Two Proposals for Causal Grammars

Thomas L. Griffiths
Department of Cognitive and Linguistic Sciences
Brown University

Joshua B. Tenenbaum
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology

1. Introduction

In the previous chapter (Tenenbaum, Griffiths, & Niyogi, this volume), we introduced a framework for thinking about the structure, function, and acquisition of intuitive theories inspired by an analogy to the research program of generative grammar in linguistics. We argued that a principal function for intuitive theories, just as for grammars for natural languages, is to generate a constrained space of hypotheses that people consider in carrying out a class of cognitively central and otherwise severely underconstrained inductive inference tasks. Linguistic grammars generate a hypothesis space of syntactic structures considered in sentence comprehension; intuitive theories generate a hypothesis space of causal network structures considered in causal induction. Both linguistic grammars and intuitive causal theories must also be reliably learnable from primary data available to people. In our view, these functional characteristics of intuitive theories should strongly constrain the content and form of the knowledge they represent, leading to representations somewhat like those used in generative grammars for language. However, until now we have not presented any specific proposals for formalizing the knowledge content or representational form of "causal grammars." That is our goal here.

Just as linguistic grammars encode the principles that implicitly underlie all grammatical utterances in a language, so do causal grammars express knowledge more abstract than any one causal network in a domain. Consequently, existing approaches for representing causal knowledge based on Bayesian networks defined over observable events, properties or variables, are not sufficient to characterize causal grammars. Causal grammars are in some sense analogous to the "framework theories" for core domains that have been studied in cognitive development (Wellman & Gelman, 1992): the domain-specific concepts and principles that allow learners to construct appropriate causal networks for reasoning about

We thank Elizabeth Baraff, Tania Lombrozo, Rebecca Saxe, and Marty Tenenbaum for helpful conversations about this work. Many of the ideas in this chapter emerged from discussions and collaborations with Sourabh Niyogi and Charles Kemp. TLG was supported by a Stanford Graduate Fellowship. JBT was supported by the Paul E. Newton Career Development Chair and a grant from NTT Communication Sciences Laboratory.

systems in a given domain, and the expectations about which causal relations are more or less likely a priori, that enable causal learning to proceed from the sparse data typically encountered.

Tenenbaum, Griffiths, and Niyogi (this volume) described a hierarchical Bayesian framework that more precisely formalizes the relationship between causal grammars and causal Bayesian networks. A learner's observations of the world are interpreted in terms of a hierarchy of increasingly abstract and general theories, with each level generating a hypothesis space and prior probability distribution for theories at the level below, thereby allowing those lower-level theories to be learned in a top-down fashion based on only sparse bottom-up input. The most specific level of intuitive theories concern cause-effect relationships between observable events, properties or variables, which can be formalized as causal Bayesian networks. Higher levels of abstraction require something like the representational powers of generative grammars, specifying categories of variables and rules for how composing those categories to construct the constrained space of causal networks that are possible in a given domain. For instance, to recall an example from our first chapter, a learner's beliefs about possible causal network structures in a simplified medical domain might be characterized by these two principles:

P1 There exist three classes of variables: *Symptoms*, *Diseases*, and *Behaviors*. These classes are open and of unspecified size, allowing the possibility that a new variable may be introduced.

P2 Causal relations between variables are constrained with respect to these classes: direct links arise only from behaviors to diseases and from diseases to symptoms. These links may be overlapping, e.g., diseases tend to have multiple effects and symptoms tend to have multiple causes.

Figure 1 shows several causal networks (Graphs 1-4) that are consistent with these principles, as well as two networks (Graphs 5 and 6) that would be impossible or "ungrammatical" under this theory.

In this chapter, we will examine in detail two proposals for formalizing causal grammars, the first based on a kind of graph grammar that we call a *graph schema*, and the second based on a typed predicate logic. We will present applications of each approach to characterizing several small-scale intuitive theories, and show how these approaches support quantitative modeling of behavioral studies on causal learning and theory acquisition with both child and adult subjects. Both proposals will be defined in a probabilistic setting, so that we can show precisely how they support causal learning and how they themselves can be learned using the hierarchical Bayesian framework of the previous chapter. For neither approach will we be able to give fully satisfying accounts of learning at both of these levels, because of an inherent tradeoff in the representational power and learnability of any grammar: to the extent that a causal grammar generates rich and subtle constraints on possible causal networks, it will be harder to acquire that grammar from observed data. Presenting two quite different proposals for causal grammars will allow us to explore this tradeoff and lay the groundwork for future attempts to give a full account of the use and origins of abstract causal knowledge.

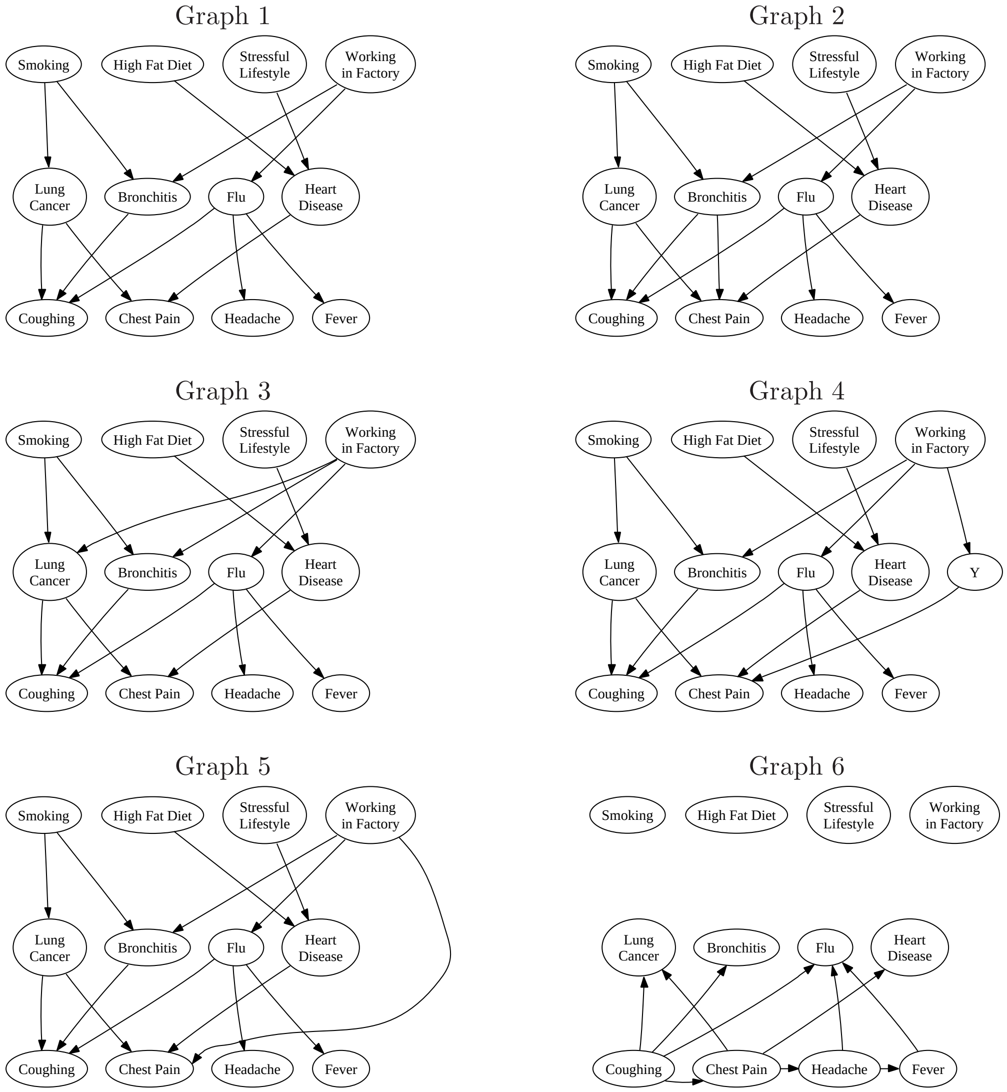


Figure 1. Causal networks illustrating different possible sets of beliefs about the relationships among behaviors, diseases, and symptoms. The same underlying causal grammar generates Graphs 1-4 but not Graphs 5 or 6.

2. Causal grammars in a hierarchical Bayesian framework

Before turning to our two proposals, we will briefly recap the necessary formal machinery for hierarchical Bayesian learning from the previous chapter. Causal Bayesian networks are identified with theories at the lowest, most concrete level of the abstraction hierarchy, level T_0. We will typically identify causal grammars with the T_1-level theories that define hypothesis spaces of T_0-level structures and assign prior probabilities to those hypotheses, thereby guiding inferences about the causal network structure T_0 mostly likely to have given rise to some observed dataset d. A Bayesian learner evaluates a causal network hypothesis T_0 based on its posterior probability,

$$P(T_0|d,T_1) = \frac{P(d|T_0)P(T_0|T_1)}{P(d|T_1)}, \tag{1}$$

where the denominator is

$$P(d|T_1) = \sum_{T_0 \in \mathcal{H}_1} P(d|T_0)P(T_0|T_1). \tag{2}$$

The causal grammar T_1 specifies a probabilistic process for generating causal-network hypotheses. The total set of networks generated by the grammar comprises the hypothesis space $\mathcal{H}_1$. The probability with which the grammar generates any particular network T_0 yields its prior probability, $P(T_0|T_1)$.

Our hierarchical Bayesian analysis also provides a framework for understanding how T_1-level theories may be inferred from data. Given a higher-level theory T_2 that specifies a prior over causal grammars, $P(T_1|T_2)$, and a collection of datasets $\mathcal{D}$ from one or more systems in the domain, the posterior probability distribution over causal grammars is

$$P(T_1|\mathcal{D},T_2) = \frac{P(\mathcal{D}|T_1)P(T_1|T_2)}{P(\mathcal{D}|T_2)}. \tag{3}$$

The denominator $P(\mathcal{D}|T_2)$ is computed in a similar fashion to Equation 2, but summing over theories at levels T_0 and T_1. In discussing our two proposals for causal grammars, one of the critical questions that will arise is how such representations could be learned. Equation 3 provides a theoretical answer to this question, but actually applying these methods to rich structures such as our causal grammars can pose significant computational challenges.

3. Theories as graph grammars

One approach to formalizing causal grammars – or higher-level causal theories – is in terms of a probabilistic graph grammar. In concrete terms, the grammar can be thought of as a machine that outputs samples from an infinite subset of labeled directed graphs, drawn from some probability distribution. Each of these graphs represents the causal structure underlying a causal Bayesian network, but the graphs are not *equivalent* to Bayesian networks: they must be supplemented with a semantic interpretation of the variable that each node represents, and a specification of how each variable depends functionally or probabilistically on its parents in the graph. Putting these complexities aside for now, a grammar for causal graphs is still a useful starting point for formalizing some aspects of abstract causal theories.

This section focuses on one elementary family of graph grammars that are sufficient to represent coarse probabilistic constraints on candidate causal network structures. We call these models *graph schemas.* They generalize an earlier proposal of Tenenbaum and Niyogi (2003). Graph schemas are clearly not adequate to express all theory-like knowledge at levels T_1 or above, but they provide a simple example of how we can begin to formalize abstract causal theories at a level beyond specific causal networks, how those theories could guide Bayesian learning of causal network structure, and how the theories may themselves be learned.

3.1 Graph schemas

A graph schema G is a probabilistic generative model for labeled directed graphs. The key components of the schema are a set of *node classes* and the *class graph*, a directed graph defined over the node classes. (In the context of causal structure learning, each node corresponds to a variable in a causal graphical model, so we will use the terms “node” and “variable” interchangeably.) Generating a graph from a graph schema involves two stages: (1) creating some number of graph nodes and assigning them to node classes; (2) creating connections between nodes in accordance with the class graph, which specifies whether a causal connection may (or must) exist from a particular variable i to a particular variable j as a function of their classes $C(i)$ and $C(j)$. A probabilistic (or deterministic) process must be defined for each of these stages, the details of which may vary from domain to domain. But the basic structures of the set of node classes and the class graph are often sufficient to characterize some important features of a domain theory.

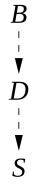
Figure 2 shows a graph schema that we refer to as G_{Dis}, which is intended to capture the constraints expressed by the principles P1 and P2 in our simplified disease domain. Consistent with P1, there are three node classes, labeled B, D and S. Corresponding lowercase letters (b, d, s) will be used to denote specific nodes in each class. All classes are *open*, meaning that the number of nodes in each class is potentially unbounded. Consistent with P2, the two arcs in the class graph specify allowed causal connections: $D \rightarrow S$ specifies that variables in class D may connect causally to variables in class S, and $B \rightarrow D$ specifies that variables in class D may connect causally to variables in class S. Both arcs are dashed to indicate that they represent laws about *possible* causal relations: links that may exist but need not. That is, any individual variable $d \in D$ may be a cause of any individual node $s \in S$, but need not be. A solid arc in the class graph (e.g., in Figure 4 or 5) indicates a *necessary* causal relation, where every node in one class is causally linked to every node in the other class.

Like a generative grammar for a language, G_{Dis} specifies abstract classes of entities (variables, instead of words) and rules about the relations (causal relations, instead of syntactic relations) that may exist between entities of various types. By analogy with linguistic grammars, we say that a graph schema G *generates* Graph i if there exists some way to partition (parse) the nodes in Graph i into the node classes of G, such that all the edges in Graph i are consistent with the possible or necessary connections specified in the class graph of G. As with a grammar for language, a graph grammar can be augmented with probabilities to enable learning and inference. A probabilistic model can be defined over a graph schema by specifying (1) a distribution over the number of nodes in the graph and the number of nodes in each open class; and (2) distributions over which specific causal

Node classes:

Class	Symbol	Status
Behavior	B	open
Disease	D	open
Symptom	S	open

Class graph:



Generative model:

1. *Generate nodes in each class.*

$$\begin{aligned} N_B &\sim \text{PowerLaw}(\alpha_B) \\ N_D &\sim \text{PowerLaw}(\alpha_D) \\ N_S &\sim \text{PowerLaw}(\alpha_S) \end{aligned}$$

2. *Generate causal relations between pairs of nodes.*

Condition	Relation	Probability
$b \in B, d \in D$	$b \rightarrow d$	β_{BD}
$d \in D, s \in S$	$d \rightarrow s$	β_{DS}

Figure 2. A graph schema G_{Dis} for networks of diseases, their causes and their effects.

links exist between nodes in classes connected in the class graph. For G_{Dis}, one way of defining these probabilities is shown in Figure 2. The number of nodes in each class follows a power law distribution, $P(N) \propto 1/N^{\alpha}$, with a class-specific exponent α. After sampling an appropriate number of nodes in each class, a causal link is generated independently at random between each pair of nodes in classes connected in the class graph, with some probability β characteristic of the parent and child classes.

The graph schema G assigns a probability $P(\text{Graph } i|G)$ to any causal network Graph i over a set of N labeled nodes in its domain. $P(\text{Graph } i|G)$ is non-zero if and only if G generates Graph i. The sizes of the graphs generated by a schema are not bounded but must be finite. The probabilities $P(\text{Graph } i|G)$ are normalized to sum to one over all labeled directed graphs with any finite number of nodes. If Graph i represents the structure of a particular causal network (T_0), then G can be thought of as those aspects of the T_1-level theory that generate a hypothesis space and prior over such structures: $P(T_0|T_1)$. Figure 3 shows two graphs sampled from $P(\text{Graph } i|G_{dis})$, each with $\alpha_B = \alpha_D = \alpha_S = 2$ and $\beta_{BD} = \beta_{DS} = 1/2$.

3.2 Examples of graph schemas in different domains

Figures 4 through 6 show schema-based graph grammars for several other domains. None of these grammars comes close to capturing all of people's abstract causal knowledge in the corresponding domain, and important details are oversimplified. The point is merely to illustrate some of the variations in abstract causal knowledge that can arise across domains and how these variations can be represented with different graph schemas. Only the qualitative structure of the graph schemas are shown, specifying the node classes and the

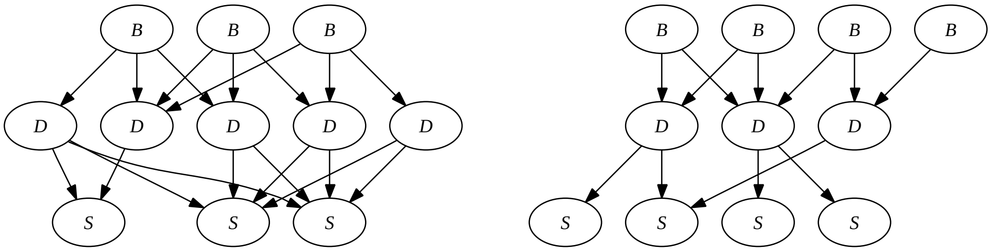


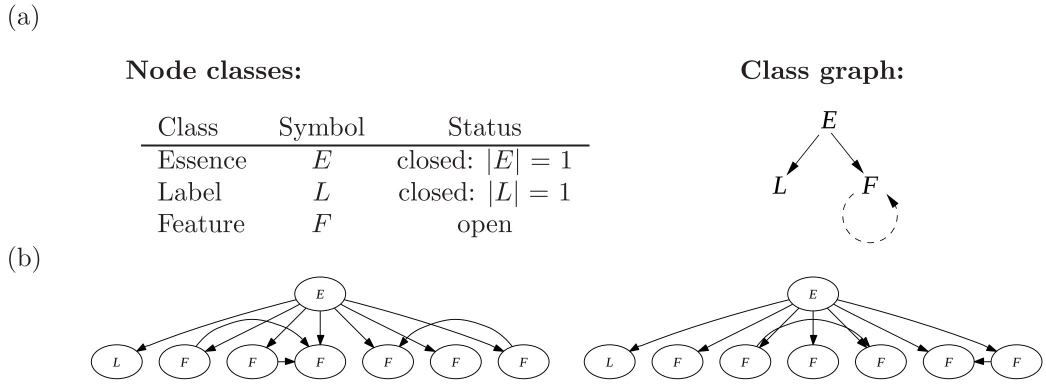
Figure 3. Causal networks sampled from G_{Dis}.

possible and necessary causal links between classes.

The "essentialist" theory, G_{Ess} (Figure 4a), generates causal networks corresponding to simple essentialist concepts for natural kinds (inspired in part by Rehder, this volume; Rehder and Burnett, in press). Different networks (e.g., Figure 4b) generated by this schema could describe different biological species, with different features or different causal relationships between features. They could also describe the same species as a learner acquires more or different beliefs about its characteristic properties and their causal connections. All of these networks place a single essence node in the same abstract causal role. The grammar captures this shared essentialist framework that underlies, supports, and constrains the infinite space of possible species concepts (Gelman, 2003). Under G_{Ess}, every species has a single essence, a single label, and one or more features. In our terminology, the essence class E and label class L are closed, but the feature class F is open. Causal relations may exist between any pair of features (represented by the dashed $F \rightarrow F$ edge in Figure 4a). The essence is also necessarily a cause of every feature (represented by the solid $E \rightarrow F$ edge); even for superficial features not directly a consequence of the essence, the causal relations that give rise to those features depend on the functioning of mechanisms that are themselves generated by the concept's essence. Finally, a causal link necessarily runs from the single essence variable to the single label variable, reflecting the lexical assumption that each concept has a single name.

The "magnetism" theory G_{Mag}, Figure 5a, generates networks appropriate for reasoning about physical causal relationships between the positions of a system of magnets (class M), magnetic objects (class T), and non-magnetic objects (class U). (Magnetic objects, such as a ball bearing, are magnetizable but not sources of magnetic force.) Different systems may have different numbers of objects in these classes (e.g., Figure 5b), but in every system, the position of every magnet causally influences the position of every magnet and every magnetic object. The schema G_{Mag} captures these abstractions by positing three open node classes and necessary causal connections from class M to itself and from M to T.[1]

[1] This graph schema may look implausible as a template for generating causal graphical models, because it generates graphs with directed cycles. However, the problem is easily remedied by imposing a simple discrete dynamics on the variables. Each variable in each node class is indexed by time step, and causal connections between nodes x and y in fact connect $x^{(t)}$, the state of variable x at time t, to $y^{(t+1)}$, the state of variable y at time $t + 1$. By default, each state variable should also depend on its value at the previous time step.

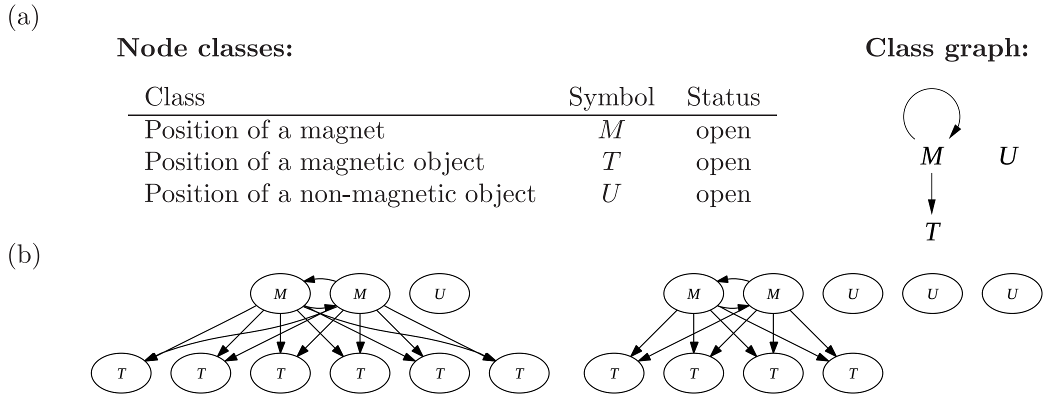


Node classes:

Class	Symbol	Status
Essence	E	closed: $\lvert E\rvert = 1$
Label	L	closed: $\lvert L\rvert = 1$
Feature	F	open

Figure 4. (a) A graph schema G_{Ess} for essentialist categories of natural kinds (c.f. Rehder, this volume). (b) Causal networks sampled from the grammar.

The "rational agent" theory, G_{Agent} (Figure 6), generates causal networks appropriate for a simple version of intuitive psychological reasoning. Different networks generated by this grammar could be appropriate for reasoning about different agents or different kinds of agents, with different specific beliefs, desires, and actions available to them. The graph schema is meant to capture the causal mental architecture that is in common across all these systems of rational agency. An agent has some set of actions A that can be produced, as well as two classes of mental states, beliefs B and desires D. Which action is chosen at a particular time depends upon the agent's beliefs and desires. Variables in class W describe relevant aspects of the state of the world. Actions may affect world states, and world states in turn affect the agent's beliefs. The agent's desires are not directly affected by the world



Node classes:

Class	Symbol	Status
Position of a magnet	M	open
Position of a magnetic object	T	open
Position of a non-magnetic object	U	open

Figure 5. (a) A graph schema G_{Mag} describing the effects of magnets on other objects. (b) Causal networks sampled from the grammar.

Node classes:

Class	Symbol	Status
World states	W	open
Beliefs	B	open
Desires	D	open
Actions	A	open

Class graph:

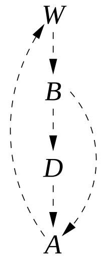


Figure 6. A graph schema G_{Agent} corresponding to a simple theory of mind for intentional agents.

but may be affected by the agent's beliefs about the world.[2] As with the graph schema G_{Dis} for the disease domain, all edges in the class graph for G_{Agent} are dashed, indicating only possible rather than necessary causal relations.

An intriguing difference between causal theories in different kinds of domains is suggested by the different patterns of necessary and possible causal relations in these graph schemas. Physical theories may be more likely to specify necessary causal links, as in G_{Mag}, in which every variable of a certain class possesses the same causal power (or lack thereof) with respect to every variable of another class. Psychological or biological theories may be more likely to specify possible causal links, as in G_{Dis}, G_{Agent}, or G_{Ess}, where a variable's ontological class may constrain its possible cause and effect relations but does not determine them necessarily. The necessary relations that characterize the essence of a natural-kind concept in G_{Ess} may be an exception that proves this rule: essentialist intuitions give rise to some of the few inviolable and all-or-none judgments about otherwise graded conceptions of natural species (Gelman, 2003). Admittedly this particular generalization is quite speculative, but some such generalizations about broad classes of domains could form the content of more abstract causal theories at higher levels of the theory hierarchy – well above the T_1 level that is our focus here.

3.3 The role of graph schemas in learning causal structure

As a model for T_1-level theories in our hierarchical Bayesian framework, probabilistic graph schemas should support the learning of causal network structures (T_0-level theories), and should themselves be learnable given a suitable hypothesis space of graph schemas (a T_2-level theory). To illustrate how graph schemas guide the learning of causal structure, consider how the schema G_{Dis} explains an inference discussed in the previous chapter: positing the existence of a new disease to explain the observation of a previously unseen correlation between a symptom (e.g., *Chest Pain*) and a behavior (e.g., *Working in Factory*).

We first need to define more precisely the probabilistic model implied by each causal network of behaviors, diseases and symptoms. In particular, we need to specify how the probability that an effect occurs depends on the presence or absence of its causes. We assume

[2] Like G_{Mag}, this graph schema oversimplifies by leaving out the dynamic nature of these state variables. But those dynamics can be included here just as we outlined for G_{Mag} in the previous footnote, by indexing each variable by a time step and unfolding all causal connections between each time step and the next.

a *noisy-OR* functional form for these cause-effect relationships (Pearl, 1988). This function is a probabilistic generalization of a logical OR gate, allowing each cause an independent opportunity to bring about the effect. If an effect E is caused by $C_1, \ldots, C_N$, then the noisy-OR states that

$$P(E = 1|c_1, \ldots, c_N) = 1 - (1 - w_0) \prod_{i=1}^{N} (1 - w_i)^{c_i} \tag{4}$$

where $E = 1$ indicates that the effect occurs, and c_i takes on the value 1 if the cause occurs, and 0 otherwise. Here, w_i is the "causal power" of cause i (c.f. Cheng, 1997) – the probability that cause i will produce the effect. The parameter w_0 represents the probability that the effect will occur in the absence of any causes. For the purpose of this demonstration, we will assume that the probability that a patient exhibits each behavior is 0.10; that behaviors cause diseases with power $w_i = 0.1$ and diseases occur spontaneously with $w_0 = 0.001$; and that diseases cause symptoms with power $w_i = 0.8$ while symptoms occur spontaneously with $w_0 = 0.001$. We will also assume that $\alpha_D = 2$.

Figure 7 shows how the graph schema G_{Dis} predicts that the posterior probabilities of five structures should change as evidence for a new correlation accumulates. For simplicity, we assume that only the first five structures shown in Figure 1 are under consideration.[3] Graph 1 is the "null hypothesis", asserting a set of relationships among behaviors, diseases, and symptoms that is consistent with our medical intuitions. Graph 2 adds an additional link from *Bronchitis* to *Chest Pain*. Graph 3 adds an additional link from *Working in Factory* to *Lung Cancer*. Graph 4 introduces a new disease, Y, which connects *Working in Factory* to *Chest Pain*. Graph 5 adds an additional link from *Working in Factory* to *Chest Pain*; this link has causal power $w_i = 0.8 \times 0.1 = 0.08$ for consistency with the assumptions of the other graphs. The dataset d consists of 1000 samples from Graph 1, together with some number of "anomalous" instances in which patients' only relevant behavior is working in a factory, and their only symptom is chest pain. For each patient, only their relevant behaviors and symptoms are observed, not their diseases.

Figure 7 (a) shows the log-likelihood, $\log P(d|\text{Graph } i)$, as a function of the number of anomalous instances observed. This quantity embodies the bottom-up influence of the data on evaluating these causal structure hypotheses, independent of the domain constraints embodied in the graph grammar. With no anomalous instances, these data are most likely under Graph 1, consistent with the fact that they were generated from this structure. As the number of anomalous instances increases, the data become more likely under structures that allow for a correlation between *Working in Factory* and *Chest Pain*. The network with a direct link between *Working in Factory* and *Chest Pain* and the network which postulates a new disease linking these conditions (Graph 5) give the highest probability to these data. The network that postulates a link from *Working in Factory* to *Lung Cancer* (Graph 3) starts off equal to those hypotheses, but declines in probability as more anomalous cases are observed (without any appearance of coughing, the other symptom associated with *Lung Cancer*).

We can compute the posterior probability of each of these graph structures by applying Bayes' rule, as in Equation 1. We want to compute $P(T_0|d, T_1)$, where T_0 refers to one of

[3] Graph 6 provides such a poor fit to the observed data that its likelihood would not show up on Figure 7.

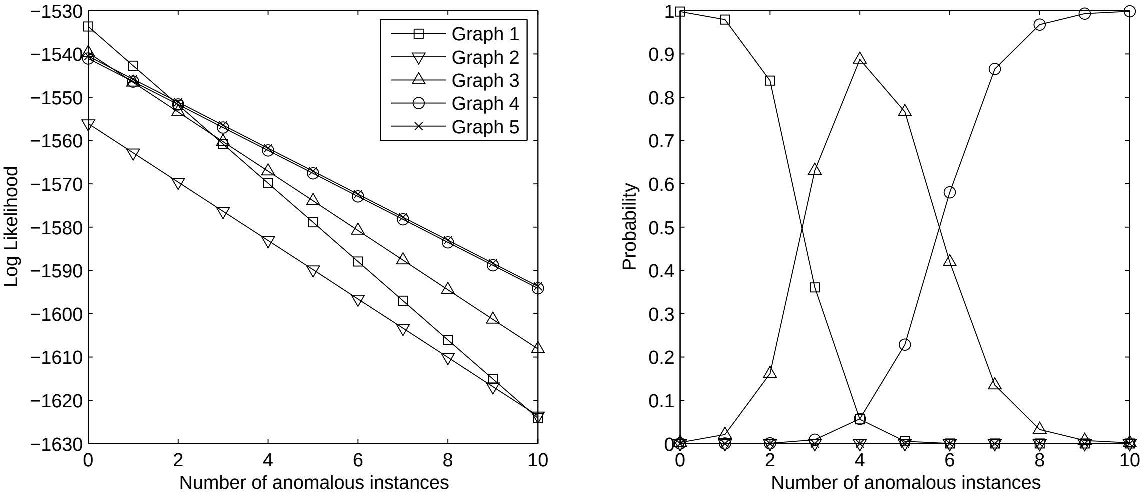


Figure 7. Learning from a correlation between working in factory and chest pain. (a) Likelihood functions for different structures as a function of the number of new instances in which *Working in Factory* and *Chest Pain* co-occur. (b) Posterior probabilities resulting from combining these likelihoods with the prior specified by G_{Dis}.

the five graphs described above and T_1 is the graph schema G_{Dis}. The prior $P(T_0|T_1)$ has both qualitative and quantitative implications for these posterior probabilities. Graph 5 is not generated by G_{Dis}, and consequently has a prior probability of zero. The remaining structures are all generated by the grammar, but with different probabilities. Graph 1, Graph 2, and Graph 3 are all approximately equally probable. Graph 4 is far less probable, for two reasons. First, it is less likely that a structure with five disease nodes will be generated than a structure with four disease nodes, since the probability of the number of nodes is proportional to $1/|D|^2$. Second, there are many more structures with five disease nodes than four, and consequently the average probability of any one of those structures is lower than the average probability of any one structure with four disease nodes.

Figure 7 (b) shows the posterior probabilities of the different causal networks. Despite receiving maximal likelihood (along with Graph 4) given three or more anomalies, Graph 5 has zero posterior probability, due to its inconsistency with G_{Dis}. As the number of anomalous instances increases, there are three discrete stages in the evolution of the posterior probabilities of the other networks. At first, Graph 1 remains favored by both the prior and the likelihood, and the apparent correlation is dismissed as just a coincidence. In the second stage, it becomes clear that the correlation between working in a factory and experiencing chest pain is genuine, and the likelihood favors the other structures. However, the prior is strongly against a new disease, so it seems most plausible that working in a factory is actually a cause of lung cancer, and it is just a coincidence that these patients do not also have the symptom of coughing associated with lung cancer. Finally, the likelihood overwhelms the prior's bias, and it becomes apparent that this pattern of data is evidence for an entirely new disease.

Node classes:

Class	Symbol	Status
Label	L	closed: $\lvert L\rvert = 1$
Feature	F	open

Class graph:

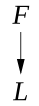


Figure 8. A graph schema, G_{Pro}, for a prototype theory of natural-kind concepts.

3.4 Learning graph schemas

To the extent that the skeletal structure of intuitive theories can be captured by graph schemas for causal networks, the development of intuitive theories may be characterized in terms of changes in those graph schemas. A theory may develop via changes in the causal relations that are necessary or possible, as well as in more radical ways – akin to what Carey (1985) calls "radical conceptual change": node classes may be added or deleted, split or merged. Often the explanatory power of a theory is deepened by adding a new class of hidden causes. For instance, the construction of the *Disease* class of unobservable intervening causes between *Behaviors* and *Symptoms* might have been an important development in medical reasoning. Similarly, Rehder (this volume) posits that essentialist concepts of natural kinds are a relatively late development. Initially, the graph schema for natural-kind concepts might look more like a prototype theory, G_{Pro} (Figure 8). There is no underlying essence node and no explicit representation of causal links between features. Concepts are simply a bundle of one or more features, each linked directly and independently to the concept label.

There are probably many ways by which knowledge at the level of graph schemas can change or grow. One mechanism could be inductive learning from known causal networks or observed patterns of cause-and-effect co-occurence. Kemp, Griffiths, and Tenenbaum (2004) have developed a computational framework for discovering class structures in relational data, which can be used to learn a version of probabilistic graph schemas. The learning algorithm takes as input one or more causal networks T_0, and automatically discovers the classes that are needed to capture the causal relationships among nodes and the probability of a relationship existing between nodes in each pair of classes. This framework does not explicitly distinguish laws for necessary or possible causal links, but treats them as special cases of a more general probabilistic model. The learning algorithm makes no a priori assumption about the number of node classes, but adopts a prior on node-class assignments that prefers to cluster most nodes into a few large classes. The learner can thus automatically discover the most parsimonious grammar, with the smallest number of classes, capable of generating the observed causal network structures.

The model defined by Kemp et al. (2004) effectively computes $P(T_1|T_0, T_2)$, the probability of a graph schema given an observed causal network generated from that grammar and some T_2-level background knowledge. It does so by defining the distributions $P(T_0|T_1)$ in Equation 2 and $P(T_1|T_2)$ in Equation 3. In order to learn a graph schema directly from observations of the variables in a causal system – that is, to compute $P(T_1|\mathcal{D}, T_2)$ – this model can be combined with the Bayesian framework for learning causal network structure described above, which specifies $P(\mathcal{D}|T_0)$.

There has been relatively little empirical work looking at how people learn abstract theories at the level of a graph schema. Tenenbaum and Niyogi (2003) found that people were able to discover a set of classes and causal laws that determined the novel causal relationships among a set of objects in a virtual world. The "objects" in their experiments consisted of blocks that could be moved around and brought into contact with other blocks. When two blocks came into contact, one or both (or neither) could light up, depending on their class memberships and the causal laws operative in the virtual world. The experiments conducted by Tenenbaum and Niyogi (2003) examined how well people learned theories corresponding to the graph schemas shown in Figure 9a. Participants found it easiest to learn laws specifying necessary causal links, such as "Every object belongs to either class A or B, and every object lights up objects in the other class, but not those in the same class." The graph schemas G_1 and G_2 have such a structure. Laws specifying possible but not necessary causal relations, such as G_3 and G_4, were more difficult to learn, but still learnable when the node classes played asymmetric roles, e.g., "Every object belongs to either class A or B, and objects in class A may or may not light up objects in class B". When the node classes played symmetric roles in a law specifying possible causal links – e.g., "Every object belongs to either class A or B, and any object may light up one or more objects in the other class, but not any in the same class" – the theory was most difficult (indeed, practically impossible) for participants to learn.

Kemp et al. (2004) applied their Bayesian algorithm for learning graph schemas to the same tasks, and showed that it accounts for the relative difficulty that participants had in learning these different grammars. Figure 9b shows how the evidence for the correct theory accumulates as more objects are encountered, for all four graph schemas (see Kemp et al., 2004, for details). Evidence is computed as the log ratio of the probability of the data under two T_2-level theories: one in which the causal relations between the objects are generated by a graph schema (with an unknown number of classes), and another in which each object belongs to its own class (and thus no non-trivial graph schema is appropriate). The evidence for the correct grammar-based theory increases in all cases as more objects and relations are observed, but the rate of increase varies across the four theories in accordance with their relative ease of learning. Intuitively, graph schemas that make more constrained predictions about possible causal networks should be easier to learn, because they assign higher probability to the causal networks they do generate. The empirical difficulty of learning was in accord with this principle. For instance, graph schemas specifying necessary causal relations were the easiest to learn, and they were also the most constraining, because an assignment of objects to classes uniquely specifies a single causal network that must be observed.

3.5 Extensions and limitations

The notion of a graph schema can be extended in many ways, to capture richer domain structures. One extension is to allow objects to belong to multiple classes. These classes might form a hierarchy, with each object in a set of nested classes, or a factorial structure, with each object belonging to one class from each of a number of groups. Furthermore, the grammar might depend upon the attributes of the objects, in addition to their class. Another possibility is to allow some kind of generative intermediate representations in the grammar, analogous to the non-terminals in context-free grammars for language, which

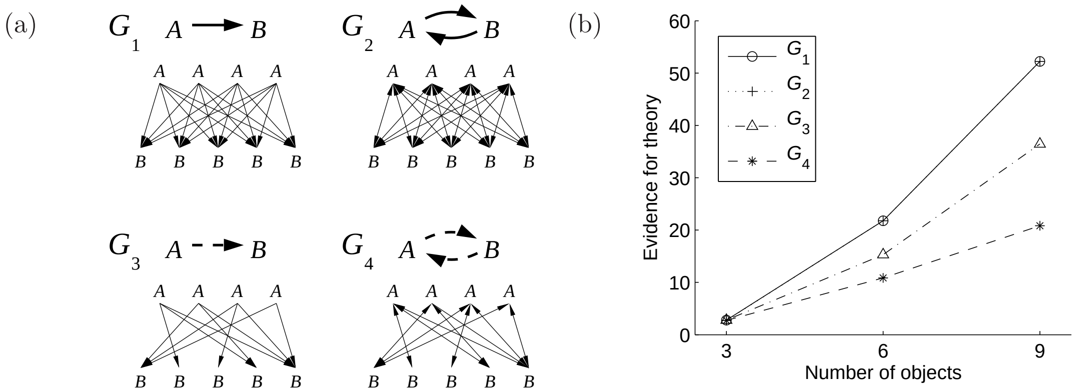


Figure 9. (a) Class graphs and sample networks representing the four graph schemas explored in the experiments of Tenenbaum and Niyogi (2003). (b) The evidence for a theory based on a graph schema increases as learners encounter more objects exhibiting causal relations consistent with that schema, but at a different rate for different graph schemas. Human learners demonstrate the same ordering in the difficulty of learning these graph schemas.

could correspond to mechanisms of transmission linking causes and effects (e.g., Shultz, 1982).

While graph schemas provide a simple way to capture some of the abstract knowledge in T_1-level theories, they leave out other knowledge that is fundamental to intuitive theories and essential for generating hypothesis spaces of causal structures. Foremost is their lack of a sufficiently expressive ontology. They take the nodes or variables of a causal network as primitive entities, without explaining how those variables – or the classes of variables represented in a graph schema – derive from knowledge about types of entities and their properties. Their representations of causal relations and the laws that generate those relations are also fundamentally limited. The class graph of a graph schema specifies which causal relationships are possible or necessary, but not what functional form those relationships take on if they exist. This knowledge of *how* effects depend on their causes should form a crucial part of both T_0- and T_1-level knowledge. At the T_0 level, it is necessary to compute the probability of an observed dataset given a causal network structure, or to make predictions about how novel interventions will affect a causal system. At the T_1 level, it provides valuable constraints on possible causal network models, and thus plays a critical role in explaining how T_0-level theories can be inferred from limited data.

4. Theories as logical grammars

Just as there are many different formalisms that one can adopt for representing linguistic grammars, varying greatly in complexity and coverage, so are there different approaches to formalizing causal grammars. Some of the shortcomings of graph grammars as accounts of T_1-level theories can be addressed by adopting a richer representational language, based on a probabilistic version of predicate logic. Logical grammars can specify more complex and realistic ontologies, in which the types of entities and predicates defined over those entities determine the space of causal Bayesian networks generated by the grammar. Unlike the graph grammars presented in the previous section, which generate only the labeled directed graph skeleton of causal networks, these logical grammars generate full T_0-level theories, each comprising a set of semantically grounded variables, a network of cause-effect relations, and the functional dependencies between causes and effects. By defining a probabilistic model over these logical grammars, analogous to the introduction of probabilities in graph grammars, we can specify a complete probabilistic generative model for T_0-level theories with a well-defined prior distribution $P(T_0|T_1)$. Probabilistic models defined over logical knowledge representations are a promising area of contemporary artificial intelligence research (e.g., Friedman, Getoor, Koller, & Pfeffer, 1999; Pasula & Russell, 2001). Our approach is closest in spirit to the Bayesian Logic framework of Milch, Marthi and Russell (2004).

The theories we consider in this section will be defined using a probabilistic typed (or many-sorted) form of predicate logic. In predicate logic, a set of abstract entities are named with constants, and the properties of those entities are stated using predicates that apply to constants.[4] We will use `typewriter` font when referring to logical notions, writing constants as lower-case letters or words and predicates as capitalized words. For example, in defining a theory of diseases, we could use `ChestPain(p)` to indicate that a particular person, represented by the constant `p`, had the property of having chest pain. In some cases, we might want to talk about a predicate without committing to a particular entity, which can be done by introducing a logical variable, which we will write as a capital letter. Quantification over logical variables can be used to define the set of entities for whom a predicate holds. For example, if we had a world containing three entities, indicated by constants $\mathtt{p}_1$, $\mathtt{p}_2$, and $\mathtt{p}_3$, we could indicate that they all suffered chest pain using the expression $\forall$`P ChestPain(P)`, where `P` is a logical variable that can take on values corresponding to each of the three entities, and $\forall$ is the "universal quantifier", indicating the truth of the proposition it concerns for all values of the variable over which it quantifies. A typed logic divides entities into types, and places constraints on the types of entities to which predicates can apply. We will use the same notation used for predicates to refer to types, since types are naturally translated into predicates (e.g., Enderton, 1972). In the case of diseases, we might want to distinguish two types of entities – `People` and `Objects` – and assert that `ChestPain` is a predicate that can only apply to entities of type `People`.

This discussion of the properties of logic already reveals one of the ways in which logical representations of theories can go beyond graph grammars: they support rich ontologies, defined in terms of types of entities and the predicates that apply to them. We will illustrate some of their other properties and show how such theories may constrain people's causal inferences via an in-depth discussion of the "blicket detector" experimental paradigm (Gopnik & Sobel, 2000; Gopnik et al., 2001; Sobel, Tenenbaum, & Gopnik, 2004; Tenenbaum, Sobel, Griffiths, & Gopnik, submitted). This paradigm showcases people's ability to make causal inferences about novel physical systems from very limited data – just one or a few observations – when guided by appropriate prior knowledge. Traditional bottom-up approaches to learning causal relationships based on rational assessments of correlation, partial correlation, or other statistical measures (e.g., Cheng, 1997; Glymour, 2001; Gopnik et al., 2004; Shanks, 1995) are not readily applicable here, because people do not observe

[4] The abstract entities referred to in a logical theory need not correspond to any kind of physical object. Logical approaches to number theory consider entities that correspond to numbers, and we will consider entities that correspond to intervals of time.

sufficient data to compute these statistics. Our framework provides a rational account of both adults' and children's causal inferences in this paradigm, as well as strong quantitative predictions with a minimum of free numerical parameters.

Relative to the graph grammar formalisms of the previous section, the added power of logical grammars comes at a price. Their richer ontologies introduce more details and greater complexity, making it harder to define satisfying theories that go beyond the simplest systems. It is also much less clear how these logical theories could be learned in full generality, although we can give analyses of several special cases in the blicket detector paradigm. We discuss extensions to our logical framework and prospects for explaining learning at the end of this section.

4.1 The blicket detector

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

Gopnik and Sobel have demonstrated various conditions under which children successfully infer the causal status of blocks from just one or a few observations (Gopnik et al., 2001; Sobel et al., 2004). Two experiments of this kind are summarized in Table 1. In these experiments, children saw two blocks, `a` and `b`, placed on the detector either together or separately across a series of trials. On each trial the blicket detector either became active or remained silent. Table 1 gives the proportion of 4-year-olds who identified `a` and `b` as blickets after several different sequences of trials, encoding contact between the blocks and the detector with the variables A and B and the detector response of the detector with the variable E. Tenenbaum, Sobel, Griffiths, & Gopnik (submitted) tested adults with a similar paradigm, obtaining quantitative judgments that could be used to evaluate the precise predictions of competing computational models. They also used stimuli that were intended to provide ambiguous evidence as to whether blocks were blickets. These data are not presented in Table 1 but are discussed below in Section 4.2.

We will explain the blicket-detector inferences that children and adults draw with reference to a T_1-level theory, expressed using probabilistic logic. This account elaborates on our earlier theory-based model of blicket-detector inferences (Tenenbaum & Griffiths, 2003), by making the theory used in that analysis explicit. The theory should embody people's expectations about how machines (and detectors) work, informed by the specific instructions and familiarization experience provided to experimental participants. For the experiments described in Table 1, the blicket detector was introduced to children as a "blicket machine", and children were told that "blickets make the machine go". In a familiarization phase prior to the critical experimental trials, children saw blocks that activated the machine identified as "blickets" and blocks that did not activate the machine identified as "not blickets". A theory expressing the relevant background knowledge is sketched in Figure 10.

This theory has three parts, specifying an ontology, prescriptions as to causal struc-

Table 1: Probability of Identification as Blickets for 4-year-old Children and Deterministic and Probabilistic Theories

		Children		Deterministic		Probabilistic	
Condition	Stimuli	`a`	`b`	`a`	`b`	`a`	`b`
one cause	$e^+ \mid a^+b^-$ $e^- \mid a^-b^+$ $2e^+ \mid a^+b^+$	0.91	0.16	**1.00**	**0.00**	0.99	0.07
two cause	$3e^+ \mid a^+b^-$ $2e^+ \mid a^-b^+$ $e^- \mid a^-b^+$	0.97	0.78	?	?	**1.00**	**0.81**
indirect screening-off	$2e^+ \mid a^+b^+$ $e^- \mid a^+b^-$	0.00	1.00	**0.00**	**1.00**	0.13	0.90
backwards blocking	$2e^+ \mid a^+b^+$ $e^+ \mid a^+b^-$	1.00	0.34	**1.00**	$\boldsymbol{\beta}$	0.93	0.41
association	$e^+ \mid a^+b^-$ $2e^+ \mid a^-b^+$	0.94	1.00	**1.00**	**1.00**	0.82	0.98
backwards blocking (rare)	$2e^+ \mid a^+b^+$ $e^+ \mid a^+b^-$	1.00	0.25	**1.00**	**0.17**	0.91	0.26
backwards blocking (common)	$2e^+ \mid a^+b^+$ $e^+ \mid a^+b^-$	1.00	0.81	**1.00**	**0.83**	0.98	0.86

Note: The *one cause* and *two cause* conditions are from Gopnik, Sobel, Schulz, and Glymour (2001, Experiment 1). The *indirect screening-off*, *backwards blocking*, *association*, *backwards blocking (rare)*, and *backwards blocking (common)* conditions are from Sobel, Tenenbaum, and Gopnik (2004, Experiments 2 and 3). Boldface indicates the predictions of the model favored by the theory selection procedure outlined in Section 4.3.

ture, and expectations about the functional form of causal relations.[5] The constraints on causal structures and functional form together constitute the "causal laws" expressed in the theory. As a generative grammar for causal Bayesian networks, the three components of the theory respectively generate the nodes of the network, the causal links between nodes, and the local conditional probability distribution for each node as a function of its causes. We describe this generative model below, but first we explain the content of the theory in more detail.

The ontology identifies the types of entities in the domain and predicates defined on those types. The types are organized hierarchically, with the first cut into `Object`, `Power`,

[5] The particular versions of those components shown in Figure 10 represent just one of many possible choices that could work here. We assume this particular theory because it is simple and fairly intuitive, not because we think it corresponds precisely to people's theories in these experiments. However, we will argue that something like the key principles expressed in this theory are critical to explain people's inferences in blicket detector tasks.

Ontology:

Types	Number
`Object`	
`Block`	$N_B \sim$ PowerLaw(α_B)
`Machine`	$N_M \sim$ PowerLaw(α_M)
`Power`	$N_P \sim$ PowerLaw(α_P)
`Trial`	$N_T \sim$ PowerLaw(α_T)

Structural predicates		
`Has(Power, Object)`	$\sim$	Bernoulli(β)
`Activates(Power, Machine)`	$\sim$	Bernoulli(γ)

Causal predicates
`Contact(Object, Object, Trial)`
`Active(Machine, Trial)`

Causal laws:

Structure:

Condition	Relation	Probability
`Has(P, O)` $\wedge$ `Activates(P, M)`	$\forall$`T` `Contact(O, M, T)` $\rightarrow$ `Active(M, T)`	1

Functional form:

`Contact(O, O′, T)` $\sim$ Bernoulli($\cdot$)
`Active(M, T)` $\sim$ Bernoulli(ν) for ν given by a noisy-OR function

Cause	Strength
(Background)	$w_0 = \epsilon$
`Contact(O, M, T)`	$w_1 = 1 - \epsilon$

Figure 10. Sketch of a probabilistic logical theory for causal induction with blicket detectors.

and `Trial`. The `Object` type further divides into `Block` and `Machine`. The predicates are divided into *structural* and *causal* predicates. The causal predicates specify the kinds of variables that will appear as nodes in causal networks (T_0-level theories) describing systems in the domain. The structural predicates concern the basic properties of the entities in the domain and determine which causal relationships can or must hold among causal predicates applied to those entities – that is, the constraints on candidate causal networks defined over grounded causal predicates.

In this case, there are two kinds of causal predicates – variables that can participate in causal relationships: `Contact(O, O′, T)` is true if objects `O` and `O′` are in contact on trial `T`. `Active(M, T)` is true if machine `M` is active on trial `T`. These predicates each apply to a particular `Trial`, representing discrete temporal intervals of the experiment. There are two structural predicates: `Has(P, O)` is true if object `O` has power `P`, e.g., if an object is a blicket, and `Activates(P, M)` is true if power `P` activates machine `M`, e.g., if a machine is a blicket detector. Under this construal, being a blicket or a blicket detector is like being an acid or a base. It is to belong to a class of causal agents or causal patients, defined by the roles that they play in certain laws of causal interaction (White, 1995).

So far, we have focused on the logical structure of the ontology. The probabilistic aspect of the ontology defines a distribution for the number of entities of each type and specifies the probability with which structural predicates hold. In Figure 10, the numbers of blocks, machines, powers, and trials are assumed to follow power-law distributions with parameters α_B, α_M, α_P, and α_T respectively. These distributions are not of consequence in the experiments we will analyze: all blocks and machines are assumed to be observed, and there is just one relevant power concept, `blicket`, that is introduced verbally at the beginning of each experiment. The probability with which each object has a particular power (e.g., is a blicket), β, will be an important variable below. Because there is only one power, `blicket`, and one machine, `d`, and `d` is explicitly called a "blicket detector", the prior probability γ that `Activates(blicket,d)` is true can be assumed to be 1.

The causal laws of a theory specify which causal relations between variables may, must, or are likely to exist, and what form they take. We divide causal laws into the aspects relevant to causal structure, and those that concern functional form. The structural prescriptions of the theory determine the probability that particular causal relationships exist. Each rule consists of a set of conditions stated in terms of structural predicates, under which a causal relationship between two causal predicates holds with some probability. The causal law in Figure 10 asserts that contact between an object and a machine on a given trial will cause the machine to be active on that trial, if the object has some power (e.g., is a blicket) and the machine is activated by that power.

The structural component of the causal laws concerns only the presence or absence of causal links between variables. The strength of those links, e.g., the probability that on any one trial, the presence of the cause will indeed lead to the presence of the effect, are determined by the functional form component of the theory, which specifies the probability distribution associated with each causal predicate. This theory posits a noisy-OR form for the conditional probability distribution of any machine activating given contact with objects that can activate it. For simplicity we reduce these noisy-OR functions to just a single parameter ϵ, representing the "error rate" of a detector – the probability of a "miss" or "false alarm". To begin with, we will assume a deterministic detector with $\epsilon = 0$. This has two important implications. First, the detector cannot activate unless a blicket is in contact with it ($w_0 = 0$). Second, placing a blicket on the detector will always activate the detector ($w_i = 1$). These two assumptions are equivalent to the "activation law" of Sobel et al. 2004): a blicket detector will be active if and only if one or more blickets are in contact with it. Because people always observe which objects are in contact on each trial, the prior probabilities for contact relations are irrelevant.

The deterministic detector theory generates a hypothesis space $\mathcal{H}_1$ of causal networks defined for any set of trials involving any number of blocks and detectors. The generative process defines a prior probability distribution over that space, indicating which causal structures are more or less likely a priori. The process by which a causal network is generated from the theory is as follows:

1. *Generate nodes.* Sample a set of entities of each type from the distribution specified in the **Ontology**. Sample the structural predicates for these entities, using the appropriate probabilities. Generate the set of *grounded causal predicates.* Each of these grounded predicates can be thought of as a binary variable that is true or false. These variables will comprise the nodes of the causal network.
2. *Generate links.* Conditioned on the values of the structural predicates, sam-

ple causal links between nodes from the distribution stated in the **Structure** component of the theory's **Causal laws**.

3. *Generate local conditional probabilities.* For each node, define a local conditional probability as specified in the **Functional form** component of the theory's **Causal laws**, and set the appropriate parameters (or sample them from some prior distribution).

The set of grounded causal predicates is obtained by applying each causal predicate to all entities that can act as its arguments. Assuming that we have two blocks `a` and `b`, a single detector `d`, a single power `blicket`, and the knowledge that `d` is activated by this power, the set of grounded predicates is as follows: `Contact(a,d,T)`, `Contact(b,d,T)`, and `Active(d,T)` for each trial `T`. These grounded predicates are the variables on which the possible causal networks (or T_0-level theories) will be defined.

Since causal relationships are constant over all trials `T`, we can express these causal networks in terms of four graph structures, as shown in Figure 11(a). For shorthand, we use the variables A and B to represent `Contact(a, d, T)` and `Contact(b, d, T)` respectively, and E to represent `Active(d, T)`. The prior probabilities of these networks $P(\text{Graph } i|T_1)$ are determined by the parameter β in the T_1 theory – that is, the prior probabilities that `Has(blicket, a)` and `Has(blicket, b)` are true – since a causal relationship between a block and a detector exists if and only if that block has the power that activates the detector.

The posterior probability distribution over the set of causal networks generated by the theory can be evaluated for each set of trials shown in Table 1, identifying the observed events as the dataset d and applying Bayes' rule as in Equation 1. In the blicket detector experiments, learners are typically asked to judge whether a block (such as `a`) is a blicket. This question asks whether `Has(blicket, a)` is true. Because `Has(blicket, a)` is logically equivalent to the existence of a causal link between `Contact(a, d, T)` and `Active(d, T)`, this question can be reduced to a Bayesian inference over causal network structures: given some observed trials with a blicket detector `d`, the probability that a block is a blicket is the probability that the causal link `Contact(b, d, T)` → `Active(d, T)` exists in the causal network describing the observed system. This can be evaluated by summing the posterior probability of the models in which such a causal relationship exists. For instance, to evaluate the probability that `a` is a blicket, we compute

$$P(A \to E|d, T_1) = \sum_{T_0 \in \mathcal{H}_1} P(A \to E|T_0)P(T_0|d, T_1). \tag{5}$$

For the simple hypothesis space shown in Figure 11(a), this is just $P(\text{Graph } 2|d, T_1) + P(\text{Graph } 3|d, T_1)$.

The predictions of the deterministic detector theory are given in Table 1. The theory's predictions correspond qualitatively with children's judgments but cannot explain all of the inferences observed. In particular, it cannot explain the *two cause* condition in Experiment 1 of Gopnik et al. (2004), which served as an associative control for the *one cause* condition. In the *two cause* condition children saw the detector activate when block `a` was placed on it (alone), on three out of three trials, and also saw the detector activate when block `b` was placed on it (alone), but only on two out of three trials. These data are not compatible with any causal network generated by the deterministic detector theory, and thus the theory's predictions are undefined (indicated by the question marks in Table 1).

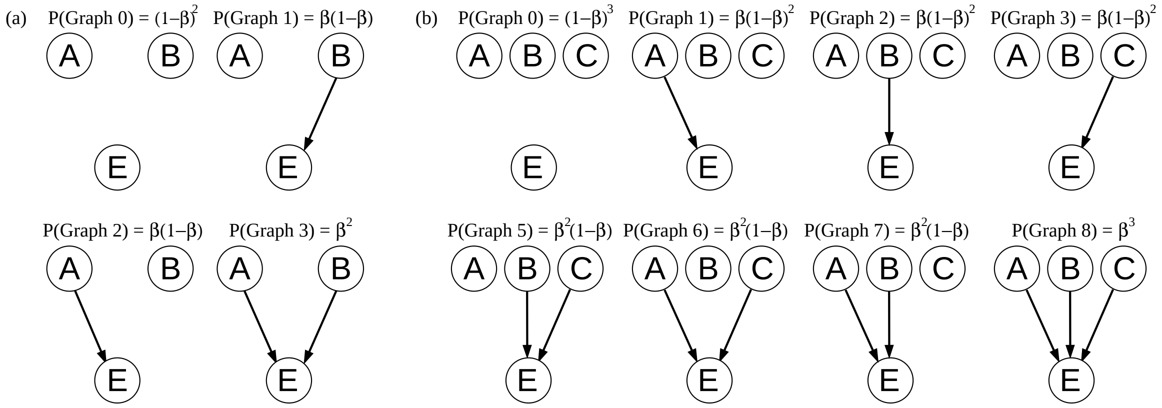


Figure 11. Graph structures generated by the causal theory for the blicket detector. (a) shows the hypothesis space for two blocks, `a` and `b`, while (b) shows the hypothesis space with three blocks, `a`, `b`, and `c`. A, B, and C denote `Contact(a,d,T)`, `Contact(b,d,T)`, and `Contact(c,d,T)` respectively, while E indicates `Active(d,T)`. These causal networks are implicitly quantified over all trials `T`.

The *two cause* dataset can be explained by relaxing one of the assumptions of the deterministic detector theory, to allow blickets to activate detectors only some of the time. We can make this change by allowing ϵ to take on some value greater than zero. This *probabilistic detector* theory gives the same predictions as the deterministic detector theory in the limit as $\epsilon \to 0$, but also predicts that both `a` and `b` are blickets with probability 1 in the *two cause* condition. Different values of ϵ give different predictions. The predictions of this theory with $\epsilon = 0.1$ and $\alpha = 1/3$ are shown in Table 1. This model captures some of the finer details of children's judgments that are not captured by the deterministic detector, such as the fact that `b` is judged less likely to be a blicket than `a` in the *two cause* condition.

4.2 Comparison with alternative accounts

Besides our theory-based Bayesian account, at least two other accounts have been proposed for how children or adults might infer causal structure in the blicket detector paradigm: (1) using a domain-general algorithm for learning causal structure based on statistical dependencies; (2) using domain-general deductive reasoning, augmented with domain-specific assumptions about the relevant class of causal mechanisms (e.g., detectors). Each of these approaches is simpler in some way than our theory-based Bayesian framework, but each is also unable to explain the full range of people's inferences in this paradigm.

Gopnik et al. (2004) advocate the first alternative, proposing that children's causal inferences can be explained by standard bottom-up algorithms for learning causal graphical models (e.g., Spirtes et al., 1993; Pearl, 2000). In particular, they argue that these algorithms will infer the same causal structure (which objects are blickets) that children do in the blicket detector experiments, given observations of the variables A, B, and E across trials. However, the Spirtes et al. (1993) and Pearl (2000) algorithms require as input the probabilistic dependence and independence relations among a set of variables, and these relations cannot be inferred with any reliability from the very small number of trials presented to human learners in the experiments. At least an order of magnitude more data – or some domain-specific assumptions about the causal mechanisms at work – would be necessary for one of these algorithms to work as a rational account of human causal learning. Gopnik et al. (2004) finesse this issue by proposing that learners assume the observed data frequencies can be safely multiplied by some large number, but this assumption is clearly unjustified in many cases. Effectively, it serves to introduce crucial aspects of the deterministic detector theory without making them explicit, because it is justified only in those domains where causal systems are deterministic and fully observable (Tenenbaum et al., submitted).

There is a clearer rational basis for accounts of children's reasoning in logical terms. An assumption that the blicket detector activates if and only there is a blicket in contact with it, plus elementary deductive reasoning capacities, is sufficient to explain all of children's inferences discussed so far (except in the *two cause* condition). However, neither this deductive model nor the Spirtes et al. (1993) or Pearl (2000) bottom-up structure learning algorithms can address another core aspect of human causal inference. Under all these alternative approaches, learners evaluate candidate causal structures in a binary fashion: each structure is either consistent or inconsistent with the data. There is no provision for representing graded degrees of belief about the existence of a causal relation, either a priori, based on expectations about which network structures are more or less plausible, or a posteriori, after observing data that is more or less compatible with multiple structures.

In contrast, our theory-based account naturally explains these gradations, through the probabilistic form of the theory and the probabilistic character of the causal inference process. For instance, after all trials have been observed in the *backwards blocking* condition, the posterior probability that block `b` is a blicket reduces to β: the prior probability that any block is a blicket (assuming the deterministic theory). This reduction to the prior occurs because, having observed that block `a` unambiguously activates the detector (and hence is definitely a blicket), the data now provide no evidence either way about `b`. More generally, even if the data do not provide unambiguous evidence about the status of any one block, they can suggest that some blocks are more likely to be blickets than others, while the prior probability β modulates the overall probability that any block is a blicket. Sobel et al. (2004) and Tenenbaum et al. (submitted) have shown that adults and children reason in accord with these graded predictions.

Tenenbaum et al. (submitted, Experiment 1) studied an analog of the *backwards blocking* condition of Sobel et al. (2004, Experiment 1) and attempted to manipulate the β parameter – the prior probability of encountering objects with the causal power to activate the detector. The experiment was performed with adults, in order to measure more precise graded judgments. They used a "superpencil" detector – rather than a blicket detector – which determined whether apparently normal pencils contained a special kind of lead called "super lead". Participants were randomly assigned to two groups, varying in how they were introduced to the notion of super lead. Both groups of participants were initially shown 12 pencils placed on the detector, one at a time. In what we will refer to as the the *rare* condition, only two of these pencils caused the detector to activate. In the *common* condition, the detector activated for 10 of the 12 pencils. It was hypothesized that learners would set the β parameter in their theories to something like the base rate of causally efficacious objects: 1/6 in the *rare* condition and 5/6 in the *common* condition.

The judgment phase had three stages. In stage one, the baseline, participants were simply shown two new pencils, `a` and `b`. In stage two, participants saw `a` and `b` placed on the detector together, and the detector activated. In stage three, just `a` was placed on the

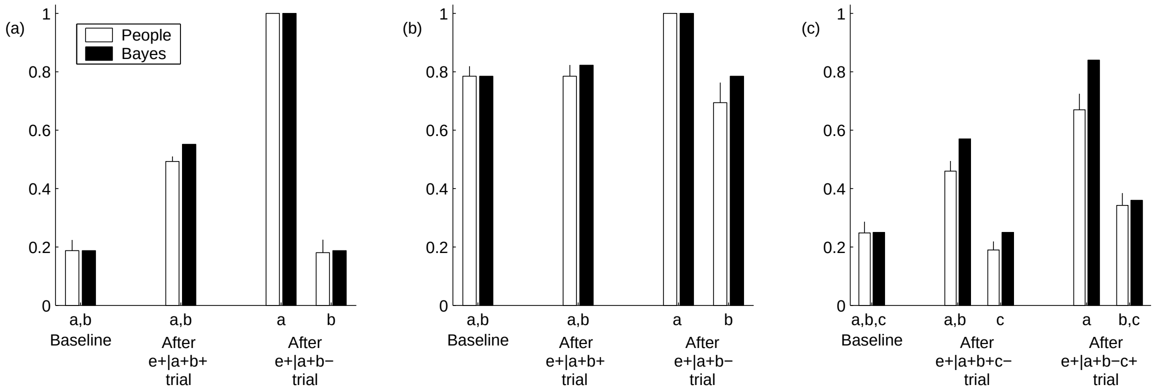


Figure 12. Adult judgments with "superpencils", an analog of the blicket detector task, from Tenenbaum, Sobel, & Gopnik (submitted). (a) and (b) show inferences from the same set of trials, but with different prior probabilities for superpencils, being rare and common respectively. (c) Inferences from ambiguous evidence.

detector, and the detector activated. After each stage, participants were asked to rate the probability that a and b were superpencils. Mean ratings after the first (baseline) stage in each condition were used to set β in our model. Then the same values of β were used to predict judgments in the remaining stages. Mean ratings in the *rare* and *common* conditions are shown in Figure 12 (a) and (b), respectively, along with our model's predictions.

Manipulating the base rate of superpencils during familiarization had the expected effect on people's baseline judgments: β was estimated at 0.19 in the *rare* condition, and 0.78 in the *common* condition. It also affected subsequent judgments as predicted by our Bayesian model under the deterministic detector theory (or the probabilistic detector theory with $\omega = 1 - \epsilon$ as $\epsilon \rightarrow 0$). The probability of a and b being superpencils increases after the first trial, and then the second trial provides unequivocal evidence that a is a superpencil while the probability that b is a superpencil returns to the prior β. Sobel et al. (2004, Experiment 3) replicated this study with 4-year-old children using the blicket detector, but collecting only binary judgments ("blicket", "not a blicket") and without the first two stages of judgment. Table 1 shows the percentage of children who labeled the a and b objects as blickets in each condition. These results showed the same effect of varying prior probabilities seen in the model predictions and adult judgments.

These results are consistent with our theory-based Bayesian account of causal inference, but they do not provide the strongest possible test of whether people's inferences are truly Bayesian. A deductive reasoning account that simply defaults to the observed base rates of causal powers when the data are ambiguous could predict people's judgments just as well. Tenenbaum et al. (submitted, Experiment 2) also asked whether people could make more subtle graded inferences from ambiguous evidence in a fashion consistent with the theory-based Bayesian account. This experiment was equivalent to the superpencil *backwards blocking (rare)* condition, except in the judgment phase. Now that phase began by introducing three new pencils, a, b, and c, and asking for baseline ratings of the probability that each pencil was a superpencil. Participants then saw a and b placed on the detector together, causing the detector to activate, and gave new ratings. Finally, they saw a and c placed on the detector together, causing the detector to activate, and were asked to rate the probability that each of the three pencils was a superpencil. The mean ratings are shown in Figure 12(c).

Model predictions are also shown in Figure 12(c), with β calibrated to the mean probability rating on the first (baseline) judgment. Figure 11(b) shows the hypothesis space $\mathcal{H}_1$ of causal network structures generated by the T_1 theory. With three blocks, there are now eight possible networks. As in Equation 5, the probability that any given block is a blicket is calculated by summing the probability of all network hypotheses in which that block's position is a cause of the detector's activation.

In this experiment, people received no unambiguous clues that a particular pencil was a superpencil: there were no trials on which a single pencil caused the detector to activate. Nonetheless, after the final trial, people were able to infer that a was likely to be a superpencil, while b and c were less likely to be superpencils, with higher judged probability than at the start of the judgment phase, but lower than the peak judgment after the first trial. These judgments are strongly in accord with our theory-based Bayesian account. Figure 12(c) shows that the Bayesian model yields four qualitatively distinct levels of belief over the course of the judgment phase, which are all matched by statistically significant differences in the corresponding ratings of participants. Qualitatively similar inferences were made by 4-year-old children in an analogous experiment with the blicket detector: after the final trial, children were most likely to say that a but not b or c were blickets (Tenenbaum et al., submitted, Experiment 3).

In sum, our theory-based Bayesian framework can explain how people make successful causal inferences about novel physical systems from just one or a few observations, as well as the gradations of judgment and the effects of prior knowledge that arise. These phenomena are not easily explained by other existing approaches to rational causal inference, based on deductive reasoning or bottom-up detection of probabilistic dependencies. Our framework also provides a strong quantitative predictive model with essentially no free numerical parameters. Qualitative assumptions were needed about the form of people's intuitive theories for how machines (or detectors) work, but we would argue that these assumptions are necessary in some form for any account that seeks to give a rational explanation of people's judgments in these scenarios.

While our discussion here has focused on the blicket detector, the same approach of Bayesian inference over logical theories provides a useful framework for understanding causal induction in a variety of settings. In Griffiths (2005) and Griffiths and Tenenbaum (in prep), we show how this approach can explain people's judgments in identifying causal structure from contingency data (Griffiths & Tenenbaum, in press), reasoning about mechanical systems (Gopnik et al., 2004), identifying causal relations and hidden causes with dynamic events (Griffiths et al., 2004), and evaluating evidence for causal relations between variables in different domains (Schulz & Gopnik, in press). The integration of Bayesian inference mechanisms with a logical theory for generating causal-network hypotheses accounts for the effects of several important dimensions along which these learning scenarios vary: the number of independent data points observed (ranging from just one or two samples up to 60-100 samples); the availability of active interventional data in addition to purely passive observational data; the possibility of and strength of evidence for hidden causes; the availability of dynamic real-time observations rather than merely discrete trials; and the a priori plausibility of a mechanism linking candidate causes and effects.

4.3 Learning logical theories

The logical theories outlined in this section are a proposal for a T_1-level representation, specifying one level of our hierarchy of theories. As with graph grammars, statistical inference can in principle be used to learn these T_1-level theories, but the greater representational expressiveness of predicate logic leads to a vastly larger hypothesis space of candidate theories – and thus a much more challenging learning problem in general.

A constrained but quite tractable form of theory learning is parameter estimation: inferring the values of numerical parameters in the theory such as those controling the number of entities of some type (e.g., the α parameters in Figure 10), the frequency with which some structural predicate holds (e.g., the β or γ parameters), or the strength of probabilistic causes (e.g., the ϵ parameter). The *rare-common* manipulation in the backwards-blocking experiments discussed above shows that adults and children can rationally adjust their beliefs about one parameter in the theory's ontology (β) to reflect the apparent abundance of a causal power (being a blicket).

More formally, in these experiments people act as if they are inferring the theory with maximum likelihood, out of all candidates in a one-dimensional hypothesis space of possible theories parameterized by β. This sort of learning is certainly less general than discovering a full theory with new classes and causal laws, as in the experiments of Tenenbaum and Niyogi (2003), but it is also more general than just learning the parameters or structure of a single causal network (learning at the T_0-level). The knowledge acquired about β exists at the T_1 level, specifying a prior distribution over possible causal networks that can be defined for any number of new entities in this domain.

In the remainder of the section, we show how similar parametric learning can take place concerning the functional form of a theory's causal laws. The blicket-detector theory in Figure 10 specifies the error rate of a detector in terms of a parameter, ϵ. We have outlined two different versions of the theory – for deterministic detectors and probabilistic detectors – that take $\epsilon = 0$ and $\epsilon > 0$ respectively. In some cases, such as the *one cause* and *two causes* experimental conditions, the probabilistic-detector theory seems to better characterize children's inferences. However, the instructions the children received suggested that the deterministic theory might be more appropriate. This raises an interesting learning question: how might a learner choose between these different theories as descriptions of a causal system? Our hierarchical Bayesian framework provides an answer, in particular Section 4.4, where we showed how the same statistical machinery used to learn causal networks could be used to make inferences about theories. In this simple case, we have just two candidate T_1 theories that differ only in the functional form of their causal laws: the deterministic theory and the probabilistic theory. We can use Bayes' rule to compute a posterior distribution over these theories, $P(T_1|\mathcal{D}, T_2)$, as shown in Equation 3.

Figure 13 shows how this process of inferring the T_1-level theory with an appropriate functional form can operate concurrently with identifying which blocks are blickets – an inference about causal networks at the T_0-level. The figure shows how the posterior distribution over the two theories – deterministic and probabilistic – evolves as the data $\mathcal{D}$ grow with each additional trial in the *two cause* condition. The bottom row shows the corresponding changes in the judged probabilities that blocks `a` and `b` are blickets, an average

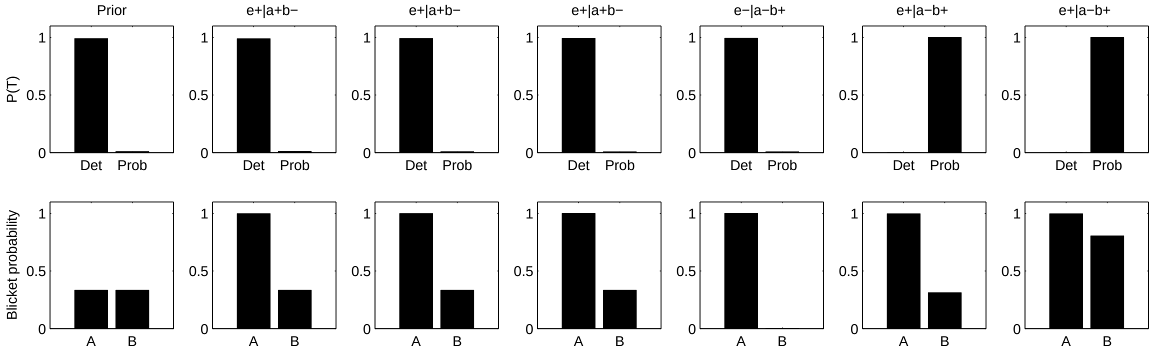


Figure 13. Learning functional form. The bar graphs along the top of the figure show the probabilities of two theories, with "Det" indicating the deterministic detector theory, and "Prob" indicating the probabilistic detector theory. The bar graphs along the bottom show the probabilities that the blocks A and B are blickets. The probabilities after successive trials are shown from left to right.

of the predictions of each T_1 theory weighted by their posterior probabilities $P(T_1|\mathcal{D}, T_2)$. The prior $P(T_1|T_2)$ assigns a probability of 0.99 to the deterministic theory, and 0.01 to the probabilistic theory, consistent with both task instructions and an intuitive bias towards determinism in mechanical systems. The base rate of blickets β is set to 1/3, and the noise level ϵ for the probabilistic theory is set to 1/10.

In the *two cause* condition, the first three trials are all $e^+|a^+b^-$: events in which block a is placed on the detector and the detector activates. This is sufficient to identify a as a blicket under either theory. The fourth trial is $e^-|a^-b^+$: b is placed on the detector and the detector does not activate. Under the deterministic theory, b would definitely not be a blicket. Under the probabilistic theory, there remains a small chance that b is a blicket, and since the probabilistic theory is still viable, the probability that b is a blicket is non-zero but extremely low. On the fifth trial, $e^+|a^-b^+$, the detector activates when b is placed upon it. The fourth and fifth trials are mutually contradictory under the deterministic theory – together they have a probability of zero – so the posterior over theories now switches suddenly to favor the probabilistic theory with probability 1. Under that theory, the data so far are uninformative about whether b is a blicket, because we assumed equal probabilities of the two types of error in the detector. It is just as likely that b is a blicket and the fourth trial was bad luck, or that b is not a blicket and the fifth trial was a fluke, so the probability reverts to the prior, β. The sixth trial provides further evidence that b is actually a blicket, $e^+|a^-b^+$. The final prediction is that a is very likely to be a blicket, while b is slightly less likely, matching the judgments of the children in Gopnik et al. (2001).

Ultimately, parametric learning of T_1-level theories is far from a complete solution to the problem of how people acquire rich representations of abstract causal knowledge. It is an open question how (and even whether) people learn T_1 theories in their full generality, not to mention theories at levels T_2 and above. Techniques of inductive logic programming (Muggleton, forthcoming) may provide one computational approach to these problems, but it is not at all clear that these techniques can scale up to human-like knowledge, or that they bear any similarity to human learning mechanisms. Formal computational frameworks for inductive learning will likely need to be extended to incorporate other cognitive capacities, such as analogy and natural language, that can provide crucial scaffolding for building appropriate hypothesis spaces of candidate theories.

5. Conclusion

This chapter has explored two proposals for formalizing the content and representational form of abstract causal theories, based on graph schemas and typed predicate logic. Each of these formalisms was cast as a probabilistic generative grammar for causal networks, inspired by an analogy between the computational problems of causal inference and natural language processing. We discussed each approach in terms of how it could account for the functional roles that abstract theories must play in a hierarchical Bayesian framework for causal inference and learning (Tenenbaum, Griffiths & Niyogi, this volume): chiefly, how the theory supports learning of causal-network structures (or lower-level theories), and how the theory could itself be learned or tuned based on observations.

We hope that readers find each of these frameworks for causal grammar intriguing but hardly satisyfing. We see them as proposals for what a causal grammar might look like rather than fully developed accounts. We close this chapter with three lessons that we have learned in the course of trying to formalize intuitive theories as causal grammars.

First, to approach human-level competence in models of intuitive theories, as in natural language grammars, it will be necessary to integrate two schools of thought that have often been treated as incommensurate or in opposition: probability and statistics on the one hand, and logical and symbolic representations on the other hand. Although this view is not yet fully accepted by researchers in generative linguistics, many computational linguists have recognized that probabilistic models defined over rule-based grammatical representations, such as stochastic finite-state grammars or context-free grammars, offer significant advantages over purely statistical or purely symbolic models while preserving the best features of both (Charniak, 1993; Jurafsky & Martin, 2000; Manning & Schütze, 1999). Logical or rule systems provide representational richness and the capacity for abstraction; probabilistic models provide the capacity for inductive inference from observed data. These same considerations motivate our proposals for expressing intuitive theories as probabilistic generative models defined over graph grammars or typed logical systems. We believe that some such integration of probability and structured rule systems will be necessary to explain how abstract causal knowledge guides the learning of new causal relations and can itself be learned from experience.

Second, in formal models of intuitive theories, as with formal models of grammar in linguistics, there will often be a tradeoff between representational capacity and learnability. For instance, hidden Markov models are much more limited than stochastic context-free grammars in terms of the syntactic regularities they can represent, but their structure can be induced from data much more readily by statistical methods. Likewise, the graph schemas we presented in Section 3 are much more limited as accounts of intuitive theories than are the typed logics we presented in Section 4, but we can give a principled and tractable algorithm for learning graph schemas (Kemp et al., 2004), while we cannot yet do that for logical theories. At this early stage it is valuable to pursue multiple approaches to formalizing theories, with the hope of ultimately converging on a framework that is both sufficiently expressive and learnable.

Finally, definitive accounts of people's intuitive theories are likely to be elusive, just as they are with natural language grammars. It is not easy to work backwards, from observations of people's judgments about linguistic utterances or cause-effect relations to formal accounts of the unobservable abstract knowledge that they bring to bear in making those judgments. In this chapter we have not attempted to claim that any particular formal model necessarily corresponds in detail to people's intuitive theory in some domain. We have merely proposed some possible models of intuitive theories that could account for aspects of people's causal inference capacities, and argued for the importance of certain general characteristics of these models. Progress on a formal account of intuitive causal theories is likely to be slow and painstaking for some time, and initially we may be able to give precise accounts only for rather small-scale domains such as the blicket detector paradigm. But if indeed there is an analogy between our project and the career of linguistics, from the early days of generative grammar through the contemporary computational era, then we can look forward to a most interesting journey.

References

Carey, S. (1985). *Conceptual change in childhood.* Cambridge, MA: MIT Press.

Charniak, E. (1993). *Statistical language learning.* Cambridge, MA: MIT Press.

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

Enderton, H. B. (1972). *A mathematical introduction to logic.* New York: Academic Press.

Friedman, N., Getoor, L., Koller, D., & Pfeffer, A. (1999). Learning probabilistic relational models. In *Proceedings of the 16th internation joint conference on artificial intelligence (IJCAI).* Stockholm, Sweden.

Gelman, S. A. (2003). *The essential child: origins of essentialism in everyday thought.* Oxford: Oxford University Press.

Glymour, C. (2001). *The mind's arrows: Bayes nets and graphical causal models in psychology.* Cambridge, MA: MIT 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. (2005). *Causes, coincidences, and theories.* Unpublished doctoral dissertation, Stanford University.

Griffiths, T. L., Baraff, E. R., & Tenenbaum, J. B. (2004). Using physical theories to infer hidden causal structure. In *Proceedings of the 26th annual meeting of the cognitive science society.*

Griffiths, T. L., & Tenenbaum, J. B. (in prep). *Theory-based causal induction.*

Griffiths, T. L., & Tenenbaum, J. B. (in press). Elemental causal induction. *Cognitive Psychology.*

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing.* Upper Saddle River, NJ: Prentice Hall.

Kemp, C., Griffiths, T. L., & Tenenbaum, J. B. (2004). *Discovering latent classes in relational data* (Tech. Rep. No. 2004-019). MIT AI Memo.

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing.* Cambridge, MA: MIT Press.

Milch, B., Marthi, B., & Russell, S. (2004). Blog: Relational modeling with unknown objects. In T. Dietterich, L. Getoor, & K. Murphy (Eds.), *Icml 2004 workshop on statistical relational learning and its connections to other fields* (p. 67-73).

Muggleton, S. H. (forthcoming). Statistical aspects of logic-based machine learning. *ACM Transactions on Computational Logic.*

Pasula, H., & Russell, S. (2001). Approximate inference for first-order probabilistic languages. In *Proceedings of the international joint conference in artificial intelligence 2001.*

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

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

Schulz, L., & Gopnik, A. (in press). Causal learning across domains. *Developmental Psychology.*

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

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

Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science*, *28*, 303-333.

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

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

Tenenbaum, J. B., & Niyogi, S. (2003). Learning causal laws. In *Proceedings of the 25th annual meeting of the cognitive science society.* Erlbaum.

Tenenbaum, J. B., Sobel, D. M., Griffiths, T. L., & Gopnik, A. (submitted). *Bayesian inference in causal learning from ambiguous data: Evidence from adults and children.*

White, P. A. (1995). *The understanding of causation and the production of action: from infancy to adulthood.* Hillsdale, NJ: Lawrence Erlbaum.