How to Grow a Mind: Statistics, Structure, and Abstraction

Joshua B. Tenenbaum,[1]* Charles Kemp,[2] Thomas L. Griffiths,[3] Noah D. Goodman[4]

In coming to understand the world—in learning concepts, acquiring language, and grasping causal relations—our minds make inferences that appear to go far beyond the data available. How do we do it? This review describes recent approaches to reverse-engineering human learning and cognitive development and, in parallel, engineering more humanlike machine learning systems. Computational models that perform probabilistic inference over hierarchies of flexibly structured representations can address some of the deepest questions about the nature and origins of human thought: How does abstract knowledge guide learning and reasoning from sparse data? What forms does our knowledge take, across different domains and tasks? And how is that abstract knowledge itself acquired?

The Challenge: How Does the Mind Get So Much from So Little?

For scientists studying how humans come to understand their world, the central challenge is this: How do our minds get so much from so little? We build rich causal models, make strong generalizations, and construct powerful abstractions, whereas the input data are sparse, noisy, and ambiguous—in every way far too limited. A massive mismatch looms between the information coming in through our senses and the ouputs of cognition.

Consider the situation of a child learning the meanings of words. Any parent knows, and scientists have confirmed (*1*, *2*), that typical 2-year-olds can learn how to use a new word such as "horse" or "hairbrush" from seeing just a few examples. We know they grasp the meaning, not just the sound, because they generalize: They use the word appropriately (if not always perfectly) in new situations. Viewed as a computation on sensory input data, this is a remarkable feat. Within the infinite landscape of all possible objects, there is an infinite but still highly constrained subset that can be called "horses" and another for "hairbrushes." How does a child grasp the boundaries of these subsets from seeing just one or a few examples of each? Adults face the challenge of learning entirely novel object concepts less often, but they can be just as good at it (Fig. 1).

Generalization from sparse data is central in learning many aspects of language, such as syntactic constructions or morphological rules (*3*). It presents most starkly in causal learning: Every statistics class teaches that correlation does not imply causation, yet children routinely infer causal links from just a handful of events (*4*), far too small a sample to compute even a reliable correlation! Perhaps the deepest accomplishment of cognitive development is the construction of larger-scale systems of knowledge: intuitive theories of physics, psychology, or biology or rule systems for social structure or moral judgment. Building these systems takes years, much longer than learning a single new word or concept, but on this scale too the final product of learning far outstrips the data observed (*5*–*7*).

Philosophers have inquired into these puzzles for over two thousand years, most famously as "the problem of induction," from Plato and Aristotle through Hume, Whewell, and Mill to Carnap, Quine, Goodman, and others in the 20th century (*8*). Only recently have these questions become accessible to science and engineering by viewing inductive learning as a species of computational problems and the human mind as a natural computer evolved for solving them.

The proposed solutions are, in broad strokes, just what philosophers since Plato have suggested. If the mind goes beyond the data given, another source of information must make up the difference. Some more abstract background knowledge must generate and delimit the hypotheses learners consider, or meaningful generalization would be impossible (*9*, *10*). Psychologists and linguists speak of "constraints;" machine learning and artificial intelligence researchers, "inductive bias;" statisticians, "priors."

This article reviews recent models of human learning and cognitive development arising at the intersection of these fields. What has come to be known as the "Bayesian" or "probabilistic" approach to reverse-engineering the mind has been heavily influenced by the engineering successes of Bayesian artificial intelligence and machine learning over the past two decades (*9*, *11*) and, in return, has begun to inspire more powerful and more humanlike approaches to machine learning.

As with "connectionist" or "neural network" models of cognition (*12*) in the 1980s (the last moment when all these fields converged on a common paradigm for understanding the mind), the labels "Bayesian" or "probabilistic" are merely placeholders for a set of interrelated principles and theoretical claims. The key ideas can be thought of as proposals for how to answer three central questions:

1) How does abstract knowledge guide learning and inference from sparse data?
2) What forms does abstract knowledge take, across different domains and tasks?
3) How is abstract knowledge itself acquired?

We will illustrate the approach with a focus on two archetypal inductive problems: learning concepts and learning causal relations. We then briefly discuss open challenges for a theory of human cognitive development and conclude with a summary of the approach's contributions.

We will also draw contrasts with two earlier approaches to the origins of knowledge: nativism and associationism (or connectionism. These approaches differ in whether they propose stronger or weaker capacities as the basis for answering the questions above. Bayesian models typically combine richly structured, expressive knowledge representations (question 2) with powerful statistical inference engines (questions 1 and 3), arguing that only a synthesis of sophisticated approaches to both knowledge representation and inductive inference can account for human intelligence. Until recently it was not understood how this fusion could work computationally. Cognitive modelers were forced to choose between two alternatives (*13*): powerful statistical learning operating over the simplest, unstructured forms of knowledge, such as matrices of associative weights in connectionist accounts of semantic cognition (*12*, *14*), or richly structured symbolic knowledge equipped with only the simplest, nonstatistical forms of learning, checks for logical inconsistency between hypotheses and observed data, as in nativist accounts of language acquisition (*15*). It appeared necessary to accept either that people's abstract knowledge is not learned or induced in a nontrivial sense from experience (hence essentially innate) or that human knowledge is not nearly as abstract or structured (as "knowledge-like") as it seems (hence simply associations). Many developmental researchers rejected this choice altogether and pursued less formal approaches to describing the growing minds of children, under the headings of "constructivism" or the "theory theory" (*5*). The potential to explain how people can genuinely learn with abstract structured knowledge may be the most distinctive feature of Bayesian models: the biggest reason for their recent popularity (*16*) and the biggest target of skepticism from their critics (*17*).

The Role of Abstract Knowledge

Over the past decade, many aspects of higher-level cognition have been illuminated by the

[1]Department of Brain and Cognitive Sciences, Computer Science and Artificial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA. [2]Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [3]Department of Psychology, University of California, Berkeley, Berkeley, CA 94720, USA. [4]Department of Psychology, Stanford University, Stanford, CA 94305, USA.

*To whom correspondence should be addressed. E-mail: jbt@mit.edu

Downloaded from www.sciencemag.org on March 10, 2011

mathematics of Bayesian statistics: our sense of similarity (*18*), representativeness (*19*), and randomness (*20*); coincidences as a cue to hidden causes (*21*); judgments of causal strength (*22*) and evidential support (*23*); diagnostic and conditional reasoning (*24*, *25*); and predictions about the future of everyday events (*26*).

The claim that human minds learn and reason according to Bayesian principles is not a claim that the mind can implement any Bayesian inference. Only those inductive computations that the mind is designed to perform well, where biology has had time and cause to engineer effective and efficient mechanisms, are likely to be understood in Bayesian terms. In addition to the general cognitive abilities just mentioned, Bayesian analyses have shed light on many specific cognitive capacities and modules that result from rapid, reliable, unconscious processing, including perception (*27*), language (*28*), memory (*29*, *30*), and sensorimotor systems (*31*). In contrast, in tasks that require explicit conscious manipulations of probabilities as numerical quantities—a recent cultural invention that few people become fluent with, and only then after sophisticated training—judgments can be notoriously biased away from Bayesian norms (*32*).

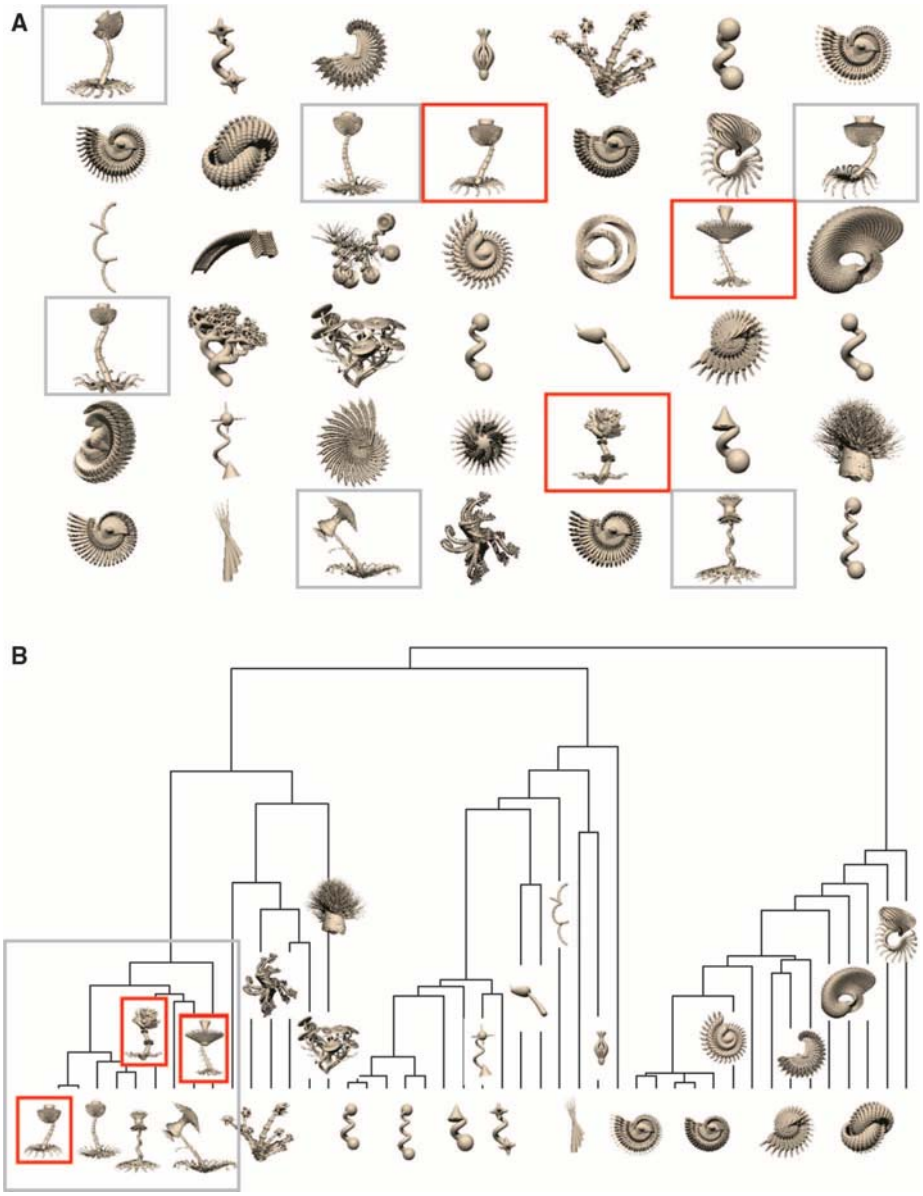

Fig. 1. Human children learning names for object concepts routinely make strong generalizations from just a few examples. The same processes of rapid generalization can be studied in adults learning names for novel objects created with computer graphics. (**A**) Given these alien objects and three examples (boxed in red) of "tufas" (a word in the alien language), which other objects are tufas? Almost everyone selects just the objects boxed in gray (*75*). (**B**) Learning names for categories can be modeled as Bayesian inference over a tree-structured domain representation (*2*). Objects are placed at the leaves of the tree, and hypotheses about categories that words could label correspond to different branches. Branches at different depths pick out hypotheses at different levels of generality (e.g., Clydesdales, draft horses, horses, animals, or living things). Priors are defined on the basis of branch length, reflecting the distinctiveness of categories. Likelihoods assume that examples are drawn randomly from the branch that the word labels, favoring lower branches that cover the examples tightly; this captures the sense of suspicious coincidence when all examples of a word cluster in the same part of the tree. Combining priors and likelihoods yields posterior probabilities that favor generalizing across the lowest distinctive branch that spans all the observed examples (boxed in gray).

At heart, Bayes's rule is simply a tool for answering question 1: How does abstract knowledge guide inference from incomplete data? Abstract knowledge is encoded in a probabilistic generative model, a kind of mental model that describes the causal processes in the world giving rise to the learner's observations as well as unobserved or latent variables that support effective prediction and action if the learner can infer their hidden state. Generative models must be probabilistic to handle the learner's uncertainty about the true states of latent variables and the true causal processes at work. A generative model is abstract in two senses: It describes not only the specific situation at hand, but also a broader class of situations over which learning should generalize, and it captures in parsimonious form the essential world structure that causes learners' observations and makes generalization possible.

Bayesian inference gives a rational framework for updating beliefs about latent variables in generative models given observed data (*33*, *34*). Background knowledge is encoded through a constrained space of hypotheses H about possible values for the latent variables, candidate world structures that could explain the observed data. Finer-grained knowledge comes in the "prior probability" $P(h)$, the learner's degree of belief in a specific hypothesis h prior to (or independent of) the observations. Bayes's rule updates priors to "posterior probabilities" $P(h|d)$ conditional on the observed data d:

$$P(h|d) = \frac{P(d|h)P(h)}{\sum_{h' \in H} P(d|h')P(h')} \propto P(d|h)P(h) \quad (1)$$

The posterior probability is proportional to the product of the prior probability and the likelihood $P(d|h)$, measuring how expected the data are under hypothesis h, relative to all other hypotheses h' in H.

To illustrate Bayes's rule in action, suppose we observe John coughing (d), and we consider three hypotheses as explanations: John has h_1, a cold; h_2, lung disease; or h_3, heartburn. Intuitively only h_1 seems compelling. Bayes's rule explains why. The likelihood favors h_1 and h_2 over h_3: only colds and lung disease cause coughing and thus elevate the probability of the data above baseline. The prior, in contrast, favors h_1 and h_3 over h_2: Colds and heartburn are much more common than lung disease. Bayes's rule weighs

Downloaded from www.sciencemag.org on March 10, 2011

hypotheses according to the product of priors and likelihoods and so yields only explanations like h_1 that score highly on both terms.

The same principles can explain how people learn from sparse data. In concept learning, the data might correspond to several example objects (Fig. 1) and the hypotheses to possible extensions of the concept. Why, given three examples of different kinds of horses, would a child generalize the word "horse" to all and only horses (h_1)? Why not h_2, "all horses except Clydesdales"; h_3, "all animals"; or any other rule consistent with the data? Likelihoods favor the more specific patterns, h_1 and h_2; it would be a highly suspicious coincidence to draw three random examples that all fall within the smaller sets h_1 or h_2 if they were actually drawn from the much larger h_3 (*18*). The prior favors h_1 and h_3, because as more coherent and distinctive categories, they are more likely to be the referents of common words in language (*1*). Only h_1 scores highly on both terms. Likewise, in causal learning, the data could be co-occurences between events; the hypotheses, possible causal relations linking the events. Likelihoods favor causal links that make the co-occurence more probable, whereas priors favor links that fit with our background knowledge of what kinds of events are likely to cause which others; for example, a disease (e.g., cold) is more likely to cause a symptom (e.g., coughing) than the other way around.

The Form of Abstract Knowledge

Abstract knowledge provides essential constraints for learning, but in what form? This is just question 2. For complex cognitive tasks such as concept learning or causal reasoning, it is impossible to simply list every logically possible hypothesis along with its prior and likelihood. Some more sophisticated forms of knowledge representation must underlie the probabilistic generative models needed for Bayesian cognition.

In traditional associative or connectionist approaches, statistical models of learning were defined over large numerical vectors. Learning was seen as estimating strengths in an associative memory, weights in a neural network, or parameters of a high-dimensional nonlinear function (*12*, *14*). Bayesian cognitive models, in contrast, have had most success defining probabilities over more structured symbolic forms of knowledge representations used in computer science and artificial intelligence, such as graphs, grammars, predicate logic, relational schemas, and functional programs. Different forms of representation are used to capture people's knowledge in different domains and tasks and at different levels of abstraction.

In learning words and concepts from examples, the knowledge that guides both children's and adults' generalizations has been well described using probabilistic models defined over tree-structured representations (Fig. 1B) (*2*, *35*). Reasoning about other biological concepts for natural kinds (e.g., given that cows and rhinos have protein X in their muscles, how likely is it that horses or squirrels do?) is also well described by Bayesian models that assume nearby objects in the tree are likely to share properties (*36*). However, trees are by no means a universal representation. Inferences about other kinds of categories or properties are best captured by using probabilistic models with different forms (Fig. 2): two-dimensional spaces or grids for reasoning about geographic properties of cities, one-dimensional orders for reasoning about values or abilities, or directed networks for causally transmitted properties of species (e.g., diseases) (*36*).

Knowledge about causes and effects more generally can be expressed in a directed graphical model (*9*, *11*): a graph structure where nodes represent variables and directed edges between nodes represent probabilistic causal links. In a medical setting, for instance (Fig. 3A), nodes might represent whether a patient has a cold, a cough, a fever or other conditions, and the presence or absence of edges indicates that colds tend to cause coughing and fever but not chest pain; lung disease tends to cause coughing and chest pain but not fever; and so on.

Such a causal map represents a simple kind of intuitive theory (*4*), but learning causal networks from limited data depends on the constraints of more abstract knowledge. For example, learning causal dependencies between medical conditions is enabled by a higher-level framework theory (*37*) specifying two classes of variables (or nodes), diseases and symptoms, and the tendency for causal relations (or graph edges) to run from diseases to symptoms, rather than within these classes or from symptoms to diseases (Fig. 3, A to C). This abstract framework can be represented by using probabilistic models defined over relational data structures such as graph schemas (*9*, *38*), templates for graphs based on types of nodes, or probabilistic graph grammars (*39*), similar in spirit to the probabilistic grammars for strings that have become standard for representing linguistic knowledge (*28*). At the most abstract level, the very concept of causality itself, in the sense of a directed relationship that supports intervention or manipulation by an external agent (*40*), can be formulated as a set of logical laws expressing constraints on the structure of directed graphs relating actions and observable events (Fig. 3D).

Each of these forms of knowledge makes different kinds of prior distributions natural to define and therefore imposes different constraints on induction. Successful generalization depends on getting these constraints right. Although inductive constraints are often graded, it is easiest to appreciate the effects of qualitative constraints that simply restrict the hypotheses learners can consider (i.e., setting priors for many logical possible hypotheses to zero). For instance, in learning concepts over a domain of n objects, there are 2^n subsets and hence 2^n logically possible hypotheses for the extension of a novel concept. Assuming concepts correspond to the branches of a specific binary tree over the objects, as in Fig. 1B, restricts this space to only $n - 1$ hypotheses. In learning a causal network over 16 variables, there are roughly 10^{46} logically possible hypotheses (directed acyclic graphs), but a framework theory restricting hypotheses to bipartite disease-symptom graphs reduces this to roughly 10^{23} hypotheses. Knowing which variables belong to the disease and symptom classes further restricts this to roughly 10^{18} networks. The smaller the hypothesis space, the more accurately a learner can be expected to generalize, but only as long as the true structure to be learned remains within or near (in a probabilistic sense) the learner's hypothesis space (*10*). It is no coincidence then that our best accounts of people's mental representations often resemble simpler versions of how scientists represent the same domains, such as tree structures for biological species. A compact description that approximates how the grain of the world actually runs offers the most useful form of constraint on inductive learning.

The Origins of Abstract Knowledge

The need for abstract knowledge and the need to get it right bring us to question 3: How do learners learn what they need to know to make learning possible? How does a child know which tree structure is the right way to organize hypotheses for word learning? At a deeper level, how can a learner know that a given domain of entities and concepts should be represented by using a tree at all, as opposed to a low-dimensional space or some other form? Or, in causal learning, how do people come to correct framework theories such as knowledge of abstract disease and symptom classes of variables with causal links from diseases to symptoms?

The acquisition of abstract knowledge or new inductive constraints is primarily the province of cognitive development (*5*, *7*). For instance, children learning words initially assume a flat, mutually exclusive division of objects into nameable clusters; only later do they discover that categories should be organized into tree-structured hierarchies (Fig. 1B) (*41*). Such discoveries are also pivotal in scientific progress: Mendeleev launched modern chemistry with his proposal of a periodic structure for the elements. Linnaeus famously proposed that relationships between biological species are best explained by a tree structure, rather than a simpler linear order (premodern Europe's "great chain of being") or some other form.

Such structural insights have long been viewed by psychologists and philosophers of science as deeply mysterious in their mechanisms, more magical than computational. Conventional algorithms for unsupervised structure discovery in statistics and machine learning—hierarchical clustering, principal components analysis, multidimensional scaling, clique detection—assume a single fixed form of structure (*42*). Unlike human children or scientists, they cannot learn multiple forms of structure or discover new forms in novel data. Neither traditional approach to cognitive development has a fully satisfying response: Nativists have assumed that,

Downloaded from www.sciencemag.org on March 10, 2011

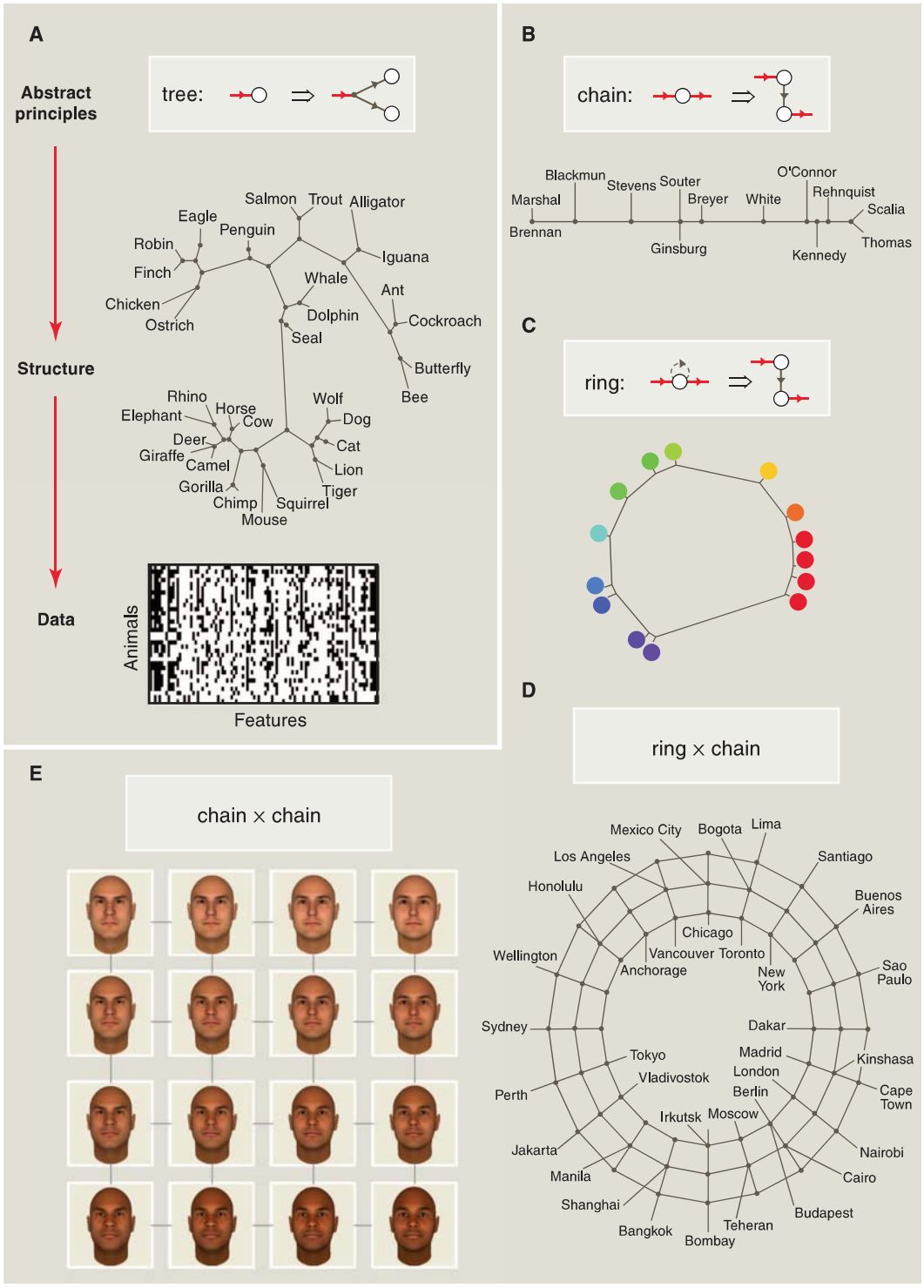

Fig. 2. Kemp and Tenenbaum (*47*) showed how the form of structure in a domain can be discovered by using a HBM defined over graph grammars. At the bottom level of the model is a data matrix *D* of objects and their properties, or similarities between pairs of objects. Each square of the matrix represents whether a given feature (column) is observed for a given object (row). One level up is the structure *S*, a graph of relations between objects that describes how the features in *D* are distributed. Intuitively, objects nearby in the graph are expected to share similar feature values; technically, the graph Laplacian parameterizes the inverse covariance of a gaussian distribution with one dimension per object, and each feature is drawn independently from that distribution. The highest level of abstract principles specifies the form *F* of structure in the domain, in terms of grammatical rules for growing a graph *S* of a constrained form out of an initial seed node. Red arrows represent $P(S|F)$ and $P(D|S)$, the conditional probabilities that each level specifies for the level below. A search algorithm attempts to find both the form *F* and the structure *S* of that form that jointly maximize the posterior probability $P(S,F|D)$, a function of the product of $P(D|S)$ and $P(S|F)$. (**A**) Given as data the features of animals, the algorithm finds a tree structure with intuitively sensible categories at multiple scales. (**B**) The same algorithm discovers that the voting patterns of U.S. Supreme Court judges are best explained by a linear "left-right" spectrum. (**C**) Subjective similarities among colors are best explained by a circular ring. (**D**) Given proximities between cities on the globe, the algorithm discovers a cylindrical representation analogous to latitude and longitude: the cross product of a ring and a ring. (**E**) Given images of realistically synthesized faces varying in two dimensions, race and masculinity, the algorithm successfully recovers the underlying two-dimensional grid structure: a cross product of two chains.

if different domains of cognition are represented in qualitatively different ways, those forms must be innate (*43*, *44*); connectionists have suggested these representations may be learned but in a generic system of associative weights that at best only approximates trees, causal networks, and other forms of structure people appear to know explicitly (*14*).

Recently cognitive modelers have begun to answer these challenges by combining the structured knowledge representations described above with state-of-the-art tools from Bayesian statistics. Hierarchical Bayesian models (HBMs) (*45*) address the origins of hypothesis spaces and priors by positing not just a single level of hypotheses to explain the data but multiple levels: hypothesis spaces of hypothesis spaces, with priors on priors. Each level of a HBM generates a probability distribution on variables at the level below. Bayesian inference across all levels allows hypotheses and priors needed for a specific learning task to themselves be learned at larger or longer time scales, at the same time as they constrain lower-level learning. In machine learning and artificial intelligence (AI), HBMs have primarily been used for transfer learning: the acquisition of inductive constraints from experience in previous related tasks (*46*). Transfer learning is critical for humans as well (SOM text and figs. S1 and S2), but here we focus on the role of HBMs in explaining how people acquire the right forms of abstract knowledge.

Kemp and Tenenbaum (*36*, *47*) showed how HBMs defined over graph- and grammar-based representations can discover the form of structure

Downloaded from www.sciencemag.org on March 10, 2011

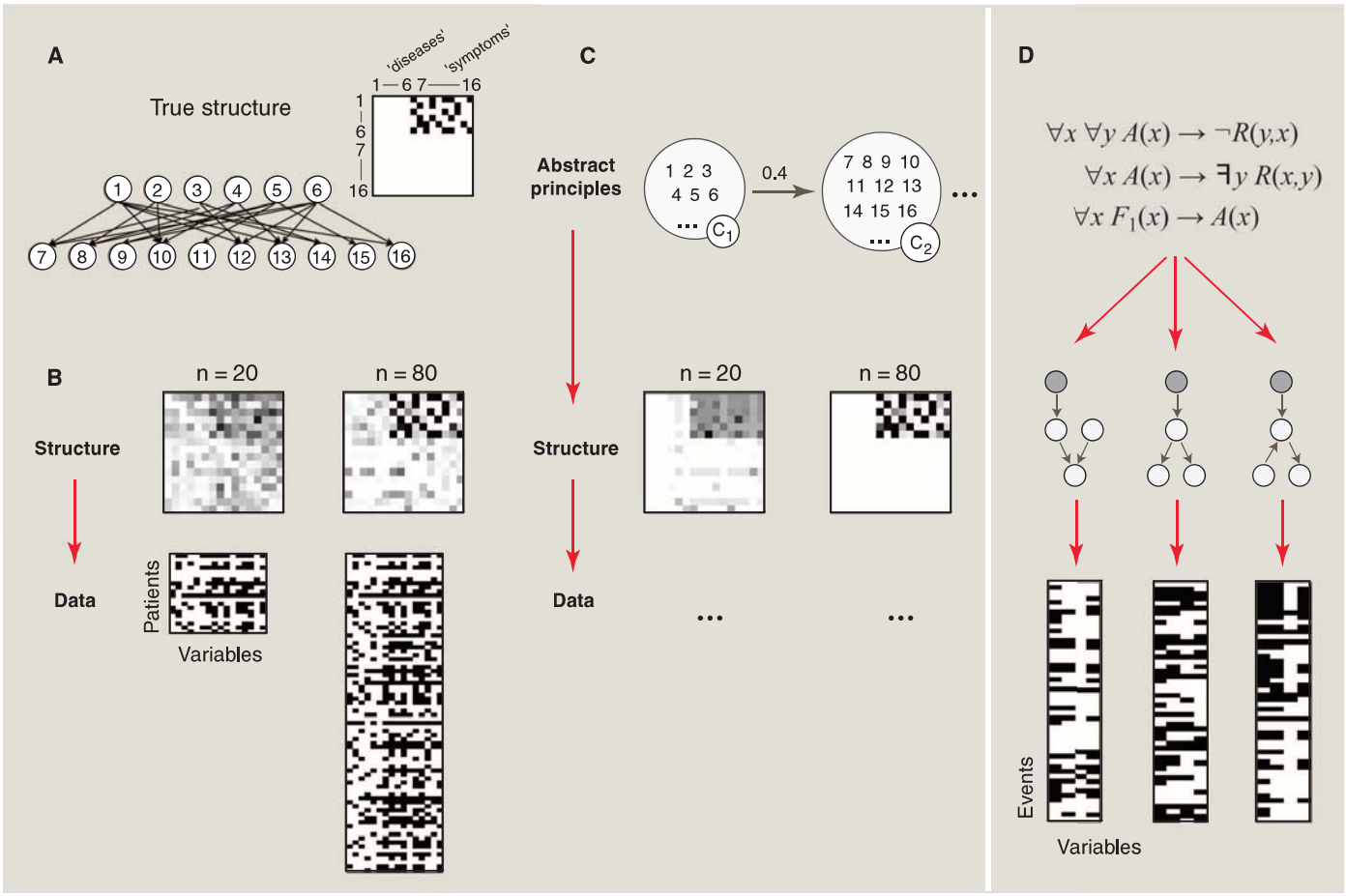

Fig. 3. HBMs defined over graph schemas can explain how intuitive theories are acquired and used to learn about specific causal relations from limited data (*38*). (**A**) A simple medical reasoning domain might be described by relations among 16 variables: The first six encode presence or absence of "diseases" (top row), with causal links to the next 10 "symptoms" (bottom row). This network can also be visualized as a matrix (top right, links shown in black). The causal learning task is to reconstruct this network based on observing data *D* on the states of these 16 variables in a set of patients. (**B**) A two-level HBM formalizes bottom-up causal learning or learning with an uninformative prior on networks. The bottom level is the data matrix *D*. The second level (structure) encodes hypothesized causal networks: a grayscale matrix visualizes the posterior probability that each pairwise causal link exists, conditioned on observing *n* patients; compare this matrix with the black-and-white ground truth matrix shown in (A). The true causal network can be recovered perfectly only from observing very many patients (n = 1000; not shown). With n = 80, spurious links (gray squares) are inferred, and with n = 20 almost none of the true structure is detected. (**C**) A three-level nonparametric HBM (*48*) adds a level of abstract principles, represented by a graph schema. The schema encodes a prior on the level below (causal network structure) that constrains and thereby accelerates causal learning. Both schema and network structure are learned from the same data observed in (B). The schema discovers the disease-symptom framework theory by assigning variables 1 to 6 to class C_1, variables 7 to 16 to class C_2, and a prior favoring only $C_1 \rightarrow C_2$ links. These assignments, along with the effective number of classes (here, two), are inferred automatically via the Bayesian Occam's razor. Although this three-level model has many more degrees of freedom than the model in (B), learning is faster and more accurate. With n = 80 patients, the causal network is identified near perfectly. Even n = 20 patients are sufficient to learn the high-level $C_1 \rightarrow C_2$ schema and thereby to limit uncertainty at the network level to just the question of which diseases cause which symptoms. (**D**) A HBM for learning an abstract theory of causality (*62*). At the highest level are laws expressed in first-order logic representing the abstract properties of causal relationships, the role of exogenous interventions in defining the direction of causality, and features that may mark an event as an exogenous intervention. These laws place constraints on possible directed graphical models at the level below, which in turn are used to explain patterns of observed events over variables. Given observed events from several different causal systems, each encoded in a distinct data matrix, and a hypothesis space of possible laws at the highest level, the model converges quickly on a correct theory of intervention-based causality and uses that theory to constrain inferences about the specific causal networks underlying the different systems at the level below.

governing similarity in a domain. Structures of different forms—trees, clusters, spaces, rings, orders, and so on—can all be represented as graphs, whereas the abstract principles underlying each form are expressed as simple grammatical rules for growing graphs of that form. Embedded in a hierarchical Bayesian framework, this approach can discover the correct forms of structure (the grammars) for many real-world domains, along with the best structure (the graph) of the appropriate form (Fig. 2). In particular, it can infer that a hierarchical organization for the novel objects in Fig. 1A (such as Fig. 1B) better fits the similarities people see in these objects, compared to alternative representations such as a two-dimensional space.

Hierarchical Bayesian models can also be used to learn abstract causal knowledge, such as the framework theory of diseases and symptoms (Fig. 3), and other simple forms of intuitive theories (*38*). Mansinghka *et al.* (*48*) showed how a graph schema representing two classes of variables, diseases and symptoms, and a preference for causal links running from disease to symptom variables can be learned from the same data that support learning causal links between specific diseases and symptoms and be learned just as fast or faster (Fig. 3, B and C). The learned schema in turn dramatically accelerates learning of specific causal relations (the

Downloaded from www.sciencemag.org on March 10, 2011

directed graph structure) at the level below. Getting the big picture first—discovering that diseases cause symptoms before pinning down any specific disease-symptom links—and then using that framework to fill in the gaps of specific knowledge is a distinctively human mode of learning. It figures prominently in children's development and scientific progress but has not previously fit into the landscape of rational or statistical learning models.

Although this HBM imposes strong and valuable constraints on the hypothesis space of causal networks, it is also extremely flexible: It can discover framework theories defined by any number of variable classes and any pattern of pairwise regularities on how variables in these classes tend to be connected. Not even the number of variable classes (two for the disease-symptom theory) need be known in advance. This is enabled by another state-of-the-art Bayesian tool, known as "infinite" or nonparametric hierarchical modeling. These models posit an unbounded amount of structure, but only finitely many degrees of freedom are actively engaged for a given data set (*49*). An automatic Occam's razor embodied in Bayesian inference trades off model complexity and fit to ensure that new structure (in this case, a new class of variables) is introduced only when the data truly require it.

The specific nonparametric distribution on node classes in Fig. 3C is a Chinese restaurant process (CRP), which has been particularly influential in recent machine learning and cognitive modeling. CRP models have given the first principled account of how people form new categories without direct supervision (*50*, *51*): As each stimulus is observed, CRP models (guided by the Bayesian Occam's razor) infer whether that object is best explained by assimilation to an existing category or by positing a previously unseen category (fig. S3). The CrossCat model extends CRPs to carve domains of objects and their properties into different subdomains or "views," subsets of properties that can all be explained by a distinct way of organizing the objects (*52*) (fig. S4). CRPs can be embedded in probabilistic models for language to explain how children discover words in unsegmented speech (*53*), learn morphological rules (*54*), and organize word meanings into hierarchical semantic networks (*55*, *56*) (fig. S5). A related but novel nonparametric construction, the Indian buffet process (IBP), explains how new perceptual features can be constructed during object categorization (*57*, *58*).

More generally, nonparametric hierarchical models address the principal challenge human learners face as knowledge grows over a lifetime: balancing constraint and flexibility, or the need to restrict hypotheses available for generalization at any moment with the capacity to expand one's hypothesis spaces, to learn new ways that the world could work. Placing nonparametric distributions at higher levels of the HBM yields flexible inductive biases for lower levels, whereas the Bayesian Occam's razor ensures the proper balance of constraint and flexibility as knowledge grows.

Across several case studies of learning abstract knowledge—discovering structural forms, causal framework theories, and other inductive constraints acquired through transfer learning—it has been found that abstractions in HBMs can be learned remarkably fast from relatively little data compared with what is needed for learning at lower levels. This is because each degree of freedom at a higher level of the HBM influences and pools evidence from many variables at levels below. We call this property of HBMs "the blessing of abstraction." It offers a top-down route to the origins of knowledge that contrasts sharply with the two classic approaches: nativism (*59*, *60*), in which abstract concepts are assumed to be present from birth, and empiricism or associationism (*14*), in which abstractions are constructed but only approximately, and only slowly in a bottom-up fashion, by layering many experiences on top of each other and filtering out their common elements. Only HBMs thus seem suited to explaining the two most striking features of abstract knowledge in humans: that it can be learned from experience, and that it can be engaged remarkably early in life, serving to constrain more specific learning tasks.

Open Questions

HBMs may answer some questions about the origins of knowledge, but they still leave us wondering: How does it all start? Developmentalists have argued that not everything can be learned, that learning can only get off the ground with some innate stock of abstract concepts such as "agent," "object," and "cause" to provide the basic ontology for carving up experience (*7*, *61*). Surely some aspects of mental representation are innate, but without disputing this Bayesian modelers have recently argued that even the most abstract concepts may in principle be learned. For instance, an abstract concept of causality expressed as logical constraints on the structure of directed graphs can be learned from experience in a HBM that generalizes across the network structures of many specific causal systems (Fig. 3D). Following the "blessing of abstraction," these constraints can be induced from only small samples of each network's behavior and in turn enable more efficient causal learning for new systems (*62*). How this analysis extends to other abstract concepts such as agent or object and whether children actually acquire these concepts in such a manner remain open questions.

Although HBMs have addressed the acquisition of simple forms of abstract knowledge, they have only touched on the hardest subjects of cognitive development: framework theories for core common-sense domains such as intuitive physics, psychology, and biology (*5*, *6*, *7*). First steps have come in explaining developing theories of mind, how children come to understand explicit false beliefs (*63*) and individual differences in preferences (*64*), as well as the origins of essentialist theories in intuitive biology and early beliefs about magnetism in intuitive physics (*39*, *38*). The most daunting challenge is that formalizing the full content of intuitive theories appears to require Turing-complete compositional representations, such as probabilistic first-order logic (*65*, *66*) and probabilistic programming languages (*67*). How to effectively constrain learning with such flexible representations is not at all clear.

Lastly, the project of reverse-engineering the mind must unfold over multiple levels of analysis, only one of which has been our focus here. Marr (*68*) famously argued for analyses that integrate across three levels: The computational level characterizes the problem that a cognitive system solves and the principles by which its solution can be computed from the available inputs in natural environments; the algorithmic level describes the procedures executed to produce this solution and the representations or data structures over which the algorithms operate; and the implementation level specifies how these algorithms and data structures are instantiated in the circuits of a brain or machine. Many early Bayesian models addressed only the computational level, characterizing cognition in purely functional terms as approximately optimal statistical inference in a given environment, without reference to how the computations are carried out (*25*, *39*, *69*). The HBMs of learning and development discussed here target a view between the computational and algorithmic levels: cognition as approximately optimal inference in probabilistic models defined over a learner's subjective and dynamically growing mental representations of the world's structure, rather than some objective and fixed world statistics.

Much ongoing work is devoted to pushing Bayesian models down through the algorithmic and implementation levels. The complexity of exact inference in large-scale models implies that these levels can at best approximate Bayesian computations, just as in any working Bayesian AI system (*9*). The key research questions are as follows: What approximate algorithms does the mind use, how do they relate to engineering approximations in probabilistic AI, and how are they implemented in neural circuits? Much recent work points to Monte Carlo or stochastic sampling–based approximations as a unifying framework for understanding how Bayesian inference may work practically across all these levels, in minds, brains, and machines (*70*–*74*). Monte Carlo inference in richly structured models is possible (*9*, *67*) but very slow; constructing more efficient samplers is a major focus of current work. The biggest remaining obstacle is to understand how structured symbolic knowledge can be represented in neural circuits. Connectionist models sidestep these challenges by denying that brains actually encode such rich knowledge, but this runs counter to the strong consensus in cognitive science and artificial intelligence that symbols and structures are essential for thought. Uncovering their neural

Downloaded from www.sciencemag.org on March 10, 2011

basis is arguably the greatest computational challenge in cognitive neuroscience more generally—our modern mind-body problem.

Conclusions

We have outlined an approach to understanding cognition and its origins in terms of Bayesian inference over richly structured, hierarchical generative models. Although we are far from a complete understanding of how human minds work and develop, the Bayesian approach brings us closer in several ways. First is the promise of a unifying mathematical language for framing cognition as the solution to inductive problems and building principled quantitative models of thought with a minimum of free parameters and ad hoc assumptions. Deeper is a framework for understanding why the mind works the way it does, in terms of rational inference adapted to the structure of real-world environments, and what the mind knows about the world, in terms of abstract schemas and intuitive theories revealed only indirectly through how they constrain generalizations.

Most importantly, the Bayesian approach lets us move beyond classic either-or dichotomies that have long shaped and limited debates in cognitive science: "empiricism versus nativism," "domain-general versus domain-specific," "logic versus probability," "symbols versus statistics." Instead we can ask harder questions of reverse-engineering, with answers potentially rich enough to help us build more humanlike AI systems. How can domain-general mechanisms of learning and representation build domain-specific systems of knowledge? How can structured symbolic knowledge be acquired through statistical learning? The answers emerging suggest new ways to think about the development of a cognitive system. Powerful abstractions can be learned surprisingly quickly, together with or prior to learning the more concrete knowledge they constrain. Structured symbolic representations need not be rigid, static, hard-wired, or brittle. Embedded in a probabilistic framework, they can grow dynamically and robustly in response to the sparse, noisy data of experience.

References and Notes

1. P. Bloom, *How Children Learn the Meanings of Words* (MIT Press, Cambridge, MA, 2000).
2. F. Xu, J. B. Tenenbaum, *Psychol. Rev.* **114**, 245 (2007).
3. S. Pinker, *Words and Rules: The Ingredients of Language* (Basic, New York, 1999).
4. A. Gopnik *et al.*, *Psychol. Rev.* **111**, 3 (2004).
5. A. Gopnik, A. N. Meltzoff, *Words, Thoughts, and Theories* (MIT Press, Cambridge, MA, 1997).
6. S. Carey, *Conceptual Change in Childhood* (MIT Press, Cambridge, MA, 1985).
7. S. Carey, *The Origin of Concepts* (Oxford Univ. Press, New York, 2009).
8. P. Godfrey-Smith, *Theory and Reality* (Univ. of Chicago Press, Chicago, 2003).
9. S. Russell, P. Norvig, *Artificial Intelligence: A Modern Approach* (Prentice Hall, Upper Saddle River, NJ, 2009).
10. D. McAllester, in *Proceedings of the Eleventh Annual Conference on Computational Learning Theory* [Association for Computing Machinery (ACM), New York, 1998], p. 234.
11. J. Pearl, *Probabilistic Reasoning in Intelligent Systems* (Morgan Kaufmann, San Francisco, CA, 1988).
12. J. McClelland, D. Rumelhart, Eds., *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (MIT Press, Cambridge, MA, 1986).
13. S. Pinker, *How the Mind Works* (Norton, New York, 1997).
14. T. Rogers, J. McClelland, *Semantic Cognition: A Parallel Distributed Processing Approach* (MIT Press, Cambridge, MA, 2004).
15. P. Niyogi, *The Computational Nature of Language Learning and Evolution* (MIT Press, Cambridge, MA, 2006).
16. T. L. Griffiths, N. Chater, C. Kemp, A. Perfors, J. B. Tenenbaum, *Trends Cogn. Sci.* **14**, 357 (2010).
17. J. L. McClelland *et al.*, *Trends Cogn. Sci.* **14**, 348 (2010).
18. J. B. Tenenbaum, T. L. Griffiths, *Behav. Brain Sci.* **24**, 629 (2001).
19. J. Tenenbaum, T. Griffiths, in *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, J. D. Moore, K. Stenning, Eds. (Erlbaum, Mahwah, NJ, 2001), pp. 1036–1041.
20. T. Griffiths, J. Tenenbaum, in *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, J. D. Moore, K. Stenning, Eds. (Erlbaum, Mahwah, NJ, 2001), pp. 370–375.
21. T. L. Griffiths, J. B. Tenenbaum, *Cognition* **103**, 180 (2007).
22. H. Lu, A. L. Yuille, M. Liljeholm, P. W. Cheng, K. J. Holyoak, *Psychol. Rev.* **115**, 955 (2008).
23. T. L. Griffiths, J. B. Tenenbaum, *Cognit. Psychol.* **51**, 334 (2005).
24. T. R. Krynski, J. B. Tenenbaum, *J. Exp. Psychol. Gen.* **136**, 430 (2007).
25. M. Oaksford, N. Chater, *Trends Cogn. Sci.* **5**, 349 (2001).
26. T. L. Griffiths, J. B. Tenenbaum, *Psychol. Sci.* **17**, 767 (2006).
27. A. Yuille, D. Kersten, *Trends Cogn. Sci.* **10**, 301 (2006).
28. N. Chater, C. D. Manning, *Trends Cogn. Sci.* **10**, 335 (2006).
29. R. M. Shiffrin, M. Steyvers, *Psychon. Bull. Rev.* **4**, 145 (1997).
30. M. Steyvers, T. L. Griffiths, S. Dennis, *Trends Cogn. Sci.* **10**, 327 (2006).
31. K. P. Körding, D. M. Wolpert, *Nature* **427**, 244 (2004).
32. A. Tversky, D. Kahneman, *Science* **185**, 1124 (1974).
33. E. T. Jaynes, *Probability Theory: The Logic of Science* (Cambridge Univ. Press, Cambridge, 2003).
34. D. J. C. Mackay, *Information Theory, Inference, and Learning Algorithms* (Cambridge Univ. Press, Cambridge, 2003).
35. F. Xu, J. B. Tenenbaum, *Dev. Sci.* **10**, 288 (2007).
36. C. Kemp, J. B. Tenenbaum, *Psychol. Rev.* **116**, 20 (2009).
37. H. M. Wellman, S. A. Gelman, *Annu. Rev. Psychol.* **43**, 337 (1992).
38. C. Kemp, J. B. Tenenbaum, S. Niyogi, T. L. Griffiths, *Cognition* **114**, 165 (2010).
39. T. L. Griffiths, J. B. Tenenbaum, in *Causal Learning: Psychology, Philosophy, and Computation*, A. Gopnik, L. Schulz, Eds. (Oxford University Press, Oxford, 2007), pp. 323–345.
40. J. Woodward, *Making Things Happen: A Theory of Causal Explanation* (Oxford Univ. Press, Oxford, 2003).
41. E. S. Markman, *Categorization and Naming in Children* (MIT Press, Cambridge, MA, 1989).
42. R. N. Shepard, *Science* **210**, 390 (1980).
43. N. Chomsky, *Rules and Representations* (Basil Blackwell, Oxford, 1980).
44. S. Atran, *Behav. Brain Sci.* **21**, 547, (1998).
45. A. Gelman, J. B. Carlin, H. S. Stern, D. B. Rubin, *Bayesian Data Analysis* (Chapman and Hall, New York, 1995).
46. C. Kemp, A. Perfors, J. B. Tenenbaum, *Dev. Sci.* **10**, 307 (2007).
47. C. Kemp, J. B. Tenenbaum, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 10687 (2008).
48. V. K. Mansinghka, C. Kemp, J. B. Tenenbaum, T. L. Griffiths, in *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, R. Dechter, T. Richardson, Eds. (AUAI Press, Arlington, VA, 2006), pp. 324–331.
49. C. Rasmussen, in *Advances in Neural Information Processing Systems* (MIT Press, Cambridge, MA, 2000), vol. 12, pp. 554–560.
50. J. R. Anderson, *Psychol. Rev.* **98**, 409 (1991).
51. T. L. Griffiths, A. N. Sanborn, K. R. Canini, D. J. Navarro, in *The Probabilistic Mind*, N. Chater, M. Oaksford, Eds. (Oxford Univ. Press, Oxford, 2008).
52. P. Shafto, C. Kemp, V. Mansinghka, M. Gordon, J. B. Tenenbaum, in *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (Erlbaum, Mahwah, NJ, 2006), pp. 2146–2151.
53. S. Goldwater, T. L. Griffiths, M. Johnson, *Cognition* **112**, 21 (2009).
54. M. Johnson, T. L. Griffiths, S. Goldwater, in *Advances in Neural Information Processing Systems* (MIT Press, Cambridge, MA, 2007), vol. 19, pp. 641–648.
55. T. L. Griffiths, M. Steyvers, J. B. Tenenbaum, *Psychol. Rev.* **114**, 211 (2007).
56. D. Blei, T. Griffiths, M. Jordan, *J. Assoc. Comput. Mach.* **57**, 1 (2010).
57. T. L. Griffiths, Z. Ghahramani, in *Advances in Neural Information Processing Systems* (MIT Press, Cambridge, MA, 2006), vol. 18, pp. 475–482.
58. J. Austerweil, T. L. Griffiths, in *Advances in Neural Information Processing Systems* (MIT Press, Cambridge, MA, 2009), vol. 21, pp. 97–104.
59. N. Chomsky, *Language and Problems of Knowledge: The Managua Lectures* (MIT Press, Cambridge, MA, 1986).
60. E. S. Spelke, K. Breinlinger, J. Macomber, K. Jacobson, *Psychol. Rev.* **99**, 605 (1992).
61. S. Pinker, *The Stuff of Thought: Language as a Window into Human Nature* (Viking, New York, 2007).
62. N. D. Goodman, T. D. Ullman, J. B. Tenenbaum, *Psychol. Rev.* **118**, 110 (2011).
63. N. Goodman *et al.*, in *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (Erlbaum, Mahwah, NJ, 2006), pp. 1382–1387.
64. C. Lucas, T. Griffiths, F. Xu, C. Fawcett, in *Advances in Neural Information Processing Systems* (MIT Press, Cambridge, MA, 2009), vol. 21, pp. 985–992.
65. B. Milch, B. Marthi, S. Russell, in *ICML 2004 Workshop on Statistical Relational Learning and Its Connections to Other Fields*, T. Dietterich, L. Getoor, K. Murphy, Eds. (Omnipress, Banff, Canada, 2004), pp. 67–73.
66. C. Kemp, N. Goodman, J. Tenenbaum, in *Proceedings of the 30th Annual Meeting of the Cognitive Science Society* (Publisher, City, Country, 2008), pp. 1606–1611.
67. N. Goodman, V. Mansinghka, D. Roy, K. Bonawitz, J. Tenenbaum, in *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence* (AUAI Press, Corvallis, OR, 2008), vol. 22, p. 23.
68. D. Marr, *Vision* (W. H. Freeman, San Francisco, CA, 1982).
69. J. B. Tenenbaum, T. L. Griffiths, in *Advances in Neural Information Processing Systems*, T. Leen, T. Dietterich, V. Tresp, Eds. (MIT Press, Cambridge, MA, 2001), vol. 13, pp. 59–65.
70. A. N. Sanborn, T. L. Griffiths, D. J. Navarro, in *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (Erlbaum, Mahwah, NJ, 2006), pp. 726–731.
71. S. D. Brown, M. Steyvers, *Cognit. Psychol.* **58**, 49 (2009).
72. R. Levy, F. Reali, T. L. Griffiths, in *Advances in Neural Information Processing Systems*, D. Koller, D. Schuurmans, Y. Bengio, L. Bottou, Eds. (MIT Press, Cambridge, MA, 2009), vol. 21, pp. 937–944.
73. J. Fiser, P. Berkes, G. Orbán, M. Lengyel, *Trends Cogn. Sci.* **14**, 119 (2010).
74. E. Vul, N. D. Goodman, T. L. Griffiths, J. B. Tenenbaum, in *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (Erlbaum, Mahwah, NJ, 2009), pp. 148–153.
75. L. Schmidt, thesis, Massachusetts Institute of Technology, Cambridge, MA (2009).

76. We gratefully acknowledge the suggestions of R. R. Saxe, M. Bernstein, and J. M. Tenenbaum on this manuscript and the collaboration of N. Chater and A. Yuille on a forthcoming joint book expanding on the methods and perspectives reviewed here. Grant support was provided by Air Force Office of Scientific Research, Office of Naval Research, Army Research Office, NSF, Defense Advanced Research Projects Agency, Nippon Telephone and Telegraph Communication Sciences Laboratories, Qualcomm, Google, Schlumberger, and the James S. McDonnell Foundation.

Supporting Online Material

www.sciencemag.org/cgi/content/full/331/6022/1279/DC1
SOM Text
Figs. S1 to S5
References

10.1126/science.1192788

Downloaded from www.sciencemag.org on March 10, 2011