Opinion

Approaches to cognitive modeling

Probabilistic models of cognition: exploring representations and inductive biases

Thomas L. Griffiths[1], Nick Chater[2], Charles Kemp[3], Amy Perfors[4] and Joshua B. Tenenbaum[5]

[1] Department of Psychology, University of California, Berkeley, 3210 Tolman Hall MC 1650, Berkeley CA 94720-1650, USA
[2] Division of Psychology and Language Sciences, University College London, Gower Street, London WC1E 6BT, UK
[3] Department of Psychology, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh PA 15213, USA
[4] School of Psychology, University of Adelaide, Level 4, Hughes Building, Adelaide, SA 5005, Australia
[5] Brain and Cognitive Sciences Department, Massachusetts Institute of Technology, Building 46-4015, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

Cognitive science aims to reverse-engineer the mind, and many of the engineering challenges the mind faces involve induction. The probabilistic approach to modeling cognition begins by identifying ideal solutions to these inductive problems. Mental processes are then modeled using algorithms for approximating these solutions, and neural processes are viewed as mechanisms for implementing these algorithms, with the result being a top-down analysis of cognition starting with the function of cognitive processes. Typical connectionist models, by contrast, follow a bottom-up approach, beginning with a characterization of neural mechanisms and exploring what macro-level functional phenomena might emerge. We argue that the top-down approach yields greater flexibility for exploring the representations and inductive biases that underlie human cognition.

Strategies for studying the mind

Most approaches to modeling human cognition agree that the mind can be studied on multiple levels. David Marr [1] defined three such levels: a 'computational' level characterizing the problem faced by the mind and how it can be solved in functional terms; an 'algorithmic' level describing the processes that the mind executes to produce this solution; and a 'hardware' level specifying how those processes are instantiated in the brain. Cognitive scientists disagree over whether explanations at all levels are useful, and on the order in which levels should be explored. Many connectionists advocate a bottom-up or 'mechanism-first' strategy (see Glossary), starting by exploring the problems that neural processes can solve. This often goes with a philosophy of 'emergentism' or 'eliminativism': higher-level explanations do not have independent validity but are at best approximations to the mechanistic truth; they describe emergent phenomena produced by lower-level mechanisms. By contrast, probabilistic models of cognition pursue a top-down or 'function-first' strategy, beginning with abstract principles that allow agents to solve problems posed by the world – the functions that minds perform – and then attempting to reduce these principles to psychological and neural processes. Understanding the lower levels does not eliminate the need for higher-level models, because the lower levels implement the functions specified at higher levels.

Glossary

Backpropagation: a gradient-descent based algorithm for estimating the weights in a multilayer perceptron, in which each weight is adjusted based on its contribution to the errors produced by the network.
Bottom-up/mechanism-first explanation: a form of explanation that starts by identifying neural or psychological mechanisms believed to be responsible for cognition, and then tries to explain behavior in those terms.
Emergentism: a scientific approach in which complex behavior is viewed as emerging from the interaction of simple elements.
Gradient-descent learning: learning algorithms based on minimizing the error of a system (or maximizing the likelihood of the observed data) by modifying the parameters of the system based on the derivative of the error.
Hypothesis space: the set of hypotheses assumed by a learner, as made explicit in Bayesian inference and potentially implicit in other learning algorithms.
Inductive biases: factors that lead a learner to favor one hypothesis over another that are independent of the observed data. When two hypotheses fit the data equally well, inductive biases are the only basis for deciding between them. In a Bayesian model, these inductive biases are expressed through the prior distribution over hypotheses.
Inductive problem: a problem in which the observed data are not sufficient to unambiguously identify the process that generated them. Inductive reasoning requires going beyond the data to evaluate different hypotheses about the generating process, while maintaining uncertainty.
Likelihood: the component of Bayes' rule that reflects the probability of the data given a hypothesis, $p(d|h)$. Intuitively, the likelihood expresses the extent to which the hypothesis fits the data.
Posterior distribution: a probability distribution over hypotheses reflecting the learner's degree of belief in each hypothesis in light of the information provided by the observed data. This is the outcome of applying Bayes' rule, $p(h|d)$.
Prior distribution: a probability distribution over hypotheses reflecting the learner's degree of belief in each hypothesis before observing data, $p(h)$. The prior captures the inductive biases of the learner, because it is a factor that contributes to the extent to which learners believe in hypotheses that is independent of the observed data.
Top-down/function-first explanation: a form of explanation that starts by considering the function that a particular aspect of cognition serves, explaining behavior in terms of performing that function.

Corresponding author: Griffiths, T.L. (tomgriffiths@berkeley.edu).

1364-6613/$ – see front matter © 2010 Elsevier Ltd. All rights reserved. doi:10.1016/j.tics.2010.05.004

Explanations at a functional level have a long history in cognitive science. Virtually all attempts to engineer human-like artificial intelligence, from the Logic Theory Machine [2] to the most successful contemporary paradigms [3], have started with computational principles rather than hardware mechanisms. The great potential of probabilistic models of cognition comes from the solutions they identify to inductive problems, which play a central role in cognitive science: Most of cognition, including acquiring a language, a concept, or a causal model, requires uncertain conjecture from partial or noisy information. A probabilistic framework lets us address key questions about these phenomena. How much information is needed? What representations subserve the inferences people make? What constraints on learning are necessary? These are computational-level questions and they are most naturally answered by computational-level theories.

Taking a top-down approach leads probabilistic models of cognition to explore a broad range of different assumptions about how people might solve inductive problems, and what representations might be involved. Representations and inductive biases are selected by considering what is needed to account for the functions the brain performs, assuming only that those functions of perception, learning, reasoning, and decision can be described as forms of probabilistic inference (Figure 1). By contrast, connectionism makes strong pre-commitments about the nature of people's representations and inductive biases based on a certain view of neural mechanisms and development: representations are graded, continuous vector spaces, lacking explicit structure, and are shaped almost exclusively by experience through gradual error-driven learning algorithms. This approach rejects a long tradition of research into knowledge representation in cognitive science, discarding notions such as rules, grammars, and logic that have proven useful in accounting for the functions of higher-level cognition.

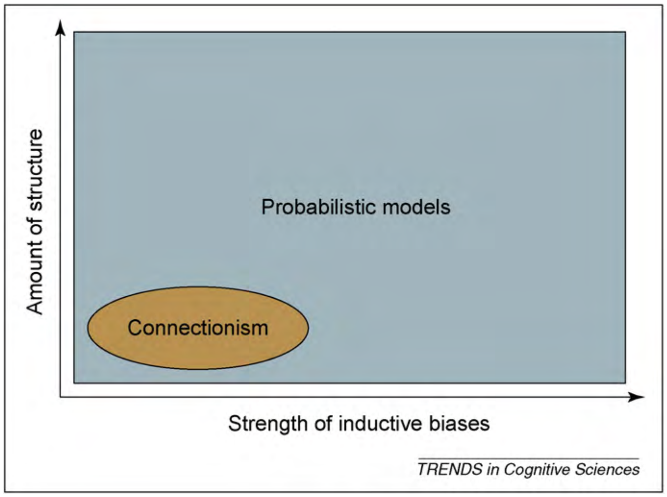

Figure 1. Theoretical commitments of connectionism and probabilistic models of cognition. Based on a certain view of brain architecture and function, connectionist models makes strong assumptions about the representations and inductive biases to be used in explaining human cognition: representations lack explicit structure and inductive biases are very weak. By contrast, probabilistic models explore a larger space of possibilities, including representations of diverse forms and degrees of structure, and inductive biases of greatly varying shapes and strength. These possibilities include highly structured representations and inductive constraints that have proven valuable – and arguably necessary – for explaining many of the functions of human cognition.

The rest of this article presents our argument for the top-down approach, focusing on the importance of representational diversity. The next section describes how structured representations of different forms can be combined with statistical learning and inference in probabilistic models of cognition, using a case study in semantic cognition that has also been the focus of recent work in the connectionist tradition [4]. We then give a broader survey, across different domains and tasks, of how probabilistic models have exploited a range of representations and inductive biases to explain different aspects of cognition that pose a challenge to accounts restricted to the limited forms of representations and weaker inductive biases assumed by connectionism. We emphasize breadth over depth of coverage because our goal is to illustrate the greater explanatory scope of probabilistic models. We then discuss how probabilistic models of cognition should be interpreted in terms of lower levels of analysis, a common point of confusion in critiques of this approach, and close with several other considerations in choosing whether to pursue a top-down, 'function-first' or bottom-up, 'mechanism-first' approach to cognitive modeling.

Knowledge representation and probabilistic models

A probabilistic model starts with a formal characterization of an inductive problem, specifying the hypotheses under consideration, the relation between these hypotheses and observable data, and the prior probability of each hypothesis (Box 1). Probabilistic models therefore provide a transparent account of the assumptions that allow a problem to be solved and make it easy to explore the consequences of different assumptions. Hypotheses can take any form, from weights in a neural network [5,6] to structured symbolic representations, as long as they specify a probability distribution over observable data. Likewise, different inductive biases can be captured by assuming different prior distributions over hypotheses. The approach makes no *a priori* commitment to any class of representations or inductive biases, but provides a framework for evaluating different proposals.

Box 1. Probabilistic inference

Probability theory provides a solution to the problem of induction, indicating how a learner should revise her degrees of belief in a set of hypotheses in light of the information provided by observed data. This solution is encapsulated in Bayes' rule: if a learner considers a set of hypotheses H that might explain observed data d, and assigns each hypothesis $h \in H$ a probability $p(h)$ before observing d (known as the 'prior' probability), then Bayes' rule indicates that the probability $p(h|d)$ assigned to h after seeing d (known as the 'posterior' probability) should be

$$p(h|d) = \frac{p(d|h)\,p(h)}{\sum_{h \in H} p(d|h)\,p(h)} \qquad (1)$$

where $p(d|h)$ is the 'likelihood', indicating the probability of observing d if h were true, and the sum in the denominator simply ensures that the posterior probabilities sum to one. Bayes' rule thus indicates that the conclusions reached by the learner will be determined by how well hypotheses cohere with prior knowledge, and how well they explain the data.

Figure 2 illustrates one way in which a probabilistic approach can illuminate the nature of mental representations. Consider a property induction problem where participants learn that horses, cows, and dolphins have a certain property then must decide whether all mammals are likely to have this property. Some researchers have proposed that inferences about novel properties of animals are supported by tree-structured representations [7], but others suggest that the underlying mental representations are closer to continuous spaces [8]. One way to resolve this debate is to define a probabilistic framework that can use either type of representation, and to see which representation best explains human inferences [9]. The results in Figure 2a suggest that a tree structure is the better of these two alternatives.

Connectionist models typically focus on a single form of knowledge – whatever can be encoded in distributed codes over layers of hidden units. Unlike the connectionist approach, the probabilistic approach is open to the idea that qualitatively different representations are used for different types of inferences. Figure 2b shows results from a property induction experiment where the items are cities and participants are told, for example, that a certain type of Native American artifact is found near Houston, Durham, and Orlando, and then asked whether this artifact is likely to be found near all major American cities. The probabilistic framework that was previously applied to the animal data (Figure 2a) now suggests that inferences about spatial relations between cities are better captured by a low-dimensional space than a tree. The same probabilistic framework also suggests how people might learn qualitatively different representations for different domains [9] (Figure 2c).

Rogers and McClelland have argued that connectionist models can implicitly capture representations like hierarchically-structured taxonomies, but some types of inferences seem to rely on explicit representations. For example, explicit representations provide a natural way to incorporate high-level semantic information provided by natural language and informed by social reasoning. To a

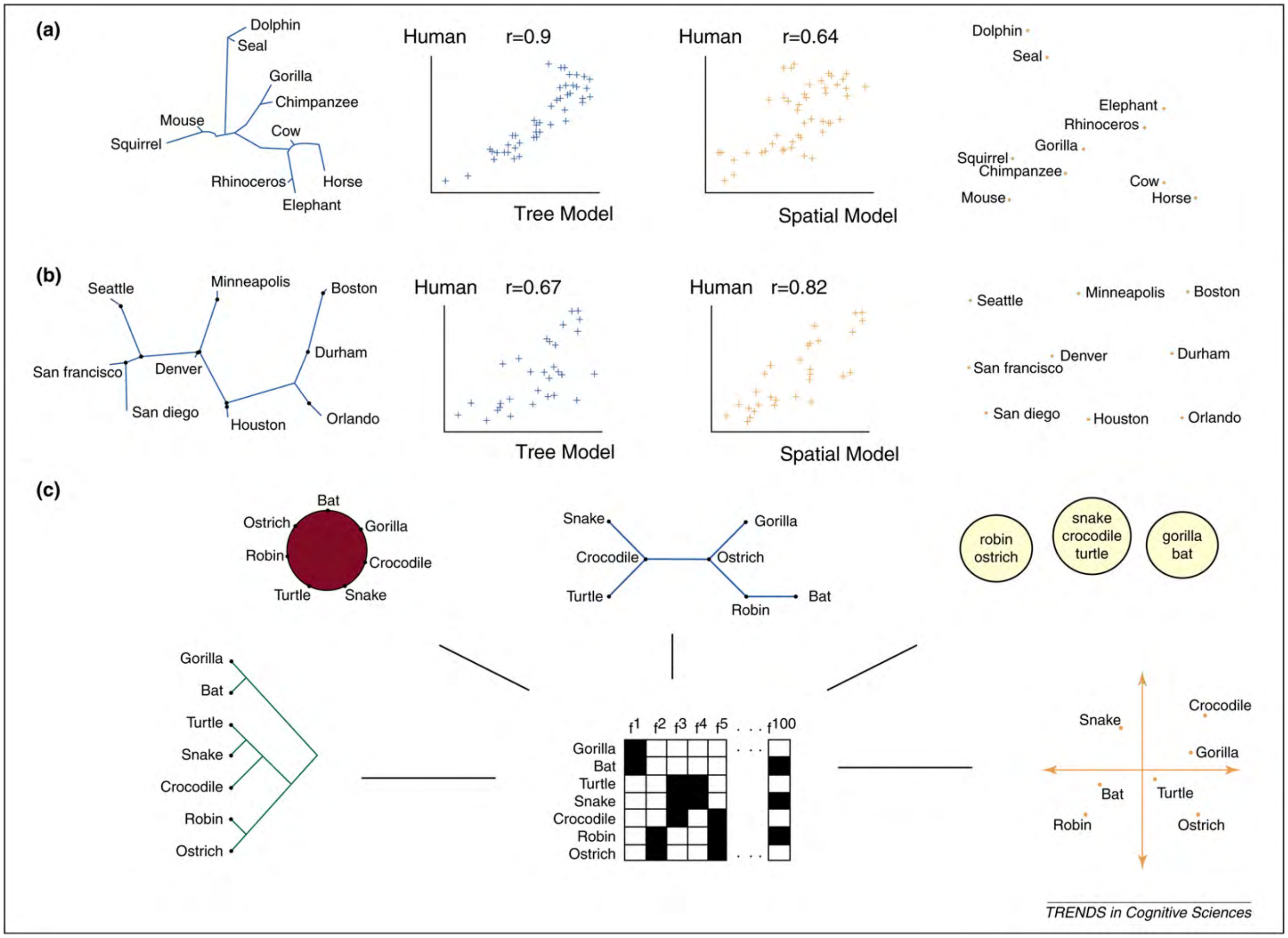

Figure 2. Qualitatively different representations are needed to account for inductive inferences about different domains (adapted from [13]). **(a)** Model predictions and human responses for a property induction task where participants learn that several animals have a property, then decide whether all animals are likely to have this property. Each point in each scatterplot corresponds to a trio of mammals, and the vertical axis indicates how strongly humans believe that all mammals have a certain property after learning that the animals in this trio have the property. The horizontal axis shows the predictions of probabilistic models which assume that nearby animals in a tree tend to have similar properties, or that nearby animals in a two dimensional space tend to have similar properties. The tree model relies on the tree structure shown and the spatial model relies on the two-dimensional space shown. **(b)** Results for a task where participants make inferences about US cities rather than animal species. The spatial model now performs better than the tree model. **(c)** Relations between biological species could be represented using a tree, a ring, a set of clusters, or a low-dimensional space, but a probabilistic model can discover that a tree best accounts for the observable features of these species.

child who believes that dolphins are fish, hearing a simple message from a knowledgeable adult (‘dolphins might look like fish but are actually mammals’) might drastically modify the inferences she makes. A learner equipped with a hierarchically structured system of categories can rearrange the hierarchy on hearing such an utterance. By contrast, a connectionist model cannot easily reconfigure itself through linguistic input. More generally, whereas both types of approaches might learn well from observing the world, only structured probabilistic approaches offer a natural route to acquiring knowledge through instruction or other forms of social communication.

Although we have focused so far on simple representations such as trees and low-dimensional spaces, many other types of representations are possible and useful. Probabilistic models defined over causal graphs, phrase structure grammars, logical rules or theories have been proposed for language, vision, and many other areas of cognition (see Figure 3, and the following section). These models inherit classic advantages of structured representations that connectionist models give up [10,11]: they generate infinite hypothesis spaces by combinatorial operations on basic elements and capture core properties of human symbolic thought, such as compositionality and recursion. Connectionists have criticized symbolic models for failing to handle exceptions or produce graded generalizations, or to account for how representations are learned [4]. Combining structured representations with probabilistic inference meets those challenges, and also explain the rich and sophisticated uses of knowledge in human cognition that appear to require symbolic forms of representation.

The advantages of representational pluralism

With their ability to operate over a broad range of candidate representations and inductive biases, probabilistic models provide a unifying framework for explaining the inferences that people make in different settings. Here we briefly summarize how probabilistic approaches have addressed several aspects of human inductive reasoning and learning that have not previously been well explained in computational terms, and in particular, that would be difficult to explain in a connectionist framework.

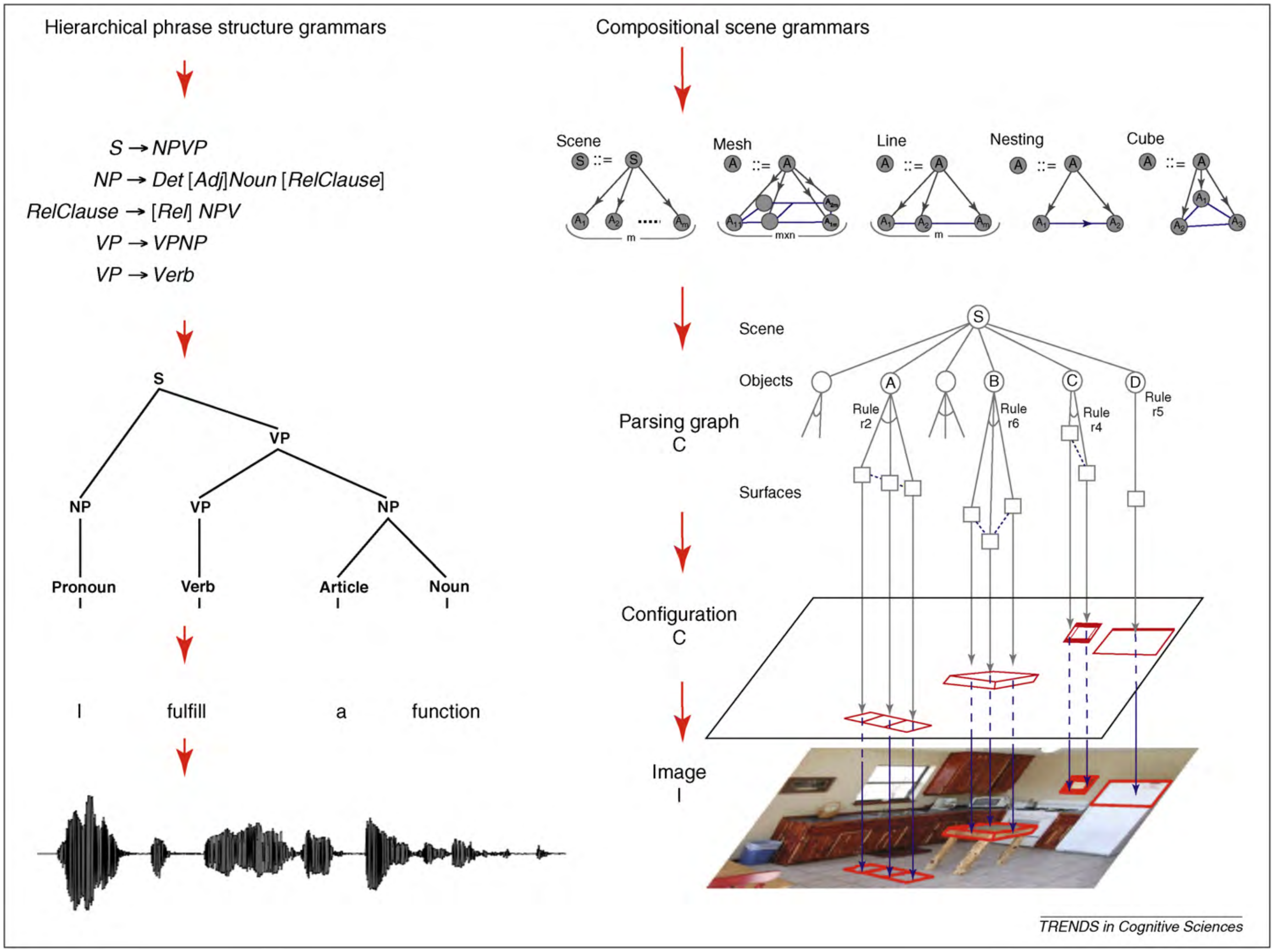

Figure 3. Structured statistical models provide a way to describe multiple levels of abstraction in a way that applies across different domains. In language, a learner needs to be able to discover how sounds are organized into words, how words are organized into sentences, and how a language is characterized by a grammar. Learning at each of these levels can be described in terms of probabilistic inference over a structured hypothesis space [36]. Analogous problems apply in vision, where grammars can be used to describe the set of objects in a scene and the surfaces that comprise those objects (figure adapted from [38]).

Rapid and flexible generalization

Human learners routinely draw successful generalizations from very limited evidence. Even young children can infer the extensions of new words or concepts, the hidden properties of objects, or the existence of causal relations from a handful of relevant observations. These abilities outstrip those of conventional machine learning algorithms, but probabilistic models have shown how rapid word learning [12], property induction [13], and causal learning [14] can be explained as Bayesian inferences. Probabilistic models have explained why people might appear to generalize differently in different contexts as a consequence of applying the same rules of optimal statistical inference over different priors [15] or knowledge representations [13] (Figure 2), and why some phenomena, such as Shepard's universal exponential law [16], might arise in an entirely representation-independent way [17]. Algorithmic-level models of generalization often posit different processes – rules to account for all-or-none generalizations, exemplar similarity to account for more graded generalizations – but probabilistic computational theories [18,19] have explained why we have these particular processes, why they work as they do, and why people use a rule-like process in some cases and a similarity process in others.

Probabilistic models have also made successful empirical predictions about novel factors that can influence children's generalizations, such as the sampling processes generating the data learners observe. Preschoolers and even infants are sensitive to whether objects exemplifying a new word or hidden property are drawn specifically from the set of positive examples ('strong sampling'), or instead from some more general or accidental process ('weak sampling'), and generalize more sharply in the former case [20,21]. Probabilistic models naturally explain these findings, giving sampling processes a central role in the statistical problem of generalization through the likelihood term of Bayes' rule [12,19]. By contrast, informative sampling was not considered in previous algorithmic models and is not easily accommodated within standard connectionist models of statistical learning.

Causal learning

Discovering the causal relations between objects and events in the environment is a basic problem of human learning. Computational-level analyses of causal learning have provided two types of insights. First, they introduce the distinction between structure and strength [22]. When scientists explore causal relations, they distinguish between questions of whether a relation exists (determining causal structure), and how strong that relation is. This distinction is blurred in associative accounts of causal learning, but is explicit when causal learning is framed as Bayesian inference over causal graphical models [23,24]. Probabilistic models based on this approach have given compelling quantitative accounts of human causal judgments [22,25–27]. Second, probabilistic inference provides a way to understand how prior knowledge is combined with statistical evidence in causal learning, characterizing the different types of constraints that prior knowledge can impose [14] and explaining how these constraints themselves could be learned [28,29].

Learning language

Children appear to be able to learn what utterances are, and are not, allowed in their native language, to some approximation, from exposure only to positive examples of the language. Learning merely from positive instances of a category has often been viewed as fundamentally problematic, sometimes leading to strong nativist conclusions. The probabilistic approach provides powerful tools, both theoretical [30] and computational [31], for exploring how much learning is possible with minimal language-specific innate biases. More broadly, because linguistic representations can be highly structured, probabilistic models provide the means to analyze what can be learned given what sort of input, and can even be used to evaluate what sorts of structures (e.g. what type of grammar or phrase structure representation) provide the best model of the data. Because all probabilistic models are couched in the common language of probability theory, they also provide a natural way to combine different sources of data (e.g. social cues and co-occurrence relations when learning the meaning of words [32]). Probabilistic models have already been applied to many problems in language development, from the acquisition of syntax [31,33,34] to word segmentation [35] to learning meanings in communicative contexts [32]. On the engineering side of natural language processing, where the same ability to learn with hierarchical, compositional or recursive representations of meaning is crucial, structured statistical models have come to dominate and reshaped the state of the art [36].

Visual perception

Probabilistic models have also revolutionized computational theories of visual perception. Models for low-level vision such as motion estimation or shape perception operate over high-dimensional continuous representations: vector fields representing motion components or depth gradients [37]. Models for higher-level visual tasks often resemble probabilistic parsing in natural language: they operate over hierarchically structured representations of objects and parts, assumed to be generated by a probabilistic grammar for natural scenes [38,39] (Figure 3).

Learning to learn

Children learn their first words slowly, but in building their initial vocabulary they also quickly acquire the ability to learn new words much more rapidly [40]. Hierarchical Bayesian models have been used to explain how humans 'learn to learn' words [41], as well as categories [42] and causal relations [28,43], by performing inference on multiple levels of abstraction. Connectionist models have explored similar phenomena [44] but have not explained how children can learn to learn so quickly, constructing abstract knowledge of the appropriate form from relatively little experience in a domain [9,43].

The psychological and neural interpretation of probabilistic models

Probabilistic models explain human learning and inductive reasoning in terms of Bayesian inference, and specify hypothesis spaces that often have symbolic structure. Critics of probabilistic models often argue that they are implausible as accounts of human cognition, pointing to the computational difficulty involved in calculating Bayesian inference as well as the requirement of specifying the hypothesis space in advance (see, e.g. the commentaries to [19]). However, all models – including connectionist models – build in hypothesis spaces; probabilistic models simply make the space explicit. Moreover, this criticism presupposes that a computational-level analysis in terms of Bayesian inference requires the algorithmic- or hardware-level analysis to take the same form. This assumption is false: using probabilistic models to provide a computational-level explanation does not require that hypothesis spaces or probability distributions be explicitly represented by the underlying psychological or neural processes, or that people learn and reason by explicitly using Bayes' rule.

To illustrate how the computational-level specification of a model can differ significantly from its realization at the algorithmic and hardware levels, it is useful to apply this approach to one of the best-known connectionist models, the multilayer perceptron. A multilayer perceptron can be characterized at the computational level as a nonlinear function approximator. Its weights parameterize an infinite, high-dimensional hypothesis space of nonlinear functions mapping input vectors to outputs. Learning involves searching this hypothesis space for a function that minimizes error on a training dataset. This can actually be cast in Bayesian terms: the error corresponds to the negative log likelihood of a hypothesis and the prior is either uniform or prefers smaller weights [5,6].

Described as Bayesian inference in an infinite, high-dimensional hypothesis space, learning the weights of a multilayer perceptron might sound implausible as a cognitive process. However, considering ways to solve this computational problem approximately but tractably suggests more plausible psychological and even neural interpretations. We can find at least a local maximum of the Bayesian posterior by computing its gradient in weight space and adjusting the weights iteratively along this gradient. Familiar gradient-descent learning algorithms such as backpropagation implement this strategy in a parallel network of neuron-like units, each computing only local functions of the activation and error signals of neighboring units. This algorithm does not require explicit enumeration or scoring of the full space of hypotheses, nor even any explicit application of Bayes' rule.

Similarly, we view the structured representations and Bayesian calculations used in probabilistic models of cognition as computational abstractions that could be implemented in the mind and brain in a variety of implicit and approximate ways. Such implementation could differ across problems, and need not look like explicit structured representations or Bayesian inference. Work on connecting probabilistic models to psychological process models (Box 2) and neural computation (Box 3) illustrates this point, and indicates a possible route towards synthesis with more bottom-up, mechanistically constrained approaches to modeling the mind (Box 4).

Box 2. Connecting to process models

The discussion in the main text shows how connections between the computational, algorithmic, and hardware levels might not be transparent. However, exploring these connections is an important part of the strategy of working through levels of analysis from the top down. One way to do so is to consider psychological processes that could approximate the computations required for probabilistic inference. Applications of statistical models in machine learning and artificial intelligence rely on such approximation algorithms, because computing exact probabilities is typically intractable for complex, real-world problems. These algorithms provide rational approximations to probabilistic inference, and thus are a potential source of 'rational process models' [45].

One class of approximation algorithms is Monte Carlo methods, in which a probability distribution is approximated with a set of samples from that distribution. One sophisticated Monte Carlo method, importance sampling, can be implemented using the same computations as the exemplar models used as process models of categorization [46,47], requiring people to store a few hypotheses in memory and activate them based on their similarity to observed data [45]. A related set of algorithms known as 'particle filters' provide a way to approximately update a probability distribution as data are observed. They have been used to model deviations from ideal performance in category learning [48], associative learning [49], detecting changes in temporal sequences [50], and sentence processing [51], and could provide a way to connect all the way to the neural level (Box 3).

Conclusion: start at the top, or at the bottom?

Top-down and bottom-up approaches to traversing levels of analysis are analogous to building a single bridge from different ends. Nonetheless, we expect that more rapid

Box 3. Probabilistic models and neural computation

Probabilistic models of cognition rarely emphasize inspiration from neuroscience, or appeal to neural plausibility. Increasingly, however, the link between probabilistic inference and neural function is drawing the attention of modelers from diverse backgrounds.

One route for linking Bayesian cognitive models to the brain uses connectionism as a mediating paradigm: many familiar connectionist algorithms for learning and inference have natural Bayesian interpretations [5,6,52], and to the extent that these algorithms are neurally plausible, they suggest how certain types of probabilistic inferences could be implemented in the brain. Several connectionist researchers have emphasized explicitly probabilistic formulations for learning and inference, while still attempting to preserve the distinctive 'connectionist style' of distributed representations arranged in hierarchical layers [53].

Another group of researchers aim to show how core computations and models from Bayesian statistics and machine learning – many of which are also central in probabilistic models of cognition – can be implemented in neurally plausible mechanisms. For instance, Pouget, Beck, Ma and colleagues have studied how to implement Bayesian parameter estimation and decision-making using probabilistic population codes in networks of spiking neurons [54]. Lee and Mumford [55] suggested that cortical hierarchies could implement a form of particle filtering, which is also a candidate for making algorithmic-level models (see main text).

Although research on the 'Bayesian brain' holds great promise, there is presently a gulf between such a research program and the Bayesian models of higher-level cognition reviewed in this article. We have argued that probabilistic inference over structured representations is crucial for explaining the use and origins of human concepts, language, or intuitive theories. Yet little is known concerning how these structured representations can be implemented in neural systems (however, see the research program of Smolensky and colleagues [56]). In our view, the single biggest challenge for theoretical neuroscience is not to understand how the brain implements probabilistic inference, but how it represents the structured knowledge over which such inference is defined.

Box 4. Outstanding questions

- What are the connections between probabilistic models at the computational level, and the psychological and neural processes involved in cognition?
- How (and to what extent) might human behavior be understood as an approximation to the 'ideal observer' behavior predicted by the probabilistic approach? To what extent can approximations built into probabilistic models implementing human-like cognitive limitations account for divergences between human and model performance?
- How might probabilistic inference and structured representations be implemented in neural hardware?
- What questions about human cognition are more naturally framed at levels lower than the computational level? Are there any phenomena for which no computation-level explanation is possible?

progress will come from attempts to reduce abstract probabilistic analyses of cognition to psychological and neural mechanisms, rather than studies of how analogous computational functions might emerge from connectionist networks (Box 5). The flexibility to explore different assumptions about representation and inductive biases, and to naturally capture inferences over rich and structured forms of knowledge, are central advantages of the top-down approach. However, there are two other important differences between these approaches.

First, the top-down strategy fits particularly well with understanding solutions to the computational problems that the mind faces. Finding engineering solutions to these problems is the type of process that typically operates top-down, from high-level specification to physical implementation. A probabilistic approach to reverse-engineering the mind forges strong connections with the latest ideas from computer science, machine learning, and statistics. Bottom-up accounts can be harder to interpret: we might simulate a complex system and find that its emergent behavior solves a cognitive problem, but that does not mean we will necessarily know how or why it solves it successfully.

Second, bottom-up accounts could be highly sensitive to details of the underlying mechanisms, and these details are either unknown or abstracted away in most current models. For instance, small differences in how neurons process information, adjust their weights, or connect with other neurons could lead to very different emergent behavior in a large neural network. These possibilities are particularly problematic given the rapidly evolving state of neuroscience research and the increasingly unclear relation between connectionist networks and biological neural circuits. Committing to a set of assumptions about the representations and inductive biases involved in human cognition thus seems premature.

Whereas the phenomena of human cognition must ultimately be analyzed at all of Marr's levels, we are far from understanding how rich knowledge structures can be implemented in neural circuits. Whether such implementations will ultimately resemble conventional connectionist models is an open question. However, when a neural-level understanding of human knowledge and its origins is eventually achieved, we predict that it will build on a deep understanding of these questions at the computational level – and that this understanding will be best framed using the concepts and principles of probabilistic inference.

Box 5. Response to McClelland *et al.*

We enjoyed reading McClelland *et al.*'s article, and hope that this comparison of probabilistic and connectionist models will continue to result in productive interactions. In this spirit, we first clarify several respects in which our position differs from its characterization in their article. Rather than treating people as optimal cognizing machines, we believe that considering optimal solutions to the computational problems people face can provide insight into human behavior. People undoubtedly approximate these optimal solutions in a variety of ways. We do not ignore mechanism and implementation, but view computational level analyses as a guide to understanding the function that those mechanisms and implementations fulfill. We do not see the fact that different mechanisms are involved when task demands differ as being inconsistent with this approach. Finally, we do not make an *a priori* commitment to particular types of representations. Probabilistic models are a tool for exploring different sets of assumptions about representations and inductive biases, making it possible for data to lead us to an account of human cognition.

McClelland *et al.* warn of the dangers of mis-specifying the computational problems that people are solving, and point to Chomskyan linguistics as an illustration. It is certainly true that any specific high-level explanation, whether probabilistic or not, can be questioned. However, McClelland *et al.* go much further than this, arguing that most if not all aspects of cognition should be explained without the 'often misleading' constructs of high-level explanation. This radical position opposes not just the probabilistic approach, but standard practice across the cognitive sciences. By contrast, we believe that greater danger lies in committing to particular incorrect low-level mechanisms – a real possibility because most connectionist networks are vastly oversimplified when compared with actual neurons. Connectionist networks are opaque, and it is typically difficult to understand what shapes their behavior and what constraints they might be implementing. This makes it hard to understand the consequences of changing the underlying mechanisms. By contrast, the transparency of probabilistic models makes it easier to understand the consequences of changing our assumptions, and thus to recover from errors of mis-specification.

Chomskyan linguistics is an interesting choice as an illustration of the dangers of the probabilistic approach. Whereas many of the ideas introduced by Chomsky have subsequently been revised or rejected, the notion of a generative grammar, once paired with the statistical principles that underlie probabilistic models, has been the basis for considerable advances in both computational linguistics and psycholinguistics (e.g. [33,34,36,51]). This marriage of structure and statistics is also at the heart of almost all modern machine learning algorithms, having become the method of choice for solving the types of real-world learning problems that people face every day. Although cognitive modeling and machine learning are two different enterprises, a basic challenge for both is to match human-level performance in domains such as language, vision, and reasoning. Of the modeling approaches that psychologists have considered, the structured statistical approach comes closest to meeting this challenge.

References

1 Marr, D. (1982) *Vision*, W.H. Freeman

2 Newell, A. and Simon, H. (1956) The logic theory machine: a complex information processing system. *IRE Trans. Information Theory* IT-2, 61–79

3 Russell, S.J. and Norvig, P. (2002) *Artificial Intelligence: a Modern Approach*, (2nd edn), Prentice Hall

4 Rogers, T. and McClelland, J. (2004) *Semantic Cognition: a Parallel Distributed Processing Approach*, MIT Press

5 Neal, R.M. (1996) *Bayesian Learning for Neural Networks (Number 118 in Lecture Notes in Statistics)*, Springer-Verlag

6 MacKay, D.J.C. (1995) Probable networks and plausible predictions – a review of practical Bayesian methods for supervised neural networks. *Network: Comput. Neural. Systems* 6, 469–505

7 Atran, S. (1998) Folk biology and the anthropology of science: cognitive universals and cultural particulars. *Behav. Brain Sci.* 21, 547–609

8 Rips, L.J. (1975) Inductive judgments about natural categories. *J. Verbal Learning Verbal Behav.* 14, 665–681

9 Kemp, C. and Tenenbaum, J.B. (2008) The discovery of structural form. *Proc. Natl. Acad. Sci. U. S. A.* 105, 10687–10692

10 Fodor, J.A. and Pylyshyn, Z.W. (1988) Connectionism and cognitive architecture: a critical analysis. *Cognition* 28, 3–71

11 Marcus, G.F. (2001) *The Algebraic Mind: Integrating Connectionism and Cognitive Science*, MIT Press

12 Xu, F. and Tenenbaum, J.B. (2007) Word learning as Bayesian inference. *Psychol. Rev.* 114, 245–272

13 Kemp, C. and Tenenbaum, J.B. (2009) Structured statistical models of inductive reasoning. *Psychol. Rev.* 116, 20–58

14 Griffiths, T.L. and Tenenbaum, J.B. (2009) Theory-based causal induction. *Psychol. Rev.* 116, 661–716

15 Griffiths, T.L. and Tenenbaum, J.B. (2006) Optimal predictions in everyday cognition. *Psychol. Sci.* 17, 767–773

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

17 Chater, N. and Vitanyi, P.M.B. (2001) The generalized universal law of generalization. *J. Math. Psychol.* 47, 346–369

18 Tenenbaum, J.B. (2000) Rules and similarity in concept learning. In *Advances in Neural Information Processing Systems* (Vol. 12) (Solla, S.A., ed.), In pp. 59–65, MIT Press

19 Tenenbaum, J.B. and Griffiths, T.L. (2001) Generalization, similarity, and Bayesian inference. *Behav. Brain Sci.* 24, 629–641

20 Xu, F. and Tenenbaum, J.B. (2007) Sensitivity to sampling in Bayesian word learning. *Dev. Sci.* 10, 288–297

21 Gweon, H. *et al.* (2010) Infants consider both the sample and the sampling process in inductive generalization. *Proc. Natl. Acad. Sci. U. S. A.* 107, 9066–9071

22 Griffiths, T.L. and Tenenbaum, J.B. (2005) Structure and strength in causal induction. *Cogn. Psychol.* 51, 354–384

23 Spirtes, P. *et al.* (1993) *Causation Prediction and Search*, Springer-Verlag

24 Pearl, J. (2000) *Causality: Models, Reasoning and Inference*, Cambridge University Press

25 Gopnik, A. *et al.* (2004) A theory of causal learning in children: causal maps and Bayes nets. *Psychol. Rev.* 111, 1–31

26 Cheng, P. (1997) From covariation to causation: a causal power theory. *Psychol. Rev.* 104, 367–405

27 Lu, H. *et al.* (2008) Bayesian generic priors for causal learning. *Psychol. Rev.* 115, 955–984

28 Kemp, C. *et al.* (2007) Learning causal schemata. In *Proceedings of the Twenty-Ninth Annual Conference of the Cognitive Science Society* (McNamara, D.S. and Trafton, J.G., eds), pp. 389–394, Cognitive Science Society

29 Lucas, C.G. and Griffiths, T.L. (2010) Learning the form of causal relationships using hierarchical Bayesian models. *Cogn. Sci.* 34, 113–147

30 Chater, N. and Vitányi, P. (2007) Ideal learning of natural language: positive results about learning from positive evidence. *J. Math. Psychol.* 51, 135–163

31 Perfors, A. *et al.* (2006) Poverty of the stimulus? A rational approach. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (Sun, R. and Miyake, N., eds), pp. 663–668, Cognitive Science Society

32 Frank, M. *et al.* (2009) Using speakers' referential intentions to model early cross-situational word learning. *Psychol. Sci* 20, 578–585

33 Klein, D. and Manning, C. (2004) Corpus-based induction of syntactic structure: models of dependency and constituency. *In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics* 486–497

34 Liang, P. et al. (2007) The infinite PCFG using hierarchical Dirichlet Processes. In *Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 688–697, Association for Computational Linguistics

35 Goldwater, S. (2009) A Bayesian framework for word segmentation: exploring the effects of context. *Cognition* 112, 21–54

36 Manning, C. and Schütze, H. (1999) *Foundations of Statistical Natural Language Processing*, MIT Press

37 Weiss, Y. *et al.* (2002) Motion illusions as optimal percepts. *Nature Neurosci.* 5, 598–604

38 Han, F. and Zhu, S.-C. (2005) Bottom-up/top-down image parsing by attribute graph grammar. In *Proceedings of the Tenth IEEE International Conference on Computer Vision*, pp. 1778–1785

39 Tu, Z. *et al.* (2005) Image parsing: unifying segmentation, detection, and recognition. *Int. J. Computer Vision* 63, 113–140

40 Smith, L.B. *et al.* (2002) Object name learning provides on-the-job training for attention. *Psychol. Sci.* 13, 13–19

41 Kemp, C. *et al.* (2007) Learning overhypotheses with hierarchical Bayesian models. *Dev. Sci.* 10, 307–321

42 Perfors, A. and Tenenbaum, J. (2009) Learning to learn categories. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (Taatgen, N. and van Rijn, H., eds), pp. 136–141, Cognitive Science Society

43 Goodman, N.D. *et al.* (2009) Learning a theory of causality. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (Taatgen, N. and van Rijn, H., eds), pp. 2188–2193, Cognitive Science Society

44 Colunga, E. and Smith, L.B. (2005) From the lexicon to expectations about kinds: a role for associative learning. *Psychol. Rev.* 112, 347–382

45 Shi, L. *et al.* (2008) Performing Bayesian inference with exemplar models. In *Proceedings of the Thirtieth Annual Conference of the Cognitive Science Society* (Love, B.C. *et al.*, eds), pp. 745–750, Cognitive Science Society

46 Medin, D.L. and Schaffer, M.M. (1978) Context theory of classification learning. *Psychol. Rev.* 85, 207–238

47 Nosofsky, R.M. (1986) Attention, similarity, and the identification-categorization relationship. *J. Exp. Psychol.: General* 115, 39–57

48 Sanborn, A.N. *et al.* (2006) A more rational model of categorization. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (Sun, R. and Miyake, N., eds), pp. 726–731, Cognitive Science Society

49 Daw, N. and Courville, A.C. (2008) The pigeon as particle filter. In *Advances in Neural Information Processing Systems* (Vol. 20) (Platt, J.C. *et al.*, eds), In MIT Press

50 Brown, S.D. and Steyvers, M. (2009) Detecting and predicting changes. *Cogn. Psychol.* 58, 49–67

51 Levy, R. *et al.* (2009) Modeling the effects of memory on human online sentence processing with particle filters. In *Advances in Neural Information Processing Systems* (Vol. 21) (Koller, D., eds), pp. 937–944

52 McClelland, J.L. (1998) Connectionist models of Bayesian inference. In *Rational Models of Cognition* (Oaksford, M. and Chater, N., eds), Oxford University Press

53 Hinton, G.E. (2007) Learning multiple layers of representation. *Trends Cogn. Sci.* 11, 428–434

54 Ma, W.J. (2008) Spiking networks for Bayesian inference and choice. *Curr. Opin. Neurobiol.* 18, 217–222

55 Lee, T-S. and Mumford, D. (2003) Hierarchical Bayesian inference in the visual cortex. *J. Opt. Soc. Am. A* 20, 1434–1448

56 Smolensky, P. and Legendre, G. (2006) *The Harmonic Mind*, MIT Press