Overview

Bayesian learning theory applied to human cognition

Robert A. Jacobs[1]* and John K. Kruschke[2]

Probabilistic models based on Bayes' rule are an increasingly popular approach to understanding human cognition. Bayesian models allow immense representational latitude and complexity. Because they use normative Bayesian mathematics to process those representations, they define optimal performance on a given task. This article focuses on key mechanisms of Bayesian information processing, and provides numerous examples illustrating Bayesian approaches to the study of human cognition. We start by providing an overview of Bayesian modeling and Bayesian networks. We then describe three types of information processing operations—inference, parameter learning, and structure learning—in both Bayesian networks and human cognition. This is followed by a discussion of the important roles of prior knowledge and of active learning. We conclude by outlining some challenges for Bayesian models of human cognition that will need to be addressed by future research. © 2010 John Wiley & Sons, Ltd. *WIREs Cogn Sci* 2011 2 8–21 DOI: 10.1002/wcs.80

INTRODUCTION

Computational modeling of human cognition has focused on a series of different formalisms over recent decades. In the 1970s, production systems were considered a methodology that would unite the studies of human and machine intelligence. In the 1980s and 1990s, connectionist networks were thought to be a key to understanding how cognitive information processing is performed by biological nervous systems. These formalisms continue to yield valuable insights into the processes of cognition. Since the 1990s, however, probabilistic models based on Bayes' rule have become increasingly popular, perhaps even dominant in the field of cognitive science. Importantly, Bayesian modeling provides a unifying framework that has made important contributions to our understanding of nearly all areas of cognition, including perception, language, motor control, reasoning, learning, memory, and development.

In this article, we describe some advantages offered by Bayesian modeling relative to previous formalisms. In particular, Bayesian models allow immense representational latitude and complexity, and use normative Bayesian mathematics to process those representations. This representational complexity contrasts with the relative simplicity of nodes in a connectionist network, or if-then rules in a production system. In this article, we focus on key mechanisms of Bayesian information processing, and we provide examples illustrating Bayesian approaches to human cognition. The article is organized as follows. We start by providing an overview of Bayesian modeling and Bayesian networks. Next, we describe three types of information processing operations found in both Bayesian networks and human cognition. We then discuss the important role of prior knowledge in Bayesian models. Finally, we describe how Bayesian models naturally address *active* learning, which is a behavior that other formalisms may not address so transparently.

BAYESIAN MODELING

Are people rational? This is a complex question whose answer depends on many factors, including the task under consideration and the definition of the word 'rational'. A common observation of cognitive scientists is that we live in an uncertain world, and rational behavior depends on the ability to process information effectively despite ambiguity or uncertainty. Cognitive scientists, therefore, need

*Correspondence to: robbie@bcs.rochester.edu

[1] Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY, USA

[2] Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN, USA

DOI: 10.1002/wcs.80

methods for characterizing information and the uncertainty in that information. Fortunately, such methods are available—probability theory provides a calculus for representing and manipulating uncertain information. An advantage of Bayesian models relative to many other types of models is that they are probabilistic.

Probability theory does not provide just any calculus for representing and manipulating uncertain information, it provides an *optimal* calculus.[1] Consequently, an advantage of Bayesian modeling is that it gives cognitive scientists a tool for defining rationality. Using Bayes' rule, Bayesian models optimally combine information based on prior beliefs with information based on observations or data. Using Bayesian decision theory, Bayesian models can use these combinations to choose actions that maximize the task performance. Owing to these optimal properties, Bayesian models perform a task as well as the task can be performed, meaning that the performance of a Bayesian model on a task defines rational behavior for that task.

Of course, the performance of a model depends on how it represents prior beliefs, observations, and task goals. That is, the representational assumptions of a model influence its performance. It is important to keep in mind that any task can be modeled in multiple ways, each using a different set of assumptions. One model may assume the use of certain dimensional perceptual representations which are combined linearly into a probabilistic choice, whereas another model may assume featural representations combined conjunctively into a probabilistic choice. But for any specific probabilistic formalization of a task, a Bayesian model specifies optimal performance given the set of assumptions made by the model.

This fact leads to an additional advantage of Bayesian modeling relative to many other approaches, namely that the assumptions underlying Bayesian models are often explicitly stated as well-defined mathematical expressions and, thus, easy to examine, evaluate, and modify. Indeed, a key reason for using the Bayesian modeling formalism is that, relative to many other computational formalisms, it allows cognitive scientists to more easily study the advantages and disadvantages of different assumptions. (Admittedly, this property is also possessed by other formalisms, especially in the mathematical psychology tradition, that have rigorous mathematical or probabilistic interpretations; see, e.g., Busemeyer and Diederich,[2] and many examples in Scarborough and Sternberg.[3] But such mathematically explicit models form only a subset of the spectrum of computational models in cognitive science; cf. Refs 4 and 5.) Through this study, scientists can ask questions about the nature or structure of a task (Marr[6] referred to the analysis of the structure of a task as a 'computational theory'), such as: What are the variables that need to be taken into account, the problems that need to be solved, and the goals that need to be achieved in order to perform a task?; Which of these problems or goals are easy or hard?; and Which assumptions are useful, necessary, and/or sufficient for performance of the task?

Let us assume that the performance of a person on a task is measured, a Bayesian model is applied to the same task, and the performances of the person and the model are equal. This provides important information to a cognitive scientist because it provides the scientist with an explanation for the person's behavior—the person is behaving optimally because he or she is using and combining all relevant information about the task in an optimal manner. In addition, this result supports (but does not prove) the hypothesis that the assumptions used by the model about prior beliefs, observations, and task goals may also be used by the person.

Alternatively, suppose that the performance of the model exceeds that of the person. This result also provides useful information. It indicates that the person is not using all relevant information or not combining this information in an optimal way. That is, it suggests that there are cognitive 'bottlenecks' preventing the person from performing better. Further experimentation can attempt to identify these bottlenecks, and training can try to ameliorate or remove these bottlenecks. Possible bottlenecks might include cognitive capacity limitations such as limits on the size of working memory or on the quantity of attentional resources. Bottlenecks might also include computational limitations. Bayesian models often perform complex computations in high-dimensional spaces. It may be that people are incapable of these complex computations and, thus, incapable of performing Bayesian calculations. Later in this article, we will address the possibility that people are not truly Bayesian, but only approximately Bayesian.

If a cognitive scientist hypothesizes that a person's performance on a task is suboptimal because of a particular bottleneck, then the scientist can develop a new model that also contains the posited bottleneck. For instance, if a scientist believes that a person performs suboptimally because of an inability to consider stimuli that occurred far in the past, the scientist can study a model that only uses recent inputs. Identical performances by the new model and the person lend support to the idea that the person, like the model, is only using recent inputs.

Finally, suppose that the performance of the person exceeds that of the model—that is, the person's performance exceeds the optimal performance defined by the model—then once again this is a useful result. It suggests that the person is using information sources or assumptions that are not currently part of the model. For instance, if a model only considers the previous word when predicting the next word in a sentence, then a person who outperforms the model is likely using more information (e.g., several previous words) when he or she makes a prediction. A cognitive scientist may consider a new model with additional inputs. As before, identical performances by the new model and the person lend support to the idea that the person, like the model, is using these additional inputs.

In some sense, the Bayesian approach to the study of human cognition might seem odd. It is based on an analysis of what level of task performance is achievable given a set of assumptions. However, it does not make a strong statement about how a person achieves that task performance—which mental representations and algorithms the person uses while performing a task. Bayesian models are not intended to provide mechanistic or process accounts of cognition.[6,7] For cognitive scientists interested in mechanistic accounts, a Bayesian model often suggests possible hypotheses about representations and algorithms that a person might use while performing a task, but these hypotheses need to be evaluated by other means. Similarly, if a person performs suboptimally relative to a Bayesian model, the model may suggest hypotheses as to which underlying mechanisms are at fault, but these hypotheses are only starting points for further investigation. Consequently, Bayesian modeling complements, but does not supplant, the use of experimental and other theoretical methodologies.

For all the reasons outlined here, Bayesian modeling has become increasingly important in the field of cognitive science.[6–16] Rather than describing particular Bayesian models in depth, a main focus of this article is on information processing operations—inference, parameter learning, and structure learning—found in Bayesian models and human cognition. Before turning to these operations, however, we first describe Bayesian networks, a formalism that makes these operations particularly easy to understand.

BAYESIAN NETWORKS

When a theorist develops a mental model of a cognitive domain, it is necessary to first identify the variables that must be taken into account. The domain can then be characterized through the joint probability distribution of these variables. Many different information processing operations can be carried out by manipulating this distribution.

Although conceptually appealing, joint distributions are often impractical to work with directly because real-world domains contain many potentially relevant variables, meaning that joint distributions can be high-dimensional. If a domain contains 100 variables, then the joint distribution is 100-dimensional. Hopefully, it is the case that some variables are independent (or conditionally independent given the values of other variables) and, thus, the joint distribution can be factored into a product of a small number of conditional distributions where each conditional distribution is relatively low-dimensional. If so, then it is more computationally efficient to work with the small set of low-dimensional conditional distributions than with the high-dimensional joint distribution.

Bayesian networks have become popular in artificial intelligence and cognitive science because they graphically express the factorization of a joint distribution.[17–19] A network contains nodes, edges, and probability distributions. Each node corresponds to a variable. Each edge corresponds to a relationship between variables. Edges go from 'parent' variables to 'child' variables, thereby indicating that the values of the parent variables directly influence the values of the child variables. Each conditional probability distribution provides the probability of a child variable taking a particular value given the values of its parent variables. The joint distribution of all variables is equal to the product of the conditional distributions. For example, we assume that the joint distribution of variables A, B, C, D, E, F, and G can be factored as follows:

$$p(A, B, C, D, E, F, G) = p(A)p(B)p(C|A)p(D|A, B) \\ p(E|B)p(F|C)p(G|D, E) \quad (1)$$

Then the Bayesian network in Figure 1 represents this joint distribution.

The parameters of a Bayesian network are the parameters underlying the conditional probability distributions. For example, suppose that the variables of the network in Figure 1 are real-valued, and suppose that each variable is distributed according to a normal distribution whose mean is equal to the weighted sum of its parent's values (plus a bias weight) and whose variance is a fixed constant [In other words, the conditional distribution of X given the values of its parents is a Normal distribution whose mean is $\sum_{i \in pa(X)} w_{Xi} V_i + w_{Xb}$ and whose variance is σ_X^2 where

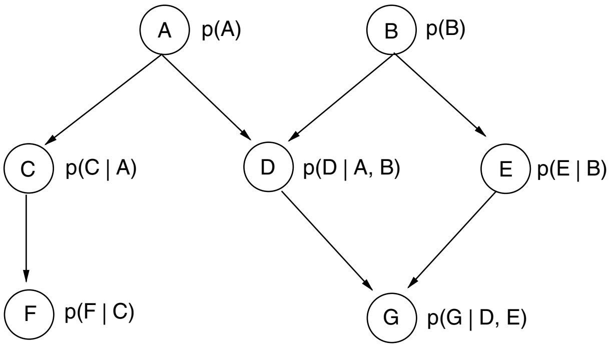

FIGURE 1 | Bayesian network representing the joint probability distribution of variables A, B, C, D, E, F, and G. A node represents the value of the variable it contains. Arrows impinging on a node indicate what other variables the value of the node depends on. The conditional probability besides each node expresses mathematically what the arrows express graphically.

X is a variable, i indexes the variables that are parents of X, V_i is the value of variable i, w_{Xi} is a weight relating the value of variable i to X, and w_{Xb} is a bias weight for X. In Figure 1, for instance, $A \sim N(w_{Ab}, \sigma_A^2)$, $C \sim N(w_{CA}A + w_{Cb}, \sigma_C^2)$, and $D \sim N(w_{DA}A + w_{DB}B + w_{Db}, \sigma_D^2)$.] Then the weights would be the network's parameters.

INFERENCE

If the values of some variables are observed, then these data can be used to update our beliefs about other variables. This process is referred to as 'inference', and can be carried out using Bayes' rule. For example, suppose that the values of F and G are observed, and we would like to update our beliefs about the unobserved values A, B, C, D, and E. Using Bayes' rule:

$$p(A, B, C, D, E|F, G) = \frac{p(F, G|A, B, C, D, E)p(A, B, C, D, E)}{p(F, G)} \quad (2)$$

where $p(A, B, C, D, E)$ is the *prior* probability of A, B, C, D, and E, $p(A, B, C, D, E|F, G)$ is the *posterior* probability of these variables given data F and G, $p(F, G|A, B, C, D, E)$ is the probability of the observed data (called the *likelihood* function of A, B, C, D, and E), and $p(F, G)$ is a normalization term referred to as the *evidence*.

Inference often requires marginalization. Suppose we are only interested in updating our beliefs about A and B (i.e., we do not care about the values of C, D, and E), this can be achieved by 'integrating out' the irrelevant variables:

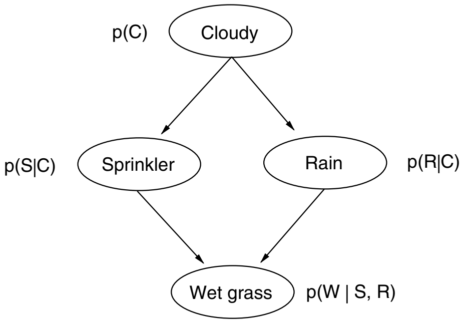

FIGURE 2 | Bayesian network characterizing a domain with four binary variables indicating whether it is cloudy, the sprinkler was recently on, it recently rained, and the grass is wet.

$$p(A, B|F, G) = \iiint p(A, B, C, D, E|F, G)dCdDdE \quad (3)$$

In some cases, inference can be carried out in a computationally efficient manner using a local message-passing algorithm.[18] In other cases, inference is computationally expensive, and approximation techniques, such as Markov chain Monte Carlo sampling, may be needed.[20]

Is Bayesian inference relevant to human cognition? We think that the answer is yes. To motivate this answer, we first consider a pattern of causal reasoning referred to as 'explaining away'. Explaining away is an instance of the 'logic of exoneration' (e.g., if one suspect confesses to a crime, then unaffiliated suspects are exonerated). In general, increasing the believability of some hypotheses necessarily decreases the believability of others.

The Bayesian network illustrated in Figure 2 characterizes a domain with four binary variables indicating whether it is cloudy, whether the sprinkler was recently on, whether it recently rained, and whether the grass is wet.[18,19] If the weather is cloudy (denoted $C = 1$), then there is a low probability that the sprinkler was recently on ($S = 1$) and a high probability that it recently rained ($R = 1$). If the weather is not cloudy, then there is a moderate probability that the sprinkler was recently on and a low probability that it recently rained. Finally, if the sprinkler was recently on, rain fell, or both, then there is a high probability that the grass is wet ($W = 1$).

You walk outside and discover that the grass is wet and that the sprinkler is on. What, if anything, can you conclude about whether it rained recently? An intuitive pattern of reasoning, and one that seems to be exhibited by people, is to conclude that the water from the sprinkler wet the grass and, thus, there is no good evidence suggesting that it rained recently. This pattern is called explaining away because the observation that the sprinkler is on explains the fact that the grass is wet, meaning that there is no reason to hypothesize another cause, such as recent rain, for why the grass is wet. This type of causal reasoning naturally emerges from Bayes' rule. Without going into the mathematical details, a comparison of the probability of recent rain given that the grass is wet, $p(R = 1|W = 1)$, with the probability of recent rain given that the grass is wet and that the sprinkler is on, $p(R = 1|W = 1, S = 1)$, would show that the latter value is significantly smaller. Thus, Bayes' rule performs explaining away.

Explaining away illustrates an important advantage of Bayesian statistics, namely, that Bayesian models maintain and update probabilities for all possible values of their variables. Because probability distributions must sum or integrate to one, if some values become more likely, then other values must become less likely. This is true even for variables whose values are not directly specified in a data set. In the example above, the variables S and R are negatively correlated given the value of C. (If it is raining, it is unlikely that a sprinkler will be on. Similarly, if a sprinkler is on, it is unlikely to be raining.) Therefore, if a Bayesian observer discovers that the grass is wet and that the sprinkler is on, the observer can reasonably accept the hypothesis that the water from the sprinkler wet the grass. In doing so, the observer simultaneously rejects the hypothesis that it rained recently and the rain water wet the grass, despite the fact that the observer did not directly obtain any information about whether it rained. Thus, a Bayesian observer can simultaneously maintain and update probabilities for multiple competing hypotheses.

Our scenario with sprinklers, rain, and wet grass provides a simple example of Bayesian inference, but cognitive scientists have studied more complicated examples. Often these examples involve prediction and, thus, inference and prediction are closely related.

Calvert et al.[21] found that auditory cortex in normal hearing individuals became activated when these individuals viewed facial movements associated with speech (e.g., lipreading) in the absence of auditory speech sounds. It is as if the individuals used their visual percepts to predict or infer what their auditory percepts would have been if auditory stimuli were present; that is, they computed p(auditory percept | visual percept).

Pirog Revill et al.[22] used brain imaging to show that people predicted the semantic properties of a word while lexical competition was in process and before the word was fully recognized. For example, the speech input /can/ is consistent with the words *can, candy*, and *candle* and, thus, *can, candy*, and *candle* are active lexical competitors when /can/ is heard. These investigators found that a neural region typically associated with visual motion processing became more activated when motion words were heard than when nonmotion words were heard. Importantly, when nonmotion words were heard, the activation of this region was modulated by whether there was a lexical competitor that was a motion word rather than another nonmotion word. It is as if the individuals used the on-going speech input to predict the meaning of the current word as the speech unfolded over time; i.e., they computed p(semantic features | on-going speech percept).

Blakemore et al.[23] used inference to explain why individuals cannot tickle themselves. A 'forward model' is a mental model that predicts the sensory consequences of a movement based on a motor command. These investigators hypothesized that when a movement is self-produced, its sensory consequences can be accurately predicted by a forward model; i.e., a person computes p(sensory consequences | motor movement), and this prediction is used to attenuate the sensory effects of the movement. In this way, the sensory effects of self-tickling are dulled.

Although now there is much experimental evidence for perceptual inference, the functional role of inference is not always obvious. For example, why is it important to predict an auditory percept based on a visual percept? In the section below titled *Parameter Learning*, we review the hypothesis that inference may play an important role in learning.

The examples above provide evidence that people infer the values of unobserved variables based on the values of observed variables. However, they do not show that this inference is quantitatively consistent with the use of Bayes' rule. Evidence suggesting that people's inferences are indeed quantitatively consistent with Bayes' rule comes from the study of sensory integration. The Bayesian network in Figure 3 characterizes a mental model of an observer who both sees and touches the objects in an environment. The nodes labeled 'scene variables' are the observer's internal variables for representing all conceivable three-dimensional scenes. As a matter of notation, let S denote the scene variables. Based on the values of the scene variables, haptic feature variables, denoted F_{H}, and visual feature variables, denoted F_{V}, are assigned values. For instance, a scene with a coffee mug gives rise to both haptic features, such as curvature and smoothness, and visual features, such as curvature

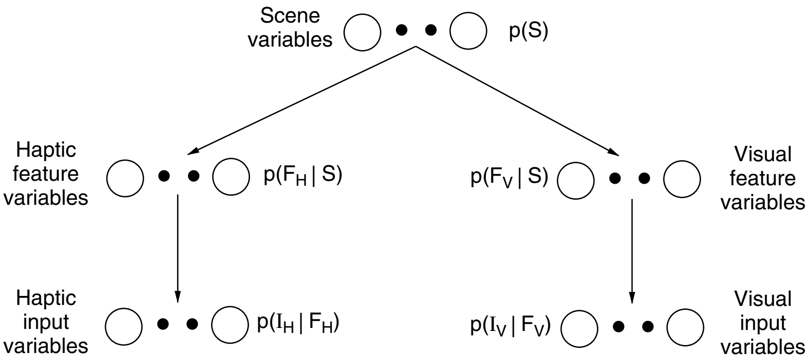

FIGURE 3 | Bayesian network characterizing a domain in which an observer both sees and touches the objects in an environment. At the top of the hierarchy, the values of scene variables determine the probabilities of *distal* haptic and visual features. The distal haptic and visual features in turn determine the probabilities of values of *proximal* haptic and visual input (sensory) variables.

and color. The haptic features influence the values of the haptic input variables, denoted I_{H}, when the observer touches the mug. Similarly, the visual features influence the values of the visual input variables, denoted I_{V}, when the observer views the mug.

The values of the input variables are 'visible' because the observer directly obtains these values as percepts arising through touch and sight. However, the feature and scene variables are not directly observable and, thus, are regarded as hidden or latent. The distribution of the latent variables may be computed by the observer from the values of the visible variables using Bayesian inference. For instance, based on the values of the haptic and visual input variables, the observer may want to infer the properties of the scene. That is, the observer may want to compute $p(S|I_{\mathrm{H}}, I_{\mathrm{V}})$.

To illustrate how an observer might compute this distribution, we consider a specific instance of sensory integration. We suppose that an observer both sees and grasps a coffee mug, and wants to infer the depth of the mug (i.e., the distance from the front of the mug to its rear). Also we can suppose that the observer's belief about the depth of the mug given its visual input has a normal distribution, denoted $N(\mu_{\mathrm{v}}, \sigma_{\mathrm{v}}^2)$. Similarly, the observer's belief about the depth given its haptic input has a normal distribution, denoted $N(\mu_{\mathrm{h}}, \sigma_{\mathrm{h}}^2)$. Then, given certain mathematical assumptions, it is easy to show (as derived in many publications; e.g., Ref 24) that the belief about the depth, given both inputs, has a normal distribution whose mean is a linear combination of μ_{v} and μ_{h}:

$$\mu_{\mathrm{v,h}} = \frac{\sigma_{\mathrm{v}}^{-2}}{\sigma_{\mathrm{v}}^{-2} + \sigma_{\mathrm{h}}^{-2}} \mu_{\mathrm{v}} + \frac{\sigma_{\mathrm{h}}^{-2}}{\sigma_{\mathrm{v}}^{-2} + \sigma_{\mathrm{h}}^{-2}} \mu_{\mathrm{h}} \quad (4)$$

and whose variance is given by:

$$\frac{1}{\sigma_{\mathrm{v,h}}^2} = \frac{1}{\sigma_{\mathrm{v}}^2} + \frac{1}{\sigma_{\mathrm{h}}^2} \quad (5)$$

At an intuitive level, this form of sensory integration is appealing for several reasons. First, the variance of the depth distribution based on both inputs $\sigma_{\mathrm{v,h}}^2$ is always less than the variances based on the individual inputs σ_{v}^2 and σ_{h}^2, meaning that depth estimates based on both inputs are more precise than estimates based on individual inputs. Second, this form of sensory integration uses information to the extent that this information is reliable or precise. This idea is illustrated in Figure 4. In the top panel, the distributions of depth given visual inputs and haptic inputs have equal variances ($\sigma_{\mathrm{v}}^2 = \sigma_{\mathrm{h}}^2$). That is, the inputs are equally precise indicators of depth. The mean of the depth distribution based on both inputs is, therefore, an equally weighted average of the means based on the individual inputs. In the bottom panel, however, the variance of the distribution based on vision is smaller than the variance based on haptics ($\sigma_{\mathrm{v}}^2 < \sigma_{\mathrm{h}}^2$), meaning that vision is a more precise indicator of depth. In this case, the mean of depth based on both inputs is also a weighted average of the means based on the individual inputs, but now the weight assigned to the visual mean is large and the weight assigned to the haptic mean is small.

Do human observers perform sensory integration in a manner consistent with this statistically optimal and intuitively appealing framework? Several studies have now shown that, in many cases, the answer is yes. Ernst and Banks[25] recorded subjects' judgments of the height of a block when they saw the block, when they grasped the block, and when they both saw and grasped the block. These investigators found that the framework accurately predicted subjects' multisensory judgments based on their unisensory judgments. This was true when visual signals were corrupted by small amounts of noise, in which case the visual signal was more reliable, and also when visual signals were corrupted by large amounts of noise, in which case the haptic signal was more reliable. Knill and Saunders[26] used the framework to predict subjects' judgments of surface slant when

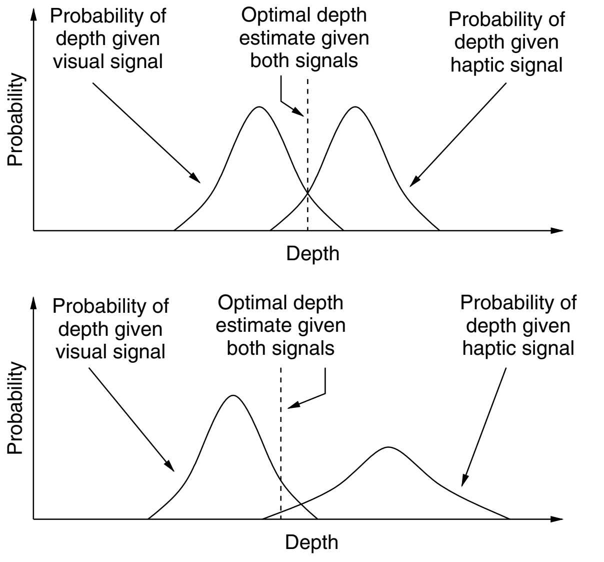

FIGURE 4 | Bayesian model of sensory integration. (Top) A situation in which visual and haptic percepts are equally good indicators of depth. (Bottom) A situation in which the visual percept is a more reliable indicator of depth.

surfaces were defined by visual stereo and texture cues based on their slant judgments when surfaces were defined by just the stereo cue or just the texture cue. It was found that the framework provided accurate predictions when surface slants were small, meaning that stereo was a more reliable cue, and when slants were large, meaning that texture was a more reliable cue. Other studies showing that people's multisensory percepts are consistent with the framework include Alais and Burr,[27] Battaglia et al.,[28] Ghahramani et al.,[29] Jacobs,[30] Körding and Wolpert,[31] Landy et al.,[32] Maloney and Landy,[33] and Young et al.[34]

PARAMETER LEARNING

Parameter learning occurs when one of the conditional distributions inside a Bayesian network is adapted. For example, in the Bayesian network on sensory integration in Figure 3, let us suppose that an observer's probability distribution of visual inputs, given its internal representation of the visual features, is a normal distribution whose mean is equal to a weighted sum of the values of the visual feature variables. If we use Bayes' rule to update the posterior distribution of the weight values based on new data, then this would be an instance of parameter learning.

Do people perform parameter learning in the same manner as Bayesian networks? Various aspects of human learning are captured by Bayesian models. Perhaps the simplest example is a behavioral phenomenon known as 'backward blocking'.[35] Suppose that in Stage 1 of an experiment, a person is repeatedly exposed to two cues, denoted C_1 and C_2, followed by an outcome O. The person will learn that each cue is at least partly predictive of the outcome. That is, the person will act as if there is a moderate association between C_1 and O and also between C_2 and O. In Stage 2, the person is repeatedly exposed to C_1 followed by O (C_2 does not appear in Stage 2). After Stage 2, the person will act as if there is a strong association between C_1 and O. This is expected because the person has consistently seen C_1 followed by O. However, the person will also act as if there is only a weak association between C_2 and O. This is surprising because it suggests that the person has retrospectively revalued C_2 by diminishing its associative strength despite the fact that C_2 did not appear in Stage 2 of the experiment.

Backward blocking and explaining away (described earlier in the article) are similar phenomena. In both cases, an unobserved variable is revalued because an observed variable co-occurs with an outcome. Explaining away is regarded as a form of inference because it takes place on a short time scale—the unobserved variable is revalued after a single co-occurrence of an observed variable and an outcome. In contrast, backward blocking is regarded as a form of learning because it might take place on a relatively longer time scale of multiple co-occurrences of an observed variable and an outcome (although it can also happen in very few exposures). Ultimately, the distinction between inference, as in explaining away, and learning, as in backward blocking, evaporates: Both inference and learning are modeled via Bayes' rule, but applied to variables with different meaning to the theorist.

It is challenging for non-Bayesian models to account for backward blocking. For example, according to the well-known Rescorla–Wagner model,[36] a learner's knowledge consists of a single weight value, referred to as an associative strength, for each cue. Learning consists of changing a cue's weight value after observing a new occurrence of the cue and outcome. Because a cue's weight value changes only on trials in which the cue occurs, the learning rule cannot account for backward blocking which seems to require a change in C_2's weight value during Stage 2 despite the fact that C_2 did not occur during Stage 2 (for references to non-Bayesian models that might address backward blocking, see Ref 37).

In contrast, Bayesian learning can account for backward blocking because a learner entertains simultaneously all possible combinations of weight values, with a degree of believability for each combination.[37–40] That is, the learner maintains a posterior probability distribution over weight combinations, given the data experienced so far. In the context of the backward blocking experiment, this distribution is denoted $p(w_1, w_2|\{\text{data}\})$. Bayesian rules can account for backward blocking as follows. After the first stage of training, in which the learner has seen cases of C_1 and C_2 together indicating the outcome, the learner has some degree of belief in a variety of weight combinations, including weight values of $w_1 = 0.5$, $w_2 = 0.5$ (both C_1 and C_2 are partially predictive of O), $w_1 = 1$, $w_2 = 0$ (C_1 is fully predictive of O but C_2 carries no information about O), and $w_1 = 0$, $w_2 = 1$ (C_2 is fully predictive of O but C_1 carries no information about O). In Stage 2, C_1 alone is repeatedly paired with O and, thus, belief increases in both the combination $w_1 = 0.5$, $w_2 = 0.5$ and $w_1 = 1$, $w_2 = 0$. Because total belief across all weight combinations must sum to one, belief in $w_1 = 0$, $w_2 = 1$ consequently drops. The learner, therefore, decreases the associative strength between C_2 and O, thereby exhibiting backward blocking.

Backward blocking is representative of only one type of learning situation, referred to as *discriminative* learning. In this situation, a learner's data consist of instances of cues and outcomes. The learner needs to accurately estimate the conditional probability distribution over outcomes given particular values of the cues. This conditional distribution is called a discriminative model, and it allows a learner to predict the likely outcomes based on the given cue values. Notice that the learner is not learning the distribution of cues. In contrast, in a *generative* learning situation, the learner learns the joint distribution of cues and outcomes. A generative model has latent variables to describe how underlying states generate the distribution of cues and outcomes simultaneously.[41] In general, learning with latent variables is a computationally difficult problem. It is often approached by using inference to determine the probability distributions of the latent variables given the values of the observed variables (e.g., cues and outcomes). These distributions over latent variables are useful because they allow a learner to adapt its parameter values underlying conditional distributions of observed variables given values of latent variables.

Consider the problem of visual learning in multisensory environments. Recall that the Bayesian network in Figure 3 represents the mental model of an observer that both sees and touches objects in a scene. Also recall that the input variables (I_V and I_H) are 'visible' because the observer obtains the values of these variables when he or she touches and views objects. However, the feature (F_V and F_H) and scene variables (S) are not directly observable and, thus, are regarded as hidden or latent. Visual learning takes place when, for example, the observer adapts the parameter values underlying $p(I_V|F_V)$, the conditional probability distribution associated with its visual input variables. To adapt these values, the observer needs information indicating how the values should be modified because F_V is an unobserved variable. How can the observer's visual system obtain this information?

Suppose that the observer touches and sees an object. The observer can infer the distribution of the visual feature variables F_V in two ways: She or he can calculate the conditional distribution of these variables given the values of both haptic and visual input variables $p(F_V|I_H, I_V)$ or given only the values of the visual input variables $p(F_V|I_V)$. The first distribution is based on more information and, thus, it can be used as a 'teaching signal'. That is, the observer can adapt its visual system so that $p(F_V|I_V)$ is closer to $p(F_V|I_H, I_V)$. This example illustrates the fact that multisensory environments are useful for visual learning because nonvisual percepts can be used by Bayes' rule to infer distributions that can be used by people when adapting their visual systems.[42]

The idea that inference provides important information used during parameter learning lies at the heart of the expectation-maximization (EM) algorithm, an optimization procedure for maximizing likelihood functions.[43] Consider the Bayesian network shown in Figure 5. The nodes in the top row correspond to latent variables whose values are unobserved, whereas the nodes in the bottom row correspond to visible variables whose values are observed. Data sets consist of multiple data items where an item is a set of values of the visible variables. According to the network, data items are generated as follows. For each data item, the values of the latent variables are sampled from their prior distributions, and then the values of the visible variables are sampled from their conditional distributions given the values of the latent variables.

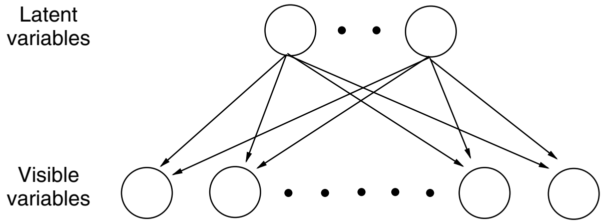

FIGURE 5 | Bayesian network characterizing a domain in which a large number of visible variables are dependent on a small number of unobserved latent variables.

During learning, a learner uses a data set to estimate the parameter values of the Bayesian network. Importantly, learning seems to require solving the following chicken-and-egg problem: In order to learn the parameter values of the conditional distributions, it seems necessary to know the values of the latent variables. But in order to infer the values of the latent variables, it seems necessary to know the parameter values of the conditional distributions. This chicken-and-egg problem is solved by the EM algorithm. The algorithm is an iterative algorithm with two steps at each iteration. During the E-step, it uses the values of the visible variables and the current estimates of the parameter values of the conditional distributions to compute the expected values of the latent variables. During the M-step, it uses the values of the visible variables and the expected values of the latent variables to re-estimate the parameter values of the conditional distributions as the values that maximize the likelihood of the data. (The reader should note that the EM algorithm is not a fully Bayesian algorithm because it finds point estimates of the parameter values as opposed to probability distributions over parameter values. Nonetheless, the algorithm has played an important role in the literature and, thus, we include it here.)

Specific instances of the use of the EM algorithm to learn the parameter values for the Bayesian network in Figure 5 are commonplace in the cognitive science literature.[44] In principal component analysis, data typically exist in a high-dimensional space (i.e., there are many visible variables), and the learner accounts for the data by hypothesizing that data items are a linear projection from a lower-dimensional space (i.e., there are relatively few latent variables). Orthogonal axes of the lower-dimensional space are discovered by analyzing the data's directions of maximum variance in the high-dimensional space.[a] Factor analysis is similar to principal component analysis, though the learner does not necessarily seek to characterize the lower-dimensional space through orthogonal axes. Instead, the learner seeks to discover latent variables that account for correlations among the visible variables. Mixture models in which the mixture components are normal distributions also fit in this framework. Here, the number of latent variables equals the number of possible clusters or categories of data items. The learner assumes that each item was generated by a single latent variable indicating the item's cluster or category membership.

STRUCTURE LEARNING

Parameter learning is not the only type of learning that can take place in a Bayesian network. Another type of learning that can occur is referred to as 'structure learning'. In the examples of Bayesian networks discussed above, the structure of a network was fixed, where the structure refers to a network's set of nodes (or set of random variables) and set of directed edges between nodes. Given certain assumptions, however, it is possible to learn a probability distribution over possible structures of a network from data.

Why are assumptions needed for structure learning? It is because the space of possible structures grows extremely fast. If there are n random variables, then the number of possible structures grows superexponentially in n [see Ref 45; specifically, the rate of growth is $O(n!2^{n!/(2!(n-2)!)})$]. Consequently, it is generally not computationally possible to search the full space of possible structures.

Because the full space of structures cannot be searched, it is common to define a small 'dictionary' of plausible structures. A researcher may decide that there are n plausible structures worth considering for a given domain. The researcher may then perform 'model comparison' by comparing the ability of each structure or model to provide an account of the data. Let $\{M_1, \ldots, M_n\}$ denote a set of plausible models. Then the researcher computes the posterior probability of each model given the data using Bayes' rule:

$$p(M_i|\{\text{data}\}) \propto p(\{\text{data}\}|M_i)p(M_i) \tag{6}$$

The likelihood term $p(\{\text{data}\}|M_i)$ can be calculated by including the parameter values for a model, denoted θ, and then by marginalizing over all possible parameter values:

$$p(\{\text{data}\}|M_i) = \int p(\{\text{data}\}|M_i, \theta)p(\theta|M_i)\mathrm{d}\theta \tag{7}$$

The distribution of the data is provided by a weighted average of each model's account of the data[46]:

$$p(\{\text{data}\}) = \sum_{i=1}^{n} p(\{\text{data}\}|M_i)p(M_i) \tag{8}$$

This approach is referred to as 'model averaging'.

Körding et al.[47] presented a model-averaging approach to sensory integration. Consider an environment containing both visual and auditory stimuli, and an observer that needs to locate events in space. A problem for the observer is to determine whether a visual stimulus and an auditory stimulus originate from the same environmental event, in which case the locations indicated by these stimuli should be integrated, or if the visual and auditory stimuli originate from two different events, in which case the locations indicated by these stimuli should not be integrated. Körding et al.[47] defined two plausible Bayesian networks, one corresponding to stimuli that originate from the same event and the other corresponding to stimuli that originate from different events. Using a weighted average of the predictions of these two networks, the overall system was able to quantitatively predict human subjects' responses in a variety of auditory–visual localization experiments.

As a second example, Griffiths et al.[48] used model averaging to account for human category learning. They noted that most computational models in the literature are either instances of prototype theory, in which a single structure characterizes the statistics of each category, or exemplar theory, in which a different structure characterizes the statistics of every exemplar. Rather than choose between these two extremes, they used an approach (known as Dirichlet process mixture models) that can be implemented using a model that starts with one structure, but then adds additional structures as needed during the course of learning. The end result is that the data indicate a probability distribution over how many structures the model will have. For some data sets, the model may tend to use very few structures, whereas it may tend to use a large number of structures for other data sets. Thus, the model resembles neither a prototype theory nor an exemplar theory, but rather a hybrid that possesses many of the best features of both types of theories.

Because model averaging can be computationally expensive, some researchers have pursued a less expensive approach: they parameterize a family of plausible structures, and then search for the parameter values with the highest posterior probability. (The reader should note that this approach is not strictly Bayesian because it uses point estimates of the parameter values instead of using the full posterior probability distribution over parameter values.) This approach essentially turns the structure learning problem into a parameter learning problem. It is typically implemented through the use of hierarchical Bayesian networks in which distributions of parameters at the top of the hierarchy govern distributions of parameters at lower levels of the hierarchy.

Perhaps the most successful example of the use of this strategy comes from Kemp and Tenenbaum.[49] These authors defined a hierarchy in which the top level determined the form of a model, where there were eight possible forms (not necessarily mutually exclusive): partitions, chains, orders, rings, hierarchies, trees, grids, and cylinders. The next level determined the structure of a particular form. For example, if a Bayesian network was to have a tree form, then this level determined the particular tree structure that the network would have. It was found that the system consistently discovered forms and structures that, intuitively, seem reasonable for a domain. For example, using a database of votes cast by Supreme Court justices, it was found that a chain structure with 'liberal' justices at one end and 'conservative' justices at the other end fit the data best. Similarly, based on the features of a large number of animals, it was found that a tree structure with categories and subcategories closely resembling the Linnaean taxonomy of species fit the data best.

Although structure learning is computationally expensive, it can be extremely powerful because it gives a system enormous representational flexibility. A system can use the data to determine a probability distribution over structures for representing that data. This representational flexibility is a key feature of Bayesian systems that is difficult to duplicate in other types of computational formalisms.

PRIOR KNOWLEDGE

Bayesian inference and learning begin with a model of observable variables (i.e., the likelihood function) and prior knowledge expressed as relative beliefs across different model structures and as the distribution of beliefs over parameter values within specific structures. The prior knowledge defines the space of all possible knowable representations, and the degree to which each representation is believed. The structure of the model of observable variables and the constraints imposed by prior knowledge strongly influence the inferences and learning that result from Bayes' rule.

Strong prior knowledge can facilitate large changes in belief from small amounts of data. For example, let us assume we have prior knowledge that a coin is a trick coin that either comes up heads almost all the time or comes up tails almost all the time. In other words, the prior belief is that intermediate biases, such as 30% heads, etc., are not possible. We flip the coin once and find it comes up heads. From this single bit of data, we infer strong posterior belief that the coin is the type that comes up heads almost always.

The ability of people to learn from small amounts of data can be addressed in a Bayesian framework by strong constraints on the prior beliefs. Let us assume, for example, that we do not yet know the meaning of the word 'dog'. If we are shown a labeled example of a dog, what should we infer is the meaning of 'dog'? Children learn the extension of the word in only a few examples but, in principle, the word could refer to many sets of objects. Does the word refer to all furry things? To all things with a tail? To all four legged things that are smaller than a pony but bigger than a mouse? A model of word learning proposed by Xu and Tenenbaum[50] is able to learn word-extension mappings very quickly, in close correspondence to human learning, largely by virtue of strong constraints on prior beliefs regarding the space of possible extensions of words. The Bayesian learning model has a prior distribution that emphasizes hierarchically structured extensions (e.g., dalmatians within dogs within animals) based on perceptual similarity. It also includes a bias toward extensions that are perceptually distinctive. Because of the limited number of word meanings consistent with the prior beliefs, only a few examples are needed to sharply narrow the possible meaning of a word.

A powerful tool for Bayesian modeling is the use of hierarchical generative models to specify prior beliefs. The highest level of the prior specifies a generic theory of the domain to be learned, brought to bear by the learner. The theory generates all possible specific model structures for the domain, and the priors within structures.[51,52] Learning simultaneously updates beliefs within and across model structures. The power of the approach is that instead of the hypothesis space being a heuristic intuited by the theorist, its assumptions are made explicit and attributed to the learner's theory regarding the domain. An example of a generative hierarchical prior was described earlier, for representing structural relations among objects.[49] That model uses a generative grammar as a theory for constructing the space of all possible structural relations among objects. The generative grammar specifies iterative rules for relational graph construction, such as varieties of node splitting and edge construction. Each rule has a probability of application, thereby implicitly defining prior probabilities on the universe of all possible relational graphs. Another example comes from Goodman et al.,[53] who used a probabilistic generative grammar to define a prior on the space of all possible disjunctive-normal-form concepts. The prior implicitly favors simpler concepts, because more complex concepts require application of additional generative steps, each of which can happen only with small probability. They showed that learning by the model captures many aspects of human concept learning.

Intuitively, it seems obvious that people bring prior knowledge to bear when learning new information. But instead of the theorist merely positing a prior and checking whether it can account for human learning, is there a way that the prior can be assayed more directly? One intriguing possibility is suggested by Kalish et al.[54] They showed that as chains of people are successively taught an input–output relationship from the previous learner's examples, the noise in the transmission and learning processes quickly caused the learned relationship to devolve to the prior. In other words, because of the accumulating uncertainty introduced by each successive learner, the prior beliefs came to dominate the acquired representation.

ACTIVE LEARNING: A NATURAL ROLE FOR BAYESIAN MODELS

Until this point in the article, we have emphasized the representational richness afforded by the Bayesian approach, with the only 'process' being the application of Bayes' rule in inference and learning. But another rich extension afforded by explicit representation of hypothesis spaces is models of active learning. In active learning, the learner can intervene upon the environment to select the next datum sampled. For example, experimental scientists select the next experiment they run to extract useful information from the world. This active probing of the environment is distinct from all the examples of learning mentioned previously in the article, which assumed the learner was a passive recipient of data chosen externally, like a sponge stuck to a rock, passively absorbing particles that happen to be sprayed over it by external forces.

One possible goal for active learning is maximal reduction in the uncertainty of beliefs. The learner should probe the environment in such a way that the information revealed is expected to shift beliefs to a relatively narrow range of hypotheses. Nelson[55] described a variety of candidate goals for an active learner. The point is that all of these goals rely on there being a space of hypotheses with a distribution of beliefs. Many aspects of human active learning have been addressed by models of uncertainty reduction. For example, the seemingly illogical choices made by people in the Wason card selection task have been shown to be optimal under certain reasonable priors.[56] The interventions people make to learn about causal networks can be modeled as optimal information gain.[57] And active choice in associative learning tasks can be addressed with different combinations of models, priors, and goals.[38]

CONCLUSIONS

As discussed above, Bayesian models can use a wide variety of assumptions about the representation of prior beliefs, observations, and task goals. Bayesian modeling has been useful to cognitive scientists because it allows these scientists to explore different sets of assumptions and their implications for rational behavior on a task. This helps cognitive scientists understand the nature of a task. When it is found that a Bayesian model closely mimics human cognition on a task, we have a useful explanation of how it is that complex cognition may be possible and why it works as it does, i.e., because it is normatively rational for that type of representational assumption and task.

We have emphasized three types of information processing operations in this article—inference, parameter learning, and structure learning—that result from applying Bayes' rule in different settings because these types of operations occur in both Bayesian models and human cognition. Does the fact that human behavior often seems to be suboptimal suggest that our emphasis on Bayesian operations is misplaced? We think that the answer is no, and there are at least two ways of justifying this answer.

First, some researchers have begun to claim that human behavior may appear to be suboptimal, but that this appearance is misleading. In fact, if cognitive limitations are taken into account, then human behavior is near-optimal.[58] For example, Daw et al.[59] argued that computational constraints may mean that people cannot use Bayes' rule to infer distributions over unobserved variables due to the complexity of the required calculations and, thus, people need to resort to approximate inference methods. Among the infinite variety of possible approximations, people use the 'best' approximations within some class of tractable approximations, where best can be defined in various reasonable ways. The use of these best approximations may help explain suboptimal behaviors. This type of approach to modeling suboptimal behaviors is currently in its infancy, though recent models show promise.[60–63]

Second, a complementary approach to the challenge of modeling suboptimal behavior is to change the level of analysis. Although it may be natural to suppose that individual persons perform Bayesian inference and learning, it is also just as plausible, in principle, that higher and lower levels of organization carry out inference and learning. For example, single neurons can be modeled as carrying out Bayesian inference and learning.[64] At the other extreme, it is not unreasonable to model corporations as inferential and learning entities, perhaps even using Bayesian formalisms. When the behavior being modeled is relatively simple, then the Bayesian computations required for the model are also relatively simple. For example, a Bayesian model of a single neuron might be tractable with a highly accurate approximation. The model remains Bayesian, with all its explanatory appeal. To capture the behavior of complex systems of neurons, what is needed is a way for these locally Bayesian agents to communicate with each other. Kruschke[65] described one reasonable heuristic for hierarchies of locally Bayesian agents to interact. The framework has been applied to an intermediate level of analysis, in which functional components of learning are purely Bayesian, although the system as a whole may exhibit what appears to be non-Bayesian behavior because of limitations on communications between components. Sanborn and Silva[62] described a way to reinterpret Kruschke's[65] scheme in terms of approximate message passing in a globally Bayesian model. Thus, what may be approximately Bayesian at one level of analysis may be Bayesian at a different level of analysis.

In summary, Bayesian models of inference and learning provide a rich domain for formulating theories of cognition. Any representational scheme is allowed, so long as it specifies prior beliefs and observable data in terms of probability distributions. Then the engine of Bayes' rule derives predictions for perception, cognition, and learning in complex situations. Moreover, by maintaining explicit representations of beliefs over plausible hypotheses, the models also permit predictions for active learning and active information foraging. By analyzing the structures of tasks—including the relationships between model assumptions and optimal task performances—cognitive scientists better understand what people's minds are trying to accomplish and what assumptions people may be making when thinking and acting in the world.

NOTES

[a] The statements here are slightly misleading. In practice, principal components are not typically identified via the EM algorithm because practitioners do not consider principal component analysis from a probabilistic viewpoint. However, if principal component analysis is given a probabilistic interpretation,[66,67] then the statements here are accurate.

ACKNOWLEDGEMENTS

We thank P. Dayan and C. Kemp for commenting on earlier versions of this manuscript. This work was supported by NSF research grant DRL-0817250.

REFERENCES

1. Cox RT. *The Algebra of Probable Inference*. Baltimore, MD: Johns Hopkins University Press; 1961.
2. Busemeyer JR, Diederich A. *Cognitive Modeling*. Los Angeles, CA: Sage Publications; 2010.
3. Scarborough D, Sternberg S. *An Invitation to Cognitive Science: Methods, Models and Conceptual Issues*, vol 4. Cambridge, MA: MIT Press; 1998.
4. Polk TA, Seifert CM. *Cognitive Modeling*. Cambridge, MA: MIT Press; 2002.
5. Sun R. *The Cambridge Handbook of Computational Psychology*. New York, NY: Cambridge University Press; 2008.
6. Marr D. *Vision*. New York, NY: Freeman; 1982.
7. Anderson JR. *The Adaptive Character of Thought*. Hillsdale, NJ: Lawrence Erlbaum Associates; 1990.
8. Barlow HB. Possible principles underlying the transformation of sensory messages. In: Rosenblith W, ed. *Sensory Communication*. Cambridge, MA: MIT Press; 1961, 217–234.
9. Chater N, Oaksford M. *The Probabilistic Mind*. Oxford, UK: Oxford University Press; 2008.
10. Geisler WS. Ideal observer analysis. In: Chalupa LM, Werner JS, eds. *The Visual Neurosciences*. Cambridge, MA: MIT Press; 2004, 825–837.
11. Green DM, Swets JA. *Signal Detection Theory and Psychophysics*. New York, NY: John Wiley & Sons; 1966.
12. Griffiths TL, Kemp C, Tenenbaum JB. Bayesian models of cognition. In: Sun R, ed. *The Cambridge Handbook of Computational Psychology*. New York, NY: Cambridge University Press; 2008, 59–100.
13. Kahneman D, Slovic P, Tversky A. *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge, UK: Cambridge University Press; 1982.
14. Knill DC, Richards W. *Perception as Bayesian Inference*. Cambridge, UK: Cambridge University Press; 1996.
15. Oaksford M, Chater N. *Rational Models of Cognition*. Oxford, UK: Oxford University Press; 1999.
16. Todorov E. Optimality principles in sensorimotor control. *Nat Neurosc* 2004, 7:907–915.
17. Neapolitan RE. *Learning Bayesian Networks*. Upper Saddle River, NJ: Pearson Prentice Hall; 2004.
18. Pearl J. *Probabilistic Reasoning in Intelligent Systems*. San Mateo, CA: Morgan Kaufmann; 1988.
19. Russell S, Norvig P. *Artificial Intelligence: A Modern Approach*. 2nd ed. Upper Saddle River, NJ: Prentice Hall; 2003.
20. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis*. 2nd ed. New York, NY: Chapman and Hall; 2003.
21. Calvert GA, Bullmore ET, Brammer MJ, Campbell R, Williams SCR, et al. Activation of auditory cortex during silent lipreading. *Science* 1997, 276:593–596.
22. Pirog Revill K, Aslin RN, Tanenhaus MK, Bavelier D. Neural correlates of partial lexical activation. *Proc Natl Acad Sci USA* 2008, 105:13110–13114.
23. Blakemore S-J, Wolpert D, Frith C. Why can't you tickle yourself? *NeuroReport* 2000, 11:11–15.
24. Yuille AL, Bülthoff HH. Bayesian theory and psychophysics. In: Knill D, Richards W, eds. *Perception as Bayesian Inference*. Cambridge, UK: Cambridge University Press; 1996, 123–161.
25. Ernst MO, Banks MS. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 2002, 415:429–433.
26. Knill DC, Saunders J. Do humans optimally integrate stereo and texture information for judgments of surface slant? *Vision Res* 2003, 43:2539–2558.
27. Alais D, Burr D. The ventriloquist effect results from near-optimal bimodal integration. *Curr Biol* 2004, 14:257–262.
28. Battaglia PW, Jacobs RA, Aslin RN. Bayesian integration of visual and auditory signals for spatial localization. *J Opt Soc Am A* 2003, 20:1391–1397.
29. Ghahramani Z, Wolpert DM, Jordan MI. Computational models of sensorimotor integration. In: Morasso PG, Sanguineti V, eds. *Self-Organization, Computational Maps, and Motor Control*. New York, NY: Elsevier Science; 1997.
30. Jacobs RA. Optimal integration of texture and motion cues to depth. *Vision Res* 1999, 39:3621–3629.
31. Körding KP, Wolpert DM. Bayesian integration in sensorimotor learning. *Nature* 2004, 427:244–247.
32. Landy MS, Maloney LT, Johnston EB, Young M. Measurement and modeling of depth cue combination: in defense of weak fusion. *Vision Res* 1995, 35:389–412.
33. Maloney LT, Landy MS. A statistical framework for robust fusion of depth information. *Visual Communications Image Processing IV*, Proceedings of the SPIE 1199, 1989, 1154–1163.
34. Young MJ, Landy MS, Maloney LT. A perturbation analysis of depth perception from combinations of texture and motion cues. *Vision Res* 1993, 33:2685–2696.
35. Shanks DR. Forward and backward blocking in human contingency judgement. *Q J Exp Psychol* 1985, 37B:1–21.
36. Rescorla RA, Wagner AR. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In: Black AH, Prokasy WF, eds. *Classical Conditioning II: Current Research and Theory*. New York, NY: Appleton-Century-Crofts; 1972.

37. Kruschke JK. Bayesian approaches to associative learning: from passive to active learning. *Learn Behav* 2008, 36:210–226.

38. Dayan P, Kakade S. Explaining away in weight space. In: Leen T, Dietterich T, Tresp V, eds. *Advances in Neural Information Processing Systems*, vol 13. Cambridge, MA: MIT Press; 2001, 451–457.

39. Sobel DM, Tenenbaum JB, Gopnik A. Children's causal inferences from indirect evidence: backwards blocking and Bayesian reasoning in preschoolers. *Cogn Sci* 2004, 28:303–333.

40. Tenenbaum JB, Griffiths TL. Theory-based causal inference. In: Becker S, Thrun S, Obermayer K, eds. *Advances in Neural Information Processing Systems*, vol 15. Cambridge, MA: MIT Press; 2003, 35–42.

41. Courville AC, Daw ND, Touretzky DS. Bayesian theories of conditioning in a changing world. *Trends Cogn Sci* 2006, 10:295–300.

42. Jacobs RA, Shams L. Visual learning in multisensory environments. *Topics Cogn Sci* 2009, 2:217–225.

43. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc B* 1977, 39:1–38.

44. Roweis S, Ghahramani Z. A unifying review of linear Gaussian models. *Neural Comput* 1999, 11:305–345.

45. Robinson RW. Counting labeled acyclic digraphs. In: Harary F, ed. *New Directions in the Theory of Graphs*. New York, NY: Academic Press; 1973, 239–273.

46. MacKay DJC. *Information Theory, Inference, and Learning Algorithms*. Cambridge, UK: Cambridge University Press; 2003.

47. Körding KP, Beierholm U, Ma WJ, Quartz S, Tenenbaum JB, et al. Causal inference in multisensory perception. *PLoS ONE* 2007, 2:e943.

48. Griffiths TL, Sanborn AN, Canini KR, Navarro DJ. Categorization as nonparametric Bayesian density estimation. In: Oaksford M, Chater N, eds. *The Probabilistic Mind: Prospects for Rational Models of Cognition*. Oxford, UK: Oxford University Press; 2009, 303–328.

49. Kemp C, Tenenbaum JB. The discovery of structural form. *Proc Natl Acad Sci USA* 2008, 105:10687–10692.

50. Xu F, Tenenbaum JB. Word learning as Bayesian inference. *Psychol Rev* 2007, 114:245–272.

51. Tenenbaum JB, Griffiths TL, Kemp C. Theory-based Bayesian models of inductive learning and reasoning. *Trends Cogn Sci* 2006, 10:309–318.

52. Tenenbaum JB, Griffiths TL, Niyogi S. Intuitive theories as grammars for causal inference. In: Gopnik A, Schulz L, eds. *Causal Learning: Psychology, Philosophy, and Computation*. Oxford, UK: Oxford University Press; 2007, 301–322.

53. Goodman ND, Tenenbaum JB, Feldman J, Griffiths TL. A rational analysis of rule-based concept learning. *Cogn Sci* 2008, 32:108–154.

54. Kalish ML, Griffiths TL, Lewandowsky S. Iterated learning: intergenerational knowledge transmission reveals inductive biases. *Psychon Bull Rev* 2007, 14:288–294.

55. Nelson JD. Finding useful questions: on Bayesian diagnosticity, probability, impact, and information gain. *Psychol Rev* 2005, 112:979–999.

56. Oaksford M, Chater N. Optimal data selection: revision, review, and reevaluation. *Psychon Bull Rev* 2003, 10:289–318.

57. Steyvers M, Tenenbaum JB, Wagenmakers E-J, Blum B. Inferring causal networks from observations and interventions. *Cogn Sci* 2003, 27:453–489.

58. Gigerenzer G, Todd PM, The ABC Research Group. *Simple Heuristics that Make Us Smart*. New York, NY: Oxford University Press; 1999.

59. Daw ND, Courville AC, Dayan P. Semi-rational models: the case of trial order. In: Chater N, Oaksford M, eds. *The Probabilistic Mind*. Oxford, UK: Oxford University Press; 2008, 431–452.

60. Brown SD, Steyvers M. Detecting and predicting changes. *Cogn Psychol* 2009, 58:49–67.

61. Sanborn A, Griffiths TL, Navarro DJ. A more rational model of categorization. *Proceedings of the 28th Annual Conference of the Cognitive Science Society*. Vancouver, Canada, 2006.

62. Sanborn A, Silva R. A machine learning perspective on the locally Bayesian model. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*. Amsterdam, The Netherlands, 2009.

63. Shi L, Feldman NH, Griffiths TL. Performing Bayesian inference with exemplar models. *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, 2008.

64. Deneve S. Bayesian spiking neurons II: learning. *Neural Comput* 2008, 20:118–145.

65. Kruschke JK. Locally Bayesian learning with applications to retrospective revaluation and highlighting. *Psychol Rev* 2006, 113:677–699.

66. Roweis S. EM algorithms for PCA and SPCA. In: Jordan MI, Kearns MJ, Solla SA, eds. *Advances in Neural Information Processing Systems 10*. Cambridge, MA: MIT Press; 1998.

67. Tipping ME, Bishop CM. Probabilistic principal component analysis. *J R Stat Soc B* 1999, 21:611–622.