Intuitive Theories as Grammars for Causal Inference

Intuitive Theories as Grammars for Causal Inference

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

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

Sourabh Niyogi
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology

1. Introduction

This chapter considers a set of questions at the interface of the study of intuitive theories, causal knowledge, and problems of inductive inference. By an intuitive theory, we mean a cognitive structure that in some important ways is analogous to a scientific theory. It is becoming broadly recognized that intuitive theories play essential roles in organizing our most basic knowledge of the world, particularly for causal structures in physical, biological, psychological or social domains (Atran, 1995; Carey, 1985a; Kelley, 1973; McCloskey, 1983; Murphy & Medin, 1985; Nichols & Stich, 2003). A principal function of intuitive theories in these domains is to support the learning of new causal knowledge: generating and constraining people's hypotheses about possible causal relations, highlighting variables, actions and observations likely to be informative about those hypotheses, and guiding people's interpretation of the data they observe (Ahn & Kalish, 2000; Pazzani, 1987; Pazzani, Dyer & Flowers, 1986; Waldmann, 1996). Leading accounts of cognitive development argue for the importance of intuitive theories in children's mental lives and frame the major transitions of cognitive development as instances of theory change (Carey, 1985a; Gopnik & Meltzoff, 1997; Inagaki & Hatano 2002; Wellman & Gelman, 1992).

Here we attempt to lay out some prospects for understanding the structure, function, and acquisition of intuitive theories from a rational computational perspective. From this viewpoint, theory-like representations are not just a convenient way of summarizing certain aspects of human knowledge. They provide crucial foundations for successful learning and reasoning, and we want to understand how they do so. With this goal in mind, we focus on

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

Table 1: Three Questions About Intuitive Theories

Q1.	What is the content and representational structure of intuitive theories?
Q2.	How do intuitive theories guide the acquisition of new causal knowledge?
Q3.	How are intuitive theories themselves acquired?

three interrelated questions (Table 1). First, what is the content of intuitive theories? What kinds of knowledge are represented and in what formats? Second, how do intuitive theories guide the acquisition of new knowledge? Theories subserve multiple cognitive functions, but their role in guiding learning is surely one of the most fundamental. Third, how are intuitive theories acquired? What if anything do mechanisms for theory-guided learning have in common with mechansisms for learning at this more abstract level – for acquiring or revising a theory itself? It goes without saying that these questions are profound and difficult ones. Our inquiry is at an early stage, and any answers we can give here will be at best preliminary.

We adopt a "reverse engineering" approach to these questions, aiming to explain what intuitive theories bring to human cognition in terms that would be valuable in designing an artificial computational system faced with the same learning and reasoning challenges (Marr, 1982; Shepard, 1987; Anderson, 1990; Oaksford & Chater, 1999). This approach proceeds in two stages. First, we identify a core set of computational problems that intuitive theories help to solve, focusing on the role of theories in learning and reasoning about causal systems. Second, we propose a formal framework, based upon the principles of Bayesian inference, for understanding how these computational problems may be solved – and thus for understanding how intuitive theories may fulfill some of their crucial functions.

There are many places one could start in characterizing the functional roles that intuitive theories play in cognition. From a reverse-engineering viewpoint, it makes sense to start with causal learning and reasoning – behaviors that have dramatic consequences for people's success and survival in the world, and where intuitive theories seem to play a critical role. Everyday causal inference operates under severe conditions, far more challenging than the scientist's preferred setting of a controlled laboratory experiment. A medic arriving at a trauma scene may need to make a snap judgment about what is wrong with the victim after seeing just a few suspicious symptoms; there is no time for exhaustive tests. A child may discover a new causal relation given only a few observations of a novel system, even in the presence of hidden variables or complex dynamics. Successful causal inferences in the presence of sparse data require strong expectations about what kinds of causal hypotheses are possible, plausible or likely a priori. In order to learn and reason about novel causal systems, these expectations must go far beyond mere records of previous experience. Intuitive theories provide the necessary glue between the inferential past and present. They specify general causal principles, abstracted from prior experience, that allow us quickly and reliably to generate appropriate spaces of hypotheses for causal inference, and to apprehend an infinite range of new causal systems.

Because causal inference can unfold on multiple levels of abstraction, intuitive theories must also be defined on multiple levels. To reason about the causes behind a specific observed event, we need intuitive theories that generate hypotheses for alternative con-

figurations of causes for that event. To learn the structure of causal relations between variables in a system, we need intuitive theories that generate hypotheses about alternative causal structures for that system. To learn such a theory itself, we need higher-order intuitive theories that generate hypotheses about theories at the next level down. The need to characterize theories at more than one level of abstraction is familiar from debates in the philosophy of science (Carnap, 1956; Lakatos, 1970; Laudan, 1977; Kuhn, 1970; Quine, 1951; see Godfrey-Smith, 2003, for a review), and has also been introduced into research on cognitive development through Wellman's distinction between "specific theories" and "framework theories" (Wellman, 1990; Wellman & Gelman, 1992). Such a hierarchy of theory representations provides a unifying approach to inferring the causes of individual events, identifying the structure of causal relations between variables in a system, and learning about the abstract structure of higher-order theories – all from finite and often sparse data.

Consideration of the role of theories in causal inference places constraints on the formalisms that can be used to represent intuitive theories. In particular, we will argue that one widely used framework for representing causal relationships, known as causal graphical models or causal Bayesian networks (Pearl, 2000; Glymour, 2001), is not sufficiently expressive to represent intuitive theories in their full generality and power. While Bayesian networks may be able to represent the lowest level of causal theories in our hierarchy, they cannot express the kind of abstract principles that are a key part of higher-level theories. In making this argument, we will draw an analogy to generative grammar in linguistics: a Bayesian network that describes the causal structure of a particular causal system is like a parse tree that describes the syntactic structure of a particular sentence. Of deeper and more general significance in linguistics is the set of abstract principles – the grammar – that generates all possible parse trees for the infinite but constrained set of grammatical sentences in a given language. So too in the study of causal inference should our focus be on theories at this more abstract level: *causal grammars* that generate hypothesis spaces of possible causal networks in a given domain of reasoning.

Construing intuitive theories as causal grammars helps to clarify the computational problems that a formal account of theories must address, as each of these problems has a direct analogue in linguistics. The analogy also suggests how such problems can be solved. The second stage of our reverse engineering of intuitive theories consists of formalizing the inferences involved in learning and reasoning about causal systems in a Bayesian framework. Any Bayesian inference requires a space of candidate hypotheses and a prior probability distribution over that hypothesis space. We cast intuitive theories as *hypothesis space generators*, systems of knowledge that generate the hypothesis spaces that make Bayesian causal inference possible. Drawing upon the idea that theories are defined at multiple levels, we adopt a hierarchical Bayesian framework in which intuitive theories defined at each level of the hierarchy generate hypothesis spaces for the more specific level below. This hierarchical Bayesian proposal specifies precise functional roles for intuitive theories in causal learning, and offers an approach to answering our second and third questions from Table 1: how theories guide the acquisition of new causal knowledge, and how theories themselves can be learned.

Approaching the computational problems posed by intuitive theories from the perspective of Bayesian inference ultimately provides us with the opportunity to assess answers

to our first question – what is the knowledge content of intuitive causal theories? – in terms of how well they function in this formal framework. Many possible representational structures for causal knowledge could be interpreted as theories in our hierarchical Bayesian framework, and they may coexist at different levels of the hierarchy. In this chapter we will have little to say about the precise nature of these representations, beyond the argument that causal Bayesian networks are too limited to capture the content of higher-level intuitive theories. A detailed discussion of two more promising approaches for representing higher-level theories, or causal grammars, is the subject of a companion chapter (Griffiths & Tenenbaum, this volume).

2. Intuitive theories as causal networks

While computational accounts of intuitive theories have not been readily forthcoming, significant progress has been made recently in the related area of causal network modeling. By a causal network, we mean a set of causal relations that hold among variables representing states of affairs in the world, which may or may not be observable. The tools of causal graphical models, causal Bayesian networks, and functional causal models (Heckerman, 1998; Pearl, 2000; Spirtes, Glymour, & Scheines, 1993; see also this volume), provide formal frameworks for representing, reasoning about, and learning causal relationships. These approaches explicate the connection between causality and statistical dependence: they distinguish causality from mere correlation or association, and they show how and under what circumstances causal relations can be induced from observations of the statistical dependencies between variables.

Causal networks have already received some attention in the cognitive science literature, as rational accounts of adult and child behavior in causal learning experiments (Danks & McKenzie, under revision; Glymour, 2001; Gopnik, Glymour, Sobel, Schulz, Kushnir, & Danks, 2004; Gopnik & Glymour, 2002; Gopnik & Schulz, 2004; Griffiths, Baraff, & Tenenbaum, 2004; Griffiths & Tenenbaum, in press; Lagnado & Sloman, 2004; Sloman, Lagnado and Waldmann, this volume; Sobel, Tenenbaum, & Gopnik, 2004; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003; Tenenbaum, Sobel, Griffiths & Gopnik, submitted; Tenenbaum & Griffiths, 2001; 2003; Waldmann, 1996). These applications have been fairly small-scale: subjects typically learn about one or a few causal relations from a small number of observations. The successful application of causal networks in these cases raises the question of whether some of the same computational tools could be applicable to larger-scale problems of cognitive development, in particular to elucidating the structure and origins of our intuitive causal theories.

The most direct line of attack is simply to identify intuitive theories with causal networks. This is how we read the proposal of Gopnik and colleagues (Gopnik & Glymour, 2002; Gopnik & Schulz, 2004), and it is related to Rehder's proposal for modeling "theory-based" categorization, or categorization based on "theoretical knowledge", using causal networks (Rehder, 2003; this volume). An appealing feature of this proposal is that it suggests a set of ready answers to our three guiding questions about the structure and function of intuitive theories (Table 1). What are intuitive theories? They are (something like) causal graphical models. How are theories formed? Using (something like) the existing learning algorithms in the graphical models literature (Pearl, 2000; Spirtes et al., 1993). How are theories used to guide learning of new causal relations? By providing constraints

for causal model learning algorithms based on the structure of previously learned causal relations.[1] In short, the proposal to model intuitive theories as causal networks promises to fill in the missing foundations of a computational account of cognitive development, by drawing on already established and well-understood formal tools.

This proposition is tempting – there is clearly something "theory-like" about causal graphical models. Yet these models are also fundamentally limited in ways that intuitive theories are not. Most accounts of intuitive theories in cognitive development emphasize the importance of abstract concepts and causal laws, in terms of which people can construct causal explanations for the phenomena in some domain (Carey, 1985b; Wellman, 1990). Causal graphical models may often be useful for representing the causal explanations that an intuitive theory generates, but they do not and cannot represent the abstract concepts and causal laws that are the core of the theory, and that set the terms in which those causal explanations are constructed.

To illustrate the strengths and weaknesses of viewing theories as causal graphical models, consider Graph 1, shown in Figure 1. This network might represent some aspects of a person's knowledge about several common diseases, their effects (symptoms), and causes (risky behaviors). It can support probabilistic causal inferences (as a Bayesian network), if we assign to each variable a probability distribution conditioned on its parents (direct causes) in the network (Pearl, 1988, 2000). Such a representation is "theory-like" in several ways. Most fundamentally, it permits causal inferences to be made from sparse data. Given one or more observed symptoms in a sick individual, the network suggests a constrained set of causal explanations: the presence of one or more diseases causally linked to those symptoms. The network also assigns relative probabilities to those hypotheses. If some of the patient's relevant behaviors are observed as well, those probabilities over the hidden disease variables will change to reflect the most probable routes from observed behaviors to observed symptoms. For instance, if a person is coughing that suggests they might suffer from bronchitis or flu, but provides no indication of heart disease. Observing that they also suffer from a headache would increase the probability of flu, while observing that they habitually smoke would increase the probability of bronchitis.

What this network description misses is theoretical knowledge of a more abstract kind: knowledge about classes of causal variables, and laws governing the causal relations between those classes. For instance, there appears to be a common domain theory underlying Graphs $1-4$, but not Graph 5 or Graph 6. Graphs $2-4$ differ from Graph 1 in the precise causal links they posit: Graph 2 posits that smoking causes flu but not lung cancer; Graph 3 represents only a subset of the conditions that Graph 1 does, but includes all the same causal links defined on that subset; Graph 4 posits a novel unnamed disease linking working in a factory with chest pain. Yet Graphs $1-4$ all express the same abstract regularities, which could be characterized in terms of 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, such as the new disease in Graph 4.

[1] For example, suppose that a correlation is observed between A and B, and it is known that no direct causal connection exists between A and B. If a third variable V is known to be a cause of A, then this knowledge suggests two simple hypotheses for interpreting the correlation between A and B: V may be a cause of B, or B may be a cause of V.

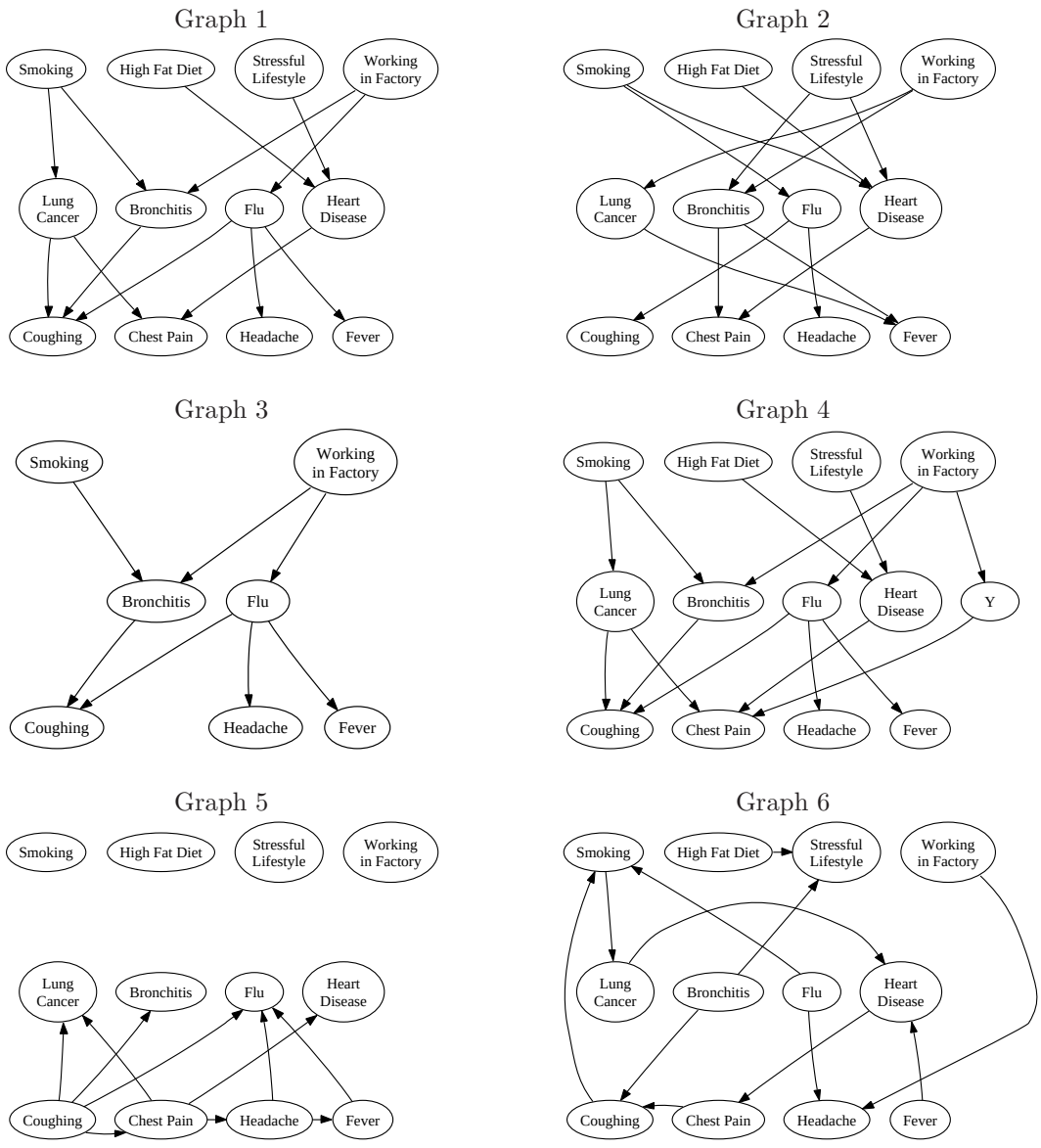

Figure 1. Causal networks illustrating different possible sets of beliefs about the relationships among behaviors, diseases, and symptoms.

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.

Principles P1 and P2 are not explicitly represented in Graphs 1-4, although they are instantiated in those networks. No single causal network defined over particular behaviors, diseases, and symptoms (e.g., *Smoking*, *Bronchitis*, *Coughing*) could capture these principles. Rather, P1 and P2 specify a large (potentially infinite) but constrained class of causal networks, which includes Graphs 1-4 but excludes Graph 5 and Graph 6. We view this more abstract level of knowledge as a core component of intuitive domain theories. While knowledge of a causal network structure like Graph 1 may support reasoning from effects to causes in specific situations, it is knowledge of abstract principles like P1 and P2 – transcending any specific network – that allows people to formulate appropriate hypotheses for new causal structures in a given domain, and thereby to learn and reason about novel causal relations or causal systems so effectively.

3. Framework theories and specific theories

Although not the focus of contemporary research on causal learning and reasoning, abstract causal knowledge at the level of principles P1 and P2 has traditionally been recognized as critical in both scientific and intuitive theories. Twentieth-century philosophers of science often distinguished the day-to-day level of theorizing from a more abstract "framework" level of theoretical knowledge – principles, concepts, or terms that shape the possible specific theories a scientist can construct (Godfrey-Smith, 2003). Such an abstract level of knowledge appears in Carnap's (1956) "linguistic frameworks", Kuhn's (1970) "paradigms", Lakatos' (1970) "research programs", and Laudan's (1977) "research traditions".

Inspired by this line of thinking, Wellman and Gelman (1992) formulated a distinction between "specific" and "framework" theories that they argued would be useful for understanding children's intuitive theories of the world:

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

Although she does not explicitly distinguish these two levels of theory structure, Carey (1985b) clearly seems to have framework-level knowledge in mind when she characterizes a child's theory as follows:

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

Traditionally, both philosophers of science and cognitive developmentalists have considered framework-level theories to be in some sense deeper and more fundamental than specific theories. A framework expresses the abstract causal principles that hold across all

systems in a broad domain, providing a language for constructing specific theories of those systems. Specific theories, though they carry much of the burden for everyday prediction, explanation, and planning, thus cannot be acquired or even formulated without the machinery of framework theories. The most dramatic instances of theory change are thought to take place at the level of frameworks, as in Kuhn's "paradigm shifts", or the conceptual revolutions of childhood studied by Carey (1985a), Wellman (1990), and others. At the same time, the role of specific theories and their interaction with framework-level knowledge cannot be ignored. Framework theories typically come into contact with the raw data of experience only through the specific theories that they generate. A framework is only as good as the specific theories it supports.

In sum, if our ultimate goal is a computational understanding of intuitive theories and their place in causal inference, we need to develop formal tools for representing both framework and specific theories, and formal tools for inference and learning that account for how specific theories support predictions and explanations about specific events, how framework theories support the construction and acquisition of specific theories in their domain, and how framework theories themselves may be acquired. Clearly our current state of understanding is far from meeting these requirements. We are in a position, however, to make progress on a more constrained version of this program: developing formal tools that allow us to represent abstract causal knowledge like principles P1 and P2, to understand the role of this knowledge in learning and reasoning about specific causal networks like Graph 1, and to explain how such knowledge itself could be acquired. This will be our goal for the remainder of this chapter and the next (Griffiths & Tenenbaum, this volume).

The relationship between principles P1 and P2 and causal graphical models is analogous to the relationship between framework and specific theories in several ways. Like a specific theory, Graph 1 spells out the causal relationships that hold among a delimited set of variables. The network does not explicitly represent any framework-level knowledge – anything that resembles an ontology, or causal laws defined over the entities identified within that ontology. Nor does the network define "a coherent form of reasoning" for the disease domain, which would extend beyond the particular variables already specified in the network to learning about novel diseases, symptoms or behaviors.

Relative to a specific causal network like Graph 1, the abstract principles P1 and P2 provide something more like framework-level knowledge. These principles specify an ontology of kinds of causally relevant variables (P1) and the basic causal laws (P2) that can be used to construct causal networks like Graphs $1-4$. Just as framework theories provide the explanatory principles from which specific theories in a domain are built, the principles P1 and P2 identify the relationships from which causal networks can be built in the disease domain. If someone tells you about a new disease Y, P1 and P2 lead you to expect that Y will have some symptoms and also some behavioral causes, and that these causes and effects may overlap with one or more familiar diseases. If you observe a novel combination of familiar symptoms in a sick individual, P1 and P2 suggest that a possible explanation is the existence of a new hidden variable – a new disease causally linked to those symptoms – rather than a web of new connections between the symptoms themselves.

A change in an individual's framework theory may fundamentally alter the specific theories they can construct (e.g., Wellman, 1990), or even the concepts they can be said to possess (Carey, 1985a; Gopnik & Meltzoff, 1997). Likewise, a change in the principles

P1 and P2 would lead a learner to construct qualitatively different kinds of causal network structures and to reason about diseases in fundamentally different ways – perhaps even to the point where we would no longer say they had the same concept *Disease*. Graph 5 appears to derive from the same ontology as Graph 1 (i.e., P1), but instead of P2 follows a set of causal laws that we might call P2′: symptoms cause diseases rather than the other way around, symptoms also cause other symptoms, and there are no links between behaviors and the other conditions. P1 and P2′ may reflect a logically possible alternative (if nonveridical) theoretical framework, with a coherent but different mode of reasoning than that of P1 and P2. In contrast, someone whose beliefs correspond to Graph 6 appears to lack a coherent mode of reasoning in this domain. Graph 6 is inconsistent with both P1 and P2, or seemingly with any ontology and causal laws that would give some regularity to its structure of causal links. Somebody whose beliefs are represented by Graph 6 not only has different beliefs about how particular diseases work than someone whose beliefs correspond to Graph 1, but seems not to possess the same ontological concepts of *Disease*, *Symptom* or *Behavior* – at least not in the causally relevant sense; he does not know how diseases in general work.

To clarify, we do not mean to suggest that P1 and P2 should necessarily be seen as a framework theory in Wellman and Gelman's sense, or that Graph 1 should be seen as a specific theory, but only that the relation between these two levels of causal knowledge is analogous to the relation between frameworks and specific theories. When cognitive developmentalists speak of a child's framework theory, they are typically referring to much more abstract knowledge than P1 and P2, with much broader scope sufficient to encompass a full domain of intuitive biology or intuitive psychology. Yet we see value in treating the concepts of "framework theory" and "specific theory" as relative notions, with more abstract frameworks providing constraints on more specific models across multiple levels of abstraction and scope. Relative to knowledge about a specific causal network such as Graph 1, principles such as P1 and P2 do appear to play a framework-like role. If we can develop formal tools for understanding how theoretical knowledge operates at both of these levels, and how they interact in learning and inference, we expect to have made real progress towards the larger program of a computational understanding of intuitive theories.

4. Intuitive theories as causal grammars

The proposal to identify intuitive theories with causal networks appeared promising in large part because the formal tools of causal graphical models offered ready answers to the questions we raised in the introduction (Table 1): what is the representational content of theories, how do theories support new inferences, and how are theories themselves learned? But as we have just argued, this view of intuitive theories does not address the structure, function, or acquisition of more abstract framework-like causal knowledge, such as principles P1 and P2, or the relation between these abstract principles and learning and reasoning with specific causal networks. The remainder of this chapter and the next (Griffiths & Tenenbaum, this volume) describe some initial attempts to approach these questions formally.

Our work on intuitive theories has been guided by an analogy to the linguist's project of working out generative grammars for natural languages and accounting for the use and learnability of those grammars (Chomsky, 1965, 1986). This "causal grammar" analogy

(Tenenbaum & Niyogi, 2003) has been so fruitful for us that it is worth discussing in some detail here, both to motivate the specific proposals we will offer and to provide more general suggestions for how future work on intuitive theories might proceed.

There is a long history of analogies between linguistic grammars and scientific theories, dating back at least to Chomsky's early work on generative grammar in language (Chomsky, 1956; 1962). Chomsky characterized a native speaker's knowledge of grammar as "an implicit theory of that language that he has mastered, a theory that predicts the grammatical structure of each of an infinite class of potential physical events" (Chomsky, 1962, p. 528). Chomsky (1956) explcitly speaks of an analogy between theories and grammars:

> Any scientific theory is based on a certain finite set of observations and, by establishing general laws stated in terms of certain hypothetical constructs, it attempts to account for these observations, to show how they are interrelated, and to predict an indefinite number of new phenomena.... Similarly, a grammar is based on a finite number of observed sentences... and it 'projects' this set to an infinite set of grammatical sentences by establishing general 'laws'... [framed in terms of] phonemes, words, phrases, and so on.... (p. 113).

It is striking – if not necessarily surprising – how closely Chomsky's characterization of grammatical knowledge here resembles the characterization of framework-level intuitive theories in cognitive development, as exemplified by the quotations from Carey, Wellman and Gelman in the previous section. Central to the Chomskyan program has always been an analogy between the descriptive goals of the linguist and the goals of the child learning language. Both are engaged in a form of theory-building, seeking to identify the general laws and grammatical categories that govern a language's structure, based on observations of primary linguistic data and guided by some (metatheoretic or innate) constraints on the space of candidate grammars.

Chomsky's grammars-as-theories analogy was intended to motivate hypotheses about the content and function of linguistic grammars, but here we will use the analogy in the opposite direction, to inspire models for intuitive theories based on the development of generative grammar in linguistics. Arguably, this is now the more profitable direction in which to run the analogy, as the last fifty years have seen significant progress in formal and computational models for language – but not so much progress in understanding causal theories, either intuitive or scientific.[2] We will first review some relevant ideas from generative grammar in language, and then discuss their implications for theories of causal grammar.

4.1. A bird's-eye-view of generative grammar

Figure 2 introduces the grammar analogy through several intuitive (if perhaps overly simplistic) examples. Like the sample causal networks for different disease theories shown in Figure 1, Figure 2 shows samples of hypothetical utterances and syntactic (phrase structure) analyses for several simplified languages. These examples clearly do not begin to approach the richness of natural language, any more than the examples shown in Figure 1 approach

[2]Readers familiar with the linguistics literature will recognize the questions in Table 1 as themselves based on an analogy to Chomsky's standard questions about knowledge of language (e.g., Chomsky, 1986). It is no accident where we started, given where we figured to end up.

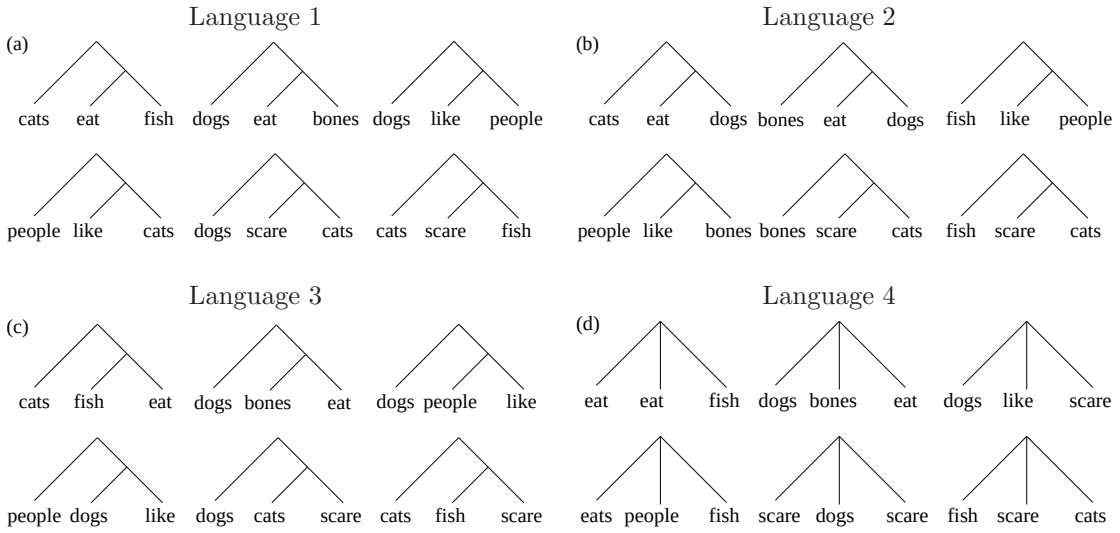

Figure 2. Example sentences and syntactic structures for several simplified languages.

the richness of our intuitive knowledge about diseases (or biology more generally). The aim is merely to illustrate how knowledge of syntactic structure in language, as with intuitive theories in causal domains, can be usefully characterized in terms of multiple interacting levels of abstraction, and to suggest parallels between the sorts of representations that could be useful in linguistic grammars and causal theories.

Figure 2a shows utterances from a simplified English-like language. Informally, each sentence consists of a subject noun followed by a verb phrase, and each verb phrase consists of a verb followed by a noun (the direct object). This phrase structure is depicted with the skeleton of a parse tree above each utterance in the figure. It is a canonical form for many simple sentences in English or other languages with SVO (Subject-Verb-Object) word ordering.

Figure 2b shows different utterances apparently in the same language, obeying the same syntactic principles. Hearing a speaker utter these sentences, we would not doubt that the individual speaks English (or a simplified version thereof), even though we might be suspicious of the particular beliefs they appear to hold. The situation is analogous to Graph 2 in Figure 1, representing the beliefs of an individual who has the standard framework-level understanding of what behaviors, diseases and symptoms are and how they are causally related, but who has different beliefs about the specific causal links that exist between particular behaviors, diseases and symptoms.

Figure 2c shows a case analogous to Graph 5 in Figure 1: an individual who appears to follow a consistent grammar defined over the same syntactic categories and the same lexical items as the speakers represented in Figures 2a and b, but with different rules prescribing how these categories can be combined to form possible syntactic structures. In particular, the utterances in Figure 2c appear to obey SOV (Subject-Object-Verb) ordering, as is characteristic of Korean, Japanese, Turkish, and many other languages, rather than the SVO ordering characteristic of English.

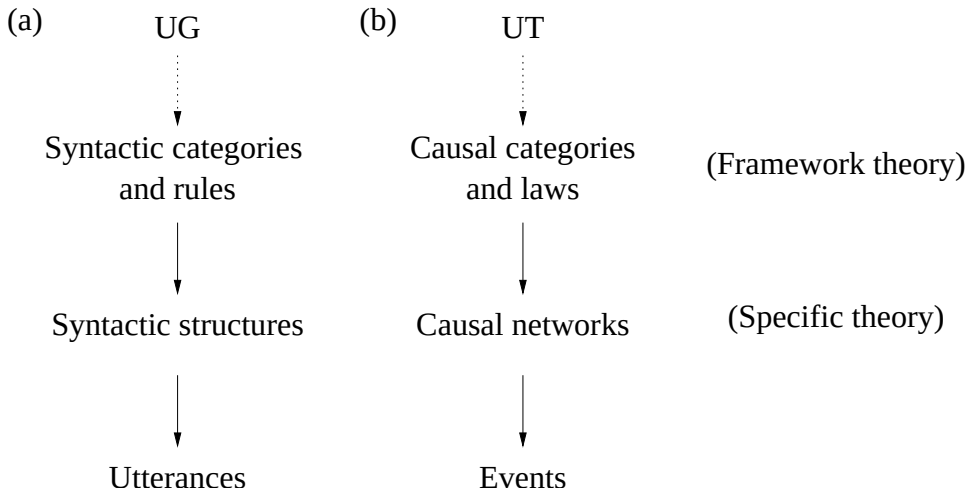

Figure 3. An analogy between multiple levels of structure in (a) knowledge of language and (b) causal knowledge. Each level generates structures at the level below, thereby establishing necessary constraints on the hypothesis space for inductive inference.

Finally, Figure 2d shows a case analogous to Graph 6 in Figure 1: an individual who appears to follow no consistent grammar, or at least no grammar that constrains the set of possible utterances based on syntactic rules or categories that are at all like those in English.

More formally, theories of generative grammar posit at least four levels of structure to knowledge of language, which may serve as a guide for how to think about corresponding levels of abstraction in intuitive theories. These levels of representation are quite distinct in their forms but are functionally interdependent: each higher level of abstraction *generates* the structures at the level below, and thus constrains the possible lower-level structures that could be encountered. Language comprehension and language acquisition – the main computations for which the language faculty is responsible – are processes of inductive inference that can be defined in terms of this representational hierarchy. In both comprehension and acquisition, the challenge is to infer some unobservable structure at an intermediate level of abstraction, by integrating observed data from a lower level generated by that structure and constraints on possible forms for that structure generated by higher levels of the abstraction hierarchy.

These four levels of structure in language can be loosely characterized as shown in Figure 3a. The lowest, most concrete level are utterances: sequences of words, spoken or written. One level up in abstraction are syntactic structures: parse trees or other hierarchical representations of phrase structure over which the meanings of utterances are defined. Language comprehension – or more precisely, syntactic comprehension or parsing – is the process of inferring the syntactic structure that gave rise to an observed utterance. This inference problem presents an inductive challenge because the set of possible syntactic structures that can be inferred for any language is, in principle, infinite in extent and complexity, and the data almost always underdetermine the true underlying structure.

To explain how people can recover an infinite set of syntactic structures from appro-

priate linguistic utterances, linguists posit a third level of knowledge more abstract than any syntactic structure. The grammar – or more precisely, the syntax – of the language generates a strongly constrained (but still infinite) space of candidate syntactic structures that could be hypothesized to explain utterances in that language. Although there is no universal consensus on the content or architecture of syntax, most theories are based on some set of abstract categories and rules for how those elements in those categories can be composed to generate allowable syntactic structures. Figure 3a labels this level of knowledge "syntactic categories and rules", but for shorthand we may refer to it simply as the "syntax" or the "grammar" of the language.

To give a concrete example, in the case of the simplified language in Figure 2a the syntax could be specified by means of a *context-free grammar*, with the categories N, V, VP, and S, and the following rewrite rules:

$$\begin{aligned} S &\rightarrow N\ VP \\ VP &\rightarrow V\ N \\ N &\rightarrow \{\ \text{dogs} \mid \text{cats} \mid \text{fish} \mid \text{people} \mid \text{bones} \mid \ \dots\ \} \\ V &\rightarrow \{\ \text{eat} \mid \text{like} \mid \text{scare} \mid \ \dots\ \}. \end{aligned} \tag{1}$$

A speaker who grasps these abstract rules of syntax, and who recognizes that the syntactic categories of nouns and verbs are open classes (capable of adding new words), can effectively produce and understand an infinite set of grammatical utterances – not just the limited sample depicted in Figure 2a. Upon hearing the novel utterance "Dogs like blickets", these principles would allow a competent listener to infer that "blickets" is in the N (noun) category , and hence that "people like blickets" or "blickets eat bones" are also grammatical (if not necessarily true) utterances.

The above grammar is sufficiently simple that there are no parsing ambiguities for the utterances in Figure 2a. Each utterance can be generated by the grammar in exactly one way. But in natural language use, syntactic ambiguity is common, which has led to the development of probabilistic gramamrs. Probabilistic grammars (see Charniak, 1993; Jurafsky & Martin, 2000; Manning & Schütze, 1999) augment the deterministic rules of traditional grammars with probabilities, so that each grammar now specifies a probability distribution over the possible syntactic structures in a language (and, typically, over possible utterances as well). Identifying the syntactic structure most likely to have given rise to a particular observed sentence then becomes a well-posed problem of statistical inference: selecting from among all syntactic structures that represent consistent parses of the sentence the structure that has highest probability under the probabilistic grammar.

Besides the problem of parsing, the other great inductive challenge in language is the problem of grammar acquisition: inferring the correct categories and rules of syntax from primary linguistic data. Like parsing, grammar acquisition also requires would-be language users to infer unobservable structures from highly underconstrained data. In principle, there is no limit to the number of grammars that could be posited to explain a given corpus of utterances. For instance, the utterances in Figure 2a could have been produced from the following grammar,

$$\begin{aligned} S &\rightarrow A\ A\ A \\ A &\rightarrow \{\ \text{dogs} \mid \text{cats} \mid \text{fish} \mid \text{people} \mid \text{bones} \mid \text{eat} \mid \text{like} \mid \text{scare} \mid \ \dots\ \}, \end{aligned} \tag{2}$$

in which there are no distinguished syntactic categories and no meaningful constraints on allowable word combinations.

To explain how children acquire the grammar of their native language, linguists have proposed a solution that is parallel to the standard account of parsing, but elevated in abstraction. Hence the highest level of structure shown in Figure 3, *Universal Grammar* or UG. UG comprises the innate knowledge that every child brings to the task of language acquisition. Just as the grammar of a language generates a constrained space of syntactic structures that could serve as hypotheses for parsing in that language, the principles of UG could be said to generate a highly constrained space of possible grammars for all human languages, thereby enabling grammar acquisition to occur in the face of what would otherwise be severly inadequate data (Nowak, Komarova & Niyogi, 2003). For instance, it may be reasonable to posit that UG rules out grammars such as (2), while allowing grammars such as (1).

As in the comprehension of syntactic structure, deterministic constraints on possible hypotheses are not sufficient to remove all ambiguities in acquisition and ensure that the correct grammar can be simply deduced from the observed data. Again, some kind of probabilistic inference is required. To illustrate why, consider the following grammar:

$$\begin{aligned} S &\rightarrow N\ VP \\ VP &\rightarrow V\ N \\ N &\rightarrow \{\ \text{dogs} \mid \text{cats} \mid \text{fish} \mid \text{people} \mid \text{bones} \mid \ \ldots\ \} \\ V &\rightarrow \{\ \text{eat} \mid \text{like} \mid \text{scare} \mid \text{fish} \mid \text{people} \mid \ \ldots\ \}. \end{aligned} \tag{3}$$

This grammar is just like 1 except that "fish" and "people" are now categorized as verbs (V) in addition to nouns (N). It is surely not in violation of the principles of UG for words to be categorized as both nouns and verbs. Indeed, many words in English bear such dual identities (including the words "fish" and "people"). Or consider another grammar,

$$\begin{aligned} S &\rightarrow N\ VP \\ VP &\rightarrow V \\ VP &\rightarrow V\ N \\ N &\rightarrow \{\ \text{dogs} \mid \text{cats} \mid \text{fish} \mid \text{people} \mid \text{bones} \mid \ \ldots\ \} \\ V &\rightarrow \{\ \text{eat} \mid \text{like} \mid \text{scare} \mid \ \ldots\ \}, \end{aligned} \tag{4}$$

which allows verbs to appear in intransitive forms, such as "cats eat", in addition to the transitive forms (e.g., "cats eat fish") shown in Figure 2a and generated by grammars 1 or 3. Again, UG should clearly allow grammars of this sort.

How could language learners infer which of these grammars is the true generative system for their language? In particular, how are they to know that certain rules should be included in the grammar, while others that seem equally plausible by the standards of UG (and that would in fact be correct in other languages) should be excluded? Probabilistic inference again provides a principled framework for answering these questions (e.g., Charniak, 1993). Probabilistic methods can identify the correct grammar underlying a corpus of utterances because the correct grammar should assign the observed utterances higher probabilities than will incorrect grammars. Under the hierarchical scheme of Figure 3, a

grammar assigns probabilities to possible utterances through a two-stage processes, by generating syntactic structures with various probabilities which in turn give rise to concrete utterances with various probabilities. The correct grammar will generate all and only the syntactic structures necessary to give rise to the observed utterances. Alternative grammar hypothesis will be hurt by *undergenerating* – failing to generate syntactic structures necessary to produce a class of observed utterances – or by *overgenerating* – generating syntactic structures that are not part of the language and that would give rise to a class of utterances not in fact observed. Either under- or overgeneration in a grammar hypothesis would lead to less accurate probabilistic expectations about the observed utterance data, and hence weaker inductive support for the grammar.

One final inference problem in language is worth noting for the sake of the causal analogy: inferences at the lowest level of Figure 3a, about partially observed utterances. Because the speech signal is inherently noisy, any individual word in isolation may be mistaken for a similar-sounding word, and listeners would be well-served if they could interpolate potentially misheard words from the context of more clearly perceived surrounding words. Because language must be processed online in real time, listeners would also be well-served if they could predict later words in an utterance from the context of earlier words. These inferences at the utterance level may be given the same treatment as inferences at higher levels of Figure 3a. Just as UG generates a constrained hypothesis space of possible grammars for a language, and just as a grammar generates a constrained hypothesis space of possible syntactic structures for an utterance, a syntactic structure generates a constrained hypothesis space of possible complete utterances that can be used to guide interpolations or predictions about missing words. For instance, if a speaker of the language in Figure 2a hears "dogs scare ...", it is a better bet that "..." should be filled in by "cats", "people" or "fish" than by "like" or "eat" (or by nothing), because the most likely syntactic structure underlying "dogs scare ..." suggests that "..." should be a noun rather than a verb or silence. This sort of inference is a central component of state-of-the-art speech-recognition systems based on probabilistic grammars (Jurafsky & Martin, 2000), and is probably important in human language processing as well.

In sum, human language users draw inductive inferences about unobserved structure at each level of the hierarchy in Figure 3a, based on data from lower levels and constraints from higher levels. Each level of structure can be viewed as a generator of hypothesis spaces for candidate structures at the next level down, and indirectly, for all levels below it. Because every level above the utterance is unobserved (and typically even the utterance-level is only partially observed), it is critical that inferences at all levels be able to proceed in parallel, based on only partial input from levels above and below. The child learning language will typically be uncertain not only about the grammar of her language, but also about the syntactic structure of many utterances she hears, as well as some of the words in each utterance. Yet somehow, after only a few years of experience, every normal child becomes an expert on all these levels. The inferential machinery underlying language learning and use must thus support a hierarchy of interlocking probabilistic inferences, operating over multiple levels of increasingly abstract representations.

4.2. Towards causal grammars

We have invested some energy here in reviewing elements of generative grammar because all of these elements – and the whole picture of language they support – have valuable parallels in the realm of intuitive causal theories. These parallels include:

- The decomposition of knowledge representation into at least four levels of increasingly abstract structure.
- The kinds of representational ingredients required at each level.
- The nature of the inductive problems to be solved at each level, and the factors that make these problems challenging.
- The manner in which levels interact, with each level generating a hypothesis space of candidates structures for the level below.
- The importance of probabilistic generative processes, which support hierarchical probabilistic inferences upwards from observed data at the lowest level to multiple higher levels of abstraction.

Of course there are other important disanalogies between the fields, and flaws even in the parallels we focus on, but still the analogy as a whole offers important lessons for how to develop formal treatments of intuitive causal theories.

Figure 3b shows a four-level decomposition of representation in causal theories, analogous to the four-level picture of linguistic knowledge in 3a. The data at the lowest, most concrete level consist of events, or instances in which the variables in a causal system take on particular values. In causal inference, these events are interpreted as having been generated from a structure one level up, a network of cause-effect relations, such as Graph 1 in the disease domain. Just as a particular linguistic utterance may be derived from an abstract syntactic structure by choosing specific words to fill the abstract categories in the structure, a particular event configuration may be generated by choosing values for each variable in a causal network conditioned on its direct causes. The formal tools of causal graphical models can be used to describe these two levels of structure and their interaction. In particular, the standard problem of inference in causal graphical models is just the problem of inferring unobserved causes or predicting future effects based on a hypothesized causal network structure – analogous to the lowest-level linguistic inferences of interpolating or predicting an incomplete utterance based on a hypothesized syntactic structure.

As we have already argued, networks of cause-effect relations such as Graph 1 are only the lowest level of structural description in a hierarchy of abstraction. Just as the specific phrase structures in a particular language are generated by a more abstract level of knowledge – the grammar or syntax of that language – so are the specific causal networks in a particular domain generated by more abstract knowledge that we can think of as a kind of "causal grammar" or "causal syntax" for that domain. Loosely speaking, in the terminology of Section 3, a causal grammar corresponds to an intuitive domain theory at the framework level, while the causal networks generated by the grammar correspond to

specific theories developed within the overarching framework theory for that domain. The real payoff of the linguistic analogy comes in its suggestions for how causal theories at this more abstract framework level may be represented, as well as how they function to guide new inferences about causal structure and how they may themselves be acquired.

Just as theories of linguistic syntax are typically framed in terms of abstract syntactic categories and rules for composing phrase structures that are defined over those categories, so can we start to formalize the syntax of a causal domain theory in terms of abstract causal categories of entities, properties, and states, and rules for composing causal-network structures defined over those categories. Principles P1 and P2 are a first attempt in this direction for a fragment of the disease domain: P1 specifies three categories of variables, and P2 specifies the rules by which variables in those categories can be connected into networks of cause-effect relations in order to generate causal networks like Graph 1, but not those like Graph 5 or 6. In the following chapter (Griffiths & Tenenbaum, this volume), we will present two more formal schemes for representing the grammars of causal framework theories and principles such as P1 and P2. These two formalisms work quite differently, but they share the basic notion of a generative syntax, with rules for constructing causal networks that are defined over abstract causal categories.

The primary functional role of a grammar for causal inference is essentially the same as the role played by grammar in language comprehension: to provide the constraints that make possible successful inductive inferences of structure at the level below. As in linguistic parsing, inferences about the causal-network structure that gave rise to a set of observed events are highly underconstrained. Many logically possible causal networks will be able to explain the sparse event data that a learner typically observes. The causal grammar reduces this problem by generating only a constrained set of causal network hypotheses that the learner need consider. The causal grammar may also be probabilistic, generating some network structures with higher probability than others, which will further help to resolve ambiguities present in the learner's data.

Some of the causal grammar's constraints on network hypotheses may be domain-general, but others will vary substantially across domains, in keeping with the crucial role of abstract theories as the frameworks on which people's distinctive understandings of different domains are built. For instance, causal grammars in many domains might assign higher probabilities to structures with fewer causal links or fewer hidden (intrinsically unobservable) causes, a la Ockham's razor. But in any one domain, a particular hypothesis that posits strictly more unobservable structure may be more likely under the causal grammar, if it accords better with the specific causal laws of that domain. For instance, consider a learner for whom Graph 1 describes her current theory of specific diseases, and P1 and P2 comprise her framework-level theory. She now observes a previously unseen correlation between a known behavior B (e.g., *Working in Factory*) and a known symptom S (e.g., *Chest Pain*) in a number of individuals. Guided by P1 and P2, she may infer that a causal chain is likely to go from B to S through some particular but undetermined disease node Y. Since no such path exists in Graph 0, she infers that most likely one of the following new structures is needed: either a new causal link from B to a known cause of S (e.g., *Heart Disease*) or a new causal link to S from a known effect of B, (e.g., *Bronchitis*). If no new link to or from an existing disease node can be added without conflicting with other knowledge, P1 and P2 suggest that a new, previously unobserved disease node Y may exist,

and that Y is causally linked to both B and S (as shown in network Graph 4). Other logically simpler hypotheses, such as inserting a single causal link directly from B to S, or from S to B, are ruled out by the ontology of P1 and the causal laws of P2.

Note that this approach to learning causal network structures from data is very different from how that problem has traditionally been approached, either in machine learning (Pearl, 2000; Glymour et al., 1993) or cognitive psychology (Cheng, 1997; Gopnik et al., 2004; Shanks, 1995), as a primarily bottom-up process of fitting or constructing a causal model that best accounts for the observed patterns of correlation among events. The causal-grammar view treats causal learning as more of a parsing operation, integrating top-down as well as bottom-up constraints in a search for the best causal model among just those candidates consistent with the learner's domain understanding. This view seems to offer more promise for explaining how people can successfully infer causal structures from so little data – sometimes just one or a few observed events, or much less than would be needed to even compute reliable correlations among events.

Finally, we turn to the problem of acquiring framework-level causal theories. Just as probabilistic grammars for languages may be learnable from a finite observed corpus of utterances, causal grammars could also be learnable via statistical methods from observations of a finite observed sample of systems in a given domain. Two aspects of this analogy are particularly worth noting. First, as with the grammar of a language, crucial constraints on causal domain theories may come from knowledge at higher levels of abstraction. Some aspects of a causal grammar may be conditioned by a truly basic (and innate) foundation, a *Universal Theory* or "UT" by analogy to UG in linguistics. But other constraints are likely to come from levels of framework-like knowledge in between the innate foundation and the frontiers of domain theories where learning typically occurs. For instance, principle P2 in the disease-domain grammar only specifes which kinds of causal links may be present; it does not require that any particular causal link necessarily exist. That may be a general quality of causal grammars in biological or social domains, where there appears (at least to most novices) to be a fair amount of arbitrariness in the causal relations that exist. In contrast, causal relations in physical domains may be more highly structured and lawful. For example, every sample of a certain kind of element or particle necessarily interacts with other elements or particles in the same way.

Also as with linguistic grammars, the empirical adequacy of hypotheses about causal theories at the framework level are evaluated only indirectly, on the success or failure of the causal networks they generate. To the extent that a causal-grammar hypothesis tends to generate causal networks that in turn generate the kinds of events a learner frequently observes, that grammar will receive inductive support. As in linguistics, a grammar may fail to predict optimally either by undergenerating or overgenerating. The peril of undergeneration should be clear: if a causal grammar generates only a small subset of the causal networks that the true grammar does, then typically there will be many systems in the domain for which that hypothetical grammar offers no reasonable description. As an example of overgeneration, consider a grammar in the disease domain that is equivalent to principles P1 and P2 except that it combines disease and symptom variables into a single class ("disymptoms") and allows causal links between any two variables in that class. This "disymptom" grammar is strictly more general than principles P1 and P2. Now suppose that we observe data produced according to Graph1. While both grammars are capable

of generating the correct generating network, the data will provide more inductive support for principles P1 and P2 than for the "disymptom" grammar, because the overly general variant generates many more competing causal-network hypotheses that are from the truth (and under which the observed data would be highly unlikely).

4.3. Summary

Viewing intuitive theories in terms of a hierarchy of increasingly abstract knowledge representations has led us to formulate problems of causal inference on three interlocking levels:

1. *Inferring causes and predicting effects.* Infer the hidden causes of an observed event, or predict its unobserved effects, given a theory at the most specific level: a network structure relating causes and effects in the relevant system.
2. *Inferring causal networks.* Infer the structure of a theory at the most specific level – a network of causal relations – that governs a system of observed variables, given more general framework-like knowledge: the principles constraining candidate causal structures in the relevant domain.
3. *Inferring causal principles.* Infer the principles that organize a set of observed causal systems, given higher-level theoretical frameworks: knowledge about a larger domain that encompasses those systems, or domain-general assumptions.

Everyday causal inference unfolds at all of these levels simultaneously, although novel inferences at higher levels may be relatively rare for adults (Gopnik & Meltzoff, 1997). This formulation of causal induction raises a significant computational challenge: explaining how all of these inference problems can be solved in concert.

In the remainder of this chapter, we will propose a response to this computational challenge that exploits the common form of all three problems: knowledge at a more abstract level generates a constrained space of candidate hypotheses to be evaluated based on data from lower levels of abstraction. The tools of Bayesian inference can be used to formulate any one of these inferences in rational statistical terms. We will propose a hierarchical Bayesian framework, in which hypotheses are defined at multiple levels of abstraction and coupled together based on the constraints that each hypothesized structure imposes on hypotheses at lower levels. This hierarchical framework unifies all three levels of inference and shows how a learner may in principle tackle them all simultaneously.

The next section introduces the technical machinery of our hierarchical Bayesian framework. If this appears to be a big step up in mathematical rigor without a clear immediate payoff, we suggest viewing it as a long-term investment. Analogous hierarchical probabilistic models have been proposed in computational linguistics for integrating language acquisition, syntactic parsing, and speech recognition (Charniak, 1993; Jurafsky & Martin, 2000; Manning & Schütze, 1999): candidate probabilistic grammars are evaluated based on how much probability they assign to the most likely parses of an observed corpus of utterances; individual word outputs from a probabilistic speech recognizer are constrained

or re-evaluated based on how well they fit with the most likely parses of the surrounding utterance. While many important aspects of representation and computation remain to be worked out, it is fair to say that the introduction of sophisticated probabilistic models with multiple levels of knowledge representation has revolutionized and reinvigorated the field of computational linguistics over the last ten to fifteen years. We have similarly high hopes for the next ten to fifteen years of research on intuitive causal theories.

5. A hierarchical Bayesian framework for causal inference

We begin with a brief review of the basics of Bayesian inference, and then show how to extend these ideas to multiple levels of inference in a hierarchy of intuitive theories, where each level functions as a hypothesis space generator for the level below.

5.1. Basic Bayes

Bayesian inference provides a general framework for how rational agents should approach problems of induction. We assume an agent who observed some data $\mathcal{D}$ and considers a space of hypotheses $\mathcal{H}$ about the processes by which that data could have been generated. The agent's a priori beliefs about the plausibility of each $h \in \mathcal{H}$, before seeing $\mathcal{D}$ but drawing upon background knowledge $\mathcal{K}$, are expressed in a *prior probability* distribution $P(h|\mathcal{K})$. The principles of Bayesian inference indicate how the agent should modify his or her beliefs in light of the data $\mathcal{D}$, computing a *posterior probability* distribution $P(h|\mathcal{D},\mathcal{K})$.

The key engine for updating beliefs in light of data is *Bayes' rule*,

$$P(h|\mathcal{D},\mathcal{K}) = \frac{P(\mathcal{D}|h,\mathcal{K})P(h|\mathcal{K})}{P(\mathcal{D}|\mathcal{K})}. \tag{5}$$

The likelihood $P(\mathcal{D}|h,\mathcal{K})$ encodes the predictions of each hypothesis h – the probability of observing $\mathcal{D}$ if h were true. The denominator, $P(\mathcal{D}|\mathcal{K})$, is an average of the predictions of all hypotheses in the hypothesis space, weighted by their prior probabilities:

$$P(\mathcal{D}|\mathcal{K}) = \sum_{h' \in \mathcal{H}} P(\mathcal{D}|h',\mathcal{K})P(h'|\mathcal{K}). \tag{6}$$

This denominator serves to normalize the terms that appear in the numerator, ensuring that the posterior $P(h|\mathcal{D},\mathcal{K})$ can be interpreted as a proper probability distribution over hypotheses.

The content of Bayes' rule can be understood intuitively by thinking about what factors make for strong arguments from observed data to hypothesized explanations in science. In order to say that some observed data $\mathcal{D}$ provide provide good reason to believe in hypothesis h, at least two conditions must hold. First, the hypothesis must predict the data. The stronger the predictions that h makes about $\mathcal{D}$, the more support h should receive from the observation of $\mathcal{D}$. Second, independent of the data, the hypothesis must be plausible given everything else we know. One can always construct some post-hoc hypothesis that is consistent with a particular experimental finding, but such a hypothesis would not be considered a good explanation for the data unless it was a well-motivated and principled consequence of our background knowledge. The combined influence of these two factors is captured in the numerator of Bayes' rule: the posterior probability assigned to some

hypothesis h upon seeing $\mathcal{D}$ is proportional to the product of the prior probability, $P(h|\mathcal{K})$, reflecting the a priori plausibility of h, and the likelihood, $P(\mathcal{D}|h, \mathcal{K})$, reflecting the extent to which $\mathcal{D}$ is predicted by h. The denominator reflects a third factor that also influences belief dynamics in science, although its rational status is not always appreciated. Data D provide better support for hypothesis h to the extent that the data are surprising: that is, either the data are unlikely given our background knowledge, or they would not be predicted under most plausible alternative hypotheses. The former condition is just equivalent to saying that $P(\mathcal{D}|\mathcal{K})$ is low, and the lower this term, the higher the posterior probability in Equation 5. The latter condition is just a different framing of the same situation, as expressed in Equation 6: $P(\mathcal{D}|\mathcal{K})$ will be low when $P(\mathcal{D}|h', \mathcal{K})$ is low for plausible alternative hypotheses (those for which $P(h'|\mathcal{K})$ is high).

In short, Bayesian inference provides a rigorous mathematical representation of a basic principle of scientific common sense: to the extent that a given hypothesis is well-motivated and strongly predictive of the observed data, and to the extent that the predicted data are surprising or otherwise unexpected, the hypothesis is more likely to be true. Our contention here is that this approach to inductive inference also offers useful insights into common-sense reasoning and learning with intuitive theories.

Bayes' rule can be applied to any problem requiring an inference about the process that produced some observed data. Different kinds of inductive problems will involve different kinds of hypothesis spaces and different kinds of data, with appropriately modified priors and likelihoods. In the next section, we formalize the three problems of causal inference identified above in Bayesian terms, identifying the hypothesis space and data used in each case and explaining how the priors and likelihoods are determined.

5.2. A hierarchical Bayesian framework

Expressing causal inference problems in Bayesian terms emphasizes the importance of constraints on which hypotheses are possible or likely a priori. Expressing multiple inference problems at different levels of abstraction in a hierarchical Bayesian framework emphasizes the coupling of these constraints across levels of inference.

In presenting our hierarchical framework we adopt the following terms and notation. A *system* is a set of causally related variables within a *domain*. For a system of N variables $\mathbf{X} = \{X_1, \ldots, X_N\}$, an *instance* is an assignment of values to these variables, $\mathbf{x} = \{x_1, \ldots, x_N\}$. We will use uppercase letters to indicate variables, lowercase to indicate their values, and boldface to indicate a set of variables or their values. For any instance, $\mathbf{x}$, a subset of variables, $\mathbf{x}_{\text{obs}}$, are *observed* (i.e., the values of those variables in that instance are known), and the remainder, $\mathbf{x}_{\text{unobs}}$ are *unobserved* (i.e., take on unknown values). A *dataset* d consists of the observed portions of M instances of some system, $d = \{\mathbf{x}_{\text{obs}}^{(1)}, \ldots, \mathbf{x}_{\text{obs}}^{(M)}\}$. Depending on the level of causal inference, the data $\mathcal{D}$ available to the learner may consist of a single observed instance of a system, a dataset d of multiple instances of the same system, or multiple datasets each from a different system in the same domain.

The three inference problems from Section 4 – inferring causes from effects, inferring causal networks from cause-effect observations, and inferring the principles underlying causal network structures in a domain – unfold at different levels of abstraction. To cast all these problems in a unified Bayesian inference framework, we define a hierarchy of increasingly abstract theories, $T_0, T_1, \ldots, T_U$, as the basis for a hierarchical generative model of the

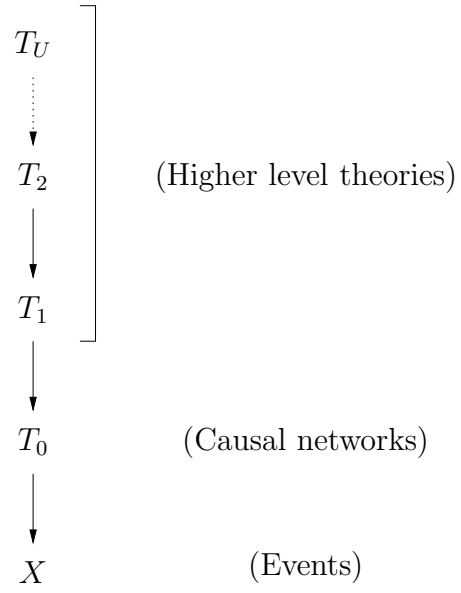

Figure 4. A hierarchical probabilistic model corresponding to the hierarchy of abstraction in causal theories shown in Figure 3. Theories at each level of abstraction define a prior probability distribution over candidate theories at the next level down, bottoming out in the observed data X. Bayesian inferences about theories at each level combine information from the observed data, propagated upwards by successful lower-level theories, with top-down constraints from higher-level theories (and ultimately perhaps some universal conceptual skeleton for all theories, T_U).

data. The subscript indicates the level of theory, with U being the highest level. Theories at each level of the hierarchy generate hypothesis spaces and prior probability distributions for Bayesian inference at the level below. The lowest-level theory, T_0, is a causal network defined upon the variables of a particular system, generating hypotheses about the values of those variables and defining a distribution $P(\mathbf{X}|T_0)$. The next level, T_1, is a set of principles that generate a hypothesis space of causal networks T_0, defining a prior distribution $P(T_0|T_1)$. This suggests a more precise definition of a domain, as the set of systems that can be generated by a theory T_i for $i > 0$. Higher-level theories are defined recursively: for any $i > 0$, a theory T_i generates a hypothesis space of theories T_{i-1} with an associated prior distribution $P(T_{i-1}|T_i)$, giving rise to a hierarchy of increasingly general theories, each with a corresponding (increasingly general) domain.

5.3. Inferring causes and predicting effects

Inferring hidden causes or predicting future effects can both be formulated as problems of inferring the values of the unobserved variables in an instance $\mathbf{x}$. In many cases where only a subset of variables are observed, $\mathbf{x}_{\text{obs}}$ are effects and $\mathbf{x}_{\text{unobs}}$ are their causes, but the problem remains the same if $\mathbf{x}_{\text{obs}}$ correspond to causes or a mixture of causes and effects. In Bayesian terms, we seek to compute the posterior distribution over $\mathbf{X}_{\text{unobs}}$ given $\mathbf{x}_{\text{obs}}$. Such an inference requires knowledge of a hypothesis space of possible values that $\mathbf{X}_{\text{unobs}}$ could take on, the prior probabilities of those values, and the probability of observing the

data $\mathbf{x}_{\text{obs}}$ conditioned on those values. A causal network T_0 can be used to generate values of $\mathbf{x}$, and consequently supplies all of these ingredients.

Taking T_0 as our background knowledge $\mathcal{K}$, we can compute the posterior distribution on $\mathbf{X}_{\text{unobs}}$ by applying Bayes' rule (Equation 5), letting $\mathbf{x}_{\text{obs}}$ play the role of the data $\mathcal{D}$, and $\mathbf{x}_{\text{unobs}}$ the role of the hypothesis h. T_0 specifies the hypothesis space $\mathcal{H} = \mathcal{H}_0$, the prior probability $P(\mathbf{x}_{\text{unobs}}|T_0)$, and the likelihood $P(\mathbf{x}_{\text{obs}}|\mathbf{x}_{\text{unobs}}, T_0)$. We thus have

$$P(\mathbf{x}_{\text{unobs}}|\mathbf{x}_{\text{obs}}, T_0) = \frac{P(\mathbf{x}_{\text{obs}}|\mathbf{x}_{\text{unobs}}, T_0)P(\mathbf{x}_{\text{unobs}}|T_0)}{P(\mathbf{x}_{\text{obs}}|T_0)} \tag{7}$$

where the denominator can be computed by summing over all values $\mathbf{x}_{\text{unobs}}$ allowed by T_0:

$$P(\mathbf{x}_{obs}|T_0) = \sum_{\mathbf{x}_{\text{unobs}} \in \mathcal{H}_0} P(\mathbf{x}_{\text{obs}}|\mathbf{x}_{\text{unobs}}, T_0)P(\mathbf{x}_{\text{unobs}}|T_0). \tag{8}$$

Evaluating Equation 7 is just the standard process of inference in a Bayesian network. The network not only sets up the hypothesis space for these computations, it also allows them to be carried out efficiently. It provides a structured representation of the joint probability distribution over all variables that enables Equations 7 and 8 to be computed by simple local computations (Pearl, 1988; Russell & Norvig, 2002).

5.4. Inferring causal networks

The problem of inferring causal network structures from cause-effect observations can be formalized as identifying the T_0-level theory that best explains a dataset d of M partially observed instances of a system. Standard "data-mining" algorithms for learning causal networks (e.g., Spirtes et al., 1993; Pearl, 2000; Heckerman, 1998) offer one approach to this problem, but for several reasons they are not promising as rational accounts of human causal learning. These algorithms require large samples to identify correlations among variables, yet human learners are willing to infer causal relationships from only a few observations, where correlations cannot be identified reliably (Gopnik, Glymour, Sobel, Schulz, Kushnir & Danks, 2004; Gopnik & Schulz, 2004; Griffiths, Baraff, & Tenenbaum, 2004; Griffiths & Tenenbaum, in press; Sobel, Tenenbaum, & Gopnik, 2004; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003; Tenenbaum, Sobel, Griffiths & Gopnik, submitted).

Human learners are able to learn causal structure from such limited data because they draw on strong prior knowledge that generic data-mining algorithms for learning causal networks are not designed to exploit. Rather than treating all variables of a causal system as equal a priori, as those algorithms do, people will typically conceive of the variables in terms of properties and relations on objects. Domain-specific theories at a more abstract level – knowledge about classes of objects and predicates, and causal laws relating these classes – will set up strong expectations about the kinds of causal network structures and functional dependencies between variables that are likely to be encountered. This is the function of principles P1 and P2 in the disease domain. The scenario in Section 4, in which a learner infers a novel hidden disease variable to explain a newly observed behavior-symptom correlation, is one example of how such domain theories may guide human learning of causal structure. People also have domain-specific knowledge about how kinds of causal mechanisms work. This knowledge may be quite skeletal (Keil, 2003), but it is often sufficient to

generate useful constraints on the nature of the functional dependency between causes and their effects. By specifying whether a hypothetical causal link – if it exists – is likely to be deterministic or probabilistic, generative or inhibitory, strong or weak, independent of other links or interacting with them, skeletal mechanism knowledge may allow learners to infer which causal relations do in fact exist, from much less data than would be required without those expectations.

This knowledge-driven approach to causal structure learning fits naturally into our hierarchical Bayesian framework. We want to compute a posterior distribution over causal networks T_0, given a dataset d and relevant background knowledge $\mathcal{K}$. The background knowledge $\mathcal{K}$ takes the form of a more abstract theory T_1, which generates a hypothesis space $\mathcal{H}_1$ of causal networks T_0, and a prior on that hypothesis space, $P(T_0|T_1)$. The probability of the dataset d under each network T_0 can be computed as follows. T_0 specifies a joint distribution over the system's variables $\mathbf{X} = \{X_1, \ldots, X_N\}$, which determines the probability $P(\mathbf{x}_{\text{obs}}^{(i)}|T_0)$ of the ith partially observed instance (Equation 8). Assuming each instance in d is sampled independently, the total probability of the dataset is

$$P(d|T_0) = \prod_{i=1}^{M} P(\mathbf{x}_{\text{obs}}^{(i)}|T_0). \tag{9}$$

We can now apply Bayes rule (Equation 5) to compute the posterior probability of a particular causal network T_0 given a dataset d and a higher-level theory T_1:

$$P(T_0|d, T_1) = \frac{P(d|T_0)P(T_0|T_1)}{P(d|T_1)}, \tag{10}$$

where the denominator is

$$P(d|T_1) = \sum_{T_0 \in \mathcal{H}_1} P(d|T_0)P(T_0|T_1). \tag{11}$$

The sum over all possible networks in Equation 11 may be computed exactly for very small systems, but in general requires some kind of stochastic sampling-based approximation scheme (e.g., Friedman & Koller, 2000).

Several cognitive scientists have proposed that human causal learning is best thought of as a knowledge-based, theory-based or top-down process (e.g., Waldmann, 1996; Lagnado & Sloman, 2004; Lagnado, Hagmayer, Sloman and Waldmann, this volume). However, these proposals have been relatively qualitative and informal. There has not been a widespread effort to propose and test principled domain-general frameworks for modeling theory-based induction of causal structure, as there has been for more bottom-up associative accounts (Rescorla & Wagner, 1972; Cheng & Novick, 1990; Cheng, 1997; Lober & Shanks, 2000; Danks, 2003). Our analysis aims to formalize the knowledge that guides causal structure learning, and to provide a rational account of how it does so. The roles of both top-down constraints from prior knowledge and bottom-up influences from observed data are reflected in the two terms in the numerator of Equation 10: the higher-order theory T_1 defines the prior probability $P(T_0|T_1)$ and delimits the set of causal networks under consideration, while the data favors some networks within this set via the likelihood $P(d|T_0)$. In a series of papers (Griffiths, Baraff, & Tenenbaum, 2004; Griffiths & Tenenbaum, in press, in preparation;

Tenenbaum & Griffiths, 2001; 2003; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003) we have shown how this theory-based Bayesian framework can be used to build rational and quantitatively accurate models of people's inferences about causal structure from very limited data.

5.5. Inferring causal principles

The machinery for theory-based inference of causal network structures T_0 can be extended up the hierarchy of theories, making it possible, in principle, to learn a theory at any level. For instance, given data $\mathcal{D}$ drawn from one or more causal systems in a domain, we can make inferences about the T_1-level principles that govern those systems (e.g., abstract classes of variables and causal laws such as principles P1 and P2 in the disease domain). We compute a posterior distribution over T_1 theories by applying Bayes' rule (Equation 5) to a hypothesis space $\mathcal{H}_2$ generated by a higher-order theory T_2.

$$P(T_1|\mathcal{D}, T_2) = \frac{P(\mathcal{D}|T_1)P(T_1|T_2)}{P(\mathcal{D}|T_2)}. \tag{12}$$

Assuming that $\mathcal{D}$ consists of L independent datasets $\{d_1, \ldots, d_L\}$, we can compute the likelihood $P(\mathcal{D}|T_1)$ as

$$P(\mathcal{D}|T_1) = \prod_{i=1}^{L} P(d_i|T_1). \tag{13}$$

Each term $P(d_i|T_1)$ corresponds to the denominator in Bayes' rule applied at the next level down (Equation 11), obtained by summing over all causal networks T_0 generated by T_1 for each of the systems represented in $\mathcal{D}$. $P(T_1|T_2)$ is the distribution over theories at level T_1 defined by the higher-level theory T_2. $P(\mathcal{D}|T_2)$ is computed in the same way as $P(d|T_1)$, except that it requires summing over all theories T_1 in $\mathcal{H}_2$, as well as all causal networks T_0 in the hypothesis space $\mathcal{H}_1$ associated with T_1. $P(\mathcal{D}|T_2)$ can be used to make inferences about T_2 given $\mathcal{D}$, and so on up the hierarchy.

This analysis shows that an ideal learner should be able to acquire higher-level causal theories from data, given an appropriate hypothesis space of candidate theories. In practice, as each new level of theory adds a whole hypothesis space of hypothesis spaces that the learner must sum over, carrying out all the required computations quickly becomes intractable. Bayesian statisticians often approximate exact inference in hierarchical models by replacing a sum over all hypotheses with a search for the most probable hypothesis, or with a sum over a sample of hypotheses generated by Markov Chain Monte Carlo techniques (e.g., Gilks, Richardson, & Spiegelhalter, 1996. An interesting open question is how the cognitive processes involved in theory change and acquisition might correspond to some of these methods for approximating Bayesian inference with complex, hierarchically structured hypothesis spaces.

There may also be processes involved in the acquisition of higher-order theories that are not so clearly evidential in nature, or that draw on kinds of evidence that are very different from direct observations of systems in the world. For instance, when a child hears an adult talking about the causal structure of a complex domain such as intuitive biology or psychology (e.g., Carey, 1985a; Gopnik & Meltzoff, 1997), invoking various hidden causes and abstract concepts, she might receive very useful evidence about the relative value of

alternative hypotheses for higher-order, framework-level theories in these domains. It is far from clear how to capture the inferences a child might make from such data in terms of our hierarchical Bayesian framework, but this remains another important open question for the research program to address.

6. Summary

We began this chapter with three guiding questions (Table 1): what is the knowledge content and representational form of intuitive theories, how are theories used to guide causal inference, and how are theories themselves acquired? Rather than stipulating arbitrarily the properties that intuitive theories should have, or trying to give a fully general account of theories, we have presented a rational analysis of causal induction and restricted ourselves to accounting for those aspects of intuitive theories necessary to explain how people perform these tasks.

The key challenge of causal induction we identified was the need to make inferences about unobservable causal relations from very sparse observed data. We argued that these inferences are made possible by strong constraints from more abstract levels of causal knowledge. These constraints often arise from domain-specific principles that run counter to simplicity or other general-purpose inductive biases, such as when a novel association between a risky behavior and a known medical symptom is attributed to an indirect link via an unknown disease, rather than to a direct causal link between the behavior and symptom.

Inspired by proposals from developmental psychology and philosophy of science, we suggested that both networks of causal relations and the more abstract causal principles that constrain them may be thought of as intuitive domain theories, but at different levels of abstraction. These levels of abstraction correspond roughly to the notions of "specific theories" and "framework theories" introduced by Wellman and Gelman (1992), but we expect there will typically be multiple levels of increasingly abstract, broad-coverage framework-like causal knowledge. Each level in this hierarchy of theories provides constraints on candidate theories at the next level down, and is itself constrained by knowledge at higher levels, perhaps ultimately grounding out in a "universal theory" of conceptual primitives underlying all intuitive domains.

Viewed from this hierarchical perspective, our initial questions about the structure, function, and acquisition of intuitive causal theories now come down to these two: how do we represent knowledge of causal structure at multiple levels of theoretical abstraction, and what processes of inference connect those knowledge levels to support learning and reasoning across the hierarchy? Existing computational formalisms based on causal Bayesian networks may be appropriate for characterizing causal theories at the most specific level, but they do not extend to the higher levels of abstraction that this hierarchical picture calls out for.

As a first step towards answering these questions, we proposed the *causal grammar* analogy: a framework for thinking about representation and inference in a hierarchy of causal theories based on parallels with some classic representational structures and inferential mechanisms that have been posited to explain language comprehension and acquisition. Just as the grammar of a natural language generates a constrained hypothesis space of syntactic structures considered in sentence comprehension, so does the set of abstract causal principles (or the framework theory) for a domain generate a constrained hypothesis space of causal network structures (or specific theories) considered in causal induction. Just as

linguistic grammars can be expressed in terms of a set of abstract syntactic categories and rules for composing instances of those categories into viable syntactic tree structures, so can higher-order causal theories – or causal grammars – be expressed in terms of a set of abstract categories of causal variables and rules for how variables in those classes can or must be related to form plausible causal network structures. Both linguistic grammars and causal grammars must also be reliably learnable, based on a combination of the primary data available to people and the constraints on possible grammars provided by more abstract, possibly innate conceptual primitives. Hypotheses about linguistic grammars or causal grammars can only be evaluated indirectly, based on how well the specific syntactic tree structures or causal network structures that they generate explain the observed primary data.

Finally, we outlined a more formal approach to learning and reasoning in a hierarchy of theories based on the tools of hierarchical Bayesian models. This analysis provides a principled and unified approach to solving causal induction problems at all levels of our hierarchy of abstraction. At its heart is the idea that intuitive theories at each level of abstraction generate hypothesis spaces for Bayesian inference about theories at lower levels, and are themselves learned via Bayesian inference over hypothesis spaces generated by higher-level theories. Thus the inductive mechanisms operating at each level of abstraction are essentially the same, and they can proceed in parallel to support coupled inferences at all levels.

This Bayesian framework provides a rational analysis of how inference and learning can operate in a hierarchy of intuitive causal theories, but it does not directly address the question of how to represent the structure and content of those theories. In terms of Table 1, we have presented a formal answer to the second and third questions, but only an incomplete answer to the first question: theories at the lowest, most specific level might be represented as causal Bayesian networks, while higher-level theories will require more expressive representations, somewhat like generative grammars. The companion to this chapter (Griffiths & Tenenbaum, this volume) examines in detail two possible representational frameworks for theories at higher levels of abstraction, based on graph schemas and predicate logic. We will show precisely how each of these two representational frameworks can fulfill the functional role of T_1-level theories in our hierarchical Bayesian picture, how they support inferences about specific cause-effect relations from sparse data and how they may themselves be learned or adjusted based on the data observed. We will identify complementary strengths and weaknesses of each representation, ultimately arguing that neither of these formalisms provides a fully adequate account of the structure and function of abstract causal theories. Still, a consideration of these alternatives lays out the challenges for future work and offers some possibilities for what the answers may look like.

References

Ahn, W., & Kalish, C. (2000). The role of covariation vs. mechanism information in causal attribution. In R. Wilson & F. Keil (Eds.), *Cognition and explanation.* Cambridge, MA: MIT Press.

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Erlbaum.

Atran, S. (1995). Classifying nature across cultures. In E. E. Smith & D. N. Osherson (Eds.), *Thinking: An invitation to cognitive science* (Vol. 3, p. 131-174). Cambridge, MA: MIT Press.

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

Carey, S. (1985b). Constraints on semantic development. In J. Mehler (Ed.), *Neonate cognition* (p. 381-398). Hillsdale, NJ: Erlbaum.

Carnap, R. (1956). Empiricism, semantics, and ontology. In (2nd ed.). Chicago: University of Chicago 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.

Cheng, P. W., & Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology*, *58*, 545-567.

Chomsky, N. (1956). Threee models for the description of language. *IRE Transactions on Information Theory*, *2*, 113-124.

Chomsky, N. (1962). Explanatory models in linguistics. In E. Nagel, P. Suppes, & A. Tarski (Eds.), *Logic, methodology, and philosophy of science* (p. 528-550). Stanford, CA: Stanford University Press.

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

Chomsky, N. (1986). *Language and problems of knowledge: The Managua lectures.* Cambridge, MA: MIT Press.

Danks, D. (2003). Equilibria of the Rescorla-Wagner Model. *Journal of Mathematical Psychology*, *47*, 109-121.

Danks, D., & McKenzie, C. R. M. (under revision). *Learning complex causal structures.*

Friedman, N., & Koller, D. (2000). Being Bayesian about network structure. In *Proceedings of the 16th annual conference on uncertainty in ai* (p. 201-210). Stanford, CA.

Gilks, W., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov chain Monte Carlo in practice.* Suffolk: Chapman and Hall.

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

Godfrey-Smith, P. (2003). *Theory and reality.* Chicago: University of Chicago Press.

Gopnik, A., & Glymour, C. (2002). Causal maps and Bayes nets: A cognitive and computational account of theory-formation. In *The cognitive basis of science.* Cambridge: Cambridge University Press.

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

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories.* Cambridge, MA: MIT Press.

Gopnik, A., & Schulz, L. (2004). Mechanisms of theory formation in young children. *Trends in Cognitive Science*, *8*, 371-377.

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 press). Elemental causal induction. *Cognitive Psychology.*

Heckerman, D. (1998). A tutorial on learning with Bayesian networks. In M. I. Jordan (Ed.), *Learning in graphical models* (p. 301-354). Cambridge, MA: MIT Press.

Inagaki, K., & Hatano, G. (2002). *Young children's thinking about biological world.* New York: Psychology press.

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

Keil, F. C. (2003). Folkscience: Coarse interpretations of a complex reality. *Trends in Cognitive Science*, *7*, 368-373.

Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist*, *28*, 107-128.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press.

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

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge.* Cambridge: Cambridge University Press.

Laudan, L. (1977). *Progress and its problems.* Berkeley, CA: University of California Press.

Lober, K., & Shanks, D. (2000). Is causal induction based on causal power? Critique of Cheng (1997). *Psychological Review*, *107*, 195-212.

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

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

McCloskey, M. (1983). Intuitive physics. *Scientific American*, *284*, 114-123.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*, 289-316.

Nichols, S., & Stich, S. (2003). *Mindreading: An integrated account of pretense, self-awareness and understanding other minds.* Oxford: Oxford University Press.

Nowak, M. A., Komarova, N. L., & Niyogi, P. (2003). Computational and evolutionary aspects of language. *Nature*, *417*, 611-617.

Oaksford, M., & Chater, N. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Science*, *3*, 57-65.

Pazzani, M. (1987). Inducing causal and social theories: A prerequisite for explanation-based learning. In *Proceedings of the fourth international workshop on machine learning* (p. 230-241). Irvine, CA: Morgan Kaufmann.

Pazzani, M., Dyer, M., & Flowers, M. (1986). The role of prior causal theries in generalization. In *Proceedings of the fifth national conference on artificial intelligence* (p. 545-550). Philadelphia, PA: Morgan Kaufmann.

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.

Quine, W. O. (1951). Two dogmas of empiricism. *Philosophical Review*, 20-43.

Rehder, B. (2003). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1141-1159.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations on the effectiveness of reinforcement and non-reinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (p. 64-99). New York: Appleton-Century-Crofts.

Russell, S. J., & Norvig, P. (2002). *Artificial intelligence: A modern approach* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

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

Shepard, R. N. (1987). Towards a universal law of generalization for psychological science. *Science*, *237*, 1317-1323.

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.

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

Tenenbaum, J. B., & Griffiths, T. L. (2003). Theory-based causal induction. In *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.*

Waldmann, M. R. (1996). Knowledge-based causal induction. In *The psychology of learning and motivation* (Vol. 34, p. 47-88). San Diego: Academic Press.

Wellman, H. M. (1990). *The child's theory of mind.* Cambridge, MA: MIT Press.

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology*, *43*, 337-375.