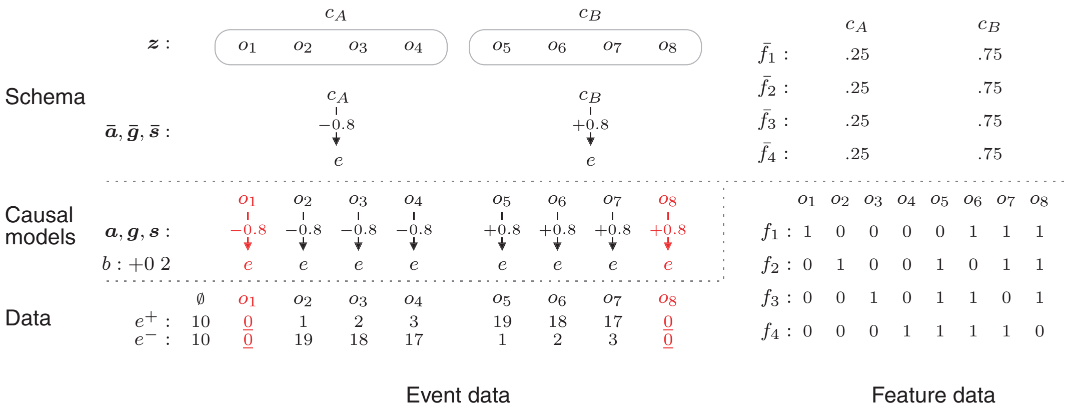


Fig. 14. Learning a schema and a set of object-level causal models given event and feature data (see Fig. 3C). Objects belonging to the same category have similar causal powers and similar features, and $\bar{f}_i$ specifies the expected value of feature f_i within each category. Note that the schema supports inferences about the causal powers of two objects (o_1 and o_8, counts underlined in red) that are very sparsely observed. The event and feature data shown are similar to the data used for Experiment 3.

sometimes pitted against causal categorization (Gopnik & Sobel, 2000). Our schema-learning model is based on the idea that real-world categories are often distinguished both by their characteristic features and their characteristic causal interactions. More often than not, one kind of information will support the categories indicated by the other, but there will also be cases where the causal data and the feature data conflict. In a later section we show how our framework can learn whether causal data or feature data provide the more reliable guide to category membership.

7. Experiment 3: Combining causal and feature data

Our first two experiments suggest that causal schemata allow causal models for novel objects to be rapidly learned, sometimes on the basis of a single causal event. Our third experiment explores whether learners can acquire a causal model for an object on the basis of its perceptual features alone. The objects in this experiment can be organized into two family resemblance categories on the basis of their perceptual features, and these two categories are associated with different causal powers. Observing the features of a novel object should allow a learner to assign it to one of these categories and to make inferences about its causal powers.

7.1. Participants

Twenty-four members of the MIT community were paid for participating in this experiment.

7.2. Procedure

Participants are initially shown an empty machine that activates on 10 of the 20 trials. Ten blocks then appear on screen, and the features of these blocks support two family resemblance categories (see Figs. 2 and 15). Before any of the blocks is placed in the machine, participants are informed that the blocks are laid out randomly, and they are encouraged to drag them around and organize them in a way that will help them predict what effect they will have on the machine. Participants then observe 20 trials for blocks o_1 through o_8, and see that blocks o_1 through o_4 activate the machine rarely, but blocks o_5 through o_8 activate the machine most of the time. After 20 trials for each block, participants respond to the same question used in Experiment 1: They imagine 100 trials involving the block and rate how likely it is that the total number of activations will fall into each of five intervals. After this training phase, participants answer the same question for test blocks o^- and o^+ without seeing *any* trials involving these blocks. Experiment 1 explored one-shot learning, and this new task might be described as zero-shot learning. After making predictions for the two test blocks, participants are asked to sort the blocks into two categories ''according to their effect on the machine'' and to explain the categories they chose.

	$\emptyset$	o^-	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o^+
$\mathbf{e}^+$:	10	0	3	2	1	2	18	18	17	19	0
$\mathbf{e}^-$:	10	0	17	18	19	18	2	2	3	1	0
f_1 :		1	0	0	0	0	0	1	1	1	1
f_2 :		0	1	0	0	0	1	0	1	1	1
f_3 :		0	0	1	0	0	1	1	0	1	1
f_4 :		0	0	0	1	0	1	1	1	0	1
f_5 :		0	0	0	0	1	1	1	1	1	0

Fig. 15. Training data for Experiment 3. The event data are consistent with two categories: The first includes objects that prevent the machine from activating, and the second includes objects that activate the machine. Features f_1 through f_5 are ''family resemblance'' features that provide noisy information about the underlying categories.

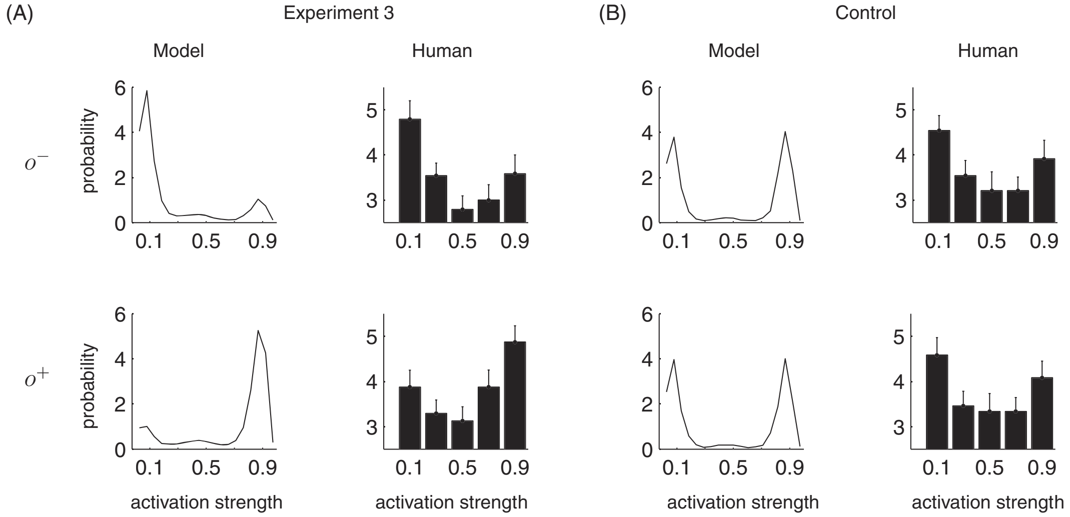


Fig. 16. Results for Experiment 3. (A) Model predictions and mean responses across 24 participants. Even though no trials are ever observed for objects o^- and o^+, participants use the features of these objects to make predictions about their causal powers. (B) Model predictions and mean responses for a control task where all objects are perceptually identical.

7.3. Model predictions

Predictions of the schema-learning model are shown in the left column of Fig. 16A. Each plot shows the probability that a test block will activate the machine on any given trial.[2] Both plots have two peaks, indicating that the model has discovered two categories but is not certain about the category assignments of the test blocks. The plots are skewed in opposite directions: based on the features of the test blocks, the model predicts that o^- will activate the machine rarely, and that o^+ will activate the machine often. The left column of Fig. 16B shows predictions about a control task that is identical to Experiment 3 except that all blocks are perceptually identical. The curves are now symmetric, indicating that the model has no basis for assigning the test blocks to one category or another.

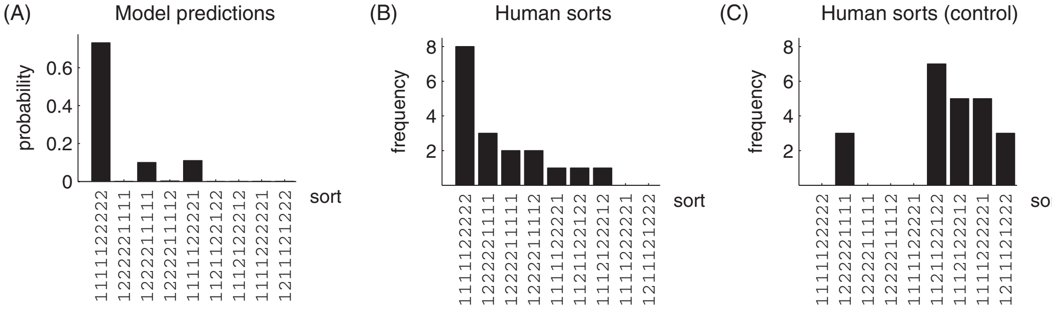


Fig. 17. Sorts for Experiment 3. (A) Relative probabilities of nine sorts according to the schema-learning model. Each sort is represented as a vector that specifies category assignments for the 10 objects in Fig. 15. The model prefers the family resemblance sort. (B) Sorts chosen by participants. Any sort not shown was chosen by at most one participant. (C) Sorts chosen in the control condition when no causal information was available.

Predictions about the sorting task are summarized in Fig. 17A. The top few sorts are included, and the most probable solution according to the model is the family resemblance sort. Although the model allows sorts with any number of categories (including one, three or more), the probabilities shown in Fig. 17A are calculated with respect to the class of all two-category solutions.

7.4. Results

Mean responses for the two test blocks are shown in the right column of Fig. 16A. Both plots are U-shaped curves, suggesting that participants realize that some blocks activate the machine rarely and others activate the machine often, but that few blocks activate the machine half the time. As predicted, the curves are skewed in opposite directions, indicating that o^+ is considered more likely to activate the machine than o^-. We ran a two-factor ANOVA which compared ratings for the first (0–20) and last (80–100) intervals across the two test blocks. There is no main effect of interval [$F(1,23) = 0.056$, $p > .5$] or of test block [$F(1,23) = 1.50$, $p > .1$], but there is a significant interaction between interval and test block [$F(1,23) = 6.90$, $p < .05$]. Follow up paired-sample t tests support the claim that both plots in Fig. 16 show skewed U-shaped curves. In the case of object o^-, the 0.1 bar is significantly greater than the 0.9 bar ($p < .05$, one-sided) and the difference between the 0.9 bar and the 0.5 bar is marginally significant ($p = .07$, one-sided). In the case of object o^-, the 0.9 bar is significantly greater than the 0.1 bar ($p < .05$, one-sided) and the difference between the 0.1 bar and the 0.5 bar is marginally significant ($p = .07$, one-sided).

We ran an additional 24 participants in a control task that was identical to Experiment 3 except that all blocks were perceptually identical. Mean responses are shown in the right column of Fig. 16B, and participants now generate U-shaped curves that are close to symmetric. The ANOVA analysis described in the previous paragraph now indicates that there is no main effect of interval or test block, and no significant interaction between interval and test block. Although both human curves in Fig. 16B are higher on the left than the right, paired-sample t tests indicate that there is no significant difference between the 0.1 bar and the 0.9 bar in either case (p > .2). The differences between the 0.9 bar and the 0.5 bar fail to reach significance in both cases, but the 0.1 bars are significantly greater than the 0.5 bars in both cases (p < .05, one-sided).

The U-shaped curves in Fig. 16A,B resolve a question left open by Experiment 1. Responses to the $f = \{0.1, 0.9\}$ condition of the first experiment did not indicate that participants had identified two categories, but the U-shaped curves in Fig. 16B suggest that participants recognized two categories of blocks. All of the blocks in Experiment 3 produce the effect sometimes, and the U-shaped curves suggest that participants can use probabilistic causal information to organize objects into categories. Two differences between Experiment 3 and the second condition of Experiment 1 seem particularly important. In Experiment 3, more blocks were observed for each category (4 rather than 3), and more trials were observed for each block (20 rather than 10). Experiment 3 therefore provides more statistical evidence that there are two categories of blocks.

Responses to the sorting task are summarized in Fig. 17B. The most popular sort organizes the blocks into the two family resemblance categories, and it is chosen by eight of the 24 participants. Studies of feature-based categorization have consistently found that family resemblance sorts are rare, and that participants prefer instead to sort objects according to a single dimension (e.g., size or color) (Medin, Wattenmaker, & Hampson, 1987). To confirm that this standard result applies in our case, we ran a control task where no causal observations were available and participants were asked to sort the blocks into categories on the basis of their perceptual features. The results in Fig. 17C show that none of 24 participants chose the family resemblance sort. A chi-square test confirms that the family resemblance sort was chosen significantly more often in the causal task than the control task (p < .01, one sided). Our results therefore suggest that the causal information provided in Experiment 3 overcomes the strong tendency to form categories based on a single perceptual dimension.

Regardless of the sort that they chose, most participants explained their response by stating that one category included ''strong activators'' or blocks that often lit up the machine, and that the other included weak activators. For example, one participant wrote that the first category ''activates approximately 10% of the time'' and the second category ''activates approximately 90% of the time.'' Although most participants seem to have realized that there were two qualitatively different kinds of blocks, only 13 of the 24 assigned the ''strong activators'' (blocks o_1 through o_4) to one category and the ''weak activators'' (blocks o_5 through o_8) to the other category. Some of the remaining participants may have deliberately chosen an alternative solution, but others gave explanations suggesting that they had lost track of the training trials. Note that the sorting task is relatively demanding, and that participants who do not organize the blocks carefully as they go along are likely to forget how many times each block activated the machine.

8. Discovering causal interactions between categories

Our approach so far captures some kinds of interactions between categories. For example, the schema in Fig. 1 captures interactions between categories of drugs and categories of people—alpha blockers tend to produce headaches, but only in A-people. This schema, however, does not capture interactions between categories of drugs, and it makes no predictions about what might happen when alpha blockers and beta blockers are simultaneously ingested. Drugs may interact in surprising ways—for example, two drugs may produce a headache when combined even though each one is innocuous on its own. We now extend our model to handle cases of this kind where each event (e.g., ingestion) can involve varying numbers of objects (e.g., drugs).

The first step is to extend our notation for domain-level problems to allow sets of objects. The domain-level problem for the drugs and headaches example now becomes

$$\texttt{ingests}(\texttt{person}, \{\texttt{drug}\}) \overset{?}{\rightarrow} \texttt{headache}(\texttt{person})$$

where each cause event now specifies that a person ingests a set of drugs. Following Novick and Cheng (2004) we will decompose each cause event into subevents, one for each subset of the set of drugs. For example, the object-level problem

$$\texttt{ingests}(\texttt{Alice}, \{\texttt{Doxazosin}, \texttt{Acebutolol}\}) \overset{?}{\rightarrow} \texttt{headache}(\texttt{Alice}) \tag{9}$$

can be viewed as a combination of four subproblems

$$\texttt{ingests}(\texttt{Alice}, []) \overset{?}{\rightarrow} \texttt{headache}(\texttt{Alice}) \tag{10a}$$

$$\texttt{ingests}(\texttt{Alice}, [\texttt{Doxazosin}]) \overset{?}{\rightarrow} \texttt{headache}(\texttt{Alice}) \tag{10b}$$

$$\texttt{ingests}(\texttt{Alice}, [\texttt{Acebutolol}]) \overset{?}{\rightarrow} \texttt{headache}(\texttt{Alice}) \tag{10c}$$

$$\texttt{ingests}(\texttt{Alice}, [\texttt{Doxazosin}, \texttt{Acebutolol}]) \overset{?}{\rightarrow} \texttt{headache}(\texttt{Alice}) \tag{10d}$$

The difference between the curly brackets in Eq. (9) and the square brackets in Eq. (10d) is significant. The subproblem in Eq. (10d) refers to a causal relationship that depends exclusively on the interaction between Doxazosin and Acebutolol. In other words, the properties of this causal relationship depend only on the causal power of the pair of drugs, not on the causal power of either drug taken in isolation. The problem in Eq. (9) refers to the overall relationship that results from combining all instances in Eqs. (10a–d). In other words, the overall effect of taking Doxazosin and Acebutolol may depend on the base rate of experiencing headaches, the effect of Doxazosin alone, the effect of Acebutolol alone, and the effect of combining the two drugs.

Building on the approach described in previous sections, we introduce a causal model for each subproblem that depends on three parameters. The first indicates whether the subevent is causally related to the effect, the second indicates the polarity of this causal relationship, and the third indicates the strength of this relationship. As before, we organize the drugs into categories, and we assume that object-level causal models are generated from category-level models that capture the causal powers of each category acting in isolation. Now, however, we introduce additional category-level models that capture interactions between categories. For instance, if Acebutolol and Atenolol are assigned to the same category, then the causal models for the subproblems

$$\texttt{ingests}(\texttt{Alice}, [\texttt{Doxazosin}, \texttt{Acebutolol}]) \xrightarrow{?} \texttt{headache}(\texttt{Alice})$$

$$\texttt{ingests}(\texttt{Alice}, [\texttt{Doxazosin}, \texttt{Atenolol}]) \xrightarrow{?} \texttt{headache}(\texttt{Alice})$$

will be generated from the same category-level model. This approach captures the intuition that members of the same category (e.g., Acebutolol and Atenolol) are expected to interact with Doxazosin in a similar way.

To formalize these ideas, we extend the Ψ in Eq. 6 to include an arrow a, a polarity g and a strength s for each combination of objects. We extend the schema in a similar fashion and include category-level models for each combination of categories. As before, the parameters for each object-level causal model are generated from the parameters ($\bar{a}$, $\bar{g}$, and $\bar{s}$) for the corresponding category-level model. For instance, Fig. 18 shows how the causal model for the $o_9 + o_{18}$ pair is generated from a category-level model that states that categories c_A and c_B interact to generate the effect.

Our main remaining task is to specify how the object-level models for the subinstances in 10 combine to influence the probability that Alice develops a headache after ingesting Doxazosin and Acebutolol. We use the sequential interaction model of Novick and Cheng (2004) and assume that subevents combine according to a network of noisy-OR and noisy-AND-NOT gates (Fig. 19). To capture the idea that the causal powers of a set of objects can be very different from the causal powers of the objects taken individually, we assume that subevents involving small sets of objects {e.g., ingests(Alice, [Doxazosin])} act first and can be overruled by subevents involving larger sets {e.g., ingests(Alice, [Doxazosin, Atenolol])}. Although the sequential interaction model seems appropriate for our purposes, the

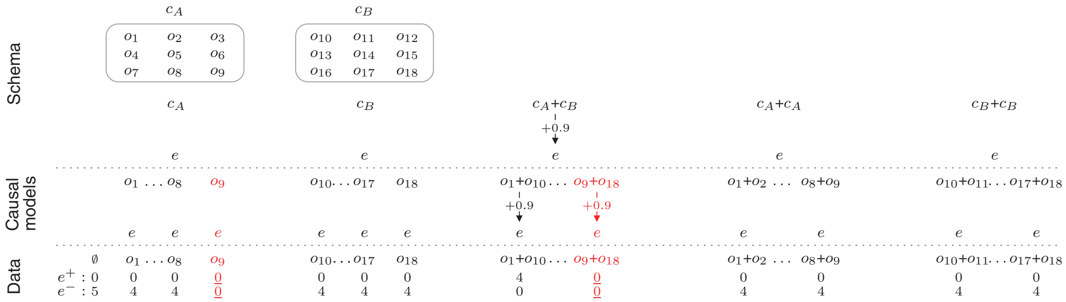

Schema
c_A
o_1 o_2 o_3 o_4 o_5 o_6 o_7 o_8 o_9
c_B
o_{10} o_{11} o_{12} o_{13} o_{14} o_{15} o_{16} o_{17} o_{18}
c_A
c_B
c_A+c_B
+0.9
c_A+c_A
c_B+c_B
e
Causal models
$o_1 \ldots o_8$ o_9
$o_{10} \ldots o_{17}$ o_{18}
o_1+$o_{10} \ldots$ o_9+o_{18}
+0.9
+0.9
o_1+$o_2 \ldots$ o_8+o_9
o_{10}+$o_{11} \ldots o_{17}$+o_{18}
Data
∅
e^+ : 0
e^- : 5
0 0 0 0 0 0 4 0 0 0 0 0
4 4 0 4 4 4 0 0 4 4 4 4

Fig. 18. Learning about interactions between objects. The schema includes category-level models for each individual category and for each pair of categories. The schema shown here has two categories: Individual objects of either category do not produce the effect, but any pair including objects from both categories will produce the effect. The collection of object-level causal models includes a model for each object and each pair of objects. Note that the schema supports inferences about sparsely observed individual objects (e.g., o_9) and about pairs that have never been observed to interact (e.g., o_9 and o_{18}, counts underlined in red).

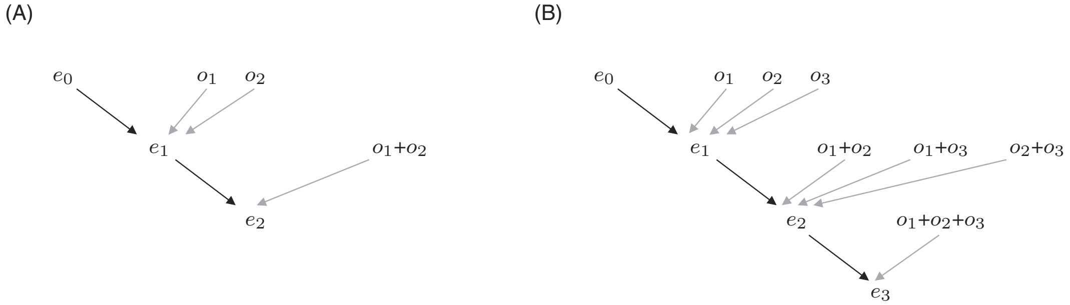


Fig. 19. The sequential interaction model of Novick and Cheng (2004). (A) Network for the case where two objects may interact to influence an effect event e. Event e_i indicates whether the effect would have occurred based on interactions of up to i objects. For example, e_o indicates whether the background cause is active, and e_1 indicates whether the effect would have occurred as a result of combining the background cause with the causal contributions of each object taken individually. Variable e_2 indicates whether the effect event actually occurs when the interaction between the two objects is taken into account. The black arrows are generative with strength 1, and the gray arrows may or may not exist, may be generative or preventive, and may have any causal strength. Once the status of each gray arrow is specified, all events in the network combine according to a noisy-OR/noisy-AND-NOT model. (B) The same general approach can handle interactions among any number of cause events. Shown here is a case where three objects may interact to influence an effect event. In this case, variable e_3 indicates whether the effect event actually occurs.

general framework we have developed allows room for accounts of schema learning that incorporate alternative models of interaction.

9. Experiment 4: Causal interactions between categories

We designed an experiment to explore schema learning in a setting where pairs of objects may interact to produce a cause. Our formal framework can now handle several kinds of data, including contingency data for single objects, contingency data for pairs of objects, and perceptual features of the objects. In real-world settings, these different kinds of data will often reinforce each other and combine to pick out a single set of categories. Here, however, we explore whether information about pairwise interactions alone is sufficient for learners to discover causal schemata.

Experiment 4 used the same scenario developed for our previous experiments, but now participants were able to place up to two blocks inside the machine. Unlike Experiments 1 through 3, the individual causal powers of the blocks were identical, and unlike Experiment 3, the blocks were perceptually indistinguishable. The blocks, however, belonged to categories, and these categories determined the pairwise interactions between blocks. In the *pairwise activation* condition, the machine never activated when it contained a single block or two blocks from the same category, but always activated whenever it contained one block from each category (Fig. 18). In the *pairwise inhibition* condition the machine always activated when it contained a single block or two blocks from the same category, but never activated when it contained one block from each category. Experiment 4 explores whether participants could infer the underlying causal categories based on pairwise interactions alone and could use this knowledge to rapidly learn causal models for novel objects.

Our experiment builds on the work of Kemp, Tenenbaum, Niyogi, and Griffiths (2010), who demonstrated that people can use relationships between objects to organize these objects into categories.[3] These authors considered interactions between objects, but their stimuli did not allow for the possibility that individual objects might produce the effect in isolation. We therefore designed a new experiment that relies on the same scenario used in Experiments 1 through 3.

9.1. Participants

Thirty-two members of the CMU community participated for pay or course credit.

9.2. Stimuli and design

Experiment 4 used the same graphical interface developed for Experiment 1. All of the blocks were perceptually indistinguishable. The experiment included two conditions and sixteen participants were assigned to each condition. In the pairwise activation condition, the machine never activated on its own and never activated when it contained a single block. The blocks, however, belonged to two categories, and the machine always activated on trials when it contained an A-block and a B-block. In the pairwise inhibition condition, the machine always activated when it contained a single block or two blocks from the same category, but it always failed to activate when it contained two blocks from different categories.

9.3. Procedure

The experiment was divided into several phases. During phase 0, participants observed five trials where the empty machine failed to activate. Three blocks were added to the screen at the start of phase 1. Unknown to the participants, two blocks were A-blocks (o_1 and o_2) and the third was a B-block (o_{10}). Participants observed four trials for each individual block, and the machine never activated (pairwise activation condition) or always activated (pairwise inhibition condition). Before observing any interactions, participants predicted what would happen when o_1 and o_{10} were simultaneously placed in the machine. The wording of the question was taken from our previous experiments: Participants imagined 100 trials when the machine contained the two blocks, and they rated the probability that the total number of activations would fall within each of five intervals. Participants then saw two trials for each pair of blocks. Phase 1 finished with a period of ‘‘free experimentation,’’ where participants were given the opportunity to carry out as many of their own trials as they wished.

Phases 2 through 6 were identical in structure. In each phase, three new blocks were added to the screen, and one of these blocks served as the ‘‘test block.’’ In some phases the test block was an A-block, and in others the test block was a B-block. Before observing any trials involving the new blocks, participants were given a pretest which required them to

Table 1
Design for Experiment 4

Phase	1	2	3	4	5	6	7
Blocks added	o_1, o_2, o_{10}	o_3, o_{11}, o_{12}	o_4, o_{13}, o_{14}	o_5, o_6, o_{15}	o_7, o_8, o_{16}	o_9, o_{17}, o_{18}	o_A, o_B
Test blocks		o_{11}	o_4	o_5	o_{16}	o_{17}	o_A, o_B
Probe blocks (i)		o_2	o_{12}	o_3	o_{15}	o_8	o_{11}, o_4
Probe blocks (ii)		o_2	o_{12}	o_3	o_{15}	o_8	o_4, o_{11}
Probe blocks (iii)		o_2	o_{12}	o_3	o_6	o_{16}	o_{11}, o_4
Probe blocks (iv)		o_2	o_{12}	o_3	o_6	o_{16}	o_4, o_{11}

Note. Blocks o_1 through o_9 belong to category c_A and blocks o_{10} through o_{18} belong to category c_B. In each pretest and posttest, participants make predictions about interactions between the test block and o_1 (an A-block) and between the test block and o_{10} (a B-block). Between each pretest and posttest, participants observe a single trial where the test block is paired with a probe block. Probe blocks for the four groups of participants are shown.

predict how the test block would interact with two of the blocks already on screen, one (o_1) from category c_A and the other (o_{10}) from category c_B. Participants then observed a single trial where the test block was paired with one of the blocks already on screen (the probe block). Armed with this single piece of information, participants completed a posttest that was identical to the pretest. The phase then finished with a period of free experimentation.

A complete specification of the design is shown in Table 1. The experiment involves 20 blocks in total: blocks o_1 through o_9 belong to category c_A, blocks o_{10} through o_{18} belong to category c_B, and there are two test blocks in the final phase (o_A and o_B). The first and second rows of Table 1 list the blocks that are added to the screen in each phase, and the block that serves as the test block in each phase. Participants were randomly assigned to one of four groups (i through iv), and the probe blocks used for each group are shown in the final four rows of Table 1. No significant differences between these groups were observed, and we will collapse across these groups when reporting our results.

The final phase was very similar to phases 2 through 6, but only two new blocks were added to the screen. One block (o_A) was an A-block, and the second (o_B) was a B-block. In phases 2 through 6 only one of the new blocks served as the test block, but in the final phase both o_A and o_B served as test blocks. In the pretest for phase 7, participants made predictions about how o_A and o_B would interact with o_1 and o_{10} before observing any pairwise trials involving the test blocks. Participants then observed a single trial involving each test block and responded to a posttest that was identical to the pretest. After providing these predictions, participants were asked to sort the blocks into two categories ''according to their effect on the machine'' and to ''describe how the blocks and machine work.''

9.4. Model predictions

Although participants made inferences during each phase of the experiment, our main question is whether they had learned a causal schema by the end of the experiment. We therefore compare inferences about the first and last phases of the experiment. Kemp et al. (2010) describe a very similar task and show learning curves that include data from all phases of the experiment.

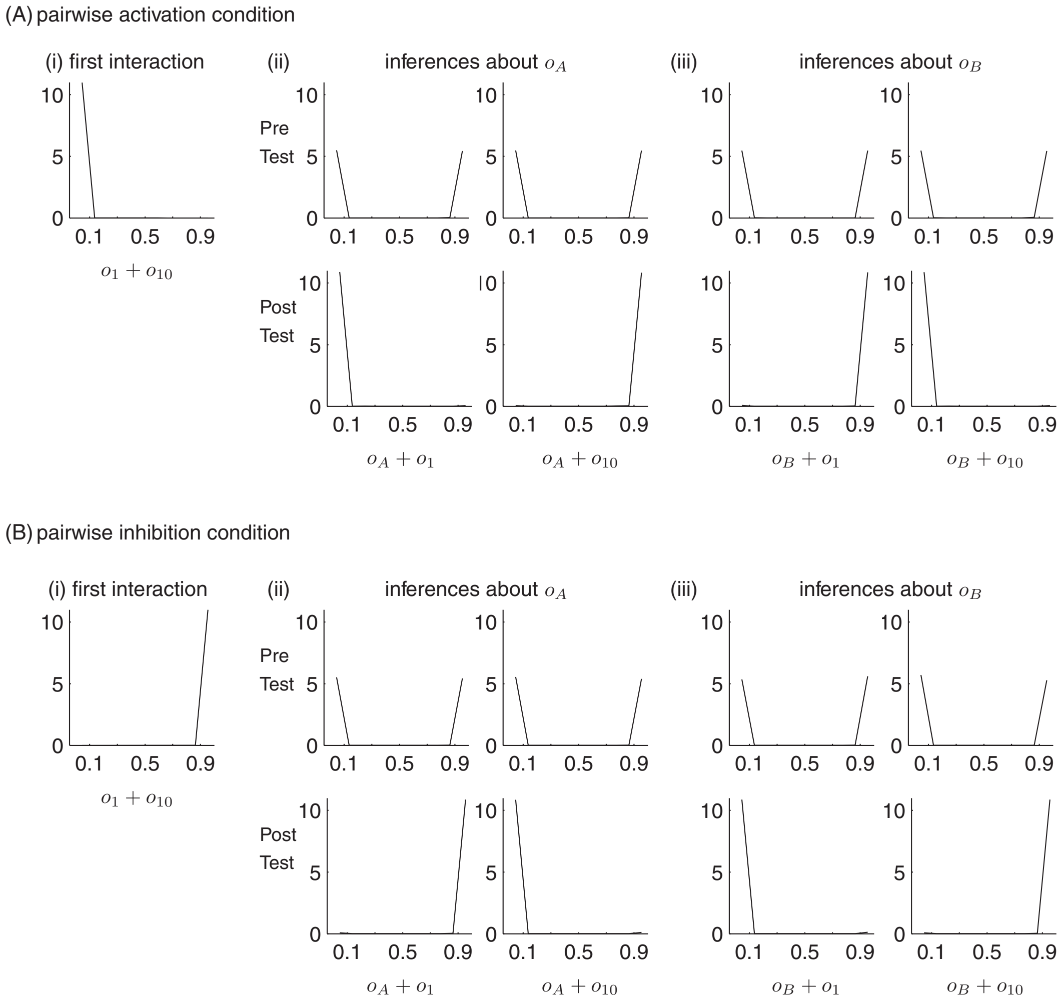


Fig. 20. Model predictions for Experiment 4. (A) Pairwise activation condition. (i) Before any pairwise trials have been observed, the model predicts that pairs of objects are unlikely to activate the machine. (ii) Inferences about test block o_A. Before observing any trials involving this block, the model is uncertain about whether it will activate the machine when paired with o_1 or o_{10}. After observing that o_A activates the machine when paired with o_{18} (a B-block), the model infers that o_A will activate the machine when paired with o_{10} but not o_1. (iii) Inferences about test block o_B show a similar pattern: The model is uncertain during the pretest, but one observation involving o_B is enough for it to make confident predictions on the posttest. (B) Pairwise inhibition condition. The prediction in (i) and the posttest predictions in (ii) and (iii) are the opposite of the corresponding predictions for the pairwise activation condition.

Figs. 20A.i, B.i show predictions about a pair of blocks before any pairwise trials have been observed. In the pairwise activation condition, the model has learned by this stage that individual blocks tend not to produce the effect, and the default expectation captured by the interaction model is that pairs of blocks will also fail to produce the effect. The model allows for several possibilities: There may or may not be a conjunctive cause corresponding to any given pair of blocks, and this conjunctive cause (if it exists) may be generative or preventive and may have high or low strength. Most of these possibilities lead to the prediction that the pair of blocks will be unlikely to activate the machine. The machine is only likely to activate if the pair of blocks corresponds to a conjunctive cause with high strength, and this possibility receives a relatively low probability compared to the combined probability assigned to all other possibilities. Similarly, in the pairwise inhibition condition the model has learned that individual blocks tend to produce the effect, and the default expectation captured by the interaction model is that pairs of blocks will also produce the effect.

After observing several pairwise interactions, the model discovers that the default expectation does not apply in all cases, and that some pairs of blocks activate the machine when combined. By the final phase of the task, the model is confident that the blocks can be organized into two categories, where blocks o_1 through o_9 belong to category c_A and blocks o_{10} through o_{18} belong to category c_B. The model, however, is initially uncertain about the category assignments of the two test blocks (blocks o_A and o_B) and cannot predict with confidence whether either block will activate the machine when paired with o_1 or o_{10} (Fig. 20ii–iii). Recall that the two categories have no distinguishing features, and that blocks o_A and o_B cannot be categorized before observing how they interact with one or more previous blocks. After observing a single trial where o_A is paired with one of the previous blocks, the model infers that o_A probably belongs to category A. In the pairwise activation condition, the model therefore predicts that the pair $\{o_A, o_{10}\}$ will probably activate the machine but that the pair $\{o_A, o_1\}$ will not (Fig. 20A.ii–iii). Similarly, in the pairwise activation condition, a single trial involving o_B is enough for the model to infer that $\{o_B, o_1\}$ will probably activate the machine although the pair $\{o_B, o_{10}\}$ will not.

9.5. Results

Figs. 21A.i, B.i show mean inferences about a pairwise interaction before any pairwise trials have been observed. As expected, participants infer that two blocks which fail to activate the machine individually will fail to activate the machine when combined (pairwise activation condition), and that two blocks which individually activate the machine will activate the machine when combined (pairwise inhibition condition). A pair of t tests indicates that the 0.1 bar is significantly greater than the 0.9 bar in Fig. 21A.i ($p < .001$, one-sided) but that the 0.9 bar is significantly greater than the 0.1 bar in Fig. 21B.i ($p < .001$, one-sided). These findings are consistent with the idea that learners assume by default that multiple causes will act independently of one another.

By the end of the experiment, participants were able to use a single trial involving a novel block to infer how this block would interact with other previously observed blocks. The mean responses in Fig. 21 match the predictions of our model and show that one-shot learning is possible even in a setting where any two blocks taken in isolation appear to have identical causal powers. A series of paired-sample t tests indicates that the difference between the 0.1 and the 0.9 bars is not significant for any of the pretest plots in Fig. 21 ($p > .3$ in all

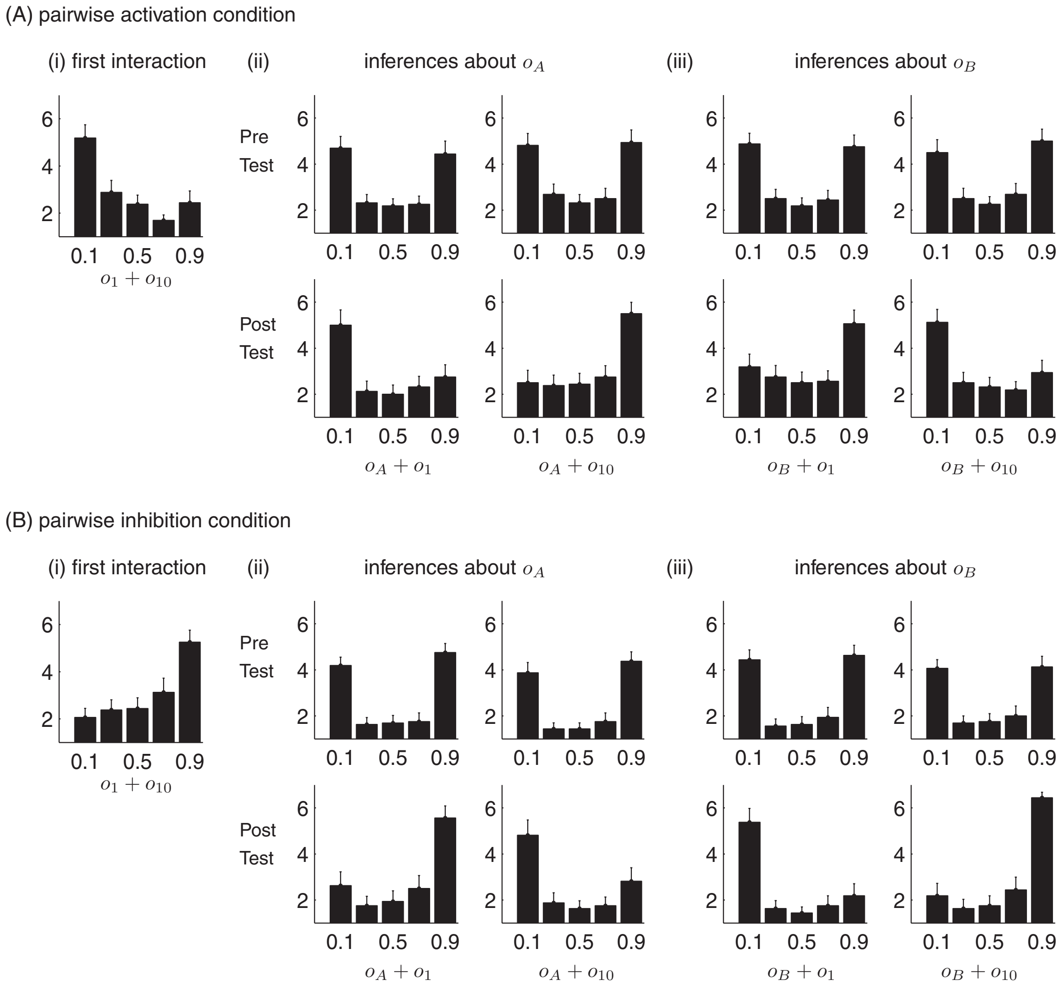


Fig. 21. Data for Experiment 4. All inferences are qualitatively similar to the model predictions in Fig. 20.

cases), but the difference between these bars is significant for each posttest plot ($p < .05$ in all cases). Although the model predictions are broadly consistent with our data, the model is often extremely confident in cases where the mean human response appears to be a U-shaped curve. In all of these cases, however, few individuals generate U-shaped curves, and the U-shaped mean is a consequence of averaging over a majority of individuals who match the model and a minority who generate curves that are skewed in the opposite direction.

Responses to the sorting task provided further evidence that participants were able to discover a causal schema based on interaction data alone. In each condition, the most common sort organized the 18 blocks into the two underlying categories. In the pairwise activation condition, five of the 16 participants chose this response, and an additional three gave responses that were within three moves of this solution. In the pairwise inhibition condition, nine of the 16 participants chose this response, and an additional two gave responses that were within three moves of this solution. The remaining sorts appeared to vary idiosyncratically, and no sort other than the most common response was chosen by more than one participant. As in Experiment 3, the sorting task is relatively challenging, and participants who did not organize the blocks as they went found it difficult to sort them into two categories at the end of the experiment. Several participants gave explanations suggesting that they had lost track of the observations they had seen.

Other explanations, however, suggested that some participants had discovered an explicit causal schema. In the pairwise activation condition, one participant sorted the blocks into categories that she called ''activators'' and ''partners,'' and wrote that ''the machine requires both an activator and a partner to work.'' In the pairwise inhibition condition, one participant wrote the following:

> The machine appears to take two different types of blocks. Any individual block turns on the machine, and any pair of blocks from the same group turns on the machine. Pairing blocks from different groups does not turn on the machine.

An approach similar to the exemplar model described earlier will account for people's inferences about test blocks o_A and o_B. For example, if o_A is observed to activate o_{18} in the pairwise activation condition, the exemplar model will assume that o_A is similar to other blocks that have previously activated o_{18}, and will therefore activate o_{11} but not o_1. Note, however, that the exemplar model assumes that learners have access to the observations made for all previous blocks, and we propose that this information can only be maintained if learners choose to sort the blocks into categories. The exemplar model also fails to explain the results of the sorting task, and the explanations that mention an underlying set of categories. Finally, Experiment 2 of Kemp et al. (2010) considers causal interactions, and it was specifically designed to compare approaches like the exemplar model with approaches that discover categories. The results of this experiment rule out the exemplar model, but they are consistent with the predictions of our schema-learning framework.

10. Children's causal knowledge and its development

We proposed that humans learn to learn causal models by acquiring abstract causal schemata, and our experiments confirm that adults are able to learn and use abstract causal knowledge. Some of the most fundamental causal schemata, however, are probably acquired early in childhood, and learning abstract schemata may itself be a key component of cognitive development. Although our experiments focused on adult learning, this section shows how our approach helps to account for children's causal learning.

Our experiments explored three learning challenges: grouping objects into categories with similar causal powers (Fig. 6 and Experiments 1 and 2), categorizing objects based on their causal powers and their perceptual features (Fig. 14 and Experiment 3), and forming categories to explain causal interactions between objects (Fig. 18 and Experiment 4). All three challenges have been explored in the developmental literature, and we consider each one in turn.

10.1. Categories and causal powers

The developmental literature on causal learning includes many studies that address the relationship between categorization and causal reasoning. Researchers have explored whether children organize objects into categories with similar causal powers, and whether their inferences rely more heavily on causal powers or perceptual features. Many studies that address these questions have used the blicket detector paradigm (Gopnik & Sobel, 2000; Nazzi & Gopnik, 2000; Sobel, Sommerville, Travers, Blumenthal, & Stoddard, 2009), and we will show how our model accounts for several results that have emerged from this paradigm.

In a typical blicket detector study, children are shown a set of blocks and a detector. Some blocks are *blickets* and will activate the detector if placed on top of it. Other blocks are inert and have no effect on the detector. Many questions can be asked using this setup, but for now we consider the case where all blocks are perceptually identical and the task is to organize these blocks into categories after observing their interactions with the detector. Gopnik and Sobel (2000) and others have established that young children can accurately infer whether a given block is a blicket given only a handful of relevant observations. For example, suppose that the detector activates when two blocks (A and B) are simultaneously placed on top of it, but fails to activate when A alone is placed on top of it. Given these outcomes, 3-year-olds correctly infer that block B must be a blicket.

Our formal approach captures many of the core ideas that motivated the original blicket detector studies, including the idea that objects have causal powers and the idea that objects with similar causal powers are organized into categories. Our work also formalizes the relationship between object categories (e.g., categories of blocks) and event data (e.g., observations of interactions between blocks and the blicket detector). In particular, we propose that children rely on an intermediate level of knowledge which specifies the causal powers of individual objects, and that they understand that the outcome of a causal event depends on the causal powers of the specific objects (e.g., blocks) involved in that event.

Several previous authors have presented Bayesian analyses of blicket-detector experiments (Gopnik, Glymour, Sobel, Schulz, Kushnir, & Danks, 2004), and it is generally accepted that the results of these experiments are consistent with a Bayesian approach. Typically, however, the Bayesian models considered do not incorporate all of the intuitions about causal kinds that are captured by our framework. A standard approach used by Gopnik et al. (2004) and others is to construct a Bayes net where there is a variable for each block indicating whether it is on the detector, an additional variable indicating whether the detector activates, and an arrow from each block variable to the detector variable only if that block is a blicket. This simple approach provides some insight but fails to capture key aspects of knowledge about the blicket detector setting. For example, if the experimenter introduces a new block and announces that it is a blicket, the network must be extended by adding a new variable that indicates whether the new block is on the detector and by drawing an arrow between this new variable and the detector variable. Knowing how to modify the network in this way is critical, but this knowledge is not captured by the original network. More precisely, the original network does not explicitly capture the idea that blocks can be organized into categories, and that there is a predictable relationship between the category membership of a block and the outcome of events involving that block.

To address these limitations of a basic Bayes net approach, Danks (2007) and Griffiths and Tenenbaum (2007) proposed formalisms that explicitly rely on distinct causal models for blickets and nonblickets. Both of these approaches assume that all blickets have the same causal strength, but our model is more flexible and allows objects in the same category to have different causal strengths. For example, in the $p = \{0, 0.5\}$ condition of Experiment 1, block o_6 activates the machine 4 times out of 10 and block o_7 activates the machine 6 times out of 10. Our model infers that o_7 has a greater causal strength than o_6, and the means of the strength distributions for these blocks are 0.49 and 0.56, respectively. Although the blocks vary in strength, the model is 90% certain that the two belong to the same category. To our knowledge, there are no developmental experiments that directly test whether children understand that blocks in the same category can have different causal strengths. This prediction of our model, however, is supported by two existing results. Kushnir and Gopnik (2005) found that 4-year-olds track the causal strengths of individual blocks, and Gopnik, Sobel, Shulz, and Glymour (2001) found that 3-year-olds will categorize two objects as blickets even if one activates the machine more often (three of three trials) than the other (two of three trials). Combining these results, it seems likely that 4-year-olds will understand that two objects have different causal strengths but recognize that the two belong to the same category.

Although most blicket detector studies present children with only a single category of interest (i.e., blickets), our model makes an additional prediction that children should be able to reason about multiple categories. In particular, our model predicts that children will distinguish between categories of objects that have similar causal powers but very different causal strengths. Consider a setting, for example, where there are three kinds of objects: blickets, wugs, and inert blocks. Each blicket activates the detector 100% of the time, and each wug activates the detector between 20% and 30% of the time. Our model predicts that young children will understand the difference between blickets and wugs, and will be able to organize novel blocks into these categories after observing their effects on the detector.

10.2. Categories, causal powers, and features

This section has focused so far on problems where the objects to be categorized are perceptually identical, but real-world object categories often vary in their perceptual properties as well as their causal powers. A central theme in the developmental literature is the relationship between perceptual categorization (i.e., categorization on the basis of perceptual properties) and conceptual or theory-based categorization (i.e., categorization on the basis of nonobservable causal or functional properties). Many researchers have compared these two kinds of categorization and have explored how the tradeoff between the two varies with age. One influential view proposes that infants initially form perceptual categories and only

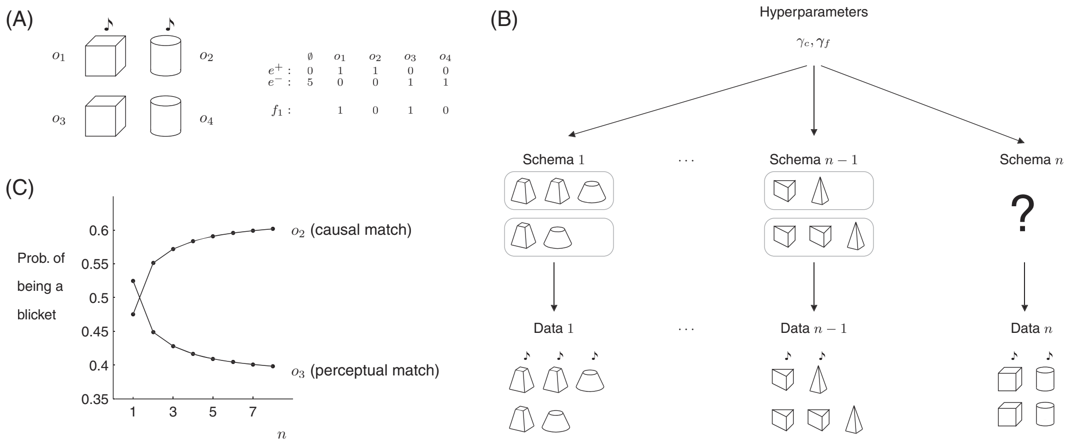


Fig. 22. Modeling the shift from perceptual to causal categorization. (A) The four objects in the Gopnik and Sobel (2000) conflict task. The two objects with the power to activate the blicket detector are marked with musical notes. Note that object o_1 could be grouped with a causal match (o_2) or a perceptual match (o_3). The table shows how the causal and perceptual data are provided as input to our model, and it includes a single feature f_1 which indicates whether the objects are cubes or cylinders. (B) Our hierarchical Bayesian framework can be extended to handle multiple systems of objects. Note that a single set of hyperparameters which specifies the relative weights of causal (γ_c) and perceptual (γ_f) information is shared across all systems. Our model observes how the objects in the first $n - 1$ systems are organized into categories, and it learns that in each case the categories are better aligned with the causal observations than the feature data. The model must now infer how the objects in the final system are organized into categories. (C) After learning that object o_1 in the final system is a blicket, the model infers whether o_2 and o_3 are likely to be blickets. Relative probabilities of these two outcomes are shown. The curves show a shift from perceptual categorization (o_3 preferred) to causal categorization (o_2 preferred).

later come to recognize categories that rely on nonobservable causal properties. Keil (1989) refers to this position as ‘‘Original Sim,’’ and he and others have explored its implications.

The blicket detector paradigm can be used to explore a simple version of the tradeoff between perceptual and causal categorization. Gopnik and Sobel (2000) considered a conflict task where the blocks to be categorized had different perceptual features, and where these perceptual features were not aligned with the causal powers of these blocks. One task used four blocks, where two blocks activated the blicket detector but two did not (Fig. 22A). Each block therefore had a causal match, and each block was also perceptually identical to exactly one other block in the set. Crucially, however, the perceptual match and the causal match for each block were different. Children were told that one of the blocks that activated the detector was a blicket and were asked to pick out the other blicket. Consistent with the ‘‘Original Sim’’ thesis, 2-year-olds preferred the perceptual match. Three- and four-year-olds relied more heavily on causal information and were equally likely to choose the perceptual and the causal match. A subsequent study by Nazzi and Gopnik (2000) used a similar task and found that 4.5-year-olds showed a small but reliable preference for the causal match. Taken together, these results provide evidence for a developmental shift from perceptual to causal categorization.

Unlike previous Bayes net models of blicket detector tasks, our approach can be applied to problems like the conflict task where causal information and perceptual information are both available. As demonstrated in our third experiment, a causal schema can specify information about the appearance and the causal powers of the members of a given category, and our schema learning model can exploit both kinds of information. In the conflict task of Gopnik and Sobel (2000), the inference made by our model will depend on the relative values of two hyperparameters: γ_c and γ_f, which specify the extent to which the blocks in a given category are expected to have different causal powers (γ_c) and different features (γ_f). For modeling our adult experiments we set γ_c to a smaller value than γ_f ($\gamma_c = 0.1$ and $\gamma_f = 0.5$), which captures the idea that adults view causal information as a more reliable guide to category membership than perceptual information. As initially configured, our model therefore aims to capture causal knowledge at a stage after the perceptual to causal shift has occurred.

A natural next step is to embed our model in a framework where the hyperparameters γ_c and γ_f are learned from experience. The resulting approach is motivated by the idea that the developmental shift from perceptual to causal categorization may be explained in part as a consequence of rational statistical inference. Given exposure to many settings where causal information provides a more reliable guide to category membership than perceptual information, a child may learn to rely on causal information in future settings. To illustrate this idea, we describe a simple simulation based on the Gopnik and Sobel (2000) conflict task.

Fig. 22B shows how our schema learning framework can be extended to handle multiple systems of objects. We consider a simple setting where each system has two causal categories and up to six objects. Fig. 22B shows that the observations for the final test system are consistent with the Gopnik and Sobel (2000) conflict task: objects o_1 and o_2 activate the detector but the remaining objects do not, and object o_1 is perceptually identical to o_3 (both have feature f_1) but not o_2 or o_4. We assume that causal and feature data are available for each previous system, that the category assignments for each previous system are observed, and that these category assignments are always consistent with the causal data rather than the feature data. Two of these previous systems are shown in Fig. 22B.

Fig. 22B indicates that the category assignments for the test system are unobserved, and that the model must decide whether o_1 is more likely to be grouped with o_2 (the causal match) or o_3 (the perceptual match). If the test system is the first system observed (i.e., if $n = 1$), Fig. 22C shows that the model infers that the perceptual match (o_3) is more likely to be a blicket. Given experience with several systems, however, the model now infers that the causal match (o_2) is more likely to be a blicket.

The developmental shift in Fig. 22C is driven by the model's ability to learn appropriate values of the hyperparameters γ_c and γ_f given the first $n - 1$ systems of objects. The hierarchy in Fig. 22B indicates that a single pair of hyperparameters is assumed to characterize all systems, and the prior distribution used for each parameter is a uniform distribution over the set $\{2^{-6}, 2^{-5}, \ldots, 2^{3}\}$. Although the model begins with a symmetric prior over these hyperparameters, it initially prefers categories that match the features rather than the causal observations. The reason is captured by Fig. 3D, which indicates that the features are directly generated from the underlying categories but that the event data are one step removed from these categories. The model assumes that causal powers rather than causal events are directly generated from the categories, and it recognizes that a small set of event data may not accurately reflect the causal powers of the objects involved. Given experience with several previous systems, however, the model infers that γ_c is smaller than γ_f, and that causal observations are a more reliable guide to category membership than perceptual features. A similar kind of learning is discussed by Kemp et al. (2007), who describe a hierarchical Bayesian model that learns that shape tends to be a more reliable guide to category membership than other perceptual features such as texture and color.

The simulation results in Fig. 22C are based on a simple artificial scenario, and the proposal that statistical inference can help to explain the perceptual to conceptual shift needs to be explored in more naturalistic settings. Ultimately, however, this proposal may help to resolve a notable puzzle in the developmental literature. Many researchers have discussed the shift from perceptual to conceptual categorization, but Mandler (2004) writes that ''no one … has shown how generalization on the basis of physical appearance gets replaced by more theory-based generalization'' (p. 173). We have suggested that this shift might be explained as a consequence of learning to learn, and that hierarchical Bayesian models like the one we developed can help to explain how this kind of learning is achieved.

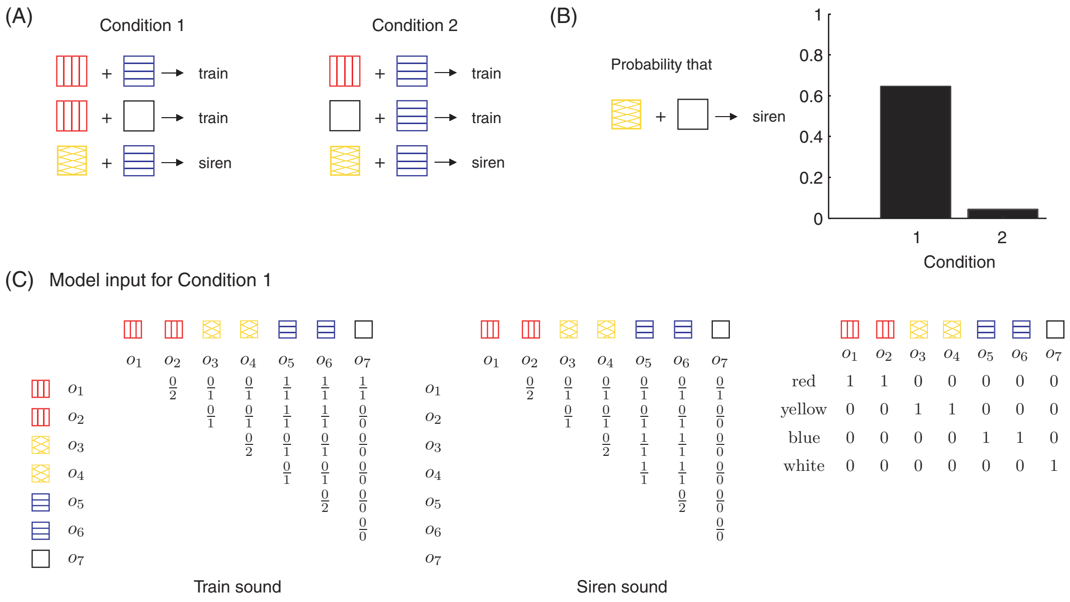
Although this section has focused on tradeoffs between perceptual and causal information, in many cases children rely on both kinds of information when organizing objects into categories. For example, children may learn that balloons and pins have characteristic features (e.g., balloons are round and pins are small and sharp) and that there is a causal relationship between these categories (pins can pop balloons). Children must also combine perceptual and causal information when acquiring the concept of animacy: Animate objects have characteristic features, including eyes (Jones, Smith & Landau, 1991), but they also share causal powers like the ability to initiate motion (Massey & Gelman, 1988). Understanding how concepts like animacy emerge over development is a challenging puzzle, but models that combine both causal and perceptual information may contribute to the solution.

10.3. Causal interactions

Children make inferences about the causal powers of individual objects but also understand how these causal powers combine when multiple objects act simultaneously. The original blicket detector studies included demonstrations where multiple objects were placed on the detector, and 4-year-olds correctly assumed that these interactions were consistent with an OR function (i.e., that the detector would activate if one or more blickets were placed on top of it). Consistent with these results, our model assumes by default that causal interactions are governed by a noisy-OR function, but Experiment 4 demonstrates that both adults and our model are able to learn about other kinds of interactions. Lucas and Griffiths (2010) present additional evidence that adults can learn about a variety of different interactions, and future studies can test the prediction that this ability is available relatively early in development.

Our modeling approach relies on the idea that causal interactions between individual objects can be predicted using abstract laws that specify how categories of objects are expected to interact. Recent work of Schulz, Goodman, Tenenbaum, and Jenkins (2008) supports the idea that young children can learn abstract laws, and they can do so on the basis of a relatively small number of observations. These authors introduced preschoolers to a set of seven blocks that included two red blocks, two yellow blocks, two blue blocks, and one white block. Some pairs of blocks produced a sound whenever they came into contact—for example, a train sound was produced whenever a red block and a blue block came into contact, and a siren sound was produced whenever a yellow block and a blue block came into contact (Fig. 23A). Other pairs of blocks produced no sound—for example, red blocks and yellow blocks never produced a sound when paired. Here we consider two conditions that differ only in the role played by the white block. In condition 1, the white block produced the train sound when paired with a red block, but in condition 2 the white block produced the train sound when paired with a blue block. No other observations involved the white block—in particular, children never observed the white block come into contact with a yellow block.

Using several dependent measures, Schulz and colleagues found that children in condition 1 expected the white block to produce the siren sound when paired with a yellow block, but that children in condition 2 did not. Our model accounts for this result. The evidence in condition 1 is consistent with the hypothesis that white blocks and blue blocks belong to the



Train sound

		▥ o_1	▥ o_2	⊠ o_3	⊠ o_4	▤ o_5	▤ o_6	□ o_7
▥	o_1		$\frac{0}{2}$	$\frac{0}{1}$	$\frac{0}{1}$	$\frac{1}{1}$	$\frac{1}{1}$	$\frac{1}{1}$
▥	o_2			$\frac{0}{1}$	$\frac{0}{1}$	$\frac{1}{1}$	$\frac{1}{1}$	$\frac{0}{0}$
⊠	o_3				$\frac{0}{2}$	$\frac{0}{1}$	$\frac{0}{1}$	$\frac{0}{0}$
⊠	o_4					$\frac{0}{1}$	$\frac{0}{1}$	$\frac{0}{0}$
▤	o_5						$\frac{0}{2}$	$\frac{0}{0}$
▤	o_6							$\frac{0}{0}$
□	o_7							

Siren sound

	▥ o_1	▥ o_2	⊠ o_3	⊠ o_4	▤ o_5	▤ o_6	□ o_7
o_1		$\frac{0}{2}$	$\frac{0}{1}$	$\frac{0}{1}$	$\frac{0}{1}$	$\frac{0}{1}$	$\frac{0}{1}$
o_2			$\frac{0}{1}$	$\frac{0}{1}$	$\frac{0}{1}$	$\frac{0}{1}$	$\frac{0}{0}$
o_3				$\frac{0}{2}$	$\frac{1}{1}$	$\frac{1}{1}$	$\frac{0}{0}$
o_4					$\frac{1}{1}$	$\frac{1}{1}$	$\frac{0}{0}$
o_5						$\frac{0}{2}$	$\frac{0}{0}$
o_6							$\frac{0}{0}$
o_7							

	▥ o_1	▥ o_2	⊠ o_3	⊠ o_4	▤ o_5	▤ o_6	□ o_7
red	1	1	0	0	0	0	0
yellow	0	0	1	1	0	0	0
blue	0	0	0	0	1	1	0
white	0	0	0	0	0	0	1

Fig. 23. (A) Evidence provided in conditions 1 and 2 of Schulz et al. (2008). (B) Model predictions about an interaction between a yellow block and a white block. Like preschoolers, the model infers that this combination is likely to produce a siren noise in condition 1 but not in condition 2. (C) Input data used to generate the model prediction for condition 1. Each entry in the first matrix shows the number of times that two blocks were touched and the number of times that the train sound was heard. For example, the red blocks came into contact twice, and the train sound was produced on neither trial. The second matrix specifies information about the siren sound, and the third matrix captures the perceptual features of the seven blocks. The input data for condition 2 are similar but not shown here.

same causal category—the category of WB blocks, say. Because the evidence suggests that yellow blocks produce the siren sound when paired with WB blocks, our model infers that the combination of a yellow block and a white block will probably produce the siren sound (Fig. 23B). In condition 2, however, the evidence supports the hypothesis that white blocks and red blocks belong to a category—the category of WR blocks. Because the evidence suggests that WR blocks and yellow blocks produce no sound when paired, the model infers that the combination of a yellow block and a white block will probably fail to produce the siren sound. The input data used to generate the model predictions for condition 1 are shown in Fig. 23C. The data include a matrix of observations for each effect (train sound and siren sound) and a matrix of perceptual features that specifies the color of each block.

The result in Fig. 23B follows directly from the observation that white blocks are just like blue blocks in condition 1, but that white blocks are just like red blocks in condition 2. This observation may seem simple, but Schulz and colleagues point out that it cannot be captured by the standard Bayes net approach to causal learning. The standard approach will learn a Bayes net defined over variables that represent events, such as a contact event involving a red block and a white block. The standard approach, however, has no basis for making predictions about novel events such as a contact event involving a yellow block and a white block. Our model overcomes this limitation by learning categories of objects and recognizing that the outcome of a novel event can be predicted given information about the category membership of the objects involved. The work of Schulz et al. suggests that young children are also able to learn causal categories from interaction data and to use these categories to make inferences about novel events.

We have now revisited three central themes addressed by our experiments—causal categorization, the tradeoff between causal and perceptual information, and causal interactions—and showed how each one is grounded in the literature on cognitive development. We described how our model can help to explain several empirical results, but future developmental experiments are needed to test our approach in more detail. Causal reasoning has received a great deal of attention from the developmental community in recent years, but there are still few studies that explore learning to learn. We hope that our approach will stimulate further work in this area, and we expect in turn that future empirical results will allow us to improve our approach as a model of children's learning.

11. Discussion

This paper developed a computational model that can handle multiple inductive tasks, and that learns rapidly about later tasks given experience with previous tasks from the same family. Our approach is motivated by the idea that learning to learn can be achieved by acquiring abstract knowledge that is relevant to all of the inductive tasks within a given family. A hierarchical Bayesian approach helps to explain how abstract knowledge can be learned after experience with the first few tasks in a family, and how this knowledge can guide subsequent learning. We illustrated this idea by developing a hierarchical Bayesian model of causal learning.

The model we described includes representations at several levels of abstraction. Near the top of the hierarchy is a schema that organizes objects into categories and specifies the causal powers and characteristic features of these categories. We showed that schemata of this kind support top-down learning and capture background knowledge that is useful when learning causal models for sparsely observed objects. Our model, however, also supports bottom-up learning, and we showed how causal schemata can be learned given perceptual features and contingency data.

Our experiments suggest that our model matches the abilities of human learners in several respects. Experiment 1 explored one-shot causal learning and suggested that people learn schemata which support confident inferences given very sparse data about a new object. Experiment 2 explored a case where people learn a causal model for an object that is qualitatively different from all previous objects. Strong inductive constraints are critical when data are sparse, but Experiment 2 showed that people (and our model) can overrule these constraints when necessary. Experiment 3 focused on ''zero-shot causal learning'' and showed that people make inferences about the causal powers of an object based purely on its perceptual features. Experiment 4 suggested that people form categories that are distinguished only by their causal interactions with other categories.

Our experiments used two general strategies to test the psychological reality of the hierarchy used by our model. One strategy focused on inferences at the bottom level of the hierarchy. Experiments 1, 3, and 4 considered one-shot or zero-shot causal learning and suggested that the upper levels of the model explain how people make confident inferences given very sparse data about a new object. A second strategy is to directly probe what people learn at the upper levels of the hierarchy. Experiments 3 and 4 asked participants to sort objects into categories, and the resulting sorts provide evidence about the representations captured by the schema level of our hierarchical model. A final strategy that we did not explore is to directly provide participants with information about the upper levels of the hierarchy, and to test whether this information guides subsequent inferences. Consider, for instance, the case of a science student who is told that ''pineapple juice is an acid, and acids turn litmus paper red.'' When participants are sensitive to abstract statements of this sort, we have additional evidence that their mental representations capture some of the same information as the hierarchies used by our model.

11.1. Related models

Our work is related to three general areas that have been explored by previous researchers: causal learning, categorization, and learning to learn. This section compares our approach to some of the formal models that have been developed in each area.

11.1.1.Learning to learn

The hierarchical Bayesian approach provides a general framework for explaining learning to learn, and it has been explored by researchers from several communities. Statisticians and machine learning researchers have explored the theoretical properties of hierarchical Bayesian models (Baxter, 1998) and have applied them to challenging real-world problems (Blei, Ng, & Jordan, 2003; Gelman, Carlin, Stern, & Rubin, 2003; Good, 1980). Psychologists have suggested that these models can help to explain human learning, and they have used them to explore how children learn to learn words (Kemp et al., 2007) and feature-based categories (Perfors & Tenenbaum, 2009).

Our work is motivated by many of the same general considerations as these previous approaches, but it represents one of the first attempts to explore learning to learn in a causal context. Our work also helps to demonstrate the flexibility of the hierarchical Bayesian approach to learning. Previous hierarchical approaches in the psychological literature often use hierarchies where the knowledge at the top level is very simple—for example, where this knowledge is captured by one or two parameters (Kemp et al., 2007). Our work illustrates that the same basic approach can explain how richer kinds of abstract knowledge can be acquired. We showed, for example, how causal schemata can be learned, where each schema is a system that includes causal categories along with category-level causal models that specify causal relationships between these categories.

11.1.2. Causal learning

Although there are few accounts of learning to learn in a causal setting, there are many previous models of causal learning and reasoning. Like many of these models (Gopnik & Glymour, 2002; Griffiths & Tenenbaum, 2005; Pearl, 2000), our work uses Bayesian networks to capture causal knowledge. For example, each object-level causal model in our framework is formalized as a causal Bayesian network. Note, however, that our approach depends critically on a level of representation that is more abstract than causal networks. We suggest that human inferences rely on causal schemata or systems of knowledge that capture expectations about object-level causal models.

11.1.3. Categorization

A causal schema groups a set of objects into categories, and our account of schema learning builds on two previous models of categorization. Our approach assumes that the category assignments of two objects will predict how they relate to each other, and the same basic assumption is made by the infinite relational model (Kemp et al., 2006), a probabilistic approach that organizes objects into categories that relate to each other in predictable ways. We also assume that objects belonging to the same category will tend to have similar features, and we formalize this assumption using the same probabilistic machinery that lies at the heart of Anderson's rational approach to categorization (Anderson, 1991). Our model can therefore be viewed as an approach that combines these two accounts of categorization with a Bayesian network account of causal reasoning. Because all of these accounts work with probabilities, it is straightforward to bring them together and create a single integrated framework for causal reasoning.

11.1.4. Categorization and causal learning

Previous authors have studied the relationship between categorization and causal reasoning (Waldmann & Hagmayer, 2006), and Lien and Cheng (2000) present a formal model that combines these two aspects of cognition. These authors consider a setting similar to our third experiment where learners must combine contingency data and perceptual features to make inferences about sparsely observed objects. Their approach assumes that the objects of interest can be organized into one or more hierarchies, and that there are perceptual features which pick out each level in each hierarchy. Each perceptual feature is assumed to be a potential cause of effect e, and the *probabilistic contrast* for each cause c with respect to the effect is $P(e^+ \mid c^+) - P(e^+ \mid c^-)$. Lien and Cheng suggest that the best explanation of the effect is the cause with maximal probabilistic contrast.

Although related to our own approach, the theoretical problem addressed by the principle of maximal contrast is different from the problem of discovering causal schemata. In our terms, Lien and Cheng assume that a learner already knows about several overlapping categories, where each category corresponds to a subtree of one of the hierarchies. They do not discuss how these categories might be discovered in the first place, but they provide a method for identifying the category that best explains a novel causal relation. We have focused on a different problem: Our schema-learning model does not assume that the underlying categories are known in advance, but it shows how a single set of nonoverlapping categories can be discovered.

Our work goes beyond the Lien and Cheng approach in several respects. Our model accounts for the results of Experiments 1, 2, and 4, which suggest that people organize perceptually identical objects into causal categories. In contrast, the Lien and Cheng model has no way to address problems where all objects are perceptually identical. In their own experiments, Lien and Cheng apply their model to several problems where causal information and perceptual features are both available, and where a subset of the perceptual features pick out the underlying causal categories. Experiment 3, however, exposes a second important difference between our model and their approach. Our model handles cases like Fig. 14 where the features provide a noisy indication of the underlying causal categories, but the Lien and Cheng approach can only handle causal categories that correlate perfectly with a perceptual feature. Experiment 3 supports our approach by demonstrating that people can discover categories in settings where perceptual features correlate roughly with the underlying categories, but where there is no single feature that perfectly distinguishes these categories.

Although the Lien and Cheng model will not account for the results of any of our experiments, it goes beyond our work in one important respect. Lien and Cheng suggest that potential causes can be organized into hierarchies—for example, ''eating cheese'' is an instance of ''eating dairy products'' which in turn is an instance of ''eating animal products.'' Different causal relationships are best described at different levels of these hierarchies—for example, a certain allergy might be caused by ''eating dairy products,'' and a vegan may feel sick at the thought of ''eating animal products.'' Our model does not incorporate the notion of a causal hierarchy—objects are grouped into categories, but these categories are not grouped into higher-level categories. As described in the next section, however, it should be possible to develop extensions of our approach where object-level causal models and features are generated over a hierarchy rather than a flat set of categories.

11.2. Learning and prior knowledge

Any inductive learner must rely on prior knowledge of some kind and our model is no exception. This section highlights the prior knowledge assumed by our approach and discusses where this knowledge might come from. Understanding the knowledge assumed by our framework is especially important when considering its developmental implications. The ultimate goal should be to situate our approach in a developmental sequence that helps to explain the origin of each of its components, and we sketch some initial steps towards this goal.

The five shaded nodes in Fig. 3D capture much of the knowledge assumed by our approach. Consider first the nodes that represent domains (e.g., people) and events (e.g., `ingests`(·,·)). Domains can be viewed as categories in their own right, and these categories might emerge as the outcome of prior learning. For example, our approach could help to explain how a learner organizes the domain of physical objects into animate and inanimate objects, and how the domain of animate objects is organized into categories like people and animals. As these examples suggest, future extensions of our approach should work with hierarchies of categories and explore how these hierarchies are learned. It may be possible, for example, to develop a model that starts with a single, general category (e.g., physical objects) and that eventually develops a hierarchy which indicates that people are animate objects and that animate objects are physical objects. There are several probabilistic approaches that work with hierarchies of categories (Kemp, Griffiths, Stromsten, & Tenenbaum, 2004; Kemp & Tenenbaum, 2008; Schmidt, Kemp, & Tenenbaum, 2006), and it should be relatively straightforward to combine one of these approaches with our causal framework.

Although our model helps to explain how categories of objects are learned, it does not explain how categories of events might emerge. There are several probabilistic approaches that explore how event categories could be learned (Buchsbaum, Griffiths, Gopnik, & Baldwin, 2009; Goodman, Mansinghka, & Tenenbaum, 2007), and it may be possible to combine these approaches with our framework. Ultimately researchers should aim for models that can learn hierarchies of event categories—for example, touching is a kind of physical contact, and physical contact is a kind of event.

The third shaded node at the top of Fig. 3D represents a domain-level problem. Our framework takes this problem for granted but could potentially learn which problems capture possible causal relationships. Given a set of domains and events, the learner could consider a hypothesis space that includes all domain-level problems constructed from these elements, and the learner could identify the problems that seem most consistent with the available data. Different domain-level problems may make different assumptions about which events are causes and which are effects, and intervention data and temporal data are likely to be especially useful for resolving this issue: Effect events can be changed by intervening on cause effects, but not vice versa, and event effects usually occur some time after cause effects.

In many cases, however, the domain-level problem will not need to be learned from data, but will be generated by inheritance over a hierarchy of events and a hierarchy of domains.

For example, suppose that a learner has formulated a domain-level problem which recognizes that acting on a physical object can affect the state of that object:

$$\texttt{action}(\texttt{physical object 1}, \texttt{physical object 2}) \overset{?}{\rightarrow} \texttt{state}(\texttt{physical object 2})$$

If the learner knows that a touching is an action, that people and toys are both physical objects, and that emitting sound is a state, then the learner can use domain and event inheritance to formulate a domain-level problem which recognizes that humans can make toys emit sound by touching them:

$$\texttt{touch}(\texttt{person}, \texttt{toy}) \overset{?}{\rightarrow} \texttt{emits\_sound}(\texttt{toy})$$

A domain-level problem identifies a causal relationship that might exist, but additional evidence is needed to learn a model which specifies whether this relationship exists in reality. The distinction between domain-level problems and causal models is therefore directly analogous to the distinction between possibility statements (this toy could be made out of wood) and truth statements (this toy is actually made out of plastic). Previous authors have suggested that possibility statements are generated by inheritance over ontological hierarchies (Keil, 1979; Sommers, 1963), and that these hierarchies can be learned (Schmidt et al., 2006). Our suggestions about the origins of domain-level problems are consistent with these previous proposals.

The final two shaded nodes in Fig. 3D represent the event and feature data that are provided as input to our framework. Like most other models, our current framework takes these inputs for granted, but it is far from clear how a learner might convert raw sensory input into a collection of events and features. We can begin to address this question by adding an additional level at the bottom of our hierarchical Bayesian model. The information observed at this level might correspond to sensory primitives, and a learner given these observations might be able to identify the events and features that our current approach takes for granted. Goodman et al. (2007) and Austerweil and Griffiths (2009) describe probabilistic models that discover events and features given low-level perceptual primitives, and the same general approach could be combined with our framework.

Even if a learner can extract events and features from the flux of sensory experience, there is still the challenge of deciding which of these events and features are relevant to the problem at hand. We minimized this challenge in our experiments by exposing our participants to simple settings where the relevant features and events were obvious. Future analyses can consider problems where many features and events are available, some of which are consistent with an underlying causal schema, but most of which are noisy. Machine learning researchers have developed probabilistic methods for feature selection that learn a weight for each feature and are able to distinguish between features that carry useful information and those that are effectively random (George & McCulloch, 1993; Neal, 1996). It should be possible to combine these methods with our framework, and the resulting model may help to explain how children and adults extract causal information from settings that are noisy and complex.

We have now discussed how several components of the framework in Fig. 3D could be learned rather than specified in advance. Although our model could be extended in several directions, note that there are fundamental questions about the origins of causal knowledge that it does not address. For example, our model suggests how a schema learner might discover the schema that accounts best for a given domain, but it does not explain how a learner might develop the ability to think about schemata in the first place. Similarly, our model can learn about the causal powers of novel objects, but it does not explain how a precausal learner might develop the ability to think about causal powers. There are two possible solutions to developmental questions like these: Either concepts like causal schema and causal power could be innate, or one or both of these concepts could emerge as a consequence of early learning. Our work is compatible with both possible solutions, and future modeling efforts may help to suggest which of the two is closer to the truth.

12. Conclusion

We developed a hierarchical Bayesian framework that addresses the problem of learning to learn. Given experience with the causal powers of an initial set of objects, our framework helps to explain how learners rapidly learn causal models for subsequent objects from the same family. Our approach relies on the acquisition and use of causal schemata, or systems of abstract causal knowledge. A causal schema organizes a set of objects into categories and specifies the causal powers and characteristic features of each categories. Once acquired, these causal schemata support rapid top-down inferences about the causal powers of novel objects.

Although we focused on causal learning, the hierarchical Bayesian approach can help to explain learning to learn in other domains, including word learning, visual learning, and social learning. The hierarchical Bayesian approach accommodates both abstract knowledge and learning, and it provides a convenient framework for exploring two fundamental questions about cognitive development: how abstract knowledge is acquired, and how this knowledge is used to support subsequent learning. Answers to both questions should help to explain how learning accelerates over the course of cognitive development, and how this accelerated learning can bridge the gap between knowledge in infancy and adulthood.

Notes

1. We will assume that g and s are defined even if $a = 0$ and there is no causal relationship between o and e. When $a = 0$, g and s could be interpreted as the polarity and strength that the causal relationship between o and e would have if this relationship actually existed. Assuming that g and s are always defined, however, is primarily a mathematical convenience.
2. Unlike Experiment 1, the background rate is nonzero, and these probability distributions are not equivalent to distributions on the causal power of a test block.

3. In particular, the pairwise activation condition of Experiment 4 is closely related to the symmetric regular condition described by Kemp et al. (2010).

Acknowledgments

An early version of this work was presented at the Twenty Ninth Annual Conference of the Cognitive Science Society. We thank Bobby Han for collecting the data for Experiment 4, and Art Markman and several reviewers for valuable suggestions. This research was supported by the William Asbjornsen Albert memorial fellowship (C. K.), the James S. McDonnell Foundation Causal Learning Collaborative Initiative (N. D. G., J. B. T), and the Paul E. Newton Chair (J. B. T.)

References

Aldous, D. (1985). Exchangeability and related topics. In P. L. Hennequin (Ed.), *École d'Été de Probabilités de Saint-Flour, XIII—1983* (pp. 1–198). Berlin: Springer.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429.

Austerweil, J., & Griffiths, T. L. (2009). Analyzing human feature learning as nonparametric Bayesian inference. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems 21* (pp. 97–104).

Baxter, J. (1998). Theoretical models of learning to learn. In S. Thrun & L. Pratt (Eds.), *Learning to learn* (pp. 71–94). Norwell, MA: Kluwer.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

Buchsbaum, D., Griffiths, T. L., Gopnik, A., & Baldwin, D. (2009). Learning from actions and their consequences: Inferring causal variables from continuous sequences of human action. In N. A. Taatgen & H. Van Rijn (Eds.), *Proceedings of the 31st annual conference of the Cognitive Science Society* (pp. 2493–2498). Austin, TX: Cognitive Science Society.

Caruana, R. (1997). Multitask learning. *Machine Learning*, *28*, 41–75.

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

Danks, D. (2007). Theory unification and graphical models in human categorization. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation*. (pp. 173–189). Oxford, England: Oxford University Press.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis* (2nd ed.). New York: Chapman & Hall.

George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, *88*, 881–889.

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. In E. M. Keramida (Ed.), *Computing science and statistics: Proceedings of the 23rd symposium interface* (pp. 156–163). Fairfax Station, VA: Interface Foundation.

Glymour, C. (2001). *The mind's arrows: Bayes nets and graphical causal models in psychology*. Cambridge, MA: MIT Press.

Good, I. J. (1980). Some history of the hierarchical Bayesian methodology. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics* (pp. 489–519). Valencia, Spain: Valencia University Press.

Goodman, N. D., Mansinghka, V. K., & Tenenbaum, J. B. (2007). Learning grounded causal models. In D. S. Mc Namara & J. G. Trafton (Eds.), *Proceedings of the 29th annual conference of the Cognitive Science Society* (pp. 305–310). Austin, TX: Cognitive Science Society.

Gopnik, A., & Glymour, C. (2002). Causal maps and Bayes nets: A cognitive and computational account of theory-formation. In P. Carruthers, S. Stich & M. Siegal (Eds.), *The cognitive basis of science* (pp. 117–132). Cambridge, England: Cambridge University Press.

Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, *111*, 1–31.

Gopnik, A., & Sobel, D. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child Development*, *71*, 1205–1222.

Gopnik, A., Sobel, D. M., Shulz, L. E., & Glymour, C. (2001). Causal learning mechanisms in very young children: Two, three, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology*, *37*, 620–629.

Griffiths, T. L., & 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. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 323–346). Oxford, England: Oxford University Press.

Harlow, H. F. (1949). The formation of learning sets. *Psychological Review*, *56*, 51–65.

Jain, S., & Neal, R. M. (2004). A split-merge Markov chain Monte Carlo procedure for the Dirichlet Process mixture model. *Journal of Computational and Graphical Statistics*, *13*, 158–182.

Jones, S. S., Smith, L. B., & Landau, B. (1991). Object properties and knowledge in early lexical learning. *Child Development*, *62*, 499–516.

Keil, F. C. (1979). *Semantic and conceptual development*. Cambridge, MA: Harvard University Press.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: MIT Press.

Kelley, H. H. (1972). Causal schemata and the attribution process. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. S. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: perceiving the causes of behavior* (pp. 151–174). Morristown, NJ: General Learning Press.

Kemp, C. (2008). *The acquisition of inductive constraints*. Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Kemp, C., Griffiths, T. L., Stromsten, S., & Tenenbaum, J. B. (2004). Semi-supervised learning with trees. In S. Thrun, L. Saul & B. Schölkopt (Eds.), *Advances in neural information processing systems 16* (pp. 257–264). Cambridge, England, MA: MIT Press.

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

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, *105*(31), 10687–10692.

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., & Ueda, N. (2006). Learning systems of concepts with an infinite relational model. In Y. Gil & R. J. Mooney (Eds.), *Proceedings of the 21st national conference on artificial intelligence* (pp. 381–388). Menlo park, CA: AAAI Press.

Kemp, C., Tenenbaum, J. B., Niyogi, S., & Griffiths, T. L. (2010). A probabilistic model of theory formation. *Cognition*, *114*(2), 165–196.

Kushnir, T., & Gopnik, A. (2005). Children infer causal strength from probabilities and interventions. *Psychological Science*, *16*, 678–683.

Lagnado, D., & Sloman, S. A. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 856–876.

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

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*, 309–332.

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

Lucas, C. G., & Griffiths, T. L. (2010). Learning the form of causal relationships using hierarchical Bayesian models. *Cognitive Science 34*, 113–147.

Mandler, J. M. (2004). *The foundations of mind: origins of conceptual thought*. New York: Oxford University Press.

Massey, C., & Gelman, R. (1988). Preschoolers' ability to decide whether pictured unfamiliar objects can move themselves. *Developmental Psychology*, *24*, 307–317.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207–238.

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness and category construction. *Cognitive Psychology*, *19*, 242–279.

Nazzi, T., & Gopnik, A. (2000). A shift in children's use of perceptual and causal cues to categorization. *Developmental Science*, *3*(4), 389–396.

Neal, R. M. (1996). *Bayesian learning for neural networks (No. 118)*. New York: Springer-Verlag.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

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

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

Perfors, A. F., & Tenenbaum, J. B. (2009). Learning to learn categories. In N. A. Taatqer & H. Van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 136–141). Austin, TX: Cognitive Science Society.

Sakamoto, Y., & Love, B. C. (2004). Schematic influences on category learning and recognition memory. *Journal of Experimental Psychology: General*, *133*(4), 534–553.

Schmidt, L. A., Kemp, C., & Tenenbaum, J. B. (2006). Nonsense and sensibility: Discovering unseen possibilities. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th annual conference of the Cognitive Science Society* (pp. 744–749). Mahwah, NJ: Erlbaum.

Schulz, L. E., & Gopnik, A. (2004). Causal learning across domains. *Developmental Psychology*, *40*(2), 162–176.

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

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.

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., Sommerville, J. A., Travers, L. V., Blumenthal, E. J., & Stoddard, E. (2009). The role of probability and intentionality in preschoolers' causal generalizations. *Journal of Cognition and Development*, *10*(4), 262–284.

Sommers, F. (1963). Types and ontology. *Philosophical Review*, *72*, 327–363.

Spelke, E. (1994). Initial knowledge: Six suggestions. *Cognition*, *50*, 431–445.

Stevenson, H. W. (1972). *Children's learning*. New York: Appleton-Century-Crofts.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E. J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*, 453–489.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Science*, *10*(7), 309–318.

Thorndike, E. L., & Woodworth, R. S. (1901). The influence of improvement in one mental function upon the efficiency of other functions. *Psychological Review*, *8*, 247–261.

Thrun, S. (1998). Lifelong learning algorithms. In S. Thrun & L. Pratt (Eds.), *Learning to learn* (pp. 181–209). Norwell, MA: Kluwer.

Thrun, S., & Pratt, L. (Eds.) (1998). *Learning to learn*. Norwell, MA: Kluwer.

Waldmann, M. R., & Hagmayer, Y. (2006). Categories and causality: The neglected direction. *Cognitive Psychology*, *53*, 27–58.

Yerkes, R. M. (1943). *Chimpanzees: A laboratory colony*. New Haven, CT: Yale University Press.

Appendix: A schema learning model

This appendix describes some of the mathematical details needed to specify our schema-learning framework in full.

Learning a single object-level causal model

Consider first the problem of learning a causal model that captures the relationship between a cause event and an effect event. We characterize this relationship using four parameters. Parameters a, g, and s indicate whether a causal relationship exists, whether it is generative, and the strength of this relationship. We assume that there is a generative background cause of strength b.

We place uniform priors on a and g, and we assume that the strength parameter s is drawn from a logistic normal distribution:

$$\begin{aligned} \text{logit}(s) &\sim \mathcal{N}(\mu, \sigma^2) \\ \mu &\sim \mathcal{N}(\eta, \tau\sigma^2) \\ \sigma^2 &\sim \text{Inv-gamma}(\alpha, \beta) \end{aligned} \tag{11}$$

The priors on μ and σ^2 are chosen to be conjugate to the Gaussian distribution on logit(s), and we set $\alpha = 2$, $\beta = 0.3$, $\eta = 1$, and $\tau = 10$. The background strength b is drawn from the same distribution as s, and all hyperparameters are set to the same values except for η which is set to -1. Setting η to these different values encourages b to be small and s to be large, which matches standard expectations about the likely values of these variables (Lu et al., 2008). As for all other hyperparameters in our model, $\alpha = 2$, $\beta = 0.3$, $\eta = 1$, and $\tau = 10$ were not tuned to fit our experimental results but were assigned to values that seemed plausible a priori. We expect that the qualitative predictions of our model are relatively insensitive to the precise values of these hyperparameters provided that they capture the expectation that b should be small and s should be large.

Learning multiple object-level causal models

Consider now the problem of simultaneously learning multiple object-level models. The example in Fig. 1A includes two sets of objects (people and drugs), but we initially consider the case where there is just one person and we are interested in problems like

$$\begin{aligned} \texttt{ingests(Alice, Doxazosin)} &\xrightarrow{?} \texttt{headache(Alice)} \\ \texttt{ingests(Alice, Prazosin)} &\xrightarrow{?} \texttt{headache(Alice)} \\ &\vdots \end{aligned}$$

which concern the effects of different drugs on Alice.

As described in the main text, our model organizes the drugs into categories and assumes that the object-level model for each drug is generated from a corresponding causal model at the category level. Our prior $P(z)$ on category assignments is induced by the Chinese Restaurant Process (CRP, Aldous, 1985). Imagine building a partition by starting with a single category including a single object, and adding objects one by one until every object is assigned to a category. Under the CRP, each category attracts new members in proportion to its size, and there is some probability that a new object will be assigned to a new category. The distribution over categories for object i, conditioned on the category assignments for objects 1 through $i - 1$ is

$$P(z_i = a \mid z_1, \ldots, z_{i-1}) = \begin{cases} \frac{n_a}{i-1+\gamma}, & n_a > 0 \\ \frac{\gamma}{i-1+\gamma}, & a \text{ is a new category} \end{cases} \tag{12}$$

where z_i is the category assignment for object i, n_a is the number of objects previously assigned to category a, and γ is a hyperparameter (we set $\gamma = 0.5$). Because the CRP prefers to assign objects to categories which already have many members, the resulting prior $P(z)$ favors partitions that use a small number of categories.

When learning causal models for multiple objects, the parameters for each model can be organized into three vectors $\boldsymbol{a}$, $\boldsymbol{g}$, and $\boldsymbol{s}$. Let Ψ be the tuple ($\boldsymbol{a}$, $\boldsymbol{g}$, $\boldsymbol{s}$, b) which includes all of these parameters along with the background strength b. Similarly, let $\bar{\Psi}$ be the tuple ($\bar{\boldsymbol{a}}$, $\bar{\boldsymbol{g}}$, $\bar{\boldsymbol{s}}$, $\bar{b}$) that specifies the parameters of the causal-models at the category level.

Our prior $P(\bar{\Psi})$ assumes that the entries in $\bar{\boldsymbol{a}}$ and $\bar{\boldsymbol{g}}$ are independently drawn from a Beta(γ_c,γ_c) distribution. Unless mentioned otherwise, we set $\gamma_c = 0.1$ in all cases. Each entry in $\bar{\boldsymbol{s}}$ is a pair that specifies a mean μ and a variance σ^2. We assume that these means and variances are independently drawn from the conjugate prior in Eq. 11 where $\eta = 1$. The remaining parameter $\bar{b}$ is a pair that specifies the mean and variance of the distribution that generates the background strength b. We assume that $\bar{b}$ is drawn from the conjugate prior specified by Eq. 11 where $\eta = -1$.

Suppose now that we are working in a setting (Fig. 1A) that includes two sets of objects—people and drugs. We introduce partitions z_{people} and z_{drugs} for both sets, and we place independent CRP priors on both partitions. We introduce a category-level causal model for each combination of a person category and a drug category, and we assume that each object-level causal model is generated from the corresponding category-level model. As before, we assume that the category-level parameters $\bar{a}$, $\bar{g}$, and $\bar{s}$ are generated independently for each category-level model. The same general strategy holds when working with problems that involve three or more sets of objects. We assume that each set is organized into a partition drawn from a CRP prior, introduce category level models for each combination of categories, and assume that the parameters for these category-level models are independently generated from the distributions already described.

Features

To apply Eq. 8 we need to specify a prior distribution $P(\bar{F})$ on the feature matrix $\bar{F}$. We assume that all entries in the matrix are independent draws from a Beta(γ_f,γ_f) distribution. Unless mentioned otherwise, we set $\gamma_f = 0.5$ in all cases. Our feature model is closely related to the Beta-Bernoulli model used by statisticians (Gelman et al., 2003) and is appropriate for problems where the features are binary. Some features, however, are categorical (i.e., they can take many discrete values), and others are continuous. Our approach can handle both cases by replacing the Beta-Bernoulli component with a Dirichlet-multinomial model, or a Gaussian model with conjugate prior.

Inference

Our model can be used to learn a schema (top level of Fig. 1), to learn a set of object-level causal models (middle level of Fig. 1), or to make predictions about future events involving a set of objects (bottom level of Fig. 1). All three kinds of inferences can be carried out using a Markov chain Monte Carlo (MCMC) sampler. Because we use conjugate priors on the model parameters at the category level ($\bar{\Psi}$ and $\bar{F}$), it is straightforward to integrate out these parameters and sample directly from $P(z,\Psi|V)$. To sample the schema assignments in z, we combined Gibbs updates with the split-merge scheme described by Jain and Neal (2004). We used Metropolis-Hasting updates on the parameters Ψ of the object-level models and found that mixing improved when the three parameters for a given object i (a_i, g_i and s_i) were updated simultaneously. To further facilitate mixing, we used Metropolis-coupled MCMC: We ran several Markov chains at different temperatures and regularly considered swaps between the chains (Geyer, 1991).

We evaluated our model by comparing two kinds of distributions against human responses. Figs. 8, 10, 16, and 20 show posterior distributions over the activation strength of a given block, and Fig. 17 shows a posterior distribution over category assignments. In all cases except Fig. 20ii,iii we computed model predictions by drawing a bag of MCMC samples from $P(z,\Psi \mid V,F)$. We found that our sampler did not mix well when directly applied to the setting in Experiment 4 and therefore used importance sampling to generate the predictions in Fig. 20ii,iii. Let a partition z be *plausible* if it assigns objects o_1 through o_9 to the same category and o_{10} through o_{18} to the same category. There are 15 plausible partitions, and we define a distribution $P_1(\cdot)$ that is uniform over these partitions:

$$P_1(z) = \begin{cases} \frac{1}{15}, & \text{if } z \text{ is plausible} \\ 0, & \text{otherwise} \end{cases}$$

For each plausible partition z we used a separate MCMC run to draw 20,000 samples from $P(\Psi \mid V,z)$. When aggregated, these results can be treated as a single large sample from a distribution $q(z,\Psi)$ where

$$q(z, \Psi) \propto P(\Psi \mid V, z) P_1(z).$$

We generated model predictions for Fig. 20ii,iii using $q(\cdot,\cdot)$ as an importance sampling distribution. The importance weights required take the form $P(z)P(V \mid z)$, where $P(z)$ is induced by Eq. 12 and $P(V \mid z) = \int P(V|\Psi,z)P(\Psi \mid z)d\Psi$ can be computed using a simple Monte Carlo approximation for each plausible z.