International Scholarly Research Network
ISRN Biomathematics
Volume 2012, Article ID 785791, 19 pages
doi:10.5402/2012/785791

Review Article

Bayesian Models of Brain and Behaviour

William Penny

Wellcome Trust Centre for Neuroimaging, University College, London WC1N 3BG, UK

Correspondence should be addressed to William Penny, w.penny@ucl.ac.uk

Received 5 August 2012; Accepted 24 September 2012

Academic Editors: C. Brown, E. V. Ignatieva, E. Pereda, L. Pezard, and A. Riva

This paper presents a review of Bayesian models of brain and behaviour. We first review the basic principles of Bayesian inference. This is followed by descriptions of sampling and variational methods for approximate inference, and forward and backward recursions in time for inference in dynamical models. The review of behavioural models covers work in visual processing, sensory integration, sensorimotor integration, and collective decision making. The review of brain models covers a range of spatial scales from synapses to neurons and population codes, but with an emphasis on models of cortical hierarchies. We describe a simple hierarchical model which provides a mathematical framework relating constructs in Bayesian inference to those in neural computation. We close by reviewing recent theoretical developments in Bayesian inference for planning and control.

1. Introduction

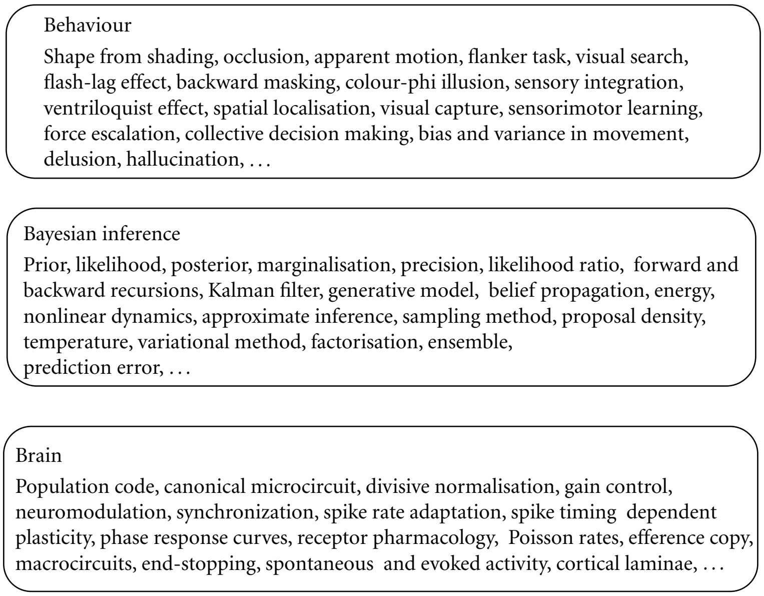
This paper presents a review of Bayesian models of brain and behaviour. Overall, the aim of the paper is to review work which relates constructs in Bayesian inference to aspects of behaviour and neural computation, as outlined in Figure 1. This is a very large research area and we refer readers to standard textbooks and other review materials [1–6].

One of the main ideas to emerge in recent years is that Bayesian inference operates at the level of cortical macrocircuits. These circuits are arranged in a hierarchy which reflects the hierarchical structure of the world around us. The idea that the brain encodes a model of the world and makes predictions about its sensory input is also known as predictive coding [7].

Consider, for example, your immediate environment. It may be populated by various objects such as desks, chairs, walls, trees, and so forth. Generic attributes of this scene and the objects in it will be represented by activity in brain regions near the top of the hierarchy. The connections from higher to lower regions then encode a model of your world, describing how scenes consist of objects, and objects by their features. If a higher level representation is activated, it will activate those lower level representations that encode the presence of, for example, configurations of oriented lines that your brain expects to receive signals about in early visual cortex.

At the lowest level of the hierarchy these predictions are compared with sensory input and the difference between them, the prediction error, is propagated back up the hierarchy. This happens simultaneously at every hierarchical level. Predictions are sent down and prediction errors back up. It is important to emphasize that this is a dynamic process. Upon entering a new environment, such as a room in a house, higher level schemas will activate the likely presence of objects or people that one expects to encounter in that room. Initially, lower-level prediction errors are likely to be large. These will change activations in higher level regions, as you find that your keys were not on the kitchen table after all. Neuronal populations that initially encoded the likely presence of a key become less active.

The overall process is expressed clearly by Fletcher and Frith [8]: "...these systems are arranged in a hierarchy so that the prediction error emitted by a lower-level system becomes the input for a higher-level system. At the same time, feedback from the higher level system provides the prior beliefs for the lower level system. In this framework, the prediction error signal is a marker that the existing model or inference has not fully accounted for the input. A readjustment at the next level in the hierarchy may increase



Behaviour

Shape from shading, occlusion, apparent motion, flanker task, visual search, flash-lag effect, backward masking, colour-phi illusion, sensory integration, ventriloquist effect, spatial localisation, visual capture, sensorimotor learning, force escalation, collective decision making, bias and variance in movement, delusion, hallucination, . . .

Bayesian inference

Prior, likelihood, posterior, marginalisation, precision, likelihood ratio, forward and backward recursions, Kalman filter, generative model, belief propagation, energy, nonlinear dynamics, approximate inference, sampling method, proposal density, temperature, variational method, factorisation, ensemble, prediction error, . . .

Brain

Population code, canonical microcircuit, divisive normalisation, gain control, neuromodulation, synchronization, spike rate adaptation, spike timing dependent plasticity, phase response curves, receptor pharmacology, Poisson rates, efference copy, macrocircuits, end-stopping, spontaneous and evoked activity, cortical laminae, . . .

Figure 1: This paper reviews work which relates constructs in Bayesian inference to those in experimental psychology and neuroscience.

the accuracy and reduce the prediction error. But if it does not, higher-level readjustments are required. Higher levels provide guidance to lower levels and ensure an internal consistency of the inferred causes of sensory input at multiple levels."

Predictive coding models comprised of multiple hierarchical levels are rather complex, however, when compared to much of the work in Bayesian modelling of brain and behaviour. We therefore structure our review to first focus on models of simple behaviours, and Bayesian models of simple computations in synapses, neurons, and neural populations before leading up to a more in depth review of Bayesian inference in cortical macrocircuits in Section 5.

Section 2 reviews concepts in Bayesian inference. This includes the basic principle underlying Bayes rule. For realistic models exact Bayesian inference is impossible, so we briefly describe two of the leading frameworks for approximate inference; sampling and variational methods. We also describe the temporal forward and backward recursions for inference in dynamical models.

Section 3 reviews behavioural models. This covers work in visual processing, sensory integration, sensorimotor integration, and collective decision making. Section 3.2 also describes how visual perceptions can depend on later sensory events, so-called postdiction [9]. It may therefore be the case that perceptions are based on both forwards and backwards inference in time.

The review of brain models in Section 4 covers a range of spatial scales from synapses to neurons and population codes. Section 5 describes models of cortical hierarchies. This is based on early work by Mumford [10], Rao and Ballard [7] and a more recent series of papers by Friston [1, 11, 12]. We describe a simple hierarchical model which provides a mathematical framework relating quantities in Bayesian inference to those in neural computation. Finally, we very briefly review recent theoretical developments in Bayesian inference for planning and control in Section 6 and close with a discussion in Section 7.

The main sections of the paper can be read in any order, so expert readers can skip to relevant sections. It is perhaps not necessary to fully understand the mathematical parts of Section 2, but they are included to provide a mathematical backbone onto which the discussion of models is later referred.

2. Bayesian Inference

It has been proposed that aspects of human behaviour are governed by statistical optimality principles, and that the brain itself is a statistical inference machine [4]. In statistics the optimal way of updating your beliefs is via Bayes rule.

Consider some quantity, x. Our beliefs about the likely values of x can be described by the probability distribution $p(x)$. If we then make a new observation y that is related to x, then we can update our belief about x using Bayesian inference.

First we need to specify the likelihood of observing y given x. This is specified by a probability distribution called the likelihood, $p(y \mid x)$. It tells us, if we know x, what are the likely values of y. Our updated belief about x, that is, after observing the new data point y is given by the posterior distribution $p(x \mid y)$. This can be computed via Bayes rule

$$p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)}. \tag{1}$$

The denominator ensures that $p(x \mid y)$ sums to 1 over all possible values of x, that is it is a probability distribution. It can be written as

$$p(y) = \int p(y \mid x')p(x')dx'. \tag{2}$$

Equations (1) and (2) describe the basic computations that underly Bayes rule. These are *multiplication*, *normalisation* (1), and *marginalisation* (2). Wolpert and Ghahramani [13] use the game of tennis to illustrate key points. Imagine that you are receiving serve. One computation you need to make before returning serve is to estimate, x, the position of the ball when it first hits the ground. This scenario is depicted in Figure 2.

It is possible to make an estimate solely on the basis of the balls trajectory, that is via the data y. We can find the value of x which maximises the likelihood, $p(y \mid x)$. This is known as Maximum Likelihood (ML) estimation. It is also possible to estimate the uncertainty in this estimate. The ML estimate and the uncertainty in it together give rise to the likelihood distribution shown in Figure 2.

But before our opponent hits the ball we may have a fair idea as to where they will serve. It may be the case, for example, that when they serve from the right the ball tends to go down the line. We can summarise this belief by the prior distribution $p(x)$ (shown in blue in Figure 2). We can then use Bayes rule to estimate the posterior distribution. This is the optimal combination of prior knowledge ("down the line") and new data (visual information from the ball's trajectory). Our final single best estimate of where the ball will land is then given by the maximum of the posterior density. This is known as MAP estimation (from "maximum a posteriori").

As we continue to see the ball coming toward us we can refine our belief as to where we think the ball will land. This can be implemented by applying Bayes rule recursively such that our belief at time point n depends only on our belief at the previous time point, $n-1$. That is

$$p(x_n \mid Y_n) = \frac{p(y_n \mid x_n)\, p(x_n \mid Y_{n-1})}{p(Y_n)}, \tag{3}$$

where $Y_n = \{y_1, y_2, \ldots, y_n\}$ denotes all observations up to time n. Our prior belief, that is, prior to observing data point y_n is simply the posterior belief after observing all data points up to time $n-1$, $p(x_n \mid Y_{n-1})$. Colloquially, we say that "today's prior is yesterday's posterior". The variable x is also referred to as a hidden variable because it is not directly observed.

If the hidden state was a discrete variable, such as whether the ball landed in or out of the service box, one can form a likelihood ratio

$$\mathrm{LR} = \frac{p(x_n = \mathrm{IN} \mid Y_n)}{p(x_n = \mathrm{OUT} \mid Y_n)}. \tag{4}$$

Decisions based on the likelihood ratio are statistically optimal in the sense of having maximum sensitivity for any given level of specificity. In contexts where LR is recursively updated these decisions correspond to a sequential likelihood ratio test [14]. There is a good deal of evidence showing that the firing rate of single neurons in the brain report evolving log LR values [15] (see section on "Neurons" below).

2.1. Gaussians. If our random variables x and y are normally distributed then Bayesian inference can be implemented exactly using simple formulae. These are most easily expressed in terms of precisions, where the precision of a random variable is its inverse variance. A precision of 10 corresponds to a variance of 0.1. We first look at inference for a single univariate measure (e.g., distance from side of tennis court). For a Gaussian prior with mean m_0 and precision λ_0, and a Gaussian likelihood with mean m_D and precision λ_D the posterior distribution is Gaussian with mean m and precision λ

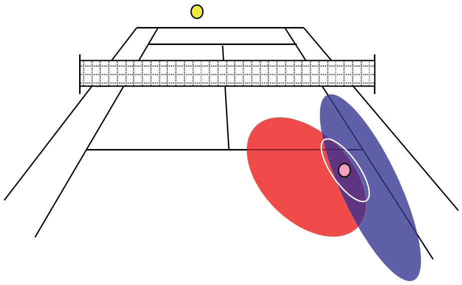


FIGURE 2: Estimating the position of the ball when it first lands. The prior is shown in blue, the likelihood distribution in red, and the posterior distribution with the white ellipse. The maximum posterior estimate is shown by the magenta ball. This estimate can be updated in light of new information about the balls trajectory (yellow). Adapted from Wolpert and Ghahramani [13].

$$\begin{aligned} \lambda &= \lambda_0 + \lambda_D, \\ m &= \frac{\lambda_0}{\lambda} m_0 + \frac{\lambda_D}{\lambda} m_D. \end{aligned} \tag{5}$$

So, precisions add and the posterior mean is the sum of the prior and data means, but each weighted by their relative precision. This relationship is illustrated in Figure 3. Though fairly simple, (5) shows how to optimally combine two sources of information. As we shall see in Section 3, various aspects of human behaviour from cue integration to instances of collective decision making have been shown to conform to this "normative model". Similar formulae exist for multivariate (instead of univariate) Gaussians [16] where we have multidimensional hidden states and observations, for example three-dimensional position of the ball and two-dimensional landing position on court surface.

2.2. Generative Models. So far we have discussed the relationship between a single hidden variable x and a single-observed variable y. More generally, we may have multiple hidden variables, for example, representing different levels of abstraction in cortical hierarchies, and multiple observed variables from different sensory modalities. These more complicated probabilistic relationships can be represented using probabilistic generative models and their associated graphical models [16, 17]. If these models do not have cycles they are referred to as Directed Acyclic Graphs (DAGs). A DAG specifies the joint probability of all variables, $x = [x_1, x_2, \ldots, x_H]$. This can be written down as

$$p(x) = \prod_{i=1}^{H} p(x_i \mid \mathrm{pa}[x_i]), \tag{6}$$

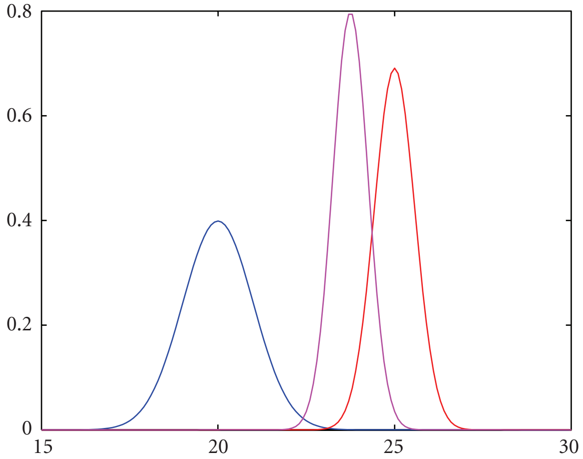


Figure 3: Bayes rule for Gaussians. For the prior $p(x)$ (blue) $m_0 = 20$, $\lambda_0 = 1$ and the likelihood $p(y \mid x)$ (red) $m_D = 25$ and $\lambda_D = 3$, the posterior $p(x \mid y)$ (magenta) shows the posterior distribution with $m = 23.75$ and $\lambda = 4$. The posterior is closer to the likelihood than the prior because the likelihood has higher precision. Bayes rule for Gaussians has been used to explain many behaviours from sensory integration to collective decision making.

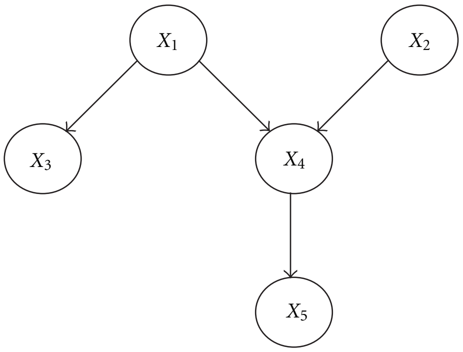


Figure 4: An example of a Directed Acyclic Graph (DAG). This tells us we can write the joint density over all variables as $p(x) = p(x_1)p(x_2)p(x_3 \mid x_1)p(x_4 \mid x_1, x_2)p(x_5 \mid x_4)$. DAGs provide a graphical shorthand for specifying Bayesian generative models.

where pa$[x_i]$ are the parents of x_i. For example, for the generative model in Figure 4 we have

$$p(x) = p(x_1)p(x_2)p(x_3 \mid x_1)p(x_4 \mid x_1, x_2)p(x_5 \mid x_4). \quad (7)$$

All other probabilities can be obtained from the joint probability via marginalisation. For example,

$$p(x_4) = \int\int\int\int p(x_1, x_2, x_3, x_4, x_5) dx_1\, dx_2\, dx_3\, dx_5. \quad (8)$$

They are therefore referred to as marginal probabilities. If one of the variables is known, for example, x_1 may be a sensory input, then the marginalisation operation will produce a posterior density

$$p(x_4 \mid x_1) = \iiint p(x_1, x_2, x_3, x_4, x_5) dx_2\, dx_3\, dx_5. \quad (9)$$

In hierarchical models of cortical macrocircuits, for example, x_4 may correspond to activity in a higher level brain region (see Section 5). The above equation then tells us how to estimate x_4 given sensory input x_1.

If multiple marginal or posterior probabilities need to be computed this is most efficiently implemented using the belief propagation algorithm [18], which effectively defines an ordering on the DAG and passes the results of marginalisations between nodes. As we shall see in Section 4, a number of researchers have proposed how belief propagation can be implemented in neural circuits [19, 20].

A central quantity in Bayesian modelling is the negative log likelihood of the joint density, which is often referred to as the energy

$$E(x) = -\log p(x). \quad (10)$$

Values of the variables x with high joint probability have low energy, and inference can be viewed as an energy minimisation process. Values with minimal energy have maximum joint probability. Because posterior densities are simply normalised joint densities then minimal energy values also have Maximum a Posterior (MAP) probability. As we shall see in Section 5, an MAP, or energy minimisation approach, has been used to derive predictive coding algorithms [7].

2.3. Approximate Inference. In most interesting models there is no way to implement exact Bayesian inference. That is, for most nonlinear and/or non-Gaussian models there are no analytic formulae for computing posterior densities. Instead we must resort to approximate inference. There are two basic approaches (i) sampling methods [21], or (ii) deterministic approximation methods [16]. The most popular deterministic methods are Laplace approximations or variational inference. Generally, deterministic methods are advantageous in being much faster but have the potential disadvantage of producing only locally optimal solutions.

As we shall see in Section 5, it has been proposed that cortical brain regions represent information at different levels of abstraction, and that top-down connections instantiate the brains generative model of the world, and bottom-up processing its algorithm for approximate inference. We now briefly review two different approximate inference methods.

2.3.1. Sampling Methods. We assume our goal is to produce samples from the multivariate posterior density $p(x \mid y)$, where y is sensory data, and x are hidden variables of interest, such as activities of neurons in a network. These samples will then provide a representation of the posterior. From this, quantities such as the posterior mean can be computed by simply taking the mean of the samples.

One of the simplest sampling methods is Gibbs sampling [21] which works as follows. We pick a variable x_i and generate a sample from the distribution $p(x_i \mid x_{\backslash i}, y)$, where $x_{\backslash i}$ are all the other variables. We then loop over i, repeat this process a large number of times, and the samples near the end of this process (typically the last half) will be from the desired posterior $p(x \mid y)$. In general, it may not be possible to easily sample from $p(x_i \mid x_{\setminus i}, y)$. This limits the applicability of the approach, but it is highly efficient for many hierarchical models [21].

A more generic procedure is Metropolis-Hastings (MH) which is a type of Markov Chain Monte Carlo (MCMC) procedure [21]. MH makes use of a proposal density $q(x'; x)$ which is dependent on the current state vector x. For symmetric q (such as a Gaussian) samples from the posterior density can be generated as follows. First, start at a point x_1 sampled from the prior, then generate a proposal x' using the density q. This proposal is then accepted with probability $\min(1, r)$, where

$$r = \frac{p(y \mid x')p(x')}{p(y \mid x)p(x)}. \tag{11}$$

If the step is accepted we set $x_{n+1} = x'$. If it is rejected we set $x_{n+1} = x_n$ (our list of samples can have duplicate entries). This procedure is guaranteed to produce samples from the posterior as long as we run it for long enough, and there are various criteria that can be used to monitor convergence [21].

Equation (11) says we should always accept a new sample if it has higher posterior probability than the last. Because it allows occasional transitions to less probable states it can avoid locally optimal solutions. To increase the likelihood of finding globally optimal solutions it is possible to run multiple chains at different temperatures and use a proposal density to switch between them [22]. We will refer to this idea again in Section 4.3.2 where we suggest that the different temperatures may be controlled in the brain via neuromodulation.

These sample-based approaches were used in early neural network models such as the Boltzmann machine and the more recent Deep Belief Networks reviewed in Section 4.4. As we shall see in Section 4.3.2 Gershmann et al. [23] have shown how MCMC can be used to account for perceptual multistability.

2.3.2. Variational Methods. If our variables comprise sensor data y and unknown hidden variables x then we can define the free energy as

$$F = -\int q(x) \log p(y, x) dx - \int q(x) \log \frac{1}{q(x)} dx, \tag{12}$$

where the first term is the average energy, and the average is taken with respect to the density $q(x)$, and the second term is the entropy of $q(x)$. Given this definition we can write the log marginal likelihood of the data as

$$\log p(y) = -F + \mathrm{KL}(q(x) \| p(x \mid y)), \tag{13}$$

where KL( ) is the Kullback-Liebler divergence measure [24]. KL is zero if the densities are equal and is otherwise positive, with larger values reflecting degree of dissimilarity. Given that the term on the left is fixed, we can minimise the KL divergence term by minimising the free energy. This will give us an approximate posterior $q(x)$ that is optimal in the sense of minimising KL divergence with the true posterior.

To obtain a practical learning algorithm we must also ensure that the integrals in (12) are tractable. One generic procedure for attaining this goal is to assume that the approximating density factorizes over groups of variables. In physics, this is known as the mean field approximation. Thus, we consider

$$q(x) = \prod_i q(x_i), \tag{14}$$

where x_i is the ith group of variables. We can also write this as

$$q(x) = q(x_i) q(x_{\setminus i}), \tag{15}$$

where $x_{\setminus i}$ denotes all variables *not* in the ith group. We then define the variational energy for the ith partition as

$$I(x_i) = -\int q(x_{\setminus i}) \log p(y, x) dx_{\setminus i} \tag{16}$$

and note that F is minimised when

$$q(x_i) = \frac{\exp[I(x_i)]}{Z}, \tag{17}$$

where Z is the normalisation factor needed to make $q(x_i)$ a valid probability distribution. This gives us a recipe for approximate inference in which we update the posteriors $q(x_i)$ in turn. This is much like Gibbs sampling, but we update sufficient statistics (e.g., mean and variance) rather than produce samples.

As we described in Section 2.2, point estimates of variables, such as the MAP estimates, can be found by minimising energy. But this does not tell us about the uncertainty in these variables. To find out this uncertainty we can find the distribution $q(x)$ that minimises the *free* energy. Out of all the distributions which minimise energy, the one that minimises free energy has maximal uncertainty (see (12)). That is, we are minimally committed to specific interpretations of sensory data, in accordance with Jaynes' principle of maximum entropy [24].

Readers can learn more about variational inference in standard tutorials [16, 25, 26]. We will later refer to variational inference in the context of the Helmholtz machine [27], in Section 4.4, and the free energy principle [12, 28] in Section 5.2.

2.4. Dynamic Models. In previous sections we have considered generative models for potentially multiple and multidimensional hidden variables and observations. Going back to the tennis example, I will receive high-dimensional visual observations from which I may wish to infer two hidden variables; the two-dimensional position on court where the ball will land and the position of my opponent.

We now consider models with an explicit dynamic component. A broad class of dynamical models are the discrete time nonlinear state-space models of the form

$$\begin{aligned} x_n &= f(x_{n-1}, u_{n-1}) + w_n, \\ y_n &= g(x_n, u_n) + e_n, \end{aligned} \tag{18}$$

where x_n are the hidden variables, y_n are the observations, u_n is a control input, w_n is state noise, and e_n is observation noise. All of these quantities are vectors. This is a Nonlinear Dynamical System (NDS) with inputs and hidden variables. The function $f()$ is a flow term which specifies the dynamics, and $g()$ specifies the mapping from hidden state to observations. The above two equations define the state transition density $p(x_n \mid x_{n-1})$ and the observation density $p(y_n \mid x_n)$ (to simplify the notation we have dropped the dependence on u_n, but this is implied).

We denote the trajectories or sequences of observations, states, and controls using $Y_n = \{y_1, y_2, \dots, y_n\}$, $X_n = \{x_1, x_2, \dots, x_n\}$, and $U_n = \{u_1, u_2, \dots, u_n\}$. Dynamical models of the above form are important for understanding, for example, Bayesian inference as applied to sensorimotor integration, as described in Section 3.3. In this context, u_n would be a copy of a motor command known as an "efference copy". The dynamical model would then allow an agent to predict the consequences of its actions.

These models can be inverted, that is, we can estimate x_n from Y_n using forward inference. This is depicted in Figure 5 and described mathematically in the following subsection. As we shall see in Section 4, Helmholtz has proposed that perception corresponds to unconscious statistical inference and this has become a working hypothesis for a modern generation of computational neuroscientists. Thus we have labelled inference about x_n as "perception" in Figure 5.

2.4.1. Forward Inference. The problem of estimating the states given current and previous observations is solved using forwards inference. This produces the marginal densities $p(x_n \mid Y_n)$. The forward inference problem can be solved in two steps. The first step is a *Time Update* or prediction step

$$p(x_n \mid Y_{n-1}) = \int p(x_n \mid x_{n-1}) p(x_{n-1} \mid Y_{n-1}) dx_{n-1}. \quad (19)$$

The second step is a *Measurement Update* or correction step

$$p(x_n \mid Y_n) = \frac{p(y_n \mid x_n) p(x_n \mid Y_{n-1})}{\int p(y_n \mid x_n) p(x_n \mid Y_{n-1}) dx_n} \quad (20)$$

which is Bayes rule with prior $p(x_n \mid Y_{n-1})$ from the time update, and likelihood $p(y_n \mid x_n)$.

For Linear Dynamical Systems (LDS), where $f()$ and $g()$ in (18) are linear, forward inference reduces to Kalman Filtering [29]. As we shall see, Beck et al. [30] have shown how Kalman filtering can be implemented using a population of spiking neurons. For Nonlinear Dynamical Systems (NDS), approximate forward inference can be instantiated using an Extended Kalman Filter (EKF). Alternative sample-based forward inference schemes can be implemented using particle filtering. Lee and Mumford have proposed how inference in visual cortical hierarchies can proceed using particle filtering [31].

2.4.2. Backward Inference. As we shall see, backward inference is important for postdiction (predictions about the past—see section on visual processing) and for planning and control (see Section 6). We define the posterior probability of state x_n given *all* observations up to time point N as

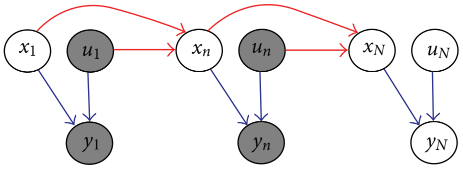


FIGURE 5: Perception as forwards inference over states. In this and subsequent figures, the gray shading indicates a known variable. Perception here corresponds to estimation of hidden state density $p(x_n \mid U_n, Y_n)$ given known motor efference copy U_n and sensory input Y_n. Here and in later figures, the red arrows indicate temporal dependencies, and U_n and Y_n indicate sequences up to time n (see main text). These dynamical models have been used to explain sensorimotor integration and sensorimotor learning.

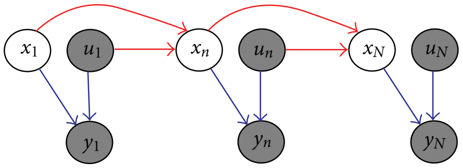


FIGURE 6: Perception as forward and backwards inference over states. Perception here corresponds to estimation of hidden state density $p(x_n \mid U_N, Y_N)$ given known motor efference copy U_N and sensory input Y_N. Here, forward estimates about previous states x_n (i.e., from forward inference) can be improved upon using more recent efference copy $u_{n+1}, \dots, u_N$ and sensory information $y_{n+1}, \dots, y_N$. These so-called postdictive estimates may be useful in, for example, visual perception.

$$\gamma(x_n) = p(x_n \mid Y_N). \quad (21)$$

This can be computed recursively using

$$\gamma(x_n) = \int p(x_n \mid x_{n+1}, Y_n) \gamma(x_{n+1}) dx_{n+1}. \quad (22)$$

The first term in the integral can be thought of as a reverse flow term and is computed using Bayes rule

$$p(x_n \mid x_{n+1}, Y_n) = \frac{p(x_{n+1} \mid x_n, Y_n) p(x_n \mid Y_n)}{\int p(x_{n+1} \mid x_n, Y_n) p(x_n \mid Y_n) dx_n}. \quad (23)$$

Importantly, this form of backward inference (the so-called gamma recursions) can be implemented without requiring storage of the observations y_n. These gamma recursions can therefore be implemented online, which is important for a potential neuronal implementation. Backward inference is represented graphically in Figure 6.

Similar backwards recursions can be derived to estimate the control signals $p(u_n \mid x_1, Y_n)$ given initial state values x_1 and desired sensory observations. This is depicted in Figure 7 and is important for planning and control as we discuss in Section 6. We envisage that backwards inference operates over short time scales for perception (tens of ms) and much longer time scales for planning and cognition.

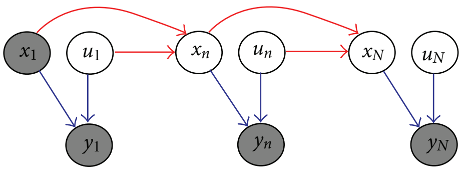


Figure 7: Planning as forward and backwards inference over states and controls. Planning can be formulated as estimation of a density over actions $p(U_N \mid x_1, Y_N)$ given current state x_1 and desired sensory consequences, Y_N.

Readers can find out more about forwards and backward inference for dynamical models in standard textbooks [16]. It is also worth noting that here we are referring to forward and backward recursions in *time*. This should not be confused with forward and backward message passing in hierarchical models as described in Section 5.

2.4.3. Parameter Estimation. Dynamical systems models also depend on unknown parameters, θ. These will parameterise the dynamical function $f()$ and the observation function $g()$. These parameters can be estimated using variational methods, for example, for LDS [32] or NDS [33] or using sampling methods [34, 35]. As we shall seee in Section 4, learning in computational models of the brain can be formulated as parameter estimation in Bayesian models.

3. Behavioural Models

An attractive feature of Bayesian models of behaviour is that they provide descriptions of what would be optimal for a given task. They are often referred to as "ideal observer" models because they quantify how much to update our beliefs in light of new evidence. Departures from these "normative models" can then be explained in terms of other constraints such as computational complexity or individual differences. One way to address individual differences is to use an Empirical Bayesian approach in which parameters of priors and their parametric forms are estimated from data. See [36] for an example of this approach in modelling visual motion processing.

What follows in this section is a review of Bayesian models of sensory integration, visual processing, sensorimotor integration, and collective decision making. As we shall see, the priors that we have about, for example, our visual world most readily show themselves in situations of stimulus ambiguity or at low signal-to-noise ratios. Much of the phenomenology of these perceptual illusions is long established [37], but Bayesian modelling provides new quantitative explanations and predictions. A more introductory review of much of this material is available in Frith's outstanding book on mind and brain [3].

3.1. Sensory Integration. Ernst and Banks [38] considered the problem of integrating information from visual and tactile (haptic) modalities. If vision v and touch t information are independent given an object x then Bayesian fusion of sensory information produces a posterior density

$$p(x \mid v, t) = \frac{p(v \mid x)p(t \mid x)p(x)}{p(v, t)}. \tag{24}$$

For a uniform prior $p(x)$ and for Gaussian likelihoods, the posterior will also be a Gaussian with precision λ_{vt}. From Bayes rule for Gaussians (5) we know that precisions add

$$\lambda_{vt} = \lambda_v + \lambda_t, \tag{25}$$

where λ_v and λ_t are the precision of visual and haptic senses alone, and the posterior mean is a relative-precision weighted combination

$$m_{vt} = \frac{\lambda_v}{\lambda_{vt}} m_v + \frac{\lambda_t}{\lambda_{vt}} m_t \tag{26}$$

or

$$m_{vt} = w_v m_v + w_t m_t \tag{27}$$

with weights w_v and w_t. Ernst and Banks [38] asked subjects which of two sequentially presented blocks was the taller. Subjects used either vision alone, touch alone, or a combination of the two. They recorded the accuracy with which discrimination could be made and plotted this as a function of difference in block height. This was repeated for each modality alone and then both together. They also used various levels of noise on the visual images. From the single modality discrimination curves they then fitted cumulative Gaussian density functions, which provided estimates of the precisions λ_t and $\lambda_v(i)$ where i indexes visual noise levels. In the dual modality experiment the weighting of visual information predicted by Bayes' rule for the ith level of visual noise is

$$\hat{w}_v(i) = \frac{\lambda_v(i)}{\lambda_v(i) + \lambda_t}. \tag{28}$$

This was found to match well with the empirically observed weighting of visual information. They observed visual capture at low levels of visual noise and haptic capture at high levels. Inference in this simple Bayesian model is consistent with standard signal detection theory [39], however, Bayesian inference is more general as it can accommodate, for example, nonuniform priors over block height.

There have been numerous studies of the potential role of Bayesian inference for integration of other senses. For example, object localisation using visual and auditory cues in the horizontal [40] and depth [41] planes has supported a Bayesian integration model with vision dominating audition in most ecologically valid contexts. This visual capture is the basis of the "ventriloquism" effect, but is rapidly degraded with visual noise. This literature has considered only simple inferences about single variables such as block height or spatial location. Nevertheless these studies have demonstrated a fundamental concept; that sensory integration is near Bayes-optimal.

3.2. Visual Processing. Kersten et al. [42] review the problem of visual object perception and argue that much of the ambiguity in visual processing, for example concerning occluded objects, can be resolved with prior knowledge. This idea is naturally embodied in a Bayesian framework [43] and has its origins in the work of Helmholtz who viewed perception as "unconscious inference." An example is how the inference of shape from shading is informed by a "light-from-above" prior. This results in circular patches which are darker at the bottom being perceived as convex. The adaptability of this prior, and subsequent perceptual experience, has been demonstrated by Adams et al. [44].

An example of such a Bayesian modelling approach is the work of Yu et al. [45] who propose a normative model for the Eriksen Flanker task. This simple decision making task was designed to probe neural and behavioural responses in the context of conflicting information. On each trial, three visual stimuli are presented and subjects are required to press a button depending on the identity of the central stimulus. The flanking stimuli are either congruent or incongruent. Yu et al. proposed a discrete time ideal observer model that qualitatively captured the dynamics of the decision making process. This used the recursive form of Bayes rule in (3). In later work, a continuum time limit of this model was derived [46]. This produced semianalytic predictions of reaction time and error rate which provided accurate numerical fits to subject behaviour. They also proposed an algorithm for how these models could be approximately implemented in a neural network [45], which we will refer to later (see Section 5).

Weiss et al. [47] propose that many motion illusions arise from the result of Bayes-optimal processing of ecologically invalid stimuli. Their model was able to reproduce a number of psychophysical effects based on the simple assumptions that measurements are noisy and the visual system has a prior which expects slower movements to be more likely than faster ones. For example, the model could predict the direction of global motion of simple objects such as rhomboids, as a function of contrast and object shape. This model was later refined [36] by showing the prior to be non-Gaussian and subject specific, and that measurement noise variance was inversely proportional to visual contrast.

Najemnik and Geisler developed an ideal Bayesian observer model of visual search for a known target embedded in a natural texture [48]. Prior beliefs in target location were updated to posterior beliefs using a likelihood term that reflected the foveated mapping properties of visual cortex. When this likelihood was matched to individual subjects discrimination ability, the resulting visual searches were nearly optimal in terms of the median number of saccades. Later work [49] showed that fixation statistics were also similar to the ideal observer.

If the world we perceive is the result of hierarchical processing in cortical networks then, because this processing may take some time (of the order of 100 ms), what is perceived to be the present could actually be the past. As this would obviously be disadvantageous for the species, it has been argued that our perceptions are based on predictive models. A 50 ms delay in processing could be accommodated by estimating the state of the world 50 ms in the future. There is much experimental evidence for this view [50]. However, a purely "predictive" account fails to accommodate recent findings in visual psychophysics. The flash-lag effect, for example, is a robust visual illusion whereby a flash and a moving object that are located in the same position are perceived to be displaced from one other. If the object stops moving at the time of the flash, no such displacement is perceived. This indicates that the position of the object after the flash affects our perception of where the flash occurred. This "postdictive" account explains the phenomenon [9], and related data where the object reverses its direction at the flash time. A simple Bayesian model has been proposed to account for the activity of V4 neurons in this task [51]. Later experimental work found evidence for a linear combination of both predictive and postdictive mechanisms [52].

Related phenomena include backward masking [53] and the colour-phi illusion [54]. Here, two coloured dots are presented one followed quickly by the other and in close spatial proximity. This gives rise to a perception of movement and of the color changing in the middle of the apparent trajectory. Because the viewer cannot know the color of the second dot until it appears, the percept attributed to the time of the trajectory must be formed in retrospect. This postdictive account motivated Dennett [55] to propose his multiple drafts theory of consciousness. However, these phenomena are perhaps more simply explained by forwards and backwards inference in dynamic Bayesian networks (see Figure 6 and Section 2.4).

3.3. Sensorimotor Integration. Wolpert et al. [56] have examined the use of dynamic Bayesian models, also referred to as forward models, for sensorimotor integration. These models are given generically by (18) where x_n is the current state, u_n is a copy of a motor command, y_n are sensory observations, and w_n and e_n are state and observation noise.

Inference in these models proceeds as described in Section 2.4.1. First, the dynamical equation describing state transitions is integrated to create an estimate of the next state. This requires as input a copy of the current motor command (so-called efference copy) and the current state. In terms of Bayesian updates in dynamical models (see earlier) this corresponds to the time update or prediction step. A prediction of sensory input can then be made based on the predicted next state and the mapping from x_n to y_n. Finally, a measurement update or correction step can be applied which updates the state estimate based on current sensory input.

Wolpert et al. cite a number of key features of dynamic Bayesian models including the following. First, they allow outcomes of actions to be predicted and acted upon before sensory feedback is available. This may be important for rapid movements. Second, they use efference copy to cancel the sensory effects of movement ("reafference"), for example, the visual world is stable despite eye movements. Third, simulation of actions allows for mental rehearsal which can potentially lead to improvements in movement accuracy.

This framework was applied to the estimation of arm position using proprioceptive feedback and a forward model based on a linear dynamical system [56]. Inference in this model was then implemented using a Kalman filter. The resulting bias and variance in estimates of arm position were shown to closely correspond to human performance, with proprioceptive input becoming more useful later on in the movement when predictions from the forward model were less accurate.

One of the core ideas behind these forward models is that, during (perceptual) inference, the sensory consequences of a movement are anticipated and used to attenuate the percepts related to these sensations. This mechanism reduces the predictable component of sensory input to self-generated stimuli, thereby enhancing the salience of sensations that have an external cause. This has many intriguing consequences. For example, it predicts that self-generated forces will be perceived as weaker than externally generated forces. This prediction was confirmed in a later experiment [57], thereby providing a neuroscientific explanation for force escalation during conflict; children trading tit-for-tat blows will often assert the other hit him harder.

Körding and Wolpert [58] have investigated learning in the sensorimotor system using a visual reaching task in which subjects moved their finger to a target and received visual feedback. This feedback provided information about target position that had an experimentally controlled bias and variance. Subjects were found to be able to learn this mapping (from vision to location) and integrate it into their behaviour, in a Bayes-optimal way.

Returning to our tennis theme, an analysis of three years of Wimbledon games has indicated that the outcome of the current point depends on the outcome of the previous point [59]. There are multiple potential sources of correlation here. It could be that a player intermittently enjoys a sweet parameter spot where his internal sensorimotor model accurately predicts body and ball position and is able to hit the ball cleanly, or perhaps a player finds a new pattern in his opponents behaviour such as body position, or previous serve, predicting current service direction.

3.4. Collective Decision Making. Sorkin et al. [60] have applied Bayes rule for Gaussians (see (5)) in their study of collective decision making. Here the optimal integration procedure involves each group members' input to the collective decision being weighted proportionally by the member's competence at the task. Mathematically, "competence" corresponds to precision. This model of group behaviour was shown to be better than a different model which assumed members made individual decisions which were then combined into a majority vote. This latter model better described collective decision making when members did not interact.

Bahrami et al. [61] investigated pairs of subjects (dyads) making collective perceptual decisions. Dyads with similarly sensitive subjects (similar precisions) were found to produce collective decisions that were close to optimal, but this was not the case for dyads with very different sensitivities. These observations were explained by a Bayes-optimal model under the assumption that subjects accurately communicated their confidence. This confidence sharing proved essential for the group decision to be better than the decision of the best subject.

4. Brain Models

We now turn to Bayesian models of the brain. As articulated by Colombo and Series [62] it could be that our behaviour is near Bayes-optimal yet the neural mechanisms underlying it are not. Current opinion on this issue is divided. According to Rust and Stocker [63] "If the system as a whole performs Bayesian inference, it seems unlikely that any one stage in this cascade represents a single component of the Bayesian model (e.g., the prior) or performs one of the mathematical operations in isolation (multiplying the prior and the likelihood)."

However, the above statement may be too heavily influenced by the simplicity of the tasks which were initially used to demonstrate near Bayes-optimal behaviour for example univariate cue integration. As we shall see, the nonlinear dynamic hierarchical models underlying predictive coding models of cortical macrocircuits (Section 5) do in fact provide a close correspondence with biology [1, 19, 64].

The structure and function of the human brain can be studied at multiple temporal and spatial scales. Research activity at the different scales effectively constitutes different scientific disciplines, although there is a good deal of work addressing integrative and unifying perspectives [2, 65, 66]. Our review of the literature proceeds through increasing spatial scale and a later section reviews work in modelling cortical macrocircuits.

4.1. Synapses and Dendrites. Most models of information processing in neural circuits require that synaptic efficacies are stable at least over seconds if not minutes or hours. However, real synapses can change strength several-fold at the time scale of a single interspike interval. This is known as Short Term Synaptic Plasticity (STP) [67]. Why do synapses change so quickly?

Pfister et al. [68] argue that neuronal membrane potentials are the primary locus of computational activity, where incoming information from thousands of presynaptic cells is integrated and analog state values, x are computed. It is then proposed that the goal of synaptic computation is to optimally reconstruct presynaptic membrane potentials, and optimal reconstructions are made possible via STP. Crudely, if a synapse has recently received a spike it increases its estimate of x and decreases it otherwise. Simple dynamic Bayesian models of this process explain empirical synaptic facilitation and depression.

Kiebel and Friston [69] propose that, through selective dendritic filtering, single neurons respond to specific sequences of presynaptic inputs. This study employs a dynamic Bayesian model of dendritic activity in which intracellular dendritic states are also viewed as predicting their presynaptic inputs. Pruning of dendritic spines then emerges as a consequence of parameter estimation in this model.

4.2. Neurons. Gold and Shadlen [15] propose that categorical decisions about sensory stimuli are based on the accumulation of information over time in the form of a log likelihood ratio (see Section 2). They review experiments in which monkeys were trained to make saccades to a target depending on the perceived direction of moving dots in the centre of a screen. Firing rates of neurons in superior colliculus and lateral intraparietal regions were seen to follow this evidence accumulation model. In follow-up experiments targets appeared on the left or right with different prior probability and initial firing rates followed these priors as predicted by the accumulation model. These models are also known as drift diffusion models and are the continuous analog of the sequential likelihood ratio test [14].

Fiorillo [70] proposed a general theory of neural computation based on prediction by single neurons. Each neuron is proposed to mirror the function of the whole system in learning to predict aspects of the world related to future reward. A neuron receives prior temporal information via nonsynaptic voltage-gated channels, and prior spatial information from a subset of its synaptic inputs. The remaining excitatory synaptic inputs provide current information about the state of the world. This would correspond to a "likelihood" term. The difference between expected and actual state is reflected as a prediction error signal encoded in the membrane potential of the cell. This proposal seems consistent with predictive coding theories that are formulated at a systems level (see Section 5).

Lengyel et al. [71] model storage and recall in an autoassociative model of hippocampal area CA3. The model treats recall as a problem of optimal probabilistic inference. Information is stored in the phase of cell firing relative to the hippocampal theta rhythm, a so-called spike-time code or phase code. Learning of these phase codes is based on Spike Timing Dependent Plasticity (STDP), such that a synapse is strengthened if the cell fires shortly after receiving a spike on that synapse. If the order of events is reversed the synapse is weakened. Synaptic changes only occur in a small time window, as described by an STDP curve. Given empirical STDP curves the Lengyel et al. model was able to predict the form of empirical Phase Response Curves (PRCs) underlying recall dynamics. These PRCs describe the synchronization properties of neurons. A refinement of their model [72] represented information in both spike timing and rate, and an approximate inference algorithm was developed using variational inference (see Section 2.3.2).

Deneve [20] shows that neurons that optimally integrate evidence about events in the world exhibit properties similar to integrate and fire neurons with spike-dependent adaptation (a gradually reducing firing rate). She proposes that neurons code for time-varying hidden variables, such as direction of motion, and the basic meaning of a spike is the occurrence of new information, and that propagation of spikes corresponds to Bayesian belief propagation (see Section 2). A companion paper [73] shows how neurons can learn to recognize dynamical patterns, and that successive layers of neurons can learn hierarchical models of sensory input. The learning that emerges is a form of STDP.

4.3. Populations

4.3.1. Probabilistic Codes. The response of a cortical neuron to sensory input is highly variable over trials, with cells showing Poisson-like distributions of firing rates. Specifically, firing rate variances grow in proportion to mean firing rates, as would be expected from a Poisson density [74]. Hoyer and Hyvarinen [75] review in vitro experiments which suggest that the variability of neuronal responses may not be a property of neurons themselves but rather emerges in intact neural circuits. This neural response variability may be a way in which neural circuits represent uncertainty.

Ma et al. [76] argue that if cells fired in the same way on every trial the brain would know exactly what the stimulus was. They suggest that the variability over a population of neurons for a single trial provides a way in which this uncertainty could be encoded in the brain, thus providing a substrate for Bayesian inference. Moreover, if the distribution of cell activities is approximately Poisson then Bayesian inference for optimal cue integration, for example, can be implemented with simple linear combinations of neural activity. They call this representation a Probabilistic Population Code (PPC). An interesting property of these codes is that sharply peaked distributions are encoded with higher firing rates (see Figure 1 in [77]). If the distribution was Gaussian this would correspond to high precision.

Ma et al. [76] concede that a deficiency of their PPC scheme is that neural activities are likely to saturate when sequential inferences are required. This can be avoided by using a nonlinearity to keep neurons within their dynamical range, which could be implemented for example using divisive normalisation [78]. This idea was taken up in later work [30] which shows how populations of cells can use PPCs to implement Kalman filtering.

4.3.2. Sampling Codes. A different interpretation of neural response variability is that populations of cells are implementing Bayesian inference by sampling from a posterior density [75] (see Section 2.3.1). They suggest that "variability over time" could be used whereby a "single neuron could represent a continuous distribution if its firing rate fluctuated in accordance with the distribution to be represented. At each instant in time, the instantaneous firing rate would be a random sample from the distribution to be represented." This interpretation is reviewed in [5, 6] and contrasted with PPCs.

This sampling perspective provides an account of bistable perception in which multiple interpretations of ambiguous input correspond to sampling from different modes of the posterior. This may occur during bistable percepts arising from, for example, binocular rivalry or the Necker cube illusion. If stimuli are modified such that one interpretation is more natural, then it becomes dominant for longer time periods. This is consistent with Bayesian sampling where more samples are taken from dominant modes [21]. The above idea was investigated empirically by placing Necker cubes against backgrounds comprised of unambiguous cubes [79]. Subjects experienced modified dominance times in line with the above predictions. In experiments on binocular rivalry, where images presented to the two eyes are different, only one of them will be perceived at a given time. A switch will then occur and the other image will be perceived. For certain stimuli, subjects tend to perceive a switch as a wave propagating across the visual field. This behaviour can be readily explained by Bayesian sampling in a Markov random field model [23].

It should be borne in mind that other proposals have been made regarding the nature of bistable perception. For example, Dayan [80] has proposed a deterministic generative and recognition model for binocular rivalry with an emphasis on competition between top-down hypotheses rather than bottom-up stimulus information. Here, switching between percepts was implemented with a simple fatigue process in which stable states slowly become unstable, resulting in perceptual oscillation.

From a computational perspective, the idea that populations of cells may be sampling from posterior densities is an attractive one. The sampling approach has become a standard method for inverting Bayesian models in statistics and engineering [21]. It is best suited, however, to low-dimensional problems, because the algorithms become very slow in high dimensions. It is popular in statistics and engineering because it is much more likely than deterministic methods to produce globally optimal posteriors. One method for encouraging this is to have a "temperature" parameter which starts off high and is gradually reduced over time, according to an annealing schedule. Annealed Importance Sampling, for example, is a gold standard method for approximating the model evidence [26]. Sampling approaches have been used in neural network models from the Boltzmann machine, to sparse hierarchical models and Deep Belief Networks (see Section 4.4).

In models with Gaussian observations the temperature corresponds to the precision of the data. As we shall see later, precisions have been proposed to be at least partly under the control of neuromodulators, so it seems reasonable to suggest that sampling based inference may be guided towards global optima via neuromodulation.

4.3.3. Spontaneous Activity. If neuronal populations encode Bayesian models of sensory data then this predicts a particular relationship between spontaneous and evoked neural activity. This has been investigated empirically by Berkes et al. [81]. If stimulus y is caused by event x then a Bayesian model will need to represent the prior distribution over the cause, $p(x)$, and update it to the posterior distribution $p(x \mid y)$. If this procedure is working properly then the average posterior (evoked) activity should be approximately equal to the prior activity. That is

$$\begin{aligned} p(x) &= \int p(x \mid y) p(y) dy \\ &\approx \sum_i p(x_i \mid y_i), \end{aligned} \tag{29}$$

where y_i are samples from the environment. Here the left hand side is the prior and the right hand side is average-evoked activity. This prediction was later confirmed by research from the same team who analysed visual cortical activity of awake ferrets during development [81]. The similarity between spontaneous and average-evoked activities, as measured using KL-divergence (see Section 2), increased with age and was specific to responses evoked by natural scenes. Fiser et al. [6] argue that the above relationship between spontaneous and average-evoked activity fits more naturally with a sampling view of neural coding.

4.4. Generative Models. This section describes macroscopic models of cortical processing either of single brain regions or of processing in hierarchical models [2, 82]. The work reviewed in this section is very closely related to that described in Section 5, the main difference being that Section 5 proposes a specific mapping onto cortical anatomy based on predictions, prediction errors, and the lamina structure of cortex.

Early models of hierarchical processing in cortex focus on feedforward processing. This transforms sensory input by static spatiotemporal filtering into more abstract representation and produces object representations that are translationally and viewpoint invariant as shown, for example, by Fukushima [83], Riesenhuber and Poggio [84], and Stringer and Rolls [85].

An alternative view on cortical processing is the idea of analysis-by-synthesis which suggests the cortex has a generative model of the world and that recognition involves inversion of this model [86]. This very general idea has also become known as predictive coding.

This idea is combined with Helmholtz's concept of perception as inference in the Helmholtz machine [27]. This is an unsupervised learning approach in which a recognition model infers a probability distribution over underlying causes of sensory input, and a separate generative model is used to train the recognition model. The approach assumes causes and inputs are binary variables. Both recognition and generative models are updated so as to minimise a variational free energy bound on the log model evidence. This implicitly minimises the Kullback-Liebler divergence between the true posterior density over causes and the approximation posterior instantiated in the recognition model (see Section 2.3.2).

Olshausen and Field [87] have proposed a sparse coding model of natural images where the likelihood is a simple linear model relating a "code" to image data, but the prior over code elements factorises and there is a sparse prior over each element. For a given image, most code elements are therefore small with a few being particularly large. This approach was applied to images of natural scenes and resulted in a bank of feature detectors that were spatially localised, oriented, and comprised a number of spatial scales, much like the simple cells in V1. A similar sparse coding approach can explain the properties of auditory nerve cells [88]. Later work [89] developed a two-layer model in which cells in the first layer were topographically organised and cells in the second layer were adapted so as to maximise the sparseness of locally pooled energies. Learning in this model produced second layer cells with large receptive fields and spatial invariance much like the complex cells in early visual cortex.

These sparse coding models have shown how responses of cells in one or two layer cortical networks can develop via learning in the appropriate generative models, but have been unable to explain how coding develops in multiple layers of cortical hierarchies. Recent progress in this area has been made using Deep Belief Networks (DBNs) [90]. These are probabilistic generative models composed of multiple layers of stochastic binary units. The top two layers have undirected, symmetric connections between them and form an associative memory, and the lower layers receive top-down directed connections from the layer above. Inference proceeds using sampling (see Section 2.3.1), and the approach allows nonlinear distributed representations to be learnt a layer at a time [91].

DBNs are based on many years of development starting with the Boltzmann machine, a network of binary stochastic units comprising hidden and visible units. This employs a type of probabilistic model called an undirected graph, where connected nodes are mutually dependent [16] (these are not DAGs). This then led to a Restricted Boltzmann Machine (RBM) where there are no connections among hidden units. DBNs can then be formed by stacking RBMs, such that hidden layer unit activities in lower level RBMs become training data for higher level RBMs. Hinton [91] notes that the key to efficient learning in these hierarchical models is the use of undirected units in their construction.

5. Cortical Hierarchies

This section describes models of Bayesian inference in cortical hierarchies by Mumford [10], Rao and Ballard [7] and a more recent series of papers by Friston [1, 11, 12]. We very briefly review the basics of cortical anatomy, describe the modelling proposals, and then provide a concrete example.

5.1. Functional Anatomy. The cortex is a thin sheet of neuronal cells which can be considered as comprising six layers, each differing in the relative density of different cell types. The relative densities of excitatory to inhibitory cells change from one cortical region to another, and these differences in "cytoarchitecture" can be used to differentiate, for example, region V1 from V2 [92, 93]. Despite these differences there are many commonalities throughout cortex. For example, layer 4 comprises mainly excitatory granule cells, and so is known as the granular layer. Other layers are also referred to as being agranular. The functional activity of a cylindrical column through the cortical sheet capturing several thousand neurons has been described in the form of canonical microcircuit [94]. This circuit is proposed to be replicated across cortex, providing a modular architecture for neural computation.

It is now well established that cortical regions are arranged in hierarchies. Felleman and van Essen [92], for example, used anatomical properties to reveal the hierarchical structure of the macaque visual system. Anatomical connections from lower to higher regions originate from superficial layer 2/3 pyramidal cells and target the granular layer [92]. Anatomical connections from higher to lower areas originate from "deep" layer 5/6 pyramidal cells and target layers 1 and 6 (agranular layers). This connectivity footprint is depicted in Figure 8. This is a generic pattern of connectivity within cortex, although it is more clearly manifested in some brain areas than others [95].

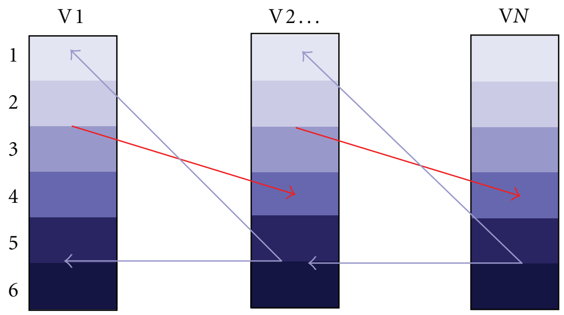


Figure 8: Anatomical connections from lower to higher regions in a serial cortical hierarchy originate from superficial layer 2/3 pyramidal cells in an ascending pathway (shown in red). Anatomical connections from higher to lower areas originate from layer 5/6 pyramidal cells and target layer 1/6 cells in lower regions (shown in purple). Adapted from Shipp [95].

Kennedy and Dehay [97] note that cortical hierarchies do not form a strict chain, for example, V1 can make a direct feedforward connection to V4 as well as indirectly through V2. They note that "hierarchical distance" can be defined in terms of laminar connectivity patterns. Long distance feedforward connections arise strictly from the supragranular layer (as Felleman and van Essen), but shorter distance ones also have contributions from infragranular layers.

Functionally, one key concept here concerning visual cortex, for example, is that there are separate "what" and "where" hierarchies although this is being challenged by recent perspectives in active vision [98]. There is a good deal of evidence showing that these higher level representations are more enduring [99]. This makes sense as more abstract causes in our sensory world exist on a longer time scale that is objects may move, they may even change shape or colour, but they are still the same object.

If sensory input is at the bottom of the hierarchy then what is at the top? One idea is that rather than there being a top and a bottom there is an "inside" and an "outside" [1, 96]. That is, there is a centre rather than a top. Brain regions around the outside receive information from different sensory modalities; vision, audition, touch. The next level in represents higher level modality specific information, such as lines and edges in the visual system or chirps and formants in the auditory system. As we progress closer to the centre, brain regions become multimodal as depicted in Figure 9.

5.2. Hierarchical Predictive Coding. Mumford [10] has proposed how Bayesian inference in hierarchical models maps onto cortical anatomy. Specifically, he proposes that top-down predictions are sent from pyramidal cells in deep layers and received by agranular layers (purple arrows in Figure 8), and that prediction errors are sent from superficial pyramidal cells and are received by stellate cells in the granular layer (red arrows in Figure 8).

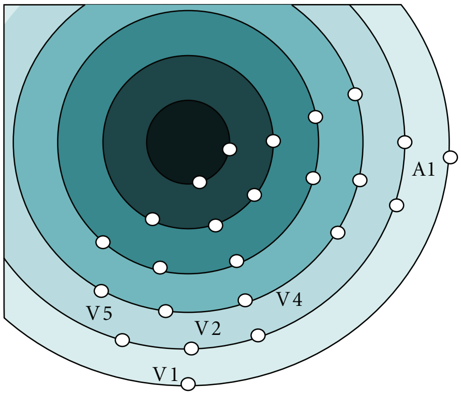


Figure 9: Cortical architecture depicting multimodal areas in the centre and unimodal sensory processing regions on the periphery, with visual regions shown at the bottom and auditory regions on the right. Adapted from Mesulam [96].

Rao and Ballard [7] describe a predictive coding model of visual cortex in which "extraclassical" receptive field properties emerge due to predictions from higher levels. When the model is presented with images of an extended bar, for example, first layer cells processing input from near the end of the bar soon stop firing as the presence of signal at that location is accurately predicted by cells in the second layer which have larger receptive fields. This "end-stopping" effect in first layer cells is explained by there being no prediction error to send up to the second layer. By this later time, cells in the second layer already know about the bar.

In related work Rao and Ballard [100] consider a similar model, but where hidden layer representations are intrinsically dynamic. Inference in this model is then implemented with an Extended Kalman Filter (see Section 2.4). These dynamics embody a nonlinear prediction step which also helps to counteract the signal propagation delays introduced by the different hierarchical levels (see Section 3.2 for a discussion of this issue).

Lee and Mumford [31] review evidence from human brain imaging and primate neurophysiology in support of the hypothesis that processing in visual cortex corresponds to inference in hierarchical Bayesian models. They describe activity in visual areas as being tightly coupled with the rest of the visual system such that long latency V1 responses reflect increasingly more global feedback from abstract high level features. This is consistent with a nonlinear hierarchical and dynamical model and they propose that inference in this model could be implemented using particle filtering (see Section 2.4.1).

George and Hawkins [19] describe a "hierarchical temporal memory" model of activity in cortical hierarchies which makes spatio-temporal predictions. Inference in this model is based on the belief propagation algorithm and detailed proposals are made regarding the mapping of various computational steps onto activity in different cortical laminae.

A series of papers by Friston [1, 11] review anatomical and functional evidence for hierarchical predictive coding, and describe implementations of Mumford's original proposal [10] with increasing levels of sophistication. These include the use of continuous-time nonlinear dynamical generative models and the use of a generalised coordinate representation of state variables. This concept from control theory provides a representation of higher order derivatives such as position, velocity, and acceleration, variables which have natural representations in the brain. Generalised coordinates effectively provide an extended time window for inference and may also provide a mechanism for postdiction (described in Section 3.2). This series of papers also describes a variational inference algorithm for estimating states (inference) and parameters (learning) and how these computations map onto cortical laminae. In later work [101] this framework was extended by expressing sensory input as a function of action, which effectively repositions an agent's sensory apparatus. The same variational inference procedures can then be used to select actions. This active inference framework is explained in recent reviews [12, 28].

5.3. Two-Level Model.

5.3. Two-Level Model. We now describe a simple Bayesian model of object recognition which illustrates many of the previously described features. This is a simplified version of the models described by Rao and Ballard [7]. We focus on perception, that is, how the beliefs regarding the hidden variables in the network can be updated. For simplicity, we focus on a hierarchical model with just two levels, although the approach can be applied to models of arbitrary depth.

The identity of an object is encoded by the variable x_2, the features of objects by the variable x_1, and a visual image by y. The model embodies the notion that x_2 causes x_1 which in turn causes y. The probabilistic dependencies in the associated generative model can be written as

$$p(y, x_1, x_2) = p(y \mid x_1)\, p(x_1 \mid x_2)\, p(x_2). \tag{30}$$

One can derive update rules for estimating the hidden variables by following the gradient of the above joint likelihood, or equivalently the log of the joint likelihood. This will produce MAP estimates of the hidden variables (see Section 2). Taking logs gives

$$\log p(y, x_1, x_2) = \log p(y \mid x_1) + \log p(x_1 \mid x_2) + \log p(x_2). \tag{31}$$

We now make the additional assumption that these distributions are Gaussian. To endow the network with sufficient flexibility of representation, for example the ability to turn features on or off, we allow nonlinear transformations, $g(\,)$, between layers. That is,

$$\begin{aligned} p(y \mid x_1) &= \mathrm{N}(y; g_1(x_1), \lambda_0 I), \\ p(x_1 \mid x_2) &= \mathrm{N}(x_1; g_2(x_2), \lambda_1 I), \\ p(x_2) &= \mathrm{N}(x_2; 0, \lambda_2 I), \end{aligned} \tag{32}$$

where $g_1(x_1)$ and $g_2(x_2)$ are top down predictions of lower level activity based on higher level representations, and λ_i are precision parameters. This can also be written as

$$\begin{aligned} y &= g_1(x_1) + e_1, \\ x_1 &= g_2(x_2) + e_2, \\ x_2 &= e_3. \end{aligned} \tag{33}$$

One can then derive the following update rules for the hidden variables [7]

$$\begin{aligned} \tau \frac{dx_1}{dt} &= \lambda_0 g_1'(x_1) e_1 + \lambda_1 e_2, \\ \tau \frac{dx_2}{dt} &= \lambda_1 g_2'(x_2) e_2 + \lambda_2 x_2, \end{aligned} \tag{34}$$

where $g'(\,)$ denotes the derivative of the nonlinearity and the prediction errors are given by

$$\begin{aligned} e_1 &= y - g_1(x_1), \\ e_2 &= x_1 - g_2(x_2). \end{aligned} \tag{35}$$

Figure 10 shows the propagation of predictions and prediction errors in this two-level network.

The parameter τ in (34) determines the time scale of perceptual inference. The degree to which the activity of a unit changes as a function input is referred to as "gain." In (34) the input is the bottom up prediction error. The gain is therefore dependent on the precision λ_i and the slope of the nonlinearity $g'(\,)$. There are therefore at least two gain control mechanisms. These will change the balance between how much network dynamics are dependent on top-down versus bottom-up information. Similar equations can be derived for how to update the parameters of the model, as shown in [7].

5.4. Gain Control. The key element of a digital computer is a voltage-gated switch, the transistor, which is turned on and off by the same sorts of currents it controls. An understanding of neuronal gain control is important to computational neuroscience [102]. Simple sensory reflexes, for example, can be turned off and replaced by responses based on higher level cognitive processing. There are a number of potential mechanisms in the brain for gain control including synchronization, neuromodulation, recurrent dynamics, and inhibition.

5.4.1. Synchronization. Equation (34) shows that the gain of a unit is dependent on the slope of the nonlinearity $g'(\,)$. If we interpret a unit as reflecting the activity of a population of cells then this slope can be increased, for example, by increasing the synchronization among cells. Highly synchronized cell populations have large gain [103]. In addition, this gain can be amplified by recurrent computation in neural networks [102, 104].

5.4.2. Neuromodulation. Equation (34) also shows that gain can be changed by manipulating the precision λ_i. It has been proposed that neuromodulators can change λ_i and so modulate the encoding of uncertainty. Neuromodulators are generated in subcortical nuclei and distributed to large regions of cortex. Different neuromodulators project to different cortical regions. For example, the highest concentrations of dopamine are found in striatum, basal ganglia, and frontal cortex. The detailed spatial specificity and temporal dynamics of neuromodulatory projections are unknown but they are thought to act as macroscopic signals [105].

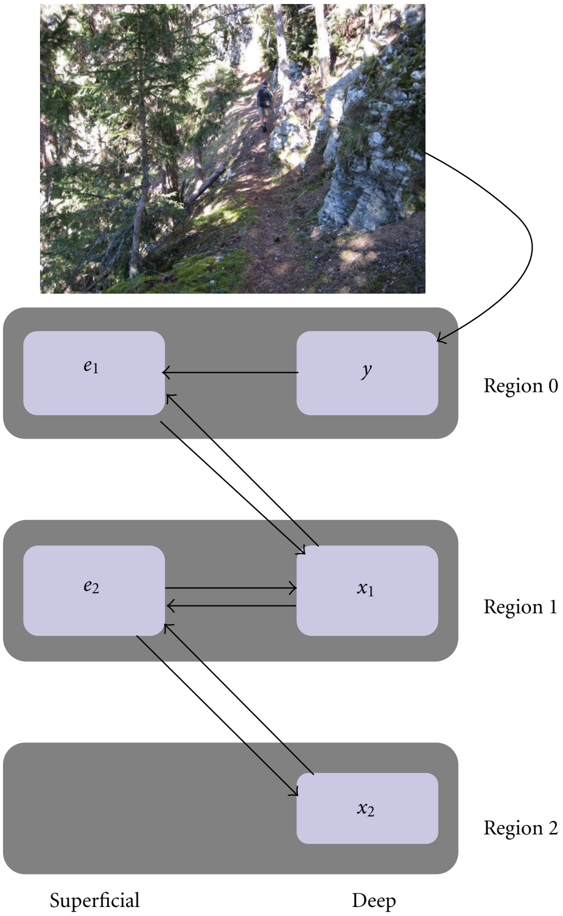


Figure 10: Predictive coding architecture for inference in hierarchical models. Each level in the hierarchy is located in a different brain region. Each region has a population of error units and a population of causal units. The error units are hypothesised to reside in superficial cortical laminae and causal units in deep laminae. Error units receive messages from the state units in the same level and the level above, whereas state units are driven by error units in the same level and the level below. The person near the centre of the image would be difficult to see without a top-down prediction that there was somebody walking along the path. This prediction may be derived from previous time steps, hence the need for dynamic models, or from higher level scene knowledge that people walk on paths.

Yu and Dayan [106] have considered the computational problem of assessing the validity of predictive cues in various contexts. Here a context reflects the set of stable statistical regularities that relate environmental entities such as objects and events to each other and to our sensory and motor systems. They propose that Acetylcholine (ACh) signals the uncertainty that is expected within a given context and that Norepinephrine (NE) signals the uncertainty associated with a change in context. Increasing levels of ACh and NE therefore downweight the strength of top-down (contextual) information and effectively upregulate bottom-up sensory input.

It has also been proposed that dopamine signals uncertainty in reward delivery [107]. This proposal has been elaborated upon by Friston et al. [108] who propose that dopamine balances the relative weighting of top-down beliefs and bottom-up sensory information when making inferences about cues that reliably signal potential actions. A dynamical model of cued sequential movements was developed in which inference proceeded using the variational approach described earlier, and the resulting simulated behaviours were examined as a function of synthetic dopamine lesions.

5.4.3. Recurrent Dynamics. Abbott [102] suggests that small differences in gain from, for example, synchronization can be amplified via dynamics in recurrent networks. Yu and Dayan [109] have used such dynamics in a model of visual attention. They developed a generative model of the Posner attentional task where a central cue predicts the location of a stimulus which then has a property (orientation) about which subjects have to make a decision, for example, press the left if the stimulus points left. Here there are two feature dimensions; spatial location and orientation. Inference in the Yu and Dayan model then shows how priors in one feature dimension (spatial) can gate inference in the other (orientation). This is consistent with electrophysiological responses whereby spatial attention has a multiplicative effect on orientation tuning of visual cortical neurons.

In the Yu et al. [45] study of the Eriksen Flanker task, referred to in Section 3.2, an approximate inference algorithm was proposed. This assumed a default assumption that the stimuli would be congruent and processing could proceed using a feedforward network in which "congruent" connections were facilitated using gain control. But upon detection of response conflict, an "incongruent" set of feedforward connections would instead be facilitated.

5.4.4. Receptor Pharmacology. Long range connections in the brain, both bottom-up and top-down, are excitatory and use the neurotransmitter glutamate. Glutamate acts on two types of postsynaptic receptor (i) AMPA receptors and (ii) NMDA receptors. NMDA receptors have a different action depending on the current level of postsynaptic potential, that is, they are voltage-gated. There is known to be a greater proportion of NMDA receptors for top-down connections which therefore provides a mechanism for top-down signals to gate bottom-up ones.

Corlett et al. [110] review the action of various drugs on psychotic effects and describe their action in terms of their receptor dynamics and inference in hierarchical Bayesian networks. Ketamine, for example, upregulates AMPA and blocks NMDA transmission. This will increase bottom-up signalling, which is AMPA-mediated, and reduce top-down signalling which is NMDA mediated. They suggest this will in turn lead to delusions, inappropriate inference of high level causes. Bayesian models of psychosis and underlying links to the pharmacology of synaptic signalling are discussed at length in [8]. See also [111] for a broader view of computational modelling for psychiatry.

6. Planning and Control

This review has briefly considered optimal decision making in terms of the likelihood ratio tests that may be reported by single neurons [15]. But as yet, we have had nothing to say about sequential decisions, planning, or control. Here, the key difference is that our decisions become actions which affect the state of the world which will in turn affect what the next optimal action would be. Because the combination of (potential) actions grows exponentially with time this is a difficult computational problem. It is usually addressed using various formalisms, from optimal control theory [112, 113] to reinforcement learning [114]. For reviews of these approaches applied to neuroscience see [115, 116].

Here we focus on recent theoretical developments in this area where research has shown how problems in optimal control theory, or "model-based" reinforcement learning, can be addressed using a purely Bayesian inference approach. For example, Attias [117] has proposed that planning problems can be solved using Bayesian inference. The central idea is to infer the control signals, u_n, conditioned on known initial state x_1 and desired goal states x_n. For example, Toussaint [118] describes the estimation of control signals using a Bayesian message passing algorithm which defaults to a classic control theoretic formulation for linear Gaussian dynamics. This framework can also be extended to accommodate desired observations, Y_N. The appropriate control signals can then be computed by estimating the density $p(u_n \mid x_1, Y_N)$ which can be implemented using backwards inference (see Section 6). This approach is currently being applied to systems level modelling of spatial cognition [119].

Similarly, Todorov has shown how control theoretic problems become linearly solvable if the cost of an action is quantified by penalising the difference between controlled and uncontrolled dynamics using Kullback-Liebler divergence [120]. Computation of optimal value functions is then equivalent to backwards inference in an equivalent dynamic Bayesian model [121] (see Section 2.4).

We refer to the above approaches using the term Planning as Inference. Planning as Inference requires the propagation of uncertainty forwards and backwards in time. This can be implemented using the forwards and backwards inference procedures described earlier. For these algorithms to be implemented in the brain we must have an online algorithm such as the gamma recursions. An advantage of considering control and planning problems as part of the same overall Bayesian inference procedure is that it becomes very natural to model the tight coupling that systems neuroscientists believe underlies action and perception [98, 122].

7. Discussion

This paper has hopefully shown that Bayesian inference provides a general theoretical framework that explains aspects of both brain activity and human behaviour. Bayesian inference can quantitatively account for results in experimental psychology on sensory integration, visual processing, sensorimotor integration, and collective decision making. It also explains the nonlinear dynamical properties of synapses, dendrites, and sensory receptive fields where neurons and neural networks are active predictors rather than passive filters of their sensory inputs.

More generally, the field is beginning to relate constructs in Bayesian inference to the underlying computational infrastructure of the brain. At the level of systems neuroscience brain imaging technologies are likely to play a key role. For example, neuroimaging modalities such as Electroencephalography (EEG) and Magnetoencephalography (MEG) are thought to mainly derive from superficial pyramidal cells. Cortical signals measured with these modalities should therefore correspond to prediction error signals in the hierarchical predictive coding models described in Section 5. Transcranial Magnetic Stimulation (TMS) can be used to knock out activity in various brain regions and therefore infer which are necessary for perceptual inference [31]. Functional Magnetic Resonance Imaging (fMRI) can be used to monitor activity in lower level regions that may be explained away by activity in higher level regions [31]. An important recent development is the use of dynamic models of brain connectivity to estimate strengths of connections between regions [123]. This allows for the quantitative assessment of changes in top-down or bottom-up signalling from brain imaging data [124, 125].

A particularly exciting recent theoretical development is the notion of Planning as Inference described in Section 6. Previously, Bayesian inference has been used to explain perception and learning. This recent research suggests how Bayesian inference may also be used to understand action and control. This closes the loop and reflects the tight coupling that systems neuroscientists believe underlies action and perception in the human brain [98, 122]. Central to this endeavour are the forwards and backwards recursions in time that are necessary to compute optimal value functions or control signals. Our review has also suggested, in Section 3.2, that they may also be necessary to model perceptual inference at a much shorter time scale.

Acknowledgments

This paper was supported by the Wellcome Trust. The author thanks Bahador Bahrami for introducing me to the literature on postdiction in visual processing and for discussing models of collective decision making.

References

[1] K. Friston, “A theory of cortical responses,” *Philosophical Transactions of the Royal Society of London Series B*, vol. 360, no. 1456, pp. 815–836, 2005.

[2] P. Dayan and L. F. Abbott, *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*, MIT Press, 2001.

[3] C. Frith, *Making Up the Mind: How the Brain Creates Our Mental World*, Wiley-Blackwell, 2007.

[4] K. Doya, S. Ishii, A. Pouget, and R. Rao, Eds., *Bayesian Brain: Probabilistic Approaches to Neural Coding*, MIT Press, 2007.

[5] T. Lochmann and S. Deneve, “Neural processing as causal inference,” *Current Opinion in Neurobiology*, vol. 21, no. 5, pp. 774–781, 2011.

[6] J. Fiser, P. Berkes, G. Orbán, and M. Lengyel, “Statistically optimal perception and learning: from behavior to neural representations,” *Trends in Cognitive Sciences*, vol. 14, no. 3, pp. 119–130, 2010.

[7] R. P. N. Rao and D. H. Ballard, “Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects,” *Nature Neuroscience*, vol. 2, no. 1, pp. 79–87, 1999.

[8] P. C. Fletcher and C. D. Frith, “Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia,” *Nature Reviews Neuroscience*, vol. 10, no. 1, pp. 48–58, 2009.

[9] D. M. Eagleman and T. J. Sejnowski, “Motion integration and postdiction in visual awareness,” *Science*, vol. 287, no. 5460, pp. 2036–2038, 2000.

[10] D. Mumford, “On the computational architecture of the neocortex—II the role of cortico-cortical loops,” *Biological Cybernetics*, vol. 66, no. 3, pp. 241–251, 1992.

[11] K. Friston, “Learning and inference in the brain,” *Neural Networks*, vol. 16, no. 9, pp. 1325–1352, 2003.

[12] K. Friston, “The free-energy principle: a unified brain theory?” *Nature Reviews Neuroscience*, vol. 11, no. 2, pp. 127–138, 2010.

[13] D. Wolpert and Z. Ghahramani, “Oxford companion to the mind,” in *Bayes Rule in Perception, Action and Cognition*, Oxford University Press, 2004.

[14] R. Bogacz, E. Brown, J. Moehlis, P. Holmes, and J. D. Cohen, “The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks,” *Psychological Review*, vol. 113, no. 4, pp. 700–765, 2006.

[15] J. I. Gold and M. N. Shadlen, “Neural computations that underlie decisions about sensory stimuli,” *Trends in Cognitive Sciences*, vol. 5, no. 1, pp. 10–16, 2001.

[16] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[17] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kauman, 1988.

[18] M. Jordan and Y. Weiss, “Graphical models: probabilistic inference,” in *The Handbook of Brain Theory and Neural Networks*, M. Arbib, Ed., MIT Press, 2002.

[19] D. George and J. Hawkins, “Towards a mathematical theory of cortical micro-circuits,” *PLoS Computational Biology*, vol. 5, no. 10, Article ID e1000532, 2009.

[20] S. Deneve, “Bayesian spiking neurons I: inference,” *Neural Computation*, vol. 20, no. 1, pp. 91–117, 2008.

[21] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, Chapman and Hall, Boca Raton, Fla, USA, 1995.

[22] A. Jasra, D. A. Stephens, and C. C. Holmes, “On population-based simulation for static inference,” *Statistics and Computing*, vol. 17, no. 3, pp. 263–279, 2007.

[23] S. Gershman E Vul J Tenenbaum, "Perceptual multistability as markov chain monte carlo inference," in *Advances in Neural Information Procesing Systems*, vol. 22, 2009.

[24] D. J. C. Mackay, *Information Theory, Inference and Learning Algorithms*, Cambridge University Press, Cambridge, UK, 2003.

[25] M. I. Jordan, Ed., *Learning in Graphical Models*, MIT Press, 1999.

[26] M. Beal, *Variational algorithms for approximate Bayesian inference [Ph.D. thesis]*, University College London, 2003.

[27] P. Dayan, G. E. Hinton, R. M. Neal, and R. S. Zemel, "The Helmholtz machine," *Neural Computation*, vol. 7, no. 5, pp. 889–904, 1995.

[28] K. Friston, "The free-energy principle: a rough guide to the brain?" *Trends in Cognitive Sciences*, vol. 13, no. 7, pp. 293–301, 2009.

[29] S. Roweis and Z. Ghahramani, "A unifying review of linear gaussian models," *Neural Computation*, vol. 11, no. 2, pp. 305–345, 1999.

[30] J. Beck, P. Latham, and A. Pouget, "Marginalization in neural circuits with divisive normalization," *The Journal of Neuroscience*, vol. 31, no. 43, pp. 15310–15319, 2011.

[31] T. S. Lee and D. Mumford, "Hierarchical Bayesian inference in the visual cortex," *Journal of the Optical Society of America A*, vol. 20, no. 7, pp. 1434–1448, 2003.

[32] Z. Ghahramani and M. J. Beal, "Propagation algorithms for Variational Bayesian learning," in *Advances in Neural Information Processing Systems*, T. K. Leen, T. Dietterich, and V. Tresp, Eds., vol. 13, MIT Press, Cambridge, Mass, USA, 2001.

[33] J. Daunizeau, K. J. Friston, and S. J. Kiebel, "Variational Bayesian identification and prediction of stochastic nonlinear dynamic causal models," *Physica D*, vol. 238, no. 21, pp. 2089–2118, 2009.

[34] A. Doucet, N. de Freitas, and N. Gordon, Eds., *Sequential Monte Carlo Methods in Practice*, Springer, 2001.

[35] M. West and J. Harrison, Eds., *Bayesian Forecasting and Dynamic Models*, Springer, 1997.

[36] A. A. Stocker and E. P. Simoncelli, "Noise characteristics and prior expectations in human visual speed perception," *Nature Neuroscience*, vol. 9, no. 4, pp. 578–585, 2006.

[37] R. Gregory, *Eye and Brain: The Psychology of Seeing*, Oxford University Press, 1998.

[38] M. O. Ernst and M. S. Banks, "Humans integrate visual and haptic information in a statistically optimal fashion," *Nature*, vol. 415, no. 6870, pp. 429–433, 2002.

[39] D. Green and J. Swets, *Signal Detection Theory and Psychophysics*, John Wiley & Sons, 1966.

[40] D. Alais and D. Burr, "The Ventriloquist effect results from near-optimal bimodal integration," *Current Biology*, vol. 14, no. 3, pp. 257–262, 2004.

[41] P. W. Battaglia, R. A. Jacobs, and R. N. Aslin, "Bayesian integration of visual and auditory signals for spatial localization," *Journal of the Optical Society of America A*, vol. 20, no. 7, pp. 1391–1397, 2003.

[42] D. Kersten, P. Mamassian, and A. Yuille, "Object perception as Bayesian inference," *Annual Review of Psychology*, vol. 55, pp. 271–304, 2004.

[43] D. Knill and W. Richards, *Perception as Bayesian Inference*, Cambridge, UK, 1996.

[44] W. J. Adams, E. W. Graf, and M. O. Ernst, "Experience can change the "light-from-above" prior," *Nature Neuroscience*, vol. 7, no. 10, pp. 1057–1058, 2004.

[45] A. J. Yu, P. Dayan, and J. D. Cohen, "Dynamics of attentional selection under conflict: toward a rational Bayesian account," *Journal of Experimental Psychology*, vol. 35, no. 3, pp. 700–717, 2009.

[46] Y. S. Liu, A. Yu, and P. Holmes, "Dynamical analysis of bayesian inferencemodels for the eriksen task," *Neural Computation*, vol. 21, no. 6, pp. 1520–1553, 2009.

[47] Y. Weiss, E. P. Simoncelli, and E. H. Adelson, "Motion illusions as optimal percepts," *Nature Neuroscience*, vol. 5, no. 6, pp. 598–604, 2002.

[48] J. Najemnik and W. S. Geisler, "Optimal eye movement strategies in visual search," *Nature*, vol. 434, no. 7031, pp. 387–391, 2005.

[49] J. Najemnik and W. S. Geisler, "Eye movement statistics in humans are consistent with an optimal search strategy," *Journal of Vision*, vol. 8, no. 3, article 4, 2008.

[50] R. Nijhawan, "Motion extrapolation in catching," *Nature*, vol. 370, no. 6487, pp. 256–257, 1994.

[51] K. A. Sundberg, M. Fallah, and J. H. Reynolds, "A motion-dependent distortion of retinotopy in area V4," *Neuron*, vol. 49, no. 3, pp. 447–457, 2006.

[52] R. Soga, R. Akaishi, and K. Sakai, "Predictive and postdictive mechanisms jointly contribute to visual awareness," *Consciousness and Cognition*, vol. 18, no. 3, pp. 578–592, 2009.

[53] T. Bachmann, *Psychophysiology of Backward Masking*, Nova Science, 1994.

[54] P. A. Kolers and M. Von Gruenau, "Shape and color in apparent motion," *Vision Research*, vol. 16, no. 4, pp. 329–335, 1976.

[55] D. Dennett, *Consciousness Explained*, Little, Brown and Company, 1991.

[56] D. M. Wolpert, Z. Ghahramani, and M. I. Jordan, "An internal model for sensorimotor integration," *Science*, vol. 269, no. 5232, pp. 1880–1882, 1995.

[57] S. S. Shergill, P. H. Bays, C. D. Frith, and D. M. Wotpert, "Two eyes for an eye: the neuroscience of force escalation," *Science*, vol. 301, no. 5630, p. 187, 2003.

[58] K. P. Körding and D. M. Wolpert, "Bayesian integration in sensorimotor learning," *Nature*, vol. 427, no. 6971, pp. 244–247, 2004.

[59] F. J. G. M. Klaassen and J. R. Magnus, "Are points in tennis independent and identically distributed? Evidence from a dynamic binary panel data model," *Journal of the American Statistical Association*, vol. 96, no. 454, pp. 500–509, 2001.

[60] R. D. Sorkin, C. J. Hays, and R. West, "Signal-detection analysis of group decision making," *Psychological Review*, vol. 108, no. 1, pp. 183–203, 2001.

[61] B. Bahrami, K. Olsen, P. E. Latham, A. Roepstorff, G. Rees, and C. D. Frith, "Optimally interacting minds," *Science*, vol. 329, no. 5995, pp. 1081–1085, 2010.

[62] M. Colombo and P. Series, "Bayes in the brain—on Bayesian modelling in neuroscience," *British Journal for the Philosophy of Science*, vol. 63, no. 3, pp. 697–723, 2012.

[63] N. C. Rust and A. A. Stocker, "Ambiguity and invariance: two fundamental challenges for visual processing," *Current Opinion in Neurobiology*, vol. 20, no. 3, pp. 382–388, 2010.

[64] R. Adams, S. Shipp, and K. Friston, "Predictions not commands: active inference in the motor system".

[65] E. Kandel, J. Schwartz, and T. Jessell, *Principals of Neural Science*, McGraw-Hill, 2000.

[66] J. Nicholls, A. R. Martin, P. Fuchs, D. Brown, M. Diamond, and D. Weisblat, *From Neuron to Brain: A Cellular and Molecular Approach to the Function of the Nervous System*, Sinaeur, 2012.

[67] L. F. Abbott and W. G. Regehr, "Synaptic computation," *Nature*, vol. 431, no. 7010, pp. 796–803, 2004.

[68] J. P. Pfister, P. Dayan, and M. Lengyel, "Synapses with short-term plasticity are optimal estimators of presynaptic membrane potentials," *Nature Neuroscience*, vol. 13, no. 10, pp. 1271–1275, 2010.

[69] S. Kiebel and K. Friston, "Free energy and dendritic self-organization," *Frontiers in Systems Neuroscience*, vol. 5, article 80, 2011.

[70] C. D. Fiorillo, "Towards a general theory of neural computation based on prediction by single neurons," *PLoS ONE*, vol. 3, no. 10, Article ID e3298, 2008.

[71] M. Lengyel, J. Kwag, O. Paulsen, and P. Dayan, "Matching storage and recall: hippocampal spike timing-dependent plasticity and phase response curves," *Nature Neuroscience*, vol. 8, no. 12, pp. 1677–1683, 2005.

[72] M. Lengyel P Dayan, "Uncertainty, phase and oscillatory hippocampal recall," in *Neural Information Processing Systems*, 2007.

[73] S. Deneve, "Bayesian spiking neurons II: learning," *Neural Computation*, vol. 20, no. 1, pp. 118–145, 2008.

[74] D. J. Tolhurst, J. A. Movshon, and A. F. Dean, "The statistical reliability of signals in single neurons in cat and monkey visual cortex," *Vision Research*, vol. 23, no. 8, pp. 775–785, 1983.

[75] P. Hoyer and A. Hyvarinen, "Interpreting neural response variability as monte carlo sampling of the posterior," in *Neural Information Processing Systems 2003*, 2003.

[76] W. J. Ma, J. M. Beck, P. E. Latham, and A. Pouget, "Bayesian inference with probabilistic population codes," *Nature Neuroscience*, vol. 9, no. 11, pp. 1432–1438, 2006.

[77] W. J. Ma, J. M. Beck, and A. Pouget, "Spiking networks for Bayesian inference and choice," *Current Opinion in Neurobiology*, vol. 18, no. 2, pp. 217–222, 2008.

[78] M. Carandini and D. Heeger, "Normalization as a canonical neural computation," *Nature Reviews Neuroscience*, vol. 13, no. 1, pp. 51–62, 2012.

[79] R. Sundareswara and P. R. Schrater, "Perceptual multistability predicted by search model for Bayesian decisions," *Journal of Vision*, vol. 8, no. 5, article 12, 2008.

[80] P. Dayan, "A hierarchical model of binocular rivalry," *Neural Computation*, vol. 10, no. 5, pp. 1119–1135, 1998.

[81] P. Berkes, G. Orbán, M. Lengyel, and J. Fiser, "Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment," *Science*, vol. 331, no. 6013, pp. 83–87, 2011.

[82] G. Hinton and T. Sejnowski, Eds., *Unsupervised Learning: Foundations of Neural Computation*, MIT Press, 1999.

[83] K. Fukushima, "Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics*, vol. 36, no. 4, pp. 193–202, 1980.

[84] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nature Neuroscience*, vol. 2, no. 11, pp. 1019–1025, 1999.

[85] S. M. Stringer and E. T. Rolls, "Invariant object recognition in the visual system with novel views of 3D objects," *Neural Computation*, vol. 14, no. 11, pp. 2585–2596, 2002.

[86] D. Mackay, *The Epistemological Problem for Automata*, Princeton University Press, 1956.

[87] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.

[88] E. C. Smith and M. S. Lewicki, "Efficient auditory coding," *Nature*, vol. 439, no. 7079, pp. 978–982, 2006.

[89] A. Hyvärinen and P. O. Hoyer, "A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images," *Vision Research*, vol. 41, no. 18, pp. 2413–2423, 2001.

[90] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[91] G. E. Hinton, "Learning multiple layers of representation," *Trends in Cognitive Sciences*, vol. 11, no. 10, pp. 428–434, 2007.

[92] D. J. Felleman and D. C. Van Essen, "Distributed hierarchical processing in the primate cerebral cortex," *Cerebral Cortex*, vol. 1, no. 1, pp. 1–47, 1991.

[93] G. Shepherd, Ed., *The Synaptic Organization of the Brain*, Oxford University Press, 2004.

[94] R. Douglas, K. Martin, and D. Whitteridge, "A canonical microcircuit for neocortex," *Neural Computation*, vol. 1, pp. 480–488, 1989.

[95] S. Shipp, "Structure and function of the cerebral cortex," *Current Biology*, vol. 17, no. 12, pp. R443–R449, 2007.

[96] M. M. Mesulam, "From sensation to cognition," *Brain*, vol. 121, no. 6, pp. 1013–1052, 1998.

[97] H. Kennedy and C. Dehay, "Self-organization and interareal networks in the primate cortex," *Progress in Brain Research*, vol. 195, pp. 341–360, 2012.

[98] P. Cisek, "Cortical mechanisms of action selection: the aordance competition hypothesis," *Philosophical Transactions of the Royal Society of London Series B*, vol. 362, no. 1485, pp. 1585–1599, 2007.

[99] S. J. Kiebel, J. Daunizeau, and K. J. Friston, "A hierarchy of time-scales and the brain," *PLoS Computational Biology*, vol. 4, no. 11, Article ID e1000209, 2008.

[100] R. P. N. Rao and D. H. Ballard, "Dynamic model of visual recognition predicts neural response properties in the visual cortex," *Neural Computation*, vol. 9, no. 4, pp. 721–763, 1997.

[101] K. Friston, J. Kilner, and L. Harrison, "A free energy principle for the brain," *Journal of Physiology Paris*, vol. 100, no. 1–3, pp. 70–87, 2006.

[102] L. Abbott, *Where Are the Switches on This Thing?*Problems in Systems Neuroscience, chapter 23, Oxford University Press, 2006.

[103] G. B. Ermentrout and D. Terman, *Mathematical Foundations of Neuroscience*, Springer, 2010.

[104] J. J. Hopfield and C. D. Brody, "What is a moment? Trasient synchrony as a collective mechanism for spatiotemporal integration," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 3, pp. 1282–1287, 2001.

[105] K. Doya, "Metalearning and neuromodulation," *Neural Networks*, vol. 15, no. 4–6, pp. 495–506, 2002.

[106] A. J. Yu and P. Dayan, "Uncertainty, neuromodulation, and attention," *Neuron*, vol. 46, no. 4, pp. 681–692, 2005.

[107] C. D. Fiorillo, P. N. Tobler, and W. Schultz, "Discrete coding of reward probability and uncertainty by dopamine neurons," *Science*, vol. 299, no. 5614, pp. 1898–1902, 2003.

[108] K. Friston, T. Shiner, T. FitzGerald et al., "Dopamine, aordance and active inference," *PLOS Computational Biology*, vol. 8, no. 1, Article ID e1002327, 2012.

[109] A. Yu and P. Dayan, "Inference, attention, and decision in a bayesian neural architecture," in *Neural Information Processing Systems*, 2005.

[110] P. R. Corlett, C. D. Frith, and P. C. Fletcher, "From drugs to deprivation: a Bayesian framework for understanding models

of psychosis," *Psychopharmacology*, vol. 206, no. 4, pp. 515–530, 2009.

[111] R. Montague, R. Dolan, K. Friston, and P. Dayan, "Computational psychiatry," *Trends in Cognitive Sciences*, vol. 16, no. 1, pp. 72–80, 2012.

[112] D. Bertsekas, *Dynamic Programming and Optimal Control*, MIT Press, 2001.

[113] A. Bryson and Y. Ho, *Applied Optimal Control*, Ginn and Company, 1969.

[114] R. Sutton and A. Barto, *Reinforcement Learning: An Intro*, MIT Press, 1998.

[115] E. Todorov, "Optimality principles in sensorimotor control," *Nature Neuroscience*, vol. 7, no. 9, pp. 907–915, 2004.

[116] J. Diedrichsen, R. Shadmehr, and R. B. Ivry, "The coordination of movement: optimal feedback control and beyond," *Trends in Cognitive Sciences*, vol. 14, no. 1, pp. 31–39, 2010.

[117] H. Attias, "Planning by probabilistic inference," in *Neural Information Processing Systems*, 2003.

[118] M. Toussaint, "Robot trajectory optimization using approximate inference," in *Proceedings of the 26th International Conference On Machine Learning (ICML '09)*, pp. 1049–1056, June 2009.

[119] W. Penny, "Forwards and backwards inference for spatial cognition".

[120] E. Todorov, "Efficient computation of optimal actions," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 28, pp. 11478–11483, 2009.

[121] E. Todorov, "General duality between optimal control and estimation," in *Proceedings of the 47th IEEE Conference on Decision and Control (CDC '08)*, vol. 47, pp. 4286–4292, December 2008.

[122] J. Findlay and I. Gilchrist, *Active Vision: The Psychology of Looking and Seeing*, Oxford, UK, 2003.

[123] K. J. Friston, L. Harrison, and W. Penny, "Dynamic causal modelling," *NeuroImage*, vol. 19, no. 4, pp. 1273–1302, 2003.

[124] M. I. Garrido, J. M. Kilner, S. J. Kiebel, and K. J. Friston, "Evoked brain responses are generated by feedback loops," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 52, pp. 20961–20966, 2007.

[125] M. Boly, M. I. Garrido, O. Gosseries et al., "Preserved feed-forward but impaired top-down processes in the vegetative state," *Science*, vol. 332, no. 6031, pp. 858–862, 2011.