Opinion

Cell PRESS

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 doi:10.1016/j.tics.2010.05.004 Trends in Cognitive Sciences 14 (2010) 357–364

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.

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.

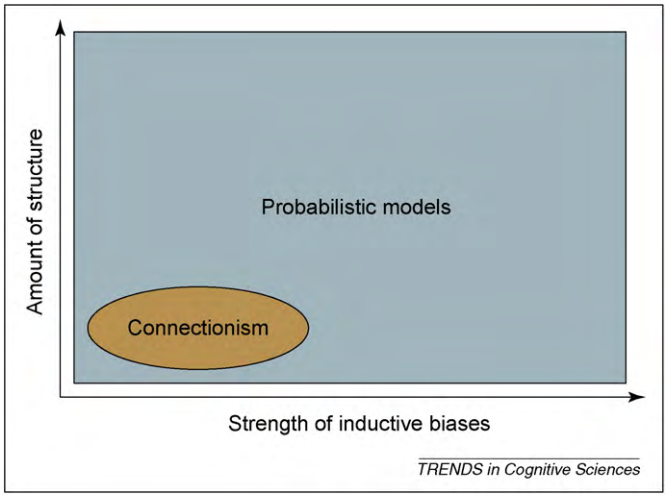

Figure 1. Theoretical commitments of connectionism and probabilistic models of cognition. Based on a certain view of brain architecture and function, connectionist models makes strong assumptions about the representations and inductive biases to be used in explaining human cognition: representations 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.

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

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

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.