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 ϑ. 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, μ, 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, μ and action, α) 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.

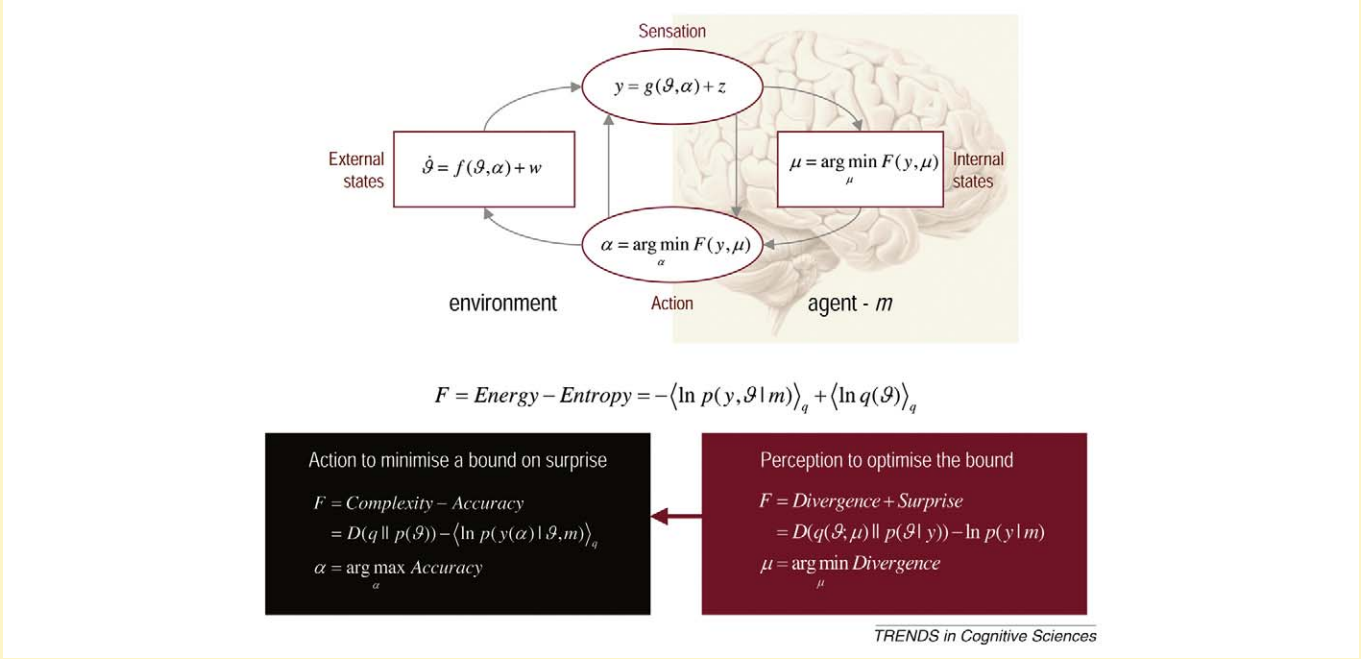

Figure I. Upper panel: schematic detailing the quantities that define free-energy. These include states of the brain μ and quantities describing exchange with the environment; sensory input $y = g(\vartheta,\alpha) + z$ and action α that changes the way the environment is sampled. The environment is described by equations of motion, $\dot{\vartheta} = f(\vartheta,\alpha) + w$, which specify the dynamics of environmental causes ϑ. 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 μ. 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.

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:

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

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 λ. 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).

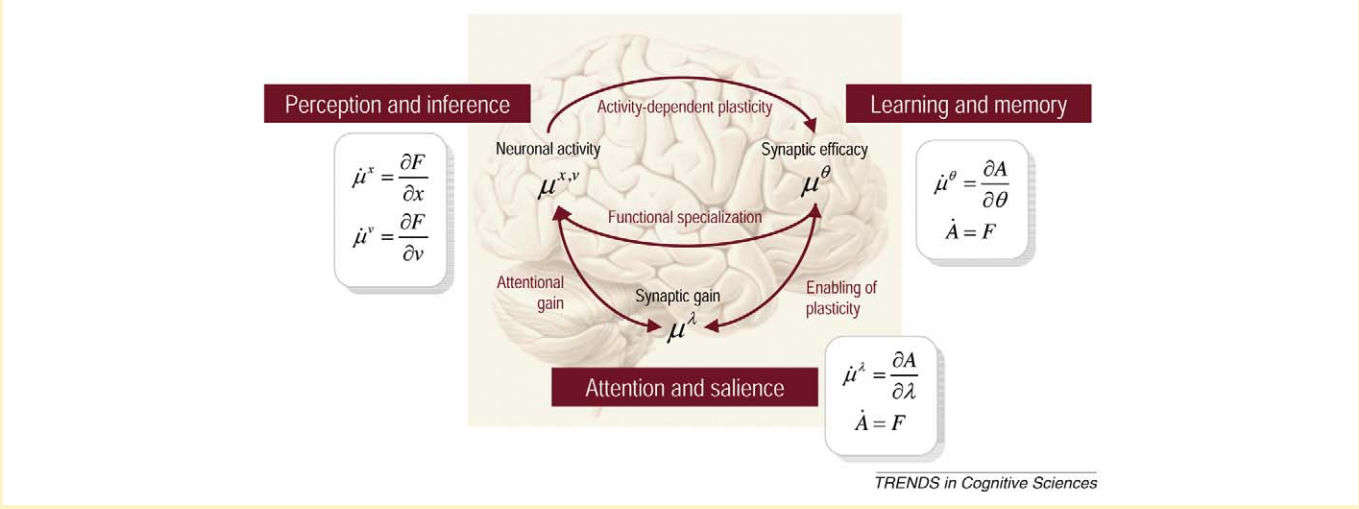

Figure I. The sufficient statistics representing a hierarchical dynamic model of the world and their recognition dynamics under the free-energy principle. The recognition density is encoded in terms of its sufficient statistics; $\mu \supset \{\mu^x, \mu^v, \mu^\theta, \mu^\lambda\}$. These representations 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; θ and precisions; λ 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.

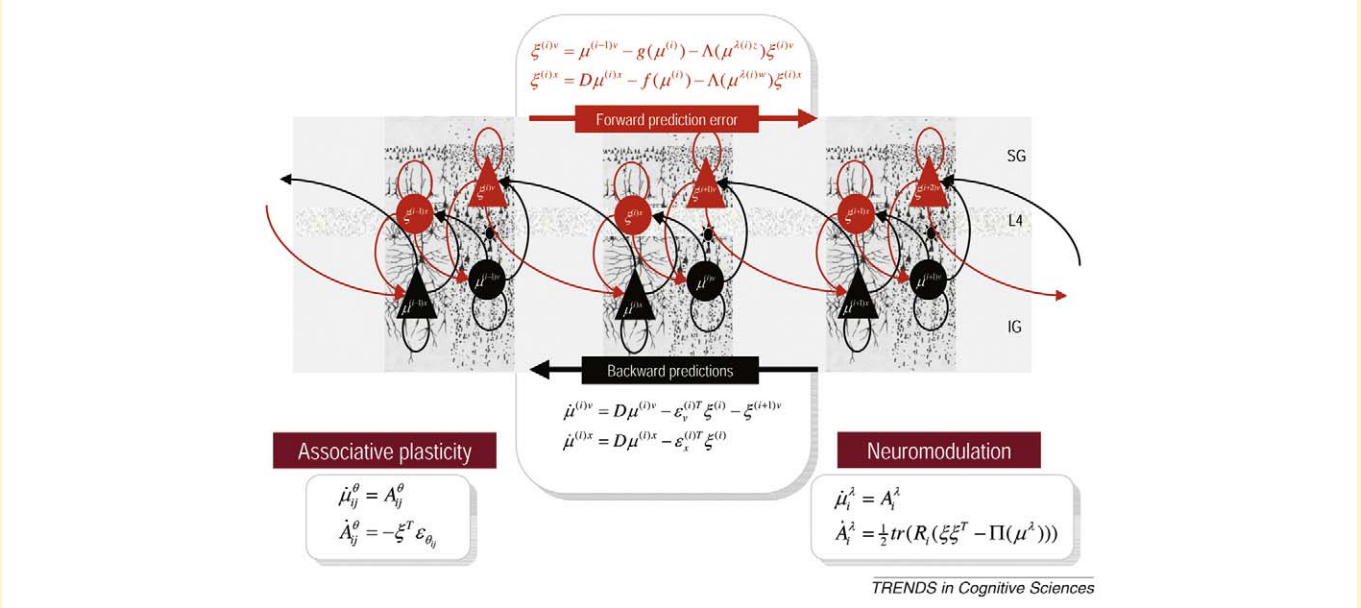

Figure I. Schematic detailing the neuronal architectures that might encode a density on the states of a hierarchical dynamic model. This shows the speculative cells of origin of forward driving connections that convey prediction error from a lower area to a higher area and the backward connections that 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?