Opinion

# The free-energy principle: a rough guide to the brain?

**Karl Friston**

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

**This article reviews a free-energy formulation that advances Helmholtz's agenda to find principles of brain function based on conservation laws and neuronal energy. It rests on advances in statistical physics, theoretical biology and machine learning to explain a remarkable range of facts about brain structure and function. We could have just scratched the surface of what this formulation offers; for example, it is becoming clear that the Bayesian brain is just one facet of the free-energy principle and that perception is an inevitable consequence of active exchange with the environment. Furthermore, one can see easily how constructs like memory, attention, value, reinforcement and salience might disclose their simple relationships within this framework.**

## Introduction

The free-energy (see Glossary) principle is a simple postulate with complicated implications. It says that any adaptive change in the brain will minimize free-energy. This minimisation could be over evolutionary time (during natural selection) or milliseconds (during perceptual synthesis). In fact, the principle applies to any biological system that resists a tendency to disorder; from single-cell organisms to social networks.

The free-energy principle is an attempt to explain the structure and function of the brain, starting from the very fact that we exist: this fact places constraints on our interactions with the world, which have been studied for years in evolutionary biology and systems theory. However, recent advances in statistical physics and machine learning point to a simple scheme that enables biological systems to comply with these constraints. If one looks at the brain as implementing this scheme (minimising a variational bound on disorder), nearly every aspect of its anatomy and physiology starts to make sense. What follows is a review of this new perspective on old ideas.

## Free-energy and self-organization

So what is free-energy? Free-energy is an information theory quantity that bounds the evidence for a model of data [1–3]. Here, the data are sensory inputs and the model is encoded by the brain. More precisely, free-energy is greater than the negative log-evidence or 'surprise' in sensory data, given a model of how they were generated. Crucially, unlike surprise itself, free-energy can be evaluated because it is a function of sensory data and brain states. In fact, under simplifying assumptions (see later), it is just the amount of prediction error.

The motivation for the free-energy principle is again simple but fundamental. It rests upon the fact that self-organising biological agents resist a tendency to disorder and therefore minimize the entropy of their sensory states [4]. Under ergodic assumptions, this entropy is:

$$H(y) = -\int p(y|m)\ln p(y|m)dy = \lim_{T\to\infty}\frac{1}{T}\int_0^T -\ln p(y|m)dt \qquad \text{(Equation 1)}$$

See Box 1 for an explanation of the variables and Ref. [5] for details. This equation (Equation 1) means that mini-

### Glossary

**[Kullback-Leibler] divergence**: information divergence, information gain, cross or relative entropy is a non-commutative measure of the difference between two probability distributions.
**Bayesian surprise**: a measure of salience based on the divergence between the recognition and prior densities. It measures the information in the data that can be recognised.
**Conditional density**: or posterior density is the probability distribution of causes or model parameters, given some data; i.e., a probabilistic mapping from observed data to causes.
**Empirical priors**: priors that are induced by hierarchical models; they provide constraints on the recognition density is the usual way but depend on the data.
**Entropy**: the average surprise of outcomes sampled from a probability distribution or density. A density with low entropy means, on average, the outcome is relatively predictable.
**Ergodic**: a process is ergodic if its long term time-average converges to its ensemble average. Ergodic processes that evolve for a long time forget their initial states.
**Free-energy**: an information theory measure that bounds (is greater than) the surprise on sampling some data, given a generative model.
**Generalised coordinates**: of motion cover the value of a variable, its motion, acceleration, jerk and higher orders of motion. A point in generalised coordinates corresponds to a path or trajectory over time.
**Generative model**: or forward model is a probabilistic mapping from causes to observed consequences (data). It is usually specified in terms of the likelihood of getting some data given their causes (parameters of a model) and priors on the parameters
**Gradient descent**: an optimisation scheme that finds a minimum of a function by changing its arguments in proportion to the negative of the gradient of the function at the current value.
**Helmholtz machine**: device or scheme that uses a generative model to furnish a recognition density. They learn hidden structure in data by optimising the parameters of generative models.
**Prior**: the probability distribution or density on the causes of data that encode beliefs about those causes prior to observing the data.
**Recognition density**: or approximating conditional density is an approximate probability distribution of the causes of data. It is the product of inference or inverting a generative model.
**Stochastic**: the successive states of stochastic processes are governed by random effects.
**Sufficient statistics**: quantities which are sufficient to parameterise a probability density (e.g., mean and covariance of a Gaussian density).
**Surprise**: or self-information is the negative log-probability of an outcome. An improbable outcome is therefore surprising.

*Corresponding author:* Friston, K. (k.friston@fil.ion.ucl.ac.uk).

1364-6613/$ – see front matter  doi:10.1016/j.tics.2009.04.005 Available online xxxxxx

**Box 1. The free-energy principle**

Free-energy is a function of a recognition density and sensory input. It comprises two terms; the energy expected under this density and its entropy. The energy is simply the surprise about the joint occurrence of sensory input $y$ and its causes $\vartheta$. The free-energy depends on two densities; one that generates sensory samples and their causes, $p(y,\vartheta)$ and a recognition density on the causes, $q(\vartheta,\mu)$. This density is specified by its sufficient statistics, $\mu$, which we assume are encoded by the brain. This means free-energy induces a generative model $m$ for any system and a recognition density over the causes or parameters of that model. Given the functional form of these densities, the free energy can always be evaluated because it is a function of sensory input and the sufficient statistics. The free-energy principle states that all quantities that can change (sufficient statistics, $\mu$ and action, $\alpha$) minimise free-energy (Figure I).

**Optimising sufficient statistics**

It is easy to show that optimizing the recognition density renders it the conditional density on environmental causes, given the sensory data. This can be seen by expressing the free-energy as surprise $-\ln p(y|m)$ plus a [Kullback Leibler] divergence between the recognition and conditional densities. Because this divergence is always positive, minimising free-energy makes the recognition density an approximation to the true posterior probability. This means the system implicitly infers or represents the causes of its sensory samples in a Bayes optimal fashion. At the same time, the free-energy becomes a tight bound on surprise, which is minimised through action.

**Optimising action**

Acting on the environment by minimising free-energy through action enforces a sampling of sensory data that is consistent with the current representation. This can be seen with a second rearrangement of the free-energy as a mixture of accuracy and complexity. Crucially, action can only affect accuracy. This means the brain will reconfigure its sensory epithelia to sample inputs that are predicted by its representations; in other words, to minimise prediction error.

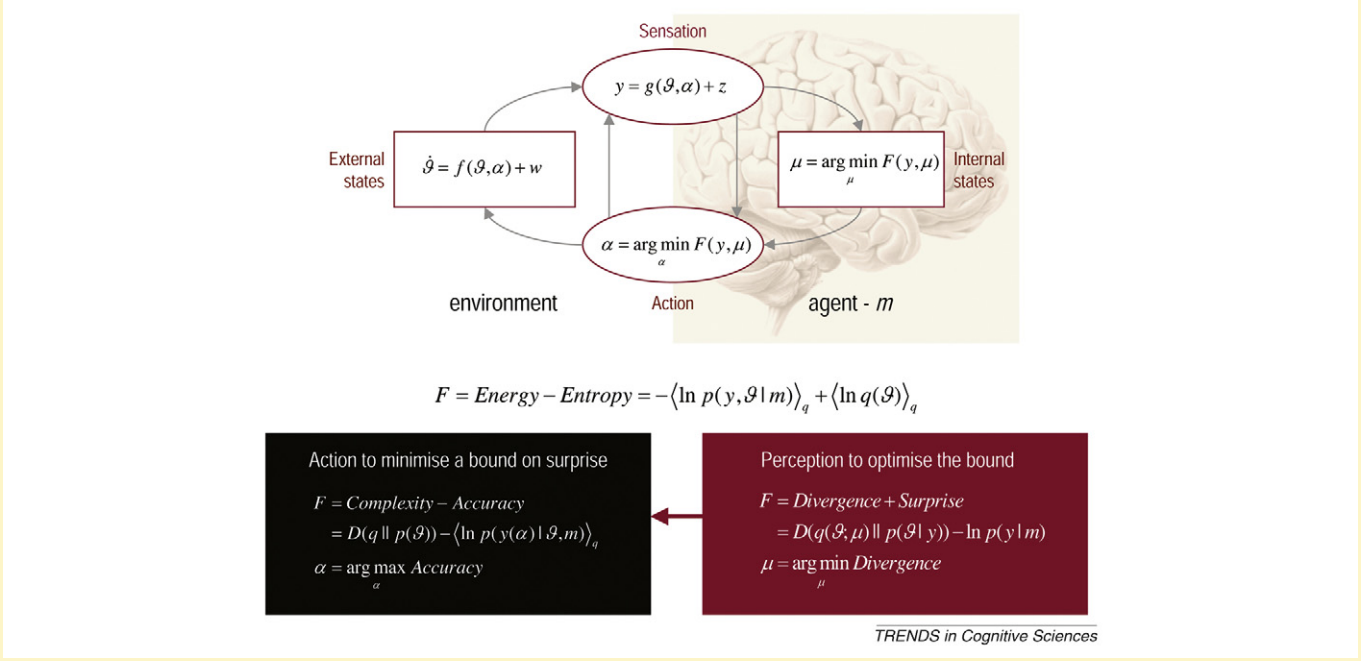


**Figure I**. Upper panel: schematic detailing the quantities that define free-energy. These include states of the brain $\mu$ and quantities describing exchange with the environment; sensory input $y = g(\vartheta,\alpha) + z$ and action $\alpha$ that changes the way the environment is sampled. The environment is described by equations of motion, $\dot{\vartheta} = f(\vartheta,\alpha) + w$, which specify the dynamics of environmental causes $\vartheta$. Brain states and action both change to minimise free-energy, which is a function of sensory input and a probabilistic representation (recognition density) $q(\vartheta,\mu)$ encoded by $\mu$. Lower panel: alternative expressions for the free-energy that show what its minimisation entails. For action, free-energy can only be suppressed by increasing the accuracy of sensory data (i.e. selectively sampling data that are predicted by the representation). Conversely, optimising brain states make the representation an approximate conditional density on the causes of sensory input. This optimisation makes the free-energy bound on surprise tighter and enables action to avoid surprising sensory encounters.

mizing entropy corresponds to suppressing surprise over time. In brief, for a well-defined agent to exist it must occupy a limited repertoire of states; for example, a fish in water. This means the equilibrium density of an ensemble of agents, describing the probability of finding an agent in a particular state, must have low entropy: a distribution with low entropy just means a small number of states are occupied most of the time. Because entropy is the long-term average of surprise, agents must avoid surprising states (e.g. a fish out of water). But there is a problem; agents cannot evaluate surprise directly; this would entail knowing all the hidden states of the world causing sensory input. However, an agent can avoid surprising exchanges with the world if it minimises its free-energy because free-energy is always bigger than surprise.

### *The Bayesian brain*

Mathematically, the difference between free-energy and surprise is the divergence between a probabilistic representation (recognition density) encoded by the agent and the true conditional distribution of the causes of sensory input (Box 1). This representation enables the brain to reduce free-energy by changing its representation, which makes the recognition density an approximate conditional density. This corresponds to Bayesian inference on unknown states of the world causing sensory data [6]. In short, the free-energy principle subsumes the Bayesian brain hypothesis; or the notion that the brain is an inference or Helmholtz machine [7–11]. Note that we have effectively shown that biological agents must engage in some form of Bayesian perception to avoid surprising exchanges with the world.

However, perception is only half the story; it makes free-energy a good proxy for surprise but it does not change the sensations themselves or their surprise.

### Beyond Bayes to active inference

To reduce surprise we have to change sensory input. This is where the free-energy principle comes into its own: it says that our actions should also minimize free-energy [5,12]. We are open systems in exchange with the environment; the environment acts on us to produce sensory impressions and we act on the environment to change its states. This exchange rests upon sensory and effector organs (like photoreceptors and oculomotor muscles). If we change the environment or our relationship to it, sensory input changes. Therefore, action can reduce free-energy (i.e. prediction errors) by changing sensory input, whereas perception reduces free-energy by changing predictions.

Mathematically, the free-energy principle requires us to sample sensory information that conforms to our expectations (Box 1). This does not mean we can simply shut down sensory channels to avoid surprise; we can only change sensory signals through action. For example, we cannot avoid pain unless we remove the noxious stimulus. In short, we sample the world to ensure our predictions become a self-fulfilling prophecy and surprises are avoided. In this view, perception is enslaved by action to provide veridical predictions (more formally, to make the free-energy a tight bound on surprise) that guides active sampling of the sensorium [5,12].

In summary, (i) agents resist a natural tendency to disorder by minimising a free-energy bound on surprise; (ii) this entails acting on the environment to avoid surprises, which (iii) rests on making Bayesian inferences about the world. In this view, the Bayesian brain ceases to be a hypothesis, it is mandated by the free-energy principle; free-energy is not used to finesse perception, perceptual inference is necessary to minimise free-energy (Box 1). This provides a principled explanation for action and perception that serve jointly to suppress surprise or prediction error; but it does not explain how the brain does this or how it encodes the representations that are optimised.

### Neuronal implementation

The free-energy principle requires the brain to represent the causes of sensory input. The nature of this representation is dictated by physiological and anatomical constraints. Irrespective of its form, the brain has to encode a recognition density with its physical attributes (e.g. synaptic activity and efficacy). These have the role of sufficient statistics, which are just numbers that specify a distribution, like mean and dispersion.

Clearly, the causes of sensory data can change at different timescales. For example, environmental states could be encoded by neuronal dynamics on a millisecond timescale, whereas causal regularities or parameters that change slowly could be encoded in connection strengths. These quantities (states and parameters) pertain to deterministic dynamics in the world. However, it is also necessary to represent random effects; for example, the amplitude of random fluctuations on states. This induces a third class of quantities (precisions or inverse variances) that generate uncertainty about states. The sufficient statistics of precision could be encoded in post-synaptic sensitivity or gain [13]; through the activity of classical neuromodulatory neurotransmitter systems (e.g. acetylcholine or dopamine; cf Ref. [14]) or synchronous interactions among neighbouring populations [15]. Precisions are an important class of representation that are induced by randomness in the world and are the focus of later sections.

According to the free-energy principle, the sufficient statistics representing all three sorts of quantities will change to minimise free-energy. This provides a principled explanation for perception, memory and attention; it accounts for perceptual inference (optimisation of synaptic activity to encode the states of the environment); perceptual learning and memory (optimisation of synaptic connections that encode contingencies and causal regularities) and attention (neuromodulatory optimisation of synaptic gain that encodes the precision of states) (Box 2).

This optimisation can be formulated as a gradient descent on free-energy to furnish differential equations, which prescribe recognition dynamics for synaptic activity, efficacy and gain. These dynamics depends on the form of the generative model employed by the brain and the sufficient statistics it encodes. If we assume the recognition density is a high-dimensional Gaussian density (the Laplace approximation), then recognition dynamics adopt plausible neuronal forms: optimising the sufficient statistics of the states looks exactly like predictive coding [10], which involves recurrent message-passing between populations encoding predictions and prediction errors. Optimising the sufficient statistics of the parameters is formally identical to associative plasticity and optimising the sufficient statistics of precision is similar to the assimilation of prediction error in reinforcement learning schemes [16].

Under the Laplace assumption, recognition dynamics become evidence accumulation schemes [17], in which changes in neuronal activity accumulate evidence (prediction error) [18]. Furthermore, one can understand the hierarchical deployment of cortical areas and the nature of message passing among cortical levels in terms of minimising prediction error under hierarchical dynamic models of the world [19,20] (Box 3). Hierarchical models are important because they are formally equivalent to empirical Bayesian models [21], in which higher levels provide empirical priors or constraints on lower levels. This allows one to interpret top-down effects in the brain as instantiating empirical priors. Under this perspective, suppressing free-energy means that each level is trying to explain away prediction errors at its own level and in the level below; leading to recurrent self-organized dynamics that converge on a self-consistent representation of sensory causes, at multiple levels of description. Recent advances in Bayesian filtering (that rest on generalised coordinates of motion) have extended the notion of empirical priors in hierarchal models to temporal hierarchies, which provide a plausible account of how we categorise sensory streams and sequences [19,22].

In summary, the free-energy principle prescribes recognition dynamics if we specify (i) the form of the generative

## Box 2. Neurobiological implementation

Generative models in the brain: to suppress free-energy one needs a probabilistic generative model of how the sensorium is caused. These models $p(y,\vartheta) = p(y|\vartheta)\,p(\vartheta)$ entail the likelihood, $p(y|\vartheta)$ of getting some data, $y$, given their causes $\vartheta \supset \{x(t),\theta,\lambda\}$ and prior beliefs $p(\vartheta)$. The models employed by the brain have to explain a world with complex dynamics on continuous states. Hierarchical dynamic models provide a general form and specify sensory data as a mixture of predictions (based on causes) and random effects:

$$\begin{aligned} y(t) &= g(x^{(1)}, v^{(1)}, \theta^{(1)}) + z^{(1)} \\ x^{(1)} &= f(x^{(1)}, v^{(1)}, \theta^{(1)}) + w^{(1)} \\ v^{(i-1)} &= g(x^{(i)}, v^{(i)}, \theta^{(i)}) + z^{(i)} \\ x^{(i)} &= f(x^{(i)}, v^{(i)}, \theta^{(i)}) + w^{(1)} \\ v^{(m)} &= \eta + z^{(m+1)} \end{aligned} \qquad \begin{bmatrix} z^{(i)} \\ w^{(i)} \end{bmatrix} \sim N(0, \Pi(\lambda^{(i)})^{-1})$$

(Euqation I)

Here (Equation I), $g^{(i)}$ and $f^{(i)}$ are continuous nonlinear functions of (hidden and causal) states, parameterised by $\theta^{(i)}$. Independent random fluctuations $z(t)^{(i)}$ and $w(t)^{(i)}$ have the role of observation noise at the first level and state-noise at higher levels. Causal states $(t)^{(i)}$ link levels, whereas hidden states $x(t)^{(i)}$ link dynamics over time and endow the model with memory. In hierarchical form, the output of one level acts as an input to the next. Top-down causes can enter the equations nonlinearly to produce quite complicated generalised convolutions of high-level causes with 'deep' (hierarchical) structure.

**Hierarchies and empirical priors**

Gaussian assumptions about the fluctuations specify the likelihood. Similarly, Gaussian assumptions about state-noise furnish empirical priors in terms of predicted motion. These assumptions are encoded by their or precision, $\Pi(\lambda)$, which depends on precision parameters $\lambda$. The conditional independence of the fluctuations means that these models have a Markov property over levels, which simplifies the architecture of attending inference schemes. In short; a hierarchical form allows models to construct their own priors. This feature is central to many inference procedures, ranging from mixed-effects analyses in classical statistics to automatic relevance determination in machine learning.

**Recognition dynamics**

Given a generative model it is relatively easy to compute the free-energy and derivatives with respect to the sufficient statistics. This enables one to write down recognition dynamics in terms of a gradient descent on the free-energy $F$ or its path-integral, $A$ (Action). Note that only time-dependent representations (i.e. expected states) minimise free-energy; all the others minimise Action. This means the recognition dynamics for states reduce to first-order differential equations of motion (evidence accumulation schemes). However, the dynamics for parameters (syntactic efficacy) and precisions (synaptic gain) are second-order and driven by terms that themselves accumulate gradients (synaptic traces or tags). Box 3 shows the form of recognition dynamics, under hierarchical dynamic models (Figure I).

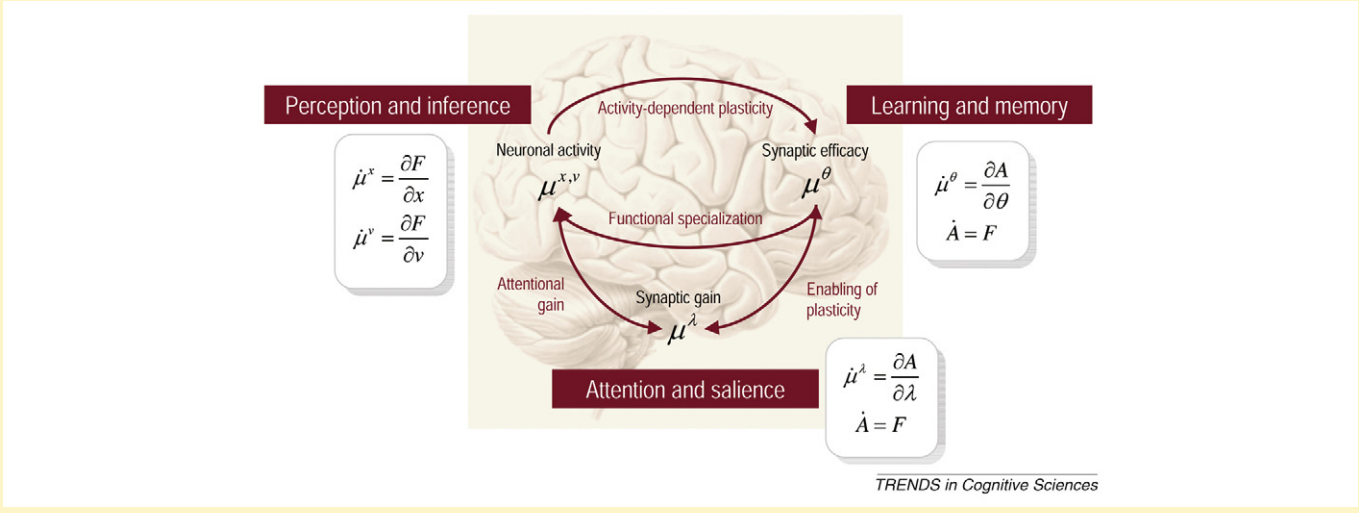


**Figure I.** The sufficient statistics representing a hierarchical dynamic model of the world and their recognition dynamics under the free-energy principle. The recognition density is encoded in terms of its sufficient statistics; $\mu \supset \{\mu^x, \mu^v, \mu^\theta, \mu^\lambda\}$. These representations or statistics change to minimise free-energy or its path-integral (i.e. Action, A). Here, we consider three sorts of representations pertaining to the states; $\{x,v\}$, parameters; $\theta$ and precisions; $\lambda$ of a hierarchical dynamic model. We suppose these are encoded by neural activity, synaptic connectivity and gain respectively. Crucially, the optimisation of any one representation depends on the others. The differential equations associated with this partition represent a gradient descent on free-energy and correspond to (i) perceptual inference on states of the world (i.e. optimising synaptic activity); (ii) perceptual learning of the parameters underlying causal regularities (i.e. optimising synaptic efficacy) and (iii) attention or optimising the expected precision of states in the face of random fluctuations and uncertainty (i.e. optimising synaptic gain).

model used by the brain, (ii) the form of the recognition density and (iii) how its sufficient statistics are optimised. The list in Table 1 assumes that (i) the brain uses a hierarchical dynamic model in generalised coordinates of motion, (ii) the recognition density is Gaussian and (iii) its expectation is optimised using gradient descent. These assumptions enable one to write down equations that predict the dynamics of synaptic activity (encoding expected states), synaptic efficacy (encoding expected parameters) and neuromodulation of synaptic gain (encoding expected precision). In Ref. [19] we consider each of these assumptions, in relation to their alternatives.

## New perspectives?

We have tried to substantiate the aforementioned formulation by explaining many empirical aspects of anatomy and physiology in terms of optimising free-energy. One can explain a remarkable range of facts; for example, the hierarchical arrangement of cortical areas, functional asymmetries between forward and backward connections, explaining away effects and many psychophysical and cognitive phenomena; see Ref. [19] and Table 1. However, we now focus on prospective issues that could offer new and possibly contentious views of constructs in neuroscience. These examples highlight the importance of

## Box 3. Recognition dynamics

**Recognition dynamics and prediction error**

If we assume that pre-synaptic activity encodes the conditional expectation of states, then a gradient descent on free-energy prescribes neuronal dynamics entailed by perception. Under the Laplace assumption (Table 2), these recognition dynamics can be expressed compactly in terms prediction errors $\varepsilon^{(i)}$ on the causal states and motion of hidden states. The ensuing equations suggest two neuronal populations that exchange messages; causal or hidden 'state-units' whose activity encodes the expected or predicted state and 'error-units' encoding precision-weighted prediction error (Figure I).

**Hierarchical message passing**

Under hierarchical models, error-units receive messages from the states in the same level and the level above; whereas state-units are driven by error-units in the same level and the level below. Crucially, inference requires only the error from the lower level $\xi^{(i)} = \Pi^{(i)}\varepsilon^{(i)} = \varepsilon^{(i)} - \Lambda^{(i)}\dot{\xi}^{(i)}$ and the level in question, $\dot{\xi}^{(i+1)}$. These provide bottom-up and lateral messages that drive conditional expectations $\mu^{(i)}$ towards better predictions to explain away prediction error. These top-down and lateral predictions correspond to $g^{(i)}$ and $f^{(i)}$. This is the essence of recurrent message passing between hierarchical levels that suppresses free-energy or prediction error. This scheme suggests that connections between error and state-units are reciprocal; the only connections that link levels are forward connections conveying prediction error to state-units and reciprocal backward connections that mediate predictions

**Functional asymmetries**

We can identify error-units with superficial pyramidal cells because the only messages that are passed up the hierarchy are prediction errors and superficial pyramidal cells originate forward connections in the brain. This is useful because these cells are primarily responsible for electroencephalographic (EEG) signals. Similarly, the only messages that are passed down the hierarchy are the predictions from state-units. The sources of backward connections are deep pyramidal cells and one might deduce that these encode the expected causes of sensory states [20]. Crucially, state-units receive a linear mixture of prediction error. This is what is observed physiologically; bottom-up driving inputs elicit obligatory responses that do not depend on other bottom-up inputs. The prediction error depends on predictions conveyed by backward connections. These embody nonlinearities in the generative model. Again, this is entirely consistent with the modulatory characteristics of backward connections.

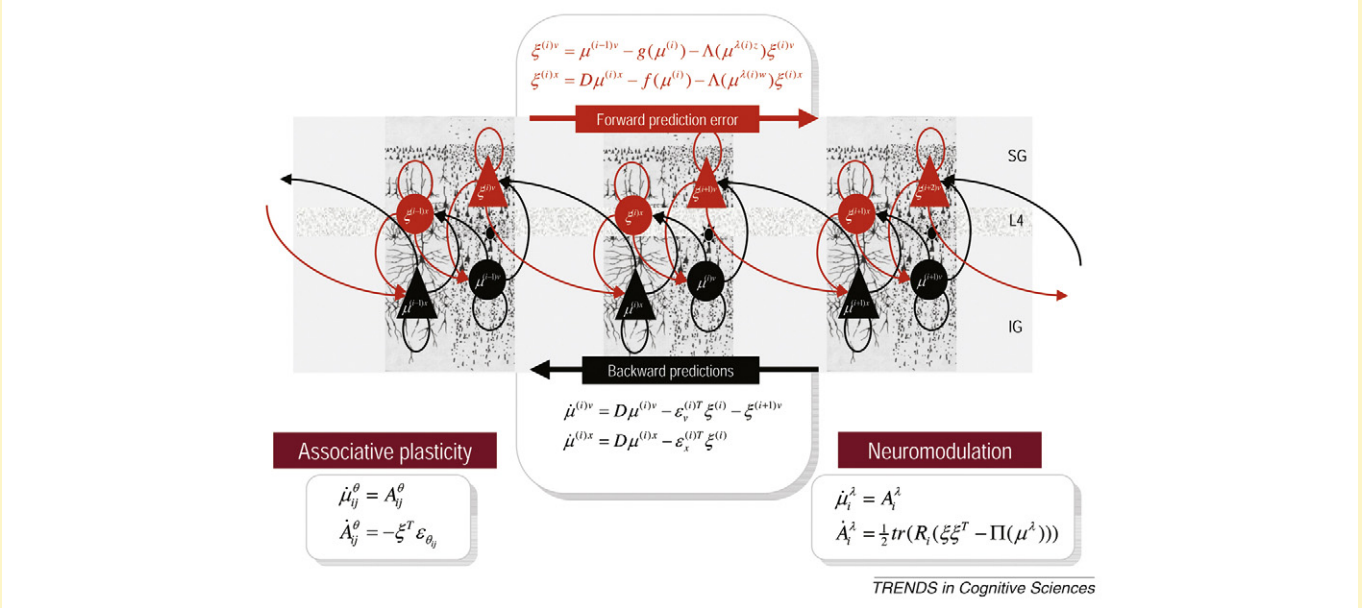


**Figure I.** Schematic detailing the neuronal architectures that might encode a density on the states of a hierarchical dynamic model. This shows the speculative cells of origin of forward driving connections that convey prediction error from a lower area to a higher area and the backward connections that construct predictions [11,20]. These predictions try to explain away prediction error in lower levels. In this scheme, the sources of forward and backward connections are superficial and deep pyramidal cells, respectively. The equations represent a gradient descent on free-energy under the hierarchical dynamic models of Box 2 (see Ref. [19] for details). State-units are in black and error-units in red. Here, neuronal populations are deployed hierarchically within three cortical areas (or macro-columns). Within each area, the cells are shown in relation to cortical layers: supra-granular (SG) granular (L4) and infra-granular (IG) layers. In this figure, subscripts denote derivatives.

representing precision (uncertainty) through neuromodulation.

### *The neural code, gain and precision*

A key implementational issue is how the brain encodes the recognition density. The free-energy principle induces this density, which has to be represented by its sufficient statistics. It is therefore a given that the brain represents probability distributions over sensory causes [23]. But what is the form of this distribution and what are the sufficient statistics that constitute the brain's probabilistic code? There are two putative forms; free-form and fixed-form. Proposals for free-form approximations include particle filtering [9] and probabilistic population codes [24]. In particle filtering, the recognition density is represented by the sample density of neuronal ensembles; whose activity encodes the location of particles in state-space. In convolution and probabilistic gain population codes [6], neuronal activity encodes the amplitude of fixed basis functions (Table 2). Fixed-form approximations are usually multinomial or Gaussian. Multinomial forms assume the world is in one of several discrete states and are usually associated with hidden Markov models [18,25]. Conversely, the Gaussian or Laplace assumption allows for continuous and correlated states.

Any scheme that optimises the sufficient statistics of these forms must conform to the free-energy principle. So why have we focussed on the Laplace approximation?

**Table 1. Structural and functional aspects of the brain that can be explained under a free-energy formulation (see Ref, [19] for details and references)**

| Domain | Predictions |
|---|---|
| **Anatomy and connectivity:** Explains the hierarchical deployment of cortical areas, recurrent architectures with functionally asymmetric forward and backward connections | Hierarchical cortical organization<br>Distinct neuronal subpopulations, encoding expected states of the world and prediction error<br>Extrinsic forward connections convey prediction error (from superficial pyramidal cells) and backward connections mediate predictions (from deep pyramidal cells)<br>Functional asymmetries in forwards (linear) and backwards (nonlinear) connections are mandated by nonlinearities in the generative model encoded by backward connections<br>Principal cells elaborating predictions (e.g. deep pyramidal cells) could show distinct (low-pass) dynamics, relative to those encoding error (e.g. superficial pyramidal cells)<br>Recurrent dynamics are intrinsically stable because they suppress prediction error (cf no strong loops) |
| **Synaptic physiology:** Explains both (short-term) neuromodulatory gain-control and (long-term) associative plasticity | Scaling of prediction errors, in proportion to their precision, affords the cortical bias or gain control seen in attention<br>Short-term modulation of synaptic gain encoding precision or uncertainty (which optimises a path-integral) must be slower than neuronal dynamics (which optimise free-energy *per se*)<br>Long-term plasticity that is formally identical to Hebbian or associative plasticity<br>Neuromodulatory factors could have a dual role in modulating postsynaptic responsiveness (e.g. through after-hyperpolarising currents) and synaptic plasticity |
| **Electrophysiology:** Accounts for (extra)-classical receptive field effects and long-latency (endogenous) components of evoked cortical responses | Event-related responses are self-limiting transients, where late components rest on top-down suppression of prediction error<br>Sensory responses are greater for surprising, unpredictable or incoherent stimuli<br>The attenuation of responses encoding prediction error, with perceptual learning, explains repetition suppression (e.g. mismatch negativity in electroencephalography) |
| **Psychophysiology:** Accounts for the behavioural correlates of these physiological phenomena | For example, priming and global precedence. In cognitive terms, it furnishes a framework in which to model and understand things like perceptual categorisation, temporal sequencing and attention |

First, free-form approximations do not scale. For example, to represent a face with thirty or so attributes, we would need to populate a thirty-dimensional perceptual state-space with more neurons (i.e. particles or basis functions) than are in the brain. The argument for free-form approximations is that they can encode complicated (e.g. multimodal) recognition densities. However, it is trivial to represent non-Gaussian forms under the Laplace approximation by using a nonlinear transformation of variables. For example, scale-parameters (e.g. precisions or rate constants) can be modelled with a log-normal distribution (this is generalised to $\varphi$-normal forms in Table 2). Furthermore, there is no electrophysiological or psychophysical evidence to suggest that the brain can encode multimodal approximations: indeed, with ambiguous figures, the fact that percepts are bistable (as opposed to bimodal and stable) suggests the recognition density is unimodal. Although multinomial approximations and hidden Markov models have an appealing simplicity (and map to empirical studies of categorization and decisions [17]) they cannot represent dependencies among states [26]. By contrast, the Laplace approximation can handle continuous and correlated states efficiently; it is particularly efficient because the recognition density is specified completely by its mean: the only other sufficient statistic (the conditional precision) can be derived from the mean and does not need to be encoded explicitly [27].

Having said this, there are some nice formal convergences among recognition schemes under different distributional approximations. For example, in hierarchical Bayesian models based upon hidden Markov models [25], belief propagation appeals to predictive coding. In

**Table 2. Probabilistic neuronal codes, based on Box 2 in Ref. [23]**

| | Code | Form[a] | Comments | Refs |
|---|---|---|---|---|
| **Free-form** population codes | **Particle** continuous | $\int (s\text{-}c)^n q(s)ds = \frac{1}{N}\sum_i (\mu_i - c)^n$ | The moments of $q(s)$ are encoded by the sample moments of $N$ 'particles' or neurons. | [9,48] |
| | **Convolution** continuous | $q(s) = \frac{1}{Z(\mu)}\sum_i \mu_i \varphi_i(S)$ | $q(s)$ is encoded as a mixture of basis functions $\varphi_i(s)$ with fixed location and form (cf, tuning curves). | [6,49] |
| | **Probabilistic** continuous | $q(s) = \frac{1}{Z(\mu)}\prod_i \frac{\exp(\varphi_i(s))\varphi_i(s)^{\mu_i}}{\mu_i!}$ | This example of a PPC or probabilistic population code assumes neuronal variability is independent and Poisson. | [24,26,50] |
| **Fixed-form** | **Explicit** discrete | $q(s = s_i) = \frac{1}{Z(\mu)}(\mu_i + c_i)$ | Multinomial code, where neuronal states are proportional to the probability of the cause encoded by each state. | [51] |
| | **Logarithmic** discrete | $q(s = s_i) = \frac{1}{Z(\mu)}(\exp(\mu_i) + c_i)$ | Multinomial code, where neuronal states represent the log-probability; this subsumes log-likelihood ratio codes. | [17,25,52] |
| | **$\varphi$ - Normal or Laplace** continuous | $q(\varphi(s)) = \frac{1}{Z(\mu)}\exp(-\frac{1}{2}\mu^T \Pi(\mu)\mu)$ | The mean is encoded explicitly and the precision $\Pi(\mu)$ implicitly, as a function of the mean. | [10,11,19] |

[a]Neuronal activity encodes an approximate conditional or recognition density, $q(s)$, on states of the world $s = \{x, v\}$, in terms of sufficient statistics, $\mu$. $Z(\mu.)$ is a partition function or normalising constant, $\varphi(s)$ is some analytic nonlinear function of the states and $c$ is a constant.

probabilistic population codes [24] precision is encoded by gain, through Poisson-like neuronal activity; whereas it is encoded by synaptic gain under the Laplace approximation [11].

### *Attention and precision*

Under hierarchical models of perception, it is necessary to optimise the relative precision of empirical (top-down) priors and (bottom-up) sensory evidence. Neurobiologically, this corresponds to modulating the gain of error-units (Box 3); in other words, synaptic gain control of the sort invoked for attention. This optimisation is crucial for inference and is like estimating the standard error (inverse precision) in a *t*-test: the importance of representing precision is even more crucial in hierarchical inference because it controls the relative influence of prior expectations at different levels. This seems a natural way to understand attention and accounts for attentional modulation of local competition [28] and contrast gain [29]. In a hierarchical context, it also accommodates functionalist perceptiveness such as feature integration [30]. Mechanistically, the role of cholinergic neurotransmission in modulating post-synaptic gain (and encoding precision) fits comfortably with its role in attention [14,31,32].

When I hear attention is 'taking possession by the mind, in clear and vivid form, . . .' [33], I think 'No it's not; attention is simply the process of optimising precision during hierarchical inference'. This might not have the poetry of Jamesian formulations but it helps understand the simplicity of attention and its necessary role in perception.

In short; attention might not be the 'selection' of sensory channels but an emergent property of 'prediction'; where high-precision prediction-errors enjoy greater gain.

### *Value-learning, motivational salience and precision*

Many treatments of behaviour and choice under uncertainty borrow from behavioural economics, control theory and dynamic programming (e.g. Refs [34–39]) to model optimal decision making and reinforcement learning. These formulations are united by the notion of loss, reward or utility that guides behaviour to maximise value or expected reward in the future. The biological substrates of value-learning have focussed on the dopaminergic system [40]. So what does free-energy bring to the table? It brings something quite fundamental; it says that loss is surprise (or a free-energy that bounds surprise) and that expected loss is expected surprise or entropy (or a path-integral of free-energy that bounds entropy). This is important because the quantities optimised by action under value-learning are exactly the same as those optimised by active sampling under the free-energy principle. This means the notion of value *per se* is redundant and that much of reinforcement and procedural learning can be recast in terms of active inference.

Recent developments provide compelling proof-of-principle that active sampling can be used to solve quite complicated problems in optimum control, without the use of value-learning (e.g. the mountain car problem [5]). The basic idea is to replace value-functions with prior expectations about sensory trajectories. Action then ensures prior expectations are met and desired states are frequented. Optimal priors are induced by perpetual learning in a training environment. This resolves two key problems with value-learning: it enables optimal control without access to hidden states of the world and circumvents the (intractable) problem of solving for the value-function. Crucially, this approach points to a central role for dopamine in both learning and prosecuting optimum behaviour. This is because action is only called on to explain away prediction errors, when predictions are precise; the absence of precise priors (low dopamine) leads to small prediction errors and poverty of action (bradykinesia seen in Parkinson's disease and with neuroleptics). (See Ref. [5] for details.)

This perspective could call for a reappraisal of the role of dopamine. If dopamine encodes precision through its classical neuromodulatory actions, how can this be reconciled with the conventional view [37,40] that it encodes prediction error on reward? The answer could be that dopamine might not encode the 'prediction error on value' but the 'value of prediction error' (the learning rate in Rescorla-

**Box 4. Questions for further research**

**What is the computational role of neuromodulation?**
Previous treatments suggest that modulatory neurotransmitters have distinct roles; for example. 'dopamine signals the error in reward prediction, serotonin controls the time scale of reward prediction, noradrenalin controls the randomness in action selection, and acetylcholine controls the speed of memory update' [53]. This contrasts with a single role in encoding precision above. Can the apparently diverse functions of these neurotransmitters be understood in terms of one role (encoding precision) in different parts of the brain?

**Can we entertain ambiguous percepts?**
Although not an integral part of the free-energy principle, we claim the brain uses unimodal recognition densities to represent one thing at a time. Although, there is compelling evidence for bimodal 'priors' in sensorimotor learning [54], people usually assume the 'recognition' density collapses to a single percept, when sensory information becomes available. The implicit challenge here is to find any electrophysiological or psychophysical evidence for multimodal recognition densities.

**Does avoiding surprise suppress salient information?**
No; a careful analysis of visual search and attention suggests that: 'only data observations which substantially affect the observer's beliefs yield (Bayesian) surprise, irrespectively of how rare or informative in Shannon's sense these observations are' [55]. This is consistent with active sampling of things we recognize (to reduce free-energy). However, it remains an interesting challenge to formally relate Bayesian surprise to the free-energy bound on (Shannon) surprise. A key issue here is whether saliency can be shown to depend on top-down perceptual expectations (P. König, personal communication).

**Which optimisation schemes does the brain use?**
We have assumed that the brain uses a deterministic gradient descent on free-energy to optimise action and perception. However, it might also use stochastic searches; sampling the sensorium randomly for a percept with low free-energy. Indeed, there is compelling evidence that our eye movements implement an optimal stochastic strategy [56]. This raises interesting questions about the role of stochastic searches; from visual search to foraging, in both perception and action.

Wagner models); where value 'is' precision or incentive salience [41,42].

In short; goal-directed behaviour might not be the 'selection' of responses but an emergent property of 'prediction'; in which high-precision predictions seem to have greater motivational salience and control over action.

If these ideas are right (Box 4), they speak to a pleasing symmetry between the role of dopamine in optimising precision in anterior (e.g. mesocortical and mesolimbic) systems trying to predict proprioceptive and interoceptive sensations (i.e. value-learning) and the role of acetylcholine in optimising hierarchical inference on extroceptive input in posterior (e.g. paralimbic and parietal) systems (i.e. attention) [43–45]. Furthermore, they sit comfortably with a gating role for dopamine [46] in selecting the percepts that guide action [47].

## Conclusion

In conclusion, the free-energy principle might provide a comprehensive account of how we represent the world and come to sample it adaptively. Furthermore, it provides a mathematical specification of 'what' the brain is doing; it is suppressing free-energy. If this uses gradient descent, one can derive differential equations that prescribe recognition dynamics that specify 'how' the brain might operate. The ensuing representations are used to elaborate prediction errors, which action tries to suppress by moving sensory epithelia to sample expected input. In this way, changes in synaptic activity, connectivity and gain can be understood as perceptual inference, learning and attention. The form of this optimisation suggests some specific attributes of neuronal responses, which look very much like empirically evoked responses, plasticity and neuromodulation.

### Acknowledgements

The Wellcome Trust funded this work. I would like to thank my colleagues at the Wellcome trust Centre for Neuroimaging and the Gatsby Computational Neuroscience unit for helpful and formative discussions.

### References

1 Feynman, R.P. (1972) *Statistical Mechanics.* Benjamin
2 MacKay, D.J.C. (1995) Free-energy minimisation algorithm for decoding and cryptoanalysis. *Electron. Lett.* 31, 445–447
3 Neal, R.M. and Hinton, G.E. (1998) A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models* (Jordan, M.I., ed.), pp. 355–368, Kluwer Academic Publishers
4 Friston, K. *et al.* (2006) A free energy principle for the brain. *J. Physiol. Paris* 100, 70–87
5 Friston, K.J. *et al.* Reinforcement-learning or active inference? *PLoS ONE.* (in press)
6 Knill, D.C. and Pouget, A. (2004) The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* 27, 712–719
7 Ballard, D.H. *et al.* (1983) Parallel visual computation. *Nature* 306, 21–26
8 Dayan, P. *et al.* (1995) The Helmholtz machine. *Neural Comput.* 7, 889–904
9 Lee, T.S. and Mumford, D. (2003) Hierarchical Bayesian inference in the visual cortex. *J. Opt. Soc. Am. A. Opt. Image Sci. Vis* 20, 1434–1448
10 Rao, R.P. and Ballard, D.H. (1999) Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive field effects. *Nat. Neurosci.* 2, 79–87
11 Friston, K.J. (2005) A theory of cortical responses. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 360, 815–836
12 Friston, K.J. and Stephan, K.E. (2007) Free energy and the brain. *Synthese* 159, 417–458
13 Abbott, L.F. *et al.* (1997) Synaptic depression and cortical gain control. *Science* 275, 220–224
14 Yu, A.J. and Dayan, P. (2005) Uncertainty, neuromodulation and attention. *Neuron* 46, 681–692
15 Womelsdorf, T. and Fries, P. (2006) Neuronal coherence during selective attentional processing and sensory-motor integration. *J. Physiol. Paris* 100, 182–193
16 Schultz, W. *et al.* (1997) A neural substrate of prediction and reward. *Science* 275, 1593–1599
17 Gold, J.I. and Shadlen, M.N. (2001) Neural computations that underlie decisions about sensory stimuli. *Trends Cogn. Sci.* 5, 10–16
18 Rao, R.P.N. (2004) Bayesian computation in recurrent neural circuits. *Neural Comput.* 16, 1–38
19 Friston, K. (2008) Hierarchical models in the brain. *PLOS Comput. Biol.* 4, e1000211
20 Mumford, D. (1992) On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biol. Cybern* 66, 241–251
21 Kass, R.E. and Steffey, D. (1989) Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *J. Am. Stat. Assoc.* 407, 717–726
22 Kiebel, S.J. *et al.* (2008) A hierarchy of time-scales and the brain. *PLoS Comput. Biol.* 4, e1000209
23 Ma, W.J. *et al.* (2008) Spiking networks for Bayesian inference and choice. *Curr. Opin. Neurobiol.* 18, 217–222
24 Ma, W.J. *et al.* (2006) Bayesian inference with probabilistic population codes. *Nat. Neurosci.* 9, 1432–1438
25 Deneve, S. (2008) Bayesian spiking neurons I: inference. *Neural Comput.* 20, 91–117
26 Sahani, M. and Dayan, P. (2003) Doubly distributional population codes: Simultaneous representation of uncertainty and multiplicity. *Neural Comput.* 15, 2255–2279
27 Friston, K.J. *et al.* (2008) DEM: A variational treatment of dynamic systems. *Neuroimage* 41, 849–885
28 Desimone, R. (1996) Neural mechanisms for visual memory and their role in attention. *Proc. Natl. Acad. Sci. U. S. A.* 93, 13494–13499
29 Maunsell, J.H. and Treue, S. (2006) Feature-based attention in visual cortex. *Trends Neurosci.* 29, 317–322
30 Treisman, A. (1998) Feature binding, attention and object perception. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 353, 1295–1306
31 Schroeder, C.E. *et al.* (2001) Determinants and mechanisms of attentional modulation of neural processing. *Front. Biosci.* 6, D672–D684
32 Hirayama, J. *et al.* (2004) Bayesian representation learning in the cortex regulated by acetylcholine. *Neural Netw.* 17, 1391–1400
33 James, W. (1890) In *The Principles of Psychology* (Vol.1), DoverPublications
34 Coricelli, G. *et al.* (2007) Brain, emotion and decision making: the paradigmatic example of regret. *Trends Cogn. Sci.* 11, 258–265
35 Daw, N.D. and Doya, K. (2006) The computational neurobiology of learning and reward. *Curr. Opin. Neurobiol.* 16, 199–204
36 Hsu, M. *et al.* (2005) Neural systems responding to degrees of uncertainty in human decision-making. *Science* 310, 1680–1683
37 Montague, P.R. *et al.* (1995) Bee foraging in uncertain environments using predictive Hebbian learning. *Nature* 377, 725–728
38 Sutton, R.S. and Barto, A.G. (1981) Toward a modern theory of adaptive networks: expectation and prediction. *Psychol. Rev.* 88, 135–170
39 Todorov, E. (2006) Linearly-solvable Markov decision problems. In *Advances in Neural Information Processing Systems* (19) (Scholkopf *et al.*, eds), In pp. 1369–1376, MIT Press
40 Schultz, W. (1998) Predictive reward signal of dopamine neurons. *J. Neurophysiol.* 80, 1–27
41 Zink, C.F. *et al.* (2004) Human striatal responses to monetary reward depend on saliency. *Neuron* 42, 509–517
42 Berridge, K.C. (2007) The debate over dopamine's role in reward: the case for incentive salience. *Psychopharmacology (Berl)* 191, 391–431
43 Grossberg, S. and Versace, M. (2008) Spikes, synchrony, and attentive learning by laminar thalamocortical circuits. *Brain Res.* 1218, 278–312
44 Disney, A.A. *et al.* (2007) Gain modulation by nicotine in macaque v1. *Neuron.* 56, 701–713

45 Furey, M.L. *et al.* (2008) Selective effects of cholinergic modulation on task performance during selective attention. *Neuropsychopharmacology* 33, 913–923
46 O'Reilly, R.C. *et al.* (2002) Prefrontal cortex and dynamic categorization tasks: representational organization and neuromodulatory control. *Cereb. Cortex.* 12, 246–257
47 Redgrave, P. *et al.* (1999) The basal ganglia: a vertebrate solution to the selection problem? *Neuroscience* 89, 1009–1023
48 Paulin, M.G. (2005) Evolution of the cerebellum as a neuronal machine for Bayesian state estimation. *J. Neural Eng.* 2, S219–S234
49 Zemel, R.S. *et al.* (1998) Probabilistic interpretation of population code. *Neural Comput.* 10, 403–430
50 Sanger, T.D. (1996) Probability density estimation for the interpretation of neural population codes. *J. Neurophysiol.* 76, 2790–2793
51 Anastasio, T.J. *et al.* (2000) Using Bayes' rule to model multisensory enhancement in the superior colliculus. *Neural Comput.* 12, 1165–1187
52 Barlow, H.B. (1969) Pattern recognition and the responses of sensory neurons. *Ann. N. Y. Acad. Sci.* 156, 872–881
53 Doya, K. (2002) Metalearning and neuromodulation. *Neural Netw.* 15, 495–506
54 Körding, K.P. and Wolpert, D.M. (2004) Bayesian integration in sensorimotor learning. *Nature.* 427, 244–247
55 Itti, L. and Baldi, P. (2008) Bayesian surprise attracts human attention. *Vision Res*, DOI: 10.1016/j.visres.2008.09.007
56 Najemnik, J. and Geisler, W.S. (2008) Eye movement statistics in humans are consistent with an optimal search strategy. *J. Vis.* 8, 4.1-14