Journal of Mathematical Psychology 81 (2017) 55–79

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Review

# The free energy principle for action and perception: A mathematical review

Christopher L. Buckley [a,*,1], Chang Sub Kim [b,1], Simon McGregor [a], Anil K. Seth [a,c]

[a] School of Engineering and Informatics, Evolutionary and Adaptive Systems Group, University of Sussex, Brighton, BN1 9QJ, UK
[b] Department of Physics, Chonnam National University, Gwangju 61186, Republic of Korea
[c] Sackler Centre for Consciousness Science, University of Sussex, Brighton, BN1 9QJ, UK

## HIGHLIGHTS

- The free energy principle (FEP) is suggested to provide a unified theory of the brain, integrating data and theory relating to action, perception, and learning.
- We provide a complete mathematical guide to a suggested biologically plausible implementation of the FEP.
- A simple agent-based model implementing perception and action under the FEP.

## ARTICLE INFO

*Article history:*
Received 27 April 2017
Received in revised form 2 August 2017
Available online 21 October 2017

*Keywords:*
Free energy principle
Perception
Action
Inference
Bayesian brain
Agent-based model

## ABSTRACT

The 'free energy principle' (FEP) has been suggested to provide a unified theory of the brain, integrating data and theory relating to action, perception, and learning. The theory and implementation of the FEP combines insights from Helmholtzian 'perception as inference', machine learning theory, and statistical thermodynamics. Here, we provide a detailed mathematical evaluation of a suggested biologically plausible implementation of the FEP that has been widely used to develop the theory. Our objectives are (i) to describe within a single article the mathematical structure of this implementation of the FEP; (ii) provide a simple but complete agent-based model utilising the FEP and (iii) to disclose the assumption structure of this implementation of the FEP to help elucidate its significance for the brain sciences.



**Contents**



* Corresponding author.
*E-mail address:* c.l.buckley@sussex.ac.uk (C.L. Buckley).
[1] Joint first author.

http://dx.doi.org/10.1016/j.jmp.2017.09.004

## 1. Introduction

The brain sciences have long searched for a 'unified brain theory' capable of integrating experimental data relating to, and disclosing the relationships among action, perception, and learning. One promising candidate theory that has emerged over recent years is the 'free energy principle' (FEP) (Friston, 2009, 2010c). The FEP is ambitious in scope and attempts to extend even beyond the brain sciences to account for adaptive biological processes spanning an enormous range of time scales, from millisecond neuronal dynamics to the tens of millions of years span covered by evolutionary theory (Friston, 2010b, c).

The FEP has an extensive historical pedigree. Some see its origins starting with Helmholtz' proposal that perceptions are extracted from sensory data by probabilistic modelling of their causes (Von Helmholtz & Southall, 2005). Helmholtz also originated the notion of thermodynamic free energy, providing a second key inspiration for the FEP.[2] These ideas have reached recent prominence in the 'Bayesian brain' and 'predictive coding' models, according to which perceptions are the results of Bayesian inversion of a causal model, and causal models are updated by processing of sensory signals according to Bayes' rule (Bubic, von Cramon, & Schubotz, 2010; Clark, 2013; Knill & Pouget, 2004b; Rao & Ballard, 1999). However, the FEP naturally accommodates and describes both action and perception within the same framework (Friston, Daunizeau, Kilner, & Kiebel, 2010), thus others see its origins in 20th-century cybernetic principles of homoeostasis and predictive control (Seth, 2015).

A recognisable precursor to the FEP as applied to brain operation was developed by Hinton and colleagues, who showed that a function resembling free energy could be used to implement a variation of the expectation–maximisation algorithm (Neal & Hinton, 1998), as well as for training autoencoders (Hinton & Zemel, 1994a) and learning neural population codes (Zemel & Hinton, 1995). Because these algorithms integrated Bayesian ideas with a notion of free energy, Hinton named them as 'Helmholtz machines' (Dayan, Hinton, Neal, & Zemel, 1995). The FEP builds on these insights to provide a global unified theory of cognition. Essentially, the FEP generalises these results by noting that all (viable) biological organisms resist a tendency to disorder as shown by their homoeostatic properties (or, more generally, their autopoietic properties), and must therefore minimise the occurrence of events which are atypical ('surprising') in their habitable environment. For example, successful fish typically find themselves surrounded by water, and very atypically find themselves out of water, since being out of water for an extended time will lead to a breakdown of homoeostatic (autopoietic) relations. Because the distribution of 'surprising' events is in general unknown and unknowable, organisms must instead minimise a tractable proxy, which according to the FEP turns out to be 'free energy'. Free energy in this context is an information-theoretic construct that (i) provides an upper bound on the extent to which sensory data is atypical ('surprising') and (ii) can be evaluated by an organism, because it depends eventually only on sensory input and an internal model of the environmental causes of sensory input. While at its most general this theory can arguably be applied to all life-processes (Friston, 2013), it provides a particularly appealing account of brain function. Specifically it describes how neuronal processes could implement free energy minimisation either by changing sensory input via action on the world, or by updating internal models via perception, with implications for understanding the dynamics of, and interactions among action, perception, and learning. These arguments have been developed in a series of papers which have appeared over the course of the last several years (Adams, Shipp, & Friston, 2013; Carhart-Harris & Friston, 2010; Friston, 2005, 2008a; Friston, Daunizeau, & Kiebel, 2009; Friston et al., 2010; Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2016; Friston & Kiebel, 2009a, b; Friston, Kilner, & Harrison, 2006; Friston, Mattout, Trujillo-Barreto, Ashburner, & Penny, 2007; Friston & Stephan, 2007; Friston, Stephan, Li, & Daunizeau, 2010; Friston, Trujillo-Barreto, & Daunizeau, 2008; Pezzulo, Rigoli, & Friston, 2015).

The FEP deserves close examination because of the claims made for its explanatory power. It has been suggested that the FEP discloses novel and straightforward relationships among fundamental psychological concepts such as memory, attention, value, reinforcement, and salience (Friston, 2009). Even more generally, the FEP is claimed to provide a "mathematical specification of 'what' the brain is doing" (Friston, 2009 p.300), to unify perception and action (Friston et al., 2010), and to provide a basis for integrating several general brain theories including the Bayesian brain hypothesis, neural Darwinism, Hebbian cell assembly theory, and optimal control and game theory (Friston, 2010c). The FEP has even been suggested to underlie Freudian constructs in psychoanalysis (Carhart-Harris & Friston, 2010).

Our purpose here is first to supply a mathematical appraisal of the FEP, which we hope will facilitate evaluation of claims such as those listed above; note that we do *not* attempt to resolve any such claims here. A mathematical appraisal is worthwhile because the FEP combines advanced concepts from several fields, particularly statistical physics, probability theory, machine learning, and theoretical neuroscience. The mathematics involved is non-trivial and has been presented over different stages of evolution and using varying notations. Here we first provide a complete technical account of the FEP, based on a history of publications through which the framework has been developed. Second we provide a complete description of a simple agent-based model working under this formulation. While we note that several other agent based models have been presented, e.g see (Friston et al., 2010), they have often made use of existing toolboxes which, while powerful, have perhaps clouded a fuller understanding of the FEP. Lastly we use our account to identify the assumption structure of the FEP, highlighting several instances in which non-obvious assumptions are required.

In the next section we provide a brief overview of the FEP followed by a detailed guide to the technical content covered in the rest of the paper.

[2] Thermodynamic free energy describes the macroscopic properties of nature, typically in thermal equilibrium where it takes minimum values, in terms of a few tractable variables.

## 2. An overview of the FEP

Broadly the FEP is an account of cognition derived from the consideration of how biological organisms maintain their state away from thermodynamic equilibrium with their ambient surroundings. The argument runs that organisms are mandated, by the very fact of their existence, to minimise the dispersion of their constituent states. The atypicality of an event can be quantified by the negative logarithm of the probability of its sensory data, which is commonly known in information theory as 'surprise' or 'self-information' and the overall atypicality of an organism's exchanges with its environment can be quantified as a total lifetime surprise (Friston, 2009, 2010c). The term surprise has caused much confusion since it is distinct from the subjective psychological phenomenon of surprise. Instead, it is a measure of how atypical a sensory exchange is. This kind of surprise can be quantified using the standard information-theoretic log-probability measure

$$-\ln p(\varphi)$$

where $p(\varphi)$ is the probability of observing some particular sensory data $\varphi$ in a typical (habitable) environment. Straightforwardly this quantity is large if the probability of the observed data is small and zero if the data is fully expected, i.e., probability 1. To avoid confusion with the common-sense meaning of the word 'surprise' we will refer to it as "surprisal" or "sensory surprisal".

### 2.1. R- and G-densities

The FEP argues that organisms cannot minimise surprisal directly, but instead minimise an upper bound called 'free energy'. To achieve this it is proposed that all (well adapted) biological organisms maintain a probabilistic model of their typical (habitable) environment (which includes their body), and attempt to minimise the occurrence of events which are atypical in such an environment as measured by this model. Two key probability densities are necessary to evaluate free energy. First it is suggested that organisms maintain an implicit representation of a "best guess" at the relevant variables that comprise their environment (i.e. those variables which cause its sensory data). This account is in the form of a probability distribution over all possible values of those variables, like a Bayesian belief; this model is instantiated, and parameterised, by physical variables in the organism's brain such as neuronal activity and synaptic strengths, respectively. When an organism receives sensory signals, it updates this distribution to better reflect the world around it, allowing it to effectively model its environment. In other words, the organism engages in a process equivalent to an approximate form of Bayesian inference regarding the state of its environment, based on sensory observations. This internal model of environmental states is called the "recognition density" or the R-density. Later in Section 3 we will assume that agents approximate the R-density as a multivariate Gaussian distribution (NB: the true posterior may be considerably more complex) where the means and variances represent an organisms best representation of the distribution of environment variables. In order to update the R-density appropriately, the organism needs some implicit assumptions about how different environmental states shape sensory input. These assumptions are presumed to be in the form of a more complicated joint probability density between sensory data and environmental variables, the "generative density", or G-density. Typically we will assume this density is also Gaussian, see Section 5.1 for the simplest example. As we will see, following a Bayesian formalism, this joint density is calculated as the product of two densities; a *likelihood* describing the probability of sensory input given some environmental state and a *prior* describing the organisms current "beliefs" of the probability distribution over environmental states.

### 2.2. Minimising free energy

Free energy is a (non-negative) quantity formed from the Kullback–Leibler divergence between the R- and G-densities. Consequently, it is not a directly measurable physical quantity: it depends on an interpretation of brain variables as encoding notional probability densities. Note: the quantity 'free energy' is distinct from thermodynamic free energy thus here we will refer to it as *variational free energy* (VFE) (referring to its role in variational Bayes, see later for details).

Minimisation of VFE has two functional consequences. First it provides an upper bound on sensory surprisal. This allows organisms to estimate the dispersion of their constituent states and is central to the interpretation of FEP as an account of life processes (Friston, 2010c). However, VFE minimisation also plays a central role in a Bayesian approximation method. Specifically ideal (exact) Bayesian inference, in general, involves evaluating difficult integrals and thus a core hypothesis of the FEP framework is that the brain implements approximate Bayesian inference in an analogous way to a method known as variational Bayes. It can be shown that minimising VFE makes the R-density a good approximation to the posterior density of environmental variables given sensory data. Under this interpretation the surprisal term in the VFE becomes more akin to the negative of *log model evidence*, see Section 3, defined in more standard implementations of variational Bayes (Hinton & Zemel, 1994b).

### 2.3. The action–perception cycle

Minimising VFE by updating the R-density provides an upper-bound on surprisal but cannot minimise it directly. The FEP suggests that organisms also act on their environment to change sensory input, and thus minimise surprisal indirectly (Friston, 2009, 2010c). The mechanism underlying this process is formally symmetric to perceptual inference, i.e., rather than inferring the cause of sensory data an organism must infer actions that best make sensory data accord with an internal representation of the environment (Friston et al., 2010). Thus, the mechanism is often referred to as *active inference*. Formally, action allows an organism to avoid the dispersion of its constituent states and is suggested to underpin a form of homoeostasis, or perhaps more precisely homeorhesis (a generalisation of homoeostasis referring to a system that is stable about a complex trajectory of states rather than around a fixed point) (Seth, 2015). However, equivalently, one can view action as satisfying hard constraints encoded in the organism's environmental model (Friston et al., 2010). Here expectations in the organism's G-density (its "beliefs" about the world) cannot be met directly by perception and thus an organism must act to satisfy them. In effect these expectations effectively encode the organism's *desires* on environmental dynamics. For example, the organism's model may prescribe that it maintains a desired local temperature; we will see an example of this in Section 7. Here action is seen as more akin to control (Seth, 2015) where behaviour arises from a process of minimising deviations between the organism's actual and a desired trajectory (Friston et al., 2010). Note: an implicit assumption here is that these constraints are conducive to the organism's survival (Friston, 2009, 2010c), perhaps arrived at by an evolutionary process. Other different roles for action within the FEP have also been suggested, e.g., actions performed to disambiguate competing models (Friston, Adams, Perrinet, & Breakspear, 2012; Seth, 2015). However, here we only consider action as a source of control (Friston et al., 2010; Seth, 2015).

### 2.4. Predictive coding

There at least two general ways to view most FEP-based research. First the central theory (Friston et al., 2006) which offers a particular explanation of cognition in terms of Bayesian inference. Second a biologically plausible *process theory* of how the relevant probability densities could be parameterised by variables in the brain (i.e. a model of what it is that brain variables encode), and how the variables should be expected to change in order to minimise VFE. The most commonly used implementation of the FEP, and the one we focus on here, is strongly analogous with the predictive coding framework (Rao & Ballard, 1999). Specifically predictive coding theory constitutes one plausible mechanism whereby an organism could update its environmental model (R-density) given a belief of how its environment works (G-density). The concept of predictive coding overturns classical notions of perception (and cognition) as a largely bottom-up process of evidence accumulation or feature detection driven by impinging sensory signals, proposing instead that perceptual content is determined by top-down predictive signals arising from multi-level generative models of the environmental causes of sensory signals, which are continually modified by bottom-up prediction error signals communicating mismatches between predicted and actual signals across hierarchical levels (see Clark, 2013 for a nice review). In the context of the FEP the R-density is updated using a hierarchical predictive coding (see Section 8). This has several theoretical benefits. Firstly, under suitable assumptions VFE becomes formally equivalent to prediction error (weighted by confidence terms), which can readily be computed in neural wetware. Hierarchical coding also provides a very generic prior which allows high-level abstract sensory features to be learned from the data, in a manner similar to deep learning nets (Hinton, 2007). Finally, the sense in which the brain models the environment can be conceptualised in a very direct way as the prediction of sensory signals. We will also see in Section 8 that this implementation suggests that we do not even need to know what environmental features the R- and G-densities constitute a model of.

### 2.5. A technical guide

In the rest of this work we review the FEP in detail but first we provide a detailed guide to each section. Most of what we present is related to standard concepts and techniques in statistical mechanics and machine learning. However, here we present these ideas in detail to make clear their role for the FEP as theory of biological systems.

In Section 3 we describe the core technical concepts of FEP including the R-density, G-density, and VFE. We show how minimising VFE has two consequences. First, it makes the R-density a better estimate of posterior beliefs about environmental state given sensory data, thus implementing approximate Bayesian inference. Second, it makes the VFE itself an upper-bound on sensory surprisal.

In Section 4 we discuss the approximations that allow the brain to explicitly instantiate the R-density and thus specify VFE. Specifically, we make the approximation that the R-density take Gaussian form, the *Laplace approximation*, and that brain states, e.g. neural activity, represent the sufficient statistics of this distribution (mean and variance). Utilising this form for the R-density and various other approximations we derive an expression for the VFE in terms of the unknown G-density only; we refer to this approximation as the *Laplace encoded energy*. The derivations in this section are done for the univariate Gaussian case, but we give an expression for the full multivariate case at the end of the section.

In Section 5 we look at different forms for the G-density. We start by specifying simple generative models which comprise the brain's model of how the world works, i.e., how sensory data is caused by environmental (including bodily) variables. We utilise these generative models to specify the brain's expectation on environmental states given sensory data in terms of a Gaussian distribution parameterised by expected means and variances (inverse precisions) on brain states. Combining this with the result of the last section allows us to write an expression for the Laplace encoded-energy as a quadratic sum of prediction errors (differences between expected and actual brain states given sensory data) modulated by expected variances (or inverse precisions), in line with predictive-coding process theories. Initially we show this for a static generative model but extend it to include dynamic generative models by introducing the concept of generalised motion. Again we derive the results for the univariate case but provide expressions for the multivariate case.

In Section 6 we show how the brain could dynamically minimise VFE. Specifically, we describe how brain states are optimised to minimise VFE through gradient descent. We discuss complications of this method when considering dynamical generative models.

Section 7 demonstrates how action can be implemented as a similar gradient descent scheme. Specifically we show how, given a suitable *inverse model*, actions are chosen to change sensation such that they minimise VFE. We ground this idea, and the mechanisms for perception described in prior sections, in a simple agent based simulation. We show how an agent with an appropriate model of the environment, can combine action and perception to minimise VFE constrained both by the environment and its own expectations on brain states.

In Section 8 we extend the formalism to include learning. Specifically we show how the brain could modify and learn the G-density. To facilitate this we describe notion of hierarchical generative models which involve empirical priors. We lastly describe a gradient descent scheme which allows the brain to infer parameters and hyperparameters of the VFE and thus allow the brain to learn environmental dynamics based on sensory data.

Finally, Section 9 summarises the FEP and discusses the implications of its assumption structure for the brain sciences.

## 3. Variational free energy

We start by considering a world that consists of a brain and its surrounding body/environment. For the rest of the presentation we refer to the body and environment as simply the *environment* and use this to refer to all processes outside of the brain. The brain is distinguished from its environment by an interface which is not necessarily a physical boundary but rather may be defined functionally; thus the boundary could reside at the sensory and motor surfaces rather than, for example, at the limits of the cranial cavity. The environment is characterised by states, denoted collectively as $\{\vartheta\}$, which include well-defined characteristics like temperature or the orientation of a joint but also unknown and uncontrollable states, all evolving according to physical laws. The environmental states, as exogenous stimuli, give rise to sensory inputs for which the symbols $\{\varphi\}$ are designated collectively. These sensory inputs are assumed to reside at the functional interface distinguishing the brain from the environment, and we assume a many-to-one (non-bijective) mapping between $\{\vartheta\}$ and $\{\varphi\}$ (Friston, 2010a). We further assume that the brain, in conjunction with the body, can perform actions to modify sensory signals.

We assume that the important states of the environment cannot be directly perceived by an organism but instead must be inferred by a process of Bayesian inference. Specifically, the goal of the agent is to determine the probability of environmental states given its sensory input. To achieve this we assume organism's encodes

prior beliefs about these states characterised by the joint density $p(\vartheta, \varphi)$ or G-density. Where the G-density can be factorised into (with respect to $\vartheta$), the *prior* $p(\vartheta)$ (corresponding to the organism's "beliefs" about the world before sensory input is received) and a *likelihood* $p(\varphi|\vartheta)$ (corresponding to the organism's assumptions about how environmental dynamics cause sensory input),

$$p(\vartheta, \varphi) = p(\varphi|\vartheta)p(\vartheta). \tag{1}$$

Given an observation, $\varphi = \phi$ (e.g. some particular sensory data), a *posterior* belief about the environment can then be written as $p(\vartheta|\varphi = \phi)$. This quantity can be calculated using the prior and likelihood using Bayes theorem as,

$$p(\vartheta|\phi) = \frac{1}{p(\varphi = \phi)} p(\phi|\vartheta)p(\vartheta) = \frac{p(\phi|\vartheta)p(\vartheta)}{\int p(\phi|\vartheta)p(\vartheta)d\vartheta}. \tag{2}$$

All the probability densities are assumed to be normalised as

$$\int d\vartheta \int d\varphi\ p(\vartheta, \varphi) = \int d\vartheta\ p(\vartheta) = \int d\varphi\ p(\varphi) = 1,$$

where $p(\vartheta)$ and $p(\varphi)$ are the reduced or marginal probability-densities conforming to

$$p(\vartheta) = \int d\varphi\ p(\vartheta, \varphi) \quad \text{and} \quad p(\varphi) = \int d\vartheta\ p(\vartheta, \varphi). \tag{3}$$

To calculate the posterior probability it is necessary to evaluate the marginal integral, $\int p(\phi|\vartheta)p(\vartheta)\vartheta$, in the denominator of Eq. (2). However, this is often difficult, if not practically intractable. For example, when continuous functions are used to approximate the likelihood and prior, the integral may be analytically intractable. Or in the discrete case, when this integral reduces to a sum, the number of calculations may grow exponentially with the number of states. *Variational Bayes* (sometimes known as 'ensemble learning') is a method for (approximately) determining $p(\vartheta|\varphi)$ which avoids the evaluation of this integral, by introducing an optimisation problem (Friston et al., 2008). Such an approach requires an auxiliary probability density, representing the current 'best guess' of the causes of sensory input. This is the *recognition density*, or R-density, introduced in the overview. Again the R-density is also normalised as:

$$\int q(\vartheta)d\vartheta = 1. \tag{4}$$

We can construct a measure of the difference between this density and the true posterior in terms of an information-theoretic divergence, e.g., the Kullback–Leibler divergence (Cover & Thomas, 1991), i.e.,

$$D_{KL}(q(\vartheta) \parallel p(\vartheta|\varphi)) = \int d\vartheta\ q(\vartheta) \ln \frac{q(\vartheta)}{p(\vartheta|\varphi)}. \tag{5}$$

An R-density that minimises this divergence would provide a good approximation to the true posterior. But obviously we cannot evaluate this quantity because we still do not know the true posterior. However, we can rewrite this equation as,

$$D_{KL}(q(\vartheta) \parallel p(\vartheta|\varphi)) = F + \ln p(\varphi) \tag{6}$$

where we have defined F as the *variational free energy* (VFE),

$$F \equiv \int d\vartheta\ q(\vartheta) \ln \frac{q(\vartheta)}{p(\vartheta, \varphi)}. \tag{7}$$

Note here we have introduced the G-density to the denominator on the right-hand side. In contrast to Eq. (5) we can evaluate VFE directly because it depends only on the R-density, which we are free to specify, and the G-density, i.e., a model of the environmental causes of sensory input. Furthermore, the second term on the right-hand side in Eq. (6) only depends on sensory input and is independent of the form of the R-density. Thus, minimising Eq. (7) with respect to the R-density will also minimise the Kullback–Leibler divergence between the R-density and the true posterior. Thus, the result of this minimisation will make the R-density approximate the true posterior.

The minimisation of VFE also suggests an indirect way to estimate surprisal. Specifically according to Jensen's inequality (Cover & Thomas, 1991), the Kullback–Leibler divergence is always greater than zero. This implies the inequality,

$$F \geqslant -\ln p(\varphi), \tag{8}$$

which means that the VFE also provides an upper bound on the *surprisal* as described in Section 1. However, note that VFE is equal to surprisal only when the R-density $q(\vartheta)$ becomes identical with the posterior density $p(\vartheta|\varphi)$; i.e., it is this condition that specifies when VFE provides a *tight bound* on surprisal (see Section 2). Furthermore, while this process furnishes the organism with an approximation of surprisal it does not minimise it. Instead the organism can minimise VFE further by minimising surprisal indirectly by acting on the environment and changing sensory input, see Section 7.

Note: formally $p(\varphi)$, which describes the agent's internal (implicit) probabilistic predictions of sensory inputs, should be written as $p(\varphi|m)$. This follows a convention in Bayesian statistics to indicate that a reasoner must begin with some arbitrary prior before it can learn anything; $p(\varphi)$ indicates the prior assigned to $p$ ab initio by agent $m$. However, this notation is unwieldy and does not change the derivations that follow thus we will omit this for the rest of the presentation.

There are several analogies between the terms in the formalism above and the formulation of Helmholtz' thermodynamic free energy. These terms can serve as useful substitutions in the derivation to come and, thus, we describe them here. Specifically when the G-density is unpacked in Eq. (7), the VFE splits into two terms,

$$F = \int d\vartheta\ q(\vartheta)E(\vartheta, \varphi) + \int d\vartheta\ q(\vartheta) \ln q(\vartheta) \tag{9}$$

where, formally speaking, the first term in Eq. (9) is an average of the quantity

$$E(\vartheta, \varphi) \equiv -\ln p(\vartheta, \varphi) \tag{10}$$

over the R-density $q(\vartheta)$ and the second term is essentially the negative entropy associated with the recognition density. By analogy with Helmholtz' thermodynamic free energy the first term in Eq. (9) is called *average energy* [Accordingly, $E(\vartheta, \varphi)$ itself may be termed the *energy*] and the second term the negative of *entropy* (Adkins, 1983).

In summary, minimising VFE with respect to the R-density, given an appropriate model for the G-density $p(\vartheta, \varphi)$ in which the sensory inputs are encapsulated, allows one to approximate the Bayesian posterior. Furthermore minimising VFE through perception also gives a upper bound on the sensory surprisal.

Table 1 provides a summary of the mathematical objects associated with the VFE.

## 4. The R-density: how the brain encodes environmental states

To implement the method described above the brain must explicitly encode the R-density. To achieve this it is suggested that neuronal quantities (e.g., neural activity) parametrise *sufficient statistics* (e.g., means and variances, see later) of a probability distribution. More precisely the neuronal variables encode a family of probability densities over environmental states, $\vartheta$. The instantaneous state of the brain $\mu$ then picks out a particular density $q(\vartheta; \mu)$ (the R-density) from this family; the semicolon in $q(\vartheta; \mu)$ indicates that $\mu$ is a parameter rather than a random variable.

**Table 1**
Mathematical objects relating to the VFA.

| Symbol | Name | Description |
|---|---|---|
| $\vartheta$ | Environmental variables | These refer to all states outside of the brain and include both environmental and bodily variables. |
| $\varphi$ | Sensory data | Signals caused by the environment (or body). |
| $q(\vartheta)$ | R-density | Organism's (implicit) probabilistic representation of environmental states which cause sensory data. |
| $p(\varphi, \vartheta)$ | G-density | Joint probability distribution relating environmental states and sensory data. This is necessary to specify the Laplace-encoded energy and is usually specified in terms of a likelihood and prior |
| $p(\vartheta)$ | Prior density | Organism's prior beliefs, encoded in the brain's state, about environmental states. |
| $p(\varphi\|\vartheta)$ | Likelihood density | Organism's implicit beliefs about how environmental states map to sensory data. |
| $p(\vartheta\|\varphi)$ | Posterior density | The inference that a perfectly rational agent (with incomplete knowledge) would make about the environment's state upon observing new sensory information, given the organism's prior assumptions. |
| $p(\varphi)$ | Sensory density | Probability density of the sensory input, encoded in the brain's state, which cannot be directly quantified given sensory data alone. |
| $-\ln p(\varphi)$ | Surprisal | *Surprise* or *self-information* in information-theory terminology, which is equal to the negative of *log model evidence* in Bayesian statistics. |
| $F(\vartheta, \varphi)$ | Variational free energy (VFE) | The quantity minimised under the FEP which forms an upper bound on surprisal allows the approximation of the posterior density. |

Finding the optimal $q(\vartheta; \mu)$ that minimises VFE in the most general case is intractable and thus further approximations about the form of this density are required. Two types of approximation are often utilised. First, an assumption that the R-density $q(\vartheta)$ can be factorised into independent sub-densities $q_1(\theta_1) \times \cdots q_N(\theta_N)$. Under this assumption the optimal R-density still cannot be expressed in closed form but an approximate solution (of general form) can be improved iteratively (Friston, 2008b). This leads to a formal solution in which the sub-densities affect each other only through mean-field quantities. Approaches that utilise this form of the R-density are often referred to an *ensemble learning*. This approach is not the focus of the work presented here but for completeness we provide a treatment of unconstrained ensemble learning in Appendix A.

A more common approximation is to assume that the R-density take Gaussian form, the so called *Laplace approximation* (Friston et al., 2008). In this scenario, the sufficient statistics of this Gaussian form become parameters which can be optimised numerically to minimise VFE. For example the R-densities take the form

$$q(\vartheta) \equiv \mathcal{N}(\vartheta; \mu, \zeta) = \frac{1}{\sqrt{2\pi\zeta}} \exp\left\{-(\vartheta - \mu)^2/(2\zeta)\right\} \tag{11}$$

where $\mu$ and $\zeta$ are the mean and variance values of a single environmental variable $\vartheta$. Substituting this form for the R-density into Eq. (7), and carrying out the integration produces a vastly simplified expression for the VFE. In the following we examine this derivation in detail. For the clarity of presentation we pursue it in the univariate case which captures all the relevant assumptions for the multivariate case. We write the formulation for the multivariate case at the end of the section. For notational ease we define

$$Z \equiv \sqrt{2\pi\zeta} \quad \text{and} \quad \mathcal{E}(\vartheta) \equiv (\vartheta - \mu)^2/(2\zeta), \tag{12}$$

to arrive at

$$q(\vartheta; \mu, \zeta) = \frac{1}{Z} e^{-\mathcal{E}(\vartheta)}, \tag{13}$$

where here we have drawn on terminology from statistical physics in which the normalisation factor $Z$ is called the *partition function* and $\mathcal{E}(\vartheta)$ the energy of the subsystem $\{\vartheta\}$ (Huang, 1987). Substituting this equation into Eq. (9) and carrying out the integration leads to a much simplified expression for VFE :

$$\begin{aligned} F &= \int d\vartheta\, q(\vartheta)(-\ln Z - \mathcal{E}) + \int d\vartheta\, q(\vartheta) E(\vartheta, \varphi) \\ &= -\ln Z - \int d\vartheta\, q(\vartheta)\mathcal{E}(\vartheta) \\ &\quad + \int d\vartheta\, q(\vartheta) E(\vartheta, \varphi) \end{aligned} \tag{14}$$

where we have used the normalisation condition, Eq. (4) in the second step. The Gaussian integration involved in the first and second terms in Eq. (14) can be evaluated straightforwardly. Specifically, utilising Eq. (12), the first term in Eq. (14) can be readily manipulated into

$$-\ln Z = -\frac{1}{2}(\ln 2\pi\zeta).$$

Using Eq. (12) the second term in Eq. (14) becomes

$$-\frac{1}{2\zeta}\int d\vartheta\, q(\vartheta)(\vartheta - \mu)^2 \to -\frac{1}{2}.$$

The final term demands further technical consideration because the energy $E(\vartheta, \varphi)$ is still unspecified. However, further simplifications can be made by assuming that the R-density, Eq. (13) is sharply peaked at its mean value $\mu$ (i.e., the Gaussian bell-shape is squeezed towards a delta function) and that $E(\vartheta, \varphi)$ is a smooth function of $\vartheta$. Under these assumptions we notice that the integration is appreciably non-zero only at the peaks. One can then use a Taylor expansion of $E(\vartheta, \varphi)$ around $\vartheta = \mu$ with respect to a small increment, $\delta\vartheta$. Note: while these assumptions permit a simple analytic model of the FEP, they have non-trivial implications for the interpretation of brain function so we return to this issue at the end of this section and in the Discussion. This assumption brings about,

$$\begin{aligned} &\int d\vartheta\, q(\vartheta) E(\vartheta, \varphi), \\ &\approx \int d\vartheta\, q(\vartheta) \left\{ E(\mu, \varphi) + \left[\frac{dE}{d\vartheta}\right]_\mu \delta\vartheta + \frac{1}{2}\left[\frac{d^2E}{\partial\vartheta^2}\right]_\mu \delta\vartheta^2 \right\}. \end{aligned}$$

Now substituting back $\delta\vartheta = \vartheta - \mu$ we get,

$$\begin{aligned} &\approx E(\mu, \varphi) + \left[\frac{\partial E}{\partial\vartheta}\right]_\mu \int d\vartheta\, q(\vartheta)(\vartheta - \mu) \\ &+ \frac{1}{2}\left[\frac{d^2E}{d\vartheta^2}\right]_\mu \int d\vartheta\, q(\vartheta)(\vartheta - \mu)^2. \end{aligned}$$

Here the second term in the first line is zero identically because the integral equates to the mean. Furthermore recognising the expression for the variance in the third term allows us to write

$$\approx E(\mu, \varphi) + \frac{1}{2}\left[\frac{d^2E}{d\vartheta^2}\right]_\mu \zeta \tag{15}$$

where we identify $E(\mu, \varphi)$ as the *Laplace-encoded energy*. Substituting all terms derived so far into Eq. (14) furnishes an approximate expression for the VFE,

$$F = E(\mu, \varphi) + \frac{1}{2}\left(\left[\frac{d^2E}{d\vartheta^2}\right]_\mu \zeta - \ln 2\pi\zeta - 1\right) \tag{16}$$

which is now written as a *function* (i.e., not a functional) of the Gaussian means and variances, and sensory inputs, i.e. $F =$

**Table 2**
Mathematical objects relating to the Laplace encoding.

| Symbol | Name | Description |
|---|---|---|
| $\mathcal{N}(\vartheta;\mu,\zeta)$ | (Gaussian) fixed-form R-density | An 'ansatz' for unknown $q(\vartheta)$ (the Laplace approximation) |
| $\mu,\zeta$ | Parameters for the R-density | Sufficient statistics (expectation and variance) of the fixed-form R-density, encoded in the brain's state. |
| $\zeta^*$ | Optimal variance | Analytically derivable optimal $\zeta$, removing an explicit dependence of $F$ on $\zeta$. |
| $p(\varphi,\mu)$ | Laplace-encoded G-density | A mathematical construct based upon the G-density that scores the surprise associated with any posterior expectation. |
| $E(\mu,\varphi)$ | Laplace-encoded energy | Mathematical construct defined to be $-\ln p(\mu,\varphi)$. |

$F(\mu,\zeta,\varphi)$. To simplify further we remove the dependence of the VFE on the variances by taking derivative of Eq. (16) with respect $\zeta$ as follows:

$$dF = \frac{1}{2}\left\{\frac{d}{d\zeta}\left(\left[\frac{d^2E}{d\vartheta^2}\right]_\mu \zeta\right) - \frac{1}{\zeta}\right\} d\zeta$$
$$= \frac{1}{2}\left\{\left[\frac{d^2E}{d\vartheta_l}\right]_\mu - \frac{1}{\zeta}\right\} d\zeta.$$

Minimising by demanding that $dF \equiv 0$ one can get

$$\zeta^* = \left[\frac{d^2E}{d\vartheta^2}\right]_\mu^{-1} \tag{17}$$

where the superscript in $\zeta^*$ indicates again that it is an optimal variance (i.e., it is the variance which optimises the VFE). Substituting Eq. (17) into Eq. (16) gives rise to the form of the VFE as

$$F = E(\mu,\varphi) - \frac{1}{2}\ln\left\{2\pi\zeta^*\right\}. \tag{18}$$

The benefit of this process has been to recast the VFE in terms of a joint density $p(\mu,\varphi)$ over sensory data $\varphi$ and the R-density's sufficient statistics $\mu$, rather than a joint density over some (unspecified) environmental features $\vartheta$. Note: this joint density amounts to an approximation of the G-density described in Eq. (1); we shall examine the implementation of this density in detail in the next section. Furthermore, under these assumptions the VFE only depends on Gaussian means (first-order Gaussian statistics) and sensory inputs, and not on variances (second-order Gaussian statistics), which considerably simplifies the expression. It is possible to pursue an analogous derivation for the full multivariate Gaussian distribution under the more general assumption that the environment states only weakly covary, i.e., both the variance of, and covariances between, variables are small. Under this assumption the full R-density distribution is still tightly peaked and the Taylor expansion employed in Eq. (15) is still valid.

To get rid of the constant variance term in Eq. (18), we write the Laplace-encoded energy for the full multivariate case, as an approximation for the full VFE as

$$E(\{\mu_\alpha\},\{\varphi_\alpha\}) = -\ln p(\{\mu_\alpha\},\{\varphi_\alpha\}), \tag{19}$$

where we define $\{\mu_\alpha\}$ and $\{\varphi_\alpha\}$ as vectors of brain states and sensory data respectively, corresponding to environmental variables $\{\vartheta_\alpha\}$ with $\alpha = 1,2,\ldots,N$ indexing N variables. This equation for the Laplace-encoded energy serves as a general approximation for the VFE which we will use in the rest of this study.

Conceptually this expression suggests the brain represents only the most likely environmental causes of sensory data and not the details of their distribution per se. However, as we will see later, the brain also encodes uncertainties through (expectations about) precisions (inverse variances) in the G-density.

Table 2 provides a glossary of mathematical objects involved in the Laplace encoding of the environmental states in the brain.

## 5. The G-density: encoding the brains beliefs about environmental causes

In the previous section we constructed an approximation of the VFE, which we called the Laplace-encoded energy, in terms of the approximate G-density $p(\mu,\varphi)$ where the environmental states $\vartheta$ have been replaced by the sufficient statistics $\mu$ of the R-density. In this section we consider how the brain could specify this G-density, and thus evaluate VFE. We start by specifying a *generative model* of the environmental causes of sensory data (informally, a description of causal dependencies in the environment and their relation to sensory signals). We then show how to move from these generative models to specification of the G-density, in terms of brain states and their expectations, and finally construct expressions for the VFE. We develop various specifications of G-densities for both static and dynamic representations of the environment and derive the different expressions for VFE they imply.

Table 3 provides a summary of the mathematical objects associated with the G-density in the simplest model and also its extension to the dynamical generative model.

### 5.1. The simplest generative model

We first consider a simplified situation corresponding to an organism that believes in an environment comprising of a single variable and a single sensory channel. To represent this environment the agent uses a single brain state $\mu$ and sensory input $\varphi$. We then write down the organism's belief about the environment directly in terms of a generative mapping between brain states and sensory data. Note these equations will have a slightly strange construction because in reality sensory data is caused by environmental, not brain, states. However, writing the organism beliefs in this way will allow us to easily construct a generative density, see below. Specifically, we assume the agent believes its sensory is generated by:

$$\varphi = g(\mu;\theta) + z \tag{20}$$

where $g$ is a linear or nonlinear function, parameterised by $\theta$ and $z$ is a random variable with zero mean and variance $\sigma_z$. Thus the organism believes its sensory data is generated as non-linear mapping between environmental states (here denoted in terms of its belief about environmental state $\mu$) with added noise. Similarly we specify the organism beliefs about how environmental state are generated as

$$\mu = \bar{\mu} + w, \tag{21}$$

where $\bar{\mu}$ is some fixed parameter and $w$ is random noise drawn from a Gaussian with zero mean and variance $\sigma_w$. In other words, the organism takes the environment's future states to be history-independent, fluctuating around some mean value $\bar{\mu}$ which is given *a priori* to the organism. There is a potential confusion here because Eq. (21) describes a distribution over the brain state variable $\mu$, which itself represents the mean of some represented environmental state $\vartheta$. Specifically, it is worth reiterating that $\bar{\mu}$

**Table 3**
Mathematical objects relating to dynamical generative models.

| Symbol | Name & description |
|---|---|
| Simple model | $p(\varphi, \mu) = p(\varphi\|\mu)p(\mu)$ |
| $g(\mu;\theta)$ | Generative mapping between the brain states $\mu$ and the observed data $\varphi$, parameterised by $\theta$ |
| $z, w$ | Random fluctuations represented by Gaussian noise |
| $\sigma_z, \sigma_w$ | The variance of these fluctuations (the inverse of precisions) |
| $p(\varphi\|\mu), p(\mu)$ | Likelihood, prior of $\mu$, which together determine $p(\varphi, \mu)$ |
| Dynamical model | $p(\varphi, \mu) = \prod_{n=0}^{\infty} p(\varphi_{[n]}\|\mu_{[n]})p(\mu_{[n+1]}\|\mu_{[n]})$ |
| $\tilde{\mu}$ | Brain states in generalised coordinates; an infinite vector whose components are given by successive time-derivatives, $\tilde{\mu} \equiv (\mu, \mu', \mu'', \ldots) \equiv (\mu_{[0]}, \mu_{[1]}, \mu_{[2]}, \ldots)$. |
| $\tilde{\varphi}$ | Sensory data, similarly defined as $\tilde{\varphi} = (\varphi, \varphi', \varphi'', \ldots)$. |
| $\varphi_{[n]} = g_{[n]} + z_{[n]}$ | Generalised mapping between the observed data $\tilde{\varphi}$ and the brain states $\tilde{\mu}$ at the dynamical order $n$ |
| $\mu_{[n+1]} = f_{[n]} + w_{[n]}$ | Generalised equations of motion of the brain state $\tilde{\mu}$ at the dynamical order $n$ |
| $g_{[n]}, f_{[n]}$ | Generative functions in the generalised coordinates |
| $p(\varphi_{[n]}\|\mu_{[n]})$ | Likelihood of the generalised state $\mu_{[n]}$, given the data $\varphi_{[n]}$ |
| $p(\mu_{[n+1]}\|\mu_{[n]})$ | Gaussian prior of the generalised state $\mu_{[n]}$ |

and $\sigma_w$ are distinct from the sufficient statistics of the R-density ($\mu$ and $\zeta$ ) see Eq. (11). The former correspond to prior beliefs about latent or hidden causes in the environment before encountering any sensory inputs. Conversely, the latter are the posterior beliefs after encountering data. This is why the R-density is referred to as a recognition density. As we will see in Section 7, there is conflict here because the organism's best estimate $\mu$ (the mean of its subjective distribution over $\vartheta$) may not be in line with its expectation $\bar{\mu}$ stemming from its model of environmental dynamics.

To construct the generative density we assume that the noise $z$ is given as Gaussian, $[1/\sqrt{2\pi\sigma_z}]\exp\left\{-z^2/(2\sigma_z)\right\}$. Then, rewriting Eq. (20) as $z = \varphi - g(\mu;\theta)$, the functional form of the likelihood $p(\varphi|\mu)$ can be written as

$$p(\varphi|\mu) = \frac{1}{\sqrt{2\pi\sigma_z}}\exp\left\{-(\varphi - g(\mu;\theta))^2/(2\sigma_z)\right\}. \tag{22}$$

Assuming similar Gaussian noise for the random deviation $w = \mu - \bar{\mu}$, in Eq. (34), the prior density $p(\mu)$ can be written as

$$p(\mu) = \frac{1}{\sqrt{2\pi\sigma_w}}\exp\left\{-(\mu - \bar{\mu})^2/(2\sigma_w)\right\} \tag{23}$$

where $\sigma_w$ is the variance.

Thus far, we have specified the likelihood and the prior of $\mu$ which together determine the G-density $p(\mu, \varphi)$ according to the identity,

$$p(\mu, \varphi) = p(\varphi, \mu) = p(\varphi|\mu)p(\mu).$$

Next, we construct the Laplace-encoded energy by substituting the likelihood and prior densities obtained above into Eq. (19) to get, up to a constant,

$$E(\mu, \varphi) = -\ln p(\varphi|\mu) - \ln p(\mu) \tag{24}$$

$$= \frac{1}{2\sigma_z}\varepsilon_z^2 + \frac{1}{2\sigma_w}\varepsilon_w^2 + \frac{1}{2}\ln(\sigma_z\sigma_w), \tag{25}$$

where the auxiliary notations have been introduced as

$$\varepsilon_z \equiv \varphi - g(\mu;\theta) \quad \text{and} \quad \varepsilon_w \equiv \mu - \bar{\mu},$$

which comprise a *residual error* or a *prediction error* in the predictive coding terminology (Rao & Ballard, 1999). The quantity $\varepsilon_z$ is a measure of the discrepancy between actual $\varphi$ and the outcome of its prediction $g(\mu;\theta)$. While $\varepsilon_w$ describes the extent to which $\mu$ itself deviates from its prior expectation $\bar{\mu}$. The former describes sensory prediction errors, $\varepsilon_z$, while the latter describes model prediction errors, $\varepsilon_w$, (i.e., how brain states deviate from their expectation). Each error term is multiplied by the inverse of variance which weights the relative influence of these terms, i.e., how they contribute to the Laplace-encoded energy. We note in other works that inverse of variance, i.e., precision, is used in these equations perhaps to highlight that these terms represent the confidence, or reliability, of the prediction. However, here we stick to more standard notation involving variances.

The above calculation can be straightforwardly extended to the multivariate case. Specifically, we represent $\{\mu_\alpha\}$ as a row vector of $N$ brain states, and write their expectations as

$$\mu_\alpha = \bar{\mu}_\alpha + w_\alpha.$$

Here $\{w_\alpha\}$ is a row vector describing correlated noise sources, thus generally the fluctuations of each variable are not independent, which all have zero mean and covariance $\Sigma_w$. We can write down a set of $N$ sensory inputs $\{\varphi_\alpha\}$ which depend on combination of these brain states in some nonlinear way such that

$$\varphi_\alpha = g_\alpha(\mu_0, \mu_1, \ldots, \mu_N) + z_\alpha. \tag{26}$$

Again $\{z_\alpha\}$ are noise sources with zero mean and covariance $\Sigma_z$ and thus each sensory input may receive statistically correlated noise. Then, the prior over brain states may be represented as the multivariate correlated Gaussian density,

$$p(\{\mu_\alpha\}) = \frac{1}{\sqrt{(2\pi)^N|\Sigma_w|}} \times \exp\left(-\frac{1}{2}\{\mu_\alpha - \bar{\mu}_\alpha\}\Sigma_w^{-1}\{\mu_\alpha - \bar{\mu}_\alpha\}^T\right), \tag{27}$$

where $\{\mu_\alpha - \bar{\mu}_\alpha\}^T$ is the transpose of vector $\{\mu_\alpha - \bar{\mu}_\alpha\}$; $|\Sigma_w|$ and $\Sigma_w^{-1}$ are the determinant and the inverse of the covariance matrix $\Sigma_w$, respectively. Similarly, we can write down the multivariate likelihood as

$$p(\{\varphi_\alpha\}|\{\mu_\alpha\}) = \frac{1}{\sqrt{(2\pi)^N|\Sigma_z|}} \times \exp\left(-\frac{1}{2}\{\varphi_\alpha - g_\alpha(\mu)\}\Sigma_z^{-1}\{\varphi_\alpha - g_\alpha(\mu)\}^T\right). \tag{28}$$

Now substituting these expressions into Eq. (19) we can get an expression of the Laplace-encoded energy as, up to an overall constant,

$$E(\{\varphi_\alpha\}, \{\mu_\alpha\}) = \frac{1}{2}\{\mu_\alpha - \bar{\mu}_\alpha\}\Sigma_w^{-1}\{\mu_\alpha - \bar{\mu}_\alpha\}^T + \frac{1}{2}\ln|\Sigma_w| + \frac{1}{2}\{\varphi_\alpha - g_\alpha(\mu)\}\Sigma_z^{-1}\{\varphi_\alpha - g_\alpha(\mu)\}^T + \frac{1}{2}\ln|\Sigma_z|. \tag{29}$$

The above Eq. (29) contains non-trivial correlations among the brain variables and sensory data. It is possible to pursue the full general case, e.g., see Bogacz (2017) for a nice tutorial on this, but we do not consider this here. Instead we can simplify on the assumption of statistical independence between environmental variables and between sensory inputs. Under this assumption the prior and likelihood are factorised into the simple forms, respectively,

$$p(\{\mu_\alpha\}) = \prod_{\alpha=1}^{N} p(\mu_\alpha), \tag{30}$$

$$p(\{\varphi_\alpha\}|\{\mu_\alpha\}) = \prod_{\alpha=1}^{N} p(\varphi_\alpha|\{\mu_\alpha\}), \tag{31}$$

where probability densities are the uncorrelated Gaussians,

$$p(\{\mu_\alpha\}) = \prod_{\alpha=1}^{N} \frac{1}{\sqrt{2\pi\sigma_z^\alpha}} \exp\left\{-[\mu_\alpha - \bar{\mu}_\alpha]^2/\left(2\sigma_w^\alpha\right)\right\},$$

$$p(\{\varphi_\alpha\}|\{\mu_\alpha\}) = \prod_{\alpha=1}^{N} \frac{1}{\sqrt{2\pi\sigma_w^\alpha}} \exp\left\{-[\varphi_\alpha - g_\alpha(\mu)]^2/\left(2\sigma_z^\alpha\right)\right\}.$$

This gives the Laplace-encoded energy as

$$E(\{\varphi_\alpha\}, \{\mu_\alpha\}) = \sum_{\alpha=1}^{N}\left[\frac{(\varepsilon_w^\alpha)^2}{2\sigma_w^\alpha} + \frac{1}{2}\ln\sigma_w^\alpha\right] + \sum_{\alpha=1}^{N}\left[\frac{(\varepsilon_z^\alpha)^2}{2\sigma_z^\alpha} + \frac{1}{2}\ln\sigma_z^\alpha\right], \tag{32}$$

where the variances $\sigma_w^\alpha$ and $\sigma_z^\alpha$ are diagonal elements of the covariance matrices $\Sigma_w$ and $\Sigma_z$, respectively. In Eq. (32) we have again used the auxiliary variables

$$\varepsilon_w^\alpha = \mu_\alpha - \bar{\mu}_\alpha,$$
$$\varepsilon_z^\alpha = \varphi_\alpha - g_\alpha.$$

The structure of Eq. (32) suggests that the Laplace-encoded energy, which is an approximation for the VFE, is a quadratic sum of the prediction-errors, modulated by the corresponding inverse variances, and an additional sum of the logarithm of the variances.

### 5.2. A dynamical generative model

In the previous section we considered a simple generative model where an organism understood the environment to be effectively static. Here we extend the formulation to dynamic generative models which have the potential to support inference in dynamically changing environments. Again we start by examining a single sensory input $\varphi$ and a univariate brain state $\mu$. Here we assume that the agent's model of environmental dynamics (again expressed in terms of brain states) follows not Eq. (21), but rather a Langevin-type equation (Zwanzig, 2001)

$$\frac{d\mu}{dt} = f(\mu) + w \tag{33}$$

where $f$ is a function of $\mu$ and $w$ is a random fluctuation. A dynamical generative model can then be obtained by combining the simple generative model, Eq. (20), with Eq. (33).

The FEP utilises the notions of *generalised coordinates* and *higher-order motion* (Friston et al., 2008) to incorporate general forms of dynamics into the G-density. Generalised coordinates involve representing the state of a dynamical system in terms of increasingly higher order derivative of its state variables. For example, generalised coordinates of a position variable may correspond to bare 'position' as well as its (unbounded) higher-order temporal derivatives (velocity, acceleration, jerk, and so on) allowing a more precise specification of a system's state (Friston et al., 2008). To obtain these coordinates we simply take recursively higher order derivatives of both Eqs. (20) and (33).

For the sensory data:

$$\begin{aligned}\varphi &= g(\mu) + z\\ \varphi' &= \frac{\partial g}{\partial \mu}\mu' + z'\\ \varphi'' &= \frac{\partial g}{\partial \mu}\mu'' + z''\\ &\vdots\end{aligned} \tag{34}$$

where we have used the notations,

$$\varphi' \equiv d\varphi/dt, \quad \mu' \equiv d\mu/dt, \quad \mu'' \equiv d^2\mu/dt^2, \quad etc.$$

and where $z, z', \ldots$ are the noises sources at each dynamic order. Here nonlinear derivative terms such as $\mu'^2$, $\mu'\mu''$, etc., have been neglected under a *local linearity assumption* (Friston et al., 2007) and only linear terms have been collected. In some treatments of the FEP it is assumed that the noise sources are correlated (Friston et al., 2008). However, here, for the clarity of the following derivations, we follow more standard state space models and assume each dynamical order receives independent noise, i.e, we assume the covariance between noise sources is zero.

Similarly, the Langevin equation, Eq. (33), is generalised as

$$\begin{aligned}\mu' &= f(\mu) + w\\ \mu'' &= \frac{\partial f}{\partial \mu}\mu' + w'\\ \mu''' &= \frac{\partial f}{\partial \mu}\mu'' + w''\\ &\vdots\end{aligned} \tag{35}$$

where again we have applied the local linearity approximation and we assume each dynamical order receives independent noise denoted as $w, w', \ldots$. Here, it is convenient to denote the multi-dimensional sensory-data $\tilde{\varphi}$ as

$$\tilde{\varphi} = (\varphi, \varphi', \varphi', \ldots) \equiv (\varphi_{[0]}, \varphi_{[1]}, \varphi_{[2]}, \ldots)$$

and states $\tilde{\mu}$ as

$$\tilde{\mu} = (\mu, \mu', \mu'', \ldots) \equiv (\mu_{[0]}, \mu_{[1]}, \mu_{[2]}, \ldots), \tag{36}$$

both being row vectors; where the $n$th-components are defined to be

$$\varphi_{[n]} \equiv \frac{d^n}{dt^n}\varphi = \varphi'_{[n-1]} \quad \text{and} \quad \mu_{[n]} \equiv \frac{d^n}{dt^n}\mu = \mu'_{[n-1]}.$$

The generalised coordinates, Eq. (36), span the generalised state-space in mathematical terms. In this state-space, a point represents an infinite-dimensional vector that encodes the instantaneous trajectory of a brain variable (Friston, 2008a). By construction, the time-derivative of the state vector $\tilde{\mu}$ becomes

$$\tilde{\mu}' \equiv D\tilde{\mu} = \frac{d}{dt}(\mu, \mu', \mu'', \ldots) = (\mu', \mu'', \mu''' \cdots) \equiv (\mu_{[1]}, \mu_{[2]}, \mu_{[3]}, \ldots).$$

The fluctuations in the generalised coordinates are written as

$$\tilde{z} = (z, z', z'', \ldots) \equiv (z_{[0]}, z_{[1]}, z_{[2]}, \ldots),$$

$$\tilde{w} = (w, w', w'', \ldots) \equiv (w_{[0]}, w_{[1]}, w_{[2]}, \ldots).$$

In addition, we denote the vectors associated with time-derivatives of the generative functions as

$$\tilde{g} \equiv (g_{[0]}, g_{[1]}, g_{[2]}, \ldots) \quad \text{and} \quad \tilde{f} \equiv (f_{[0]}, f_{[1]}, f_{[2]}, \ldots)$$

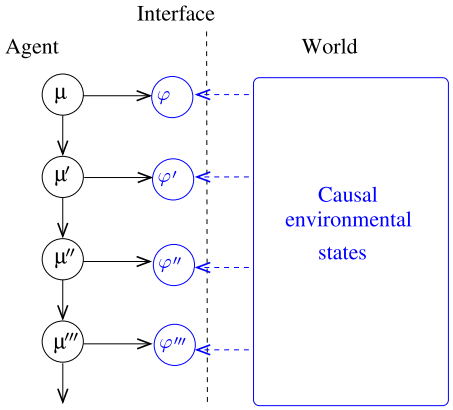


**Fig. 1.** A schematic representation of an agent comprising of a univariate dynamical generative model, see Section 5.2. Interactions between the generalised brain states, $\{\mu, \mu', \ldots\}$ (black) and sensory data, $\{\varphi, \varphi', \ldots\}$ (blue). The arrows denote where one variable (source) specifies the mean of the other (target). Solid arrows represent dependencies within the brain and dashed arrows represent incoming sensory data. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

where the components are given as $g_{[0]} \equiv g(\mu)$ and $f_{[0]} \equiv f(\mu)$, and for $n \geqslant 1$ as

$$g_{[n]} \equiv \frac{\partial g}{\partial \mu}\mu_{[n]} \quad \text{and} \quad f_{[n]} \equiv \frac{\partial f}{\partial \mu}\mu_{[n]}.$$

In terms of these constructs the infinite set of coupled Eqs. (34) and (35) can be written in a compact form as

$$\tilde{\varphi} = \tilde{g} + \tilde{z} \tag{37}$$

$$D\tilde{\mu} = \tilde{f} + \tilde{w}. \tag{38}$$

The generalised map, Eq. (37), describes how the sensory data $\tilde{\varphi}$ are inferred by the representations of their causes $\tilde{\mu}$ at each dynamical order. According to this map, the sensory data at a particular dynamical order $n$, i.e. $\varphi_{[n]}$, engages only with the same dynamical order of the brain states, i.e. $\mu_{[n]}$. The generalised equation of motion, Eq. (38), specifies the coupling between adjacent dynamical orders. A schematic picture of the interactions between these variables is provided in Fig. 1.

As before, in order to obtain the G-density we need to specify the likelihood of the sensory data $p(\tilde{\varphi}|\tilde{\mu})$ and the prior $p(\tilde{\mu})$. The statistical independence of noise at each dynamical order means that we can write the likelihood as a product of conditional densities, i.e.,

$$\begin{aligned} p(\tilde{\varphi}|\tilde{\mu}) &= p(\varphi_{[0]}, \varphi_{[1]}, \varphi_{[2]}, \cdots | \mu_{[0]}, \mu_{[1]}, \mu_{[2]}, \ldots) \\ &= \prod_{n=0}^{\infty} p(\varphi_{[n]}|\mu_{[n]}). \end{aligned} \tag{39}$$

Assuming that the fluctuations at all dynamics orders, $z_{[n]}$, are induced by Gaussian noise, the conditional likelihood-density $p(\varphi^{[n]}|\mu^{[n]})$ is specified as

$$p(\varphi_{[n]}|\mu_{[n]}) = \frac{1}{\sqrt{2\pi\sigma_{z[n]}}} \exp\left[-\left\{\varphi_{[n]} - g_{[n]}\right\}^2/\left(2\sigma_{z[n]}\right)\right].$$

Similarly, the postulate of the conditional independence of the generalised noises $w_{[n]}$ leads to a prior in the form

$$p(\tilde{\mu}) = p(\mu_{[0]}, \mu_{[1]}, \mu_{[2]}, \ldots) = \prod_{n=0}^{\infty} p(\mu_{[n+1]}|\mu_{[n]}). \tag{40}$$

The form of the prior density at dynamical order $n$ is fixed by the assumption of Gaussian noise, which is then given as

$$p(\mu_{[n+1]}|\mu_{[n]}) = \frac{1}{\sqrt{2\pi\sigma_{w[n]}}} \exp\left[-\left\{\mu_{[n+1]} - f_{[n]}\right\}^2/\left(2\sigma_{w[n]}\right)\right].$$

Utilising Eqs. (39) and (40), the G-density is specified as

$$p(\tilde{\varphi}, \tilde{\mu}) = \prod_{n=0}^{\infty} p(\varphi_{[n]}|\mu_{[n]})p(\mu_{[n+1]}|\mu_{[n]}). \tag{41}$$

Given the G-density, the Laplace-encoded energy can be calculated (Eq. (19)) to give, up to a constant,

$$\begin{aligned} E(\tilde{\mu}, \tilde{\varphi}) = &\sum_{n=0}^{\infty}\left\{\frac{1}{2\sigma_{z[n]}}[\varepsilon_{z[n]}]^2 + \frac{1}{2}\ln\sigma_{z[n]}\right\} \\ &+ \sum_{n=0}^{\infty}\left\{\frac{1}{2\sigma_{w[n]}}[\varepsilon_{w[n]}]^2 + \frac{1}{2}\ln\sigma_{w[n]}\right\} \end{aligned} \tag{42}$$

where $\varepsilon_{z[n]}$ and $\varepsilon_{w[n]}$ are $n$th component of the vectors $\tilde{\varepsilon}_z$ and $\tilde{\varepsilon}_w$, respectively, which have been defined to be

$$\varepsilon_{z[n]} \equiv \varphi_{[n]} - g_{[n]} \quad \text{and} \quad \varepsilon_{w[n]} \equiv \mu_{[n+1]} - f_{[n]}.$$

As before, the auxiliary variables, $\varepsilon_{z[n]}$ and $\varepsilon_{w[n]}$, encode *prediction errors*: $\varepsilon_{z[n]}$ is the error between the sensory data $\varphi_{[n]}$ and its prediction $g_{[n]}$ at dynamical order $n$. Likewise, $\varepsilon_{w[n]}$ measures the discrepancy between the expected higher-order output $\mu_{[n+1]}$ and its generation $f_{[n]}$ from dynamical order $n$. Typically only dynamics up to finite order are considered. This can be done by setting the highest order term to random fluctuations, i.e.,

$$\mu_{[n_{max}]} = w_{[n_{max}]}$$

where $w_{[n_{max}]}$ has large variance; thus, the corresponding error term in Eq. (42) will be close to zero and effectively eliminated from the expression for the Laplace-encoded energy. In effect it means that the order below is unconstrained, and free to change in a way that best fits the incoming sensory data. This is related to the notion of empirical priors as discussed in Section 8.1. Thus we have expressed Laplace-encoded energy for dynamics environment, which is an approximation for the VFE, is a quadratic sum of the sensory prediction-error, $\varepsilon_{w[n]}$, and model prediction errors, $\varepsilon_{w[n]}$, across different dynamical orders. Again each error term is modulated by the corresponding variances describing the degree of certainty in those predictions.

We can generalise this to the multivariate case. We set $\{\tilde{\varphi}_\alpha\}$ and $\{\tilde{\mu}_\alpha\}$ as vectors of brain states and rewrite Eqs. (37) and (38) as

$$\tilde{\varphi}_\alpha = \tilde{g}_\alpha + \tilde{z}_\alpha \tag{43}$$

$$D\tilde{\mu}_\alpha = \tilde{f}_\alpha + \tilde{w}_\alpha, \tag{44}$$

where $\alpha$ runs from 1 to $N$. Thus, Eq. (42) becomes

$$\begin{aligned} E(\{\tilde{\mu}_\alpha\}, \{\tilde{\varphi}_\alpha\}) = &\sum_{\alpha=1}^{N}\sum_{n=0}^{\infty}\left\{\frac{1}{2\sigma^\alpha_{z[n]}}[\varepsilon^\alpha_{z[n]}]^2 + \frac{1}{2}\ln\sigma^\alpha_{z[n]}\right\} \\ &+ \sum_{\alpha=1}^{N}\sum_{n=0}^{\infty}\left\{\frac{1}{2\sigma^\alpha_{w[n]}}[\varepsilon^\alpha_{w[n]}]^2 + \frac{1}{2}\ln\sigma^\alpha_{w[n]}\right\} \end{aligned} \tag{45}$$

where we have again used the auxiliary variables

$$\varepsilon^\alpha_{z[n]} \equiv \varphi_{\alpha[n]} - g_{\alpha[n]} \tag{46}$$

$$\varepsilon^\alpha_{w[n]} \equiv \mu_{\alpha[n+1]} - f_{\alpha[n]}. \tag{47}$$

Thus this constitutes an approximation of VFE for a multivariate system across arbitrary number of dynamical orders.

## 6. VFE minimisation: how organisms infer environmental states

In the previous section we demonstrated how to go from a generative model, specifying the organism's beliefs about the environment, to a generative density given expectations on brain states, and finally to an expression for the VFE. In this section we discuss how organisms could minimise VFE to make the R-density a good approximation of the posterior and thus we begin to outline a full biologically plausible process theory. In particular, here, we focus on how this minimisation could be implemented by the neuronal dynamics of the brain outlining one particular process theory.

Under the FEP it is proposed that the innate dynamics of the neural activity evolves in such a way as to minimise the VFE. Specifically, it is suggested that brain states change in such way that they implement a gradient descent scheme on VFE referred to as *recognition dynamics*. Under the proposed gradient-descent scheme, a brain state $\mu_\alpha$ is updated between two sequential steps $t$ and $t+1$ as

$$\mu_\alpha^{t+1} = \mu_\alpha^t - \kappa \hat{\mu}_\alpha \cdot \nabla_{\mu_\alpha} E(\{\mu_\alpha\}, \{\varphi_\alpha\})$$

where $\kappa$ is the learning rate and $\hat{\mu}_\alpha$ is the unit vector along $\mu_\alpha$. Note: this dot product is necessary to pick out the relevant term in the vector differential operator. This process recursively modifies brain states in a way that follows the gradient of Laplace-encoded energy. In the continuous limit the update may be converted to a differential form as

$$\mu_\alpha^{t+1} - \mu_\alpha^t \equiv \dot{\mu}_\alpha.$$

Then, the above discrete updating-scheme can be transformed into a spatio-temporal differential equation,

$$\dot{\mu}_\alpha = -\kappa \hat{\mu}_\alpha \cdot \nabla_{\mu_\alpha} E(\{\mu_\alpha\}, \{\varphi_\alpha\}). \tag{48}$$

The essence of the gradient descent method, as described in Eq. (48), is that the minima of the objective function $E$, i.e., the point where $\nabla_\mu E = 0$, occur at the stationary solution when $\dot{\mu}_\alpha$ vanishes. Thus the dynamics of the brain states settle at a point where the Laplace-encoded energy is minimised.

To update dynamical orders of the brain state $\mu_\alpha$, Eq. (48) must be further generalised to give

$$\mu_{\alpha[n]}^{t+1} - \mu_{\alpha[n]}^t = -\kappa \hat{\mu}_{\alpha[n]} \cdot \nabla_{\tilde{\mu}_\alpha} E(\{\tilde{\mu}_\alpha\}, \{\tilde{\varphi}_\alpha\})$$

where $\hat{\mu}_{\alpha[n]}$ is the unit vector along $\mu_{\alpha[n]}$, $n$th-component of the generalised brain state $\tilde{\mu}_\alpha$ (Section 5.2). Here, we face a complication because the temporal difference between dynamical orders is equal to order above *i.e.*, $\mu_{\alpha[n]}^{t+1} - \mu_{\alpha[n]}^t = \mu_{\alpha[n+1]}$. Consequently, it is not possible to make this difference vanish at any order, meaning that a gradient descent procedure equivalent to Eq. (48) is unable to construct a stationary solution at which the gradient of the Laplace-encoded energy vanishes. However, it is argued that the motion of a point (velocity), i.e. $\dot{\tilde{\mu}}_\alpha$, in the generalised state-space is distinct from the 'trajectory' encoded in the brain (flow velocity) (Friston, 2008a, b; Friston et al., 2008). The latter object is denoted by $D\tilde{\mu}_\alpha$ where $D$ implies also a time-derivative operator which, when acted on $\tilde{\mu}$, results in (see Section 5.2)

$$D\tilde{\mu}_\alpha \equiv (\mu'_{\alpha[0]}, \mu'_{\alpha[1]}, \mu'_{\alpha[2]} \cdots) \equiv (\mu'_\alpha, \mu''_\alpha, \mu'''_\alpha \cdots).$$

Note that this definition of the time derivative operator is formally distinct from the time derivative $\dot{\tilde{\mu}}_\alpha$, i.e. $\mu'_{\alpha[n]} \neq \dot{\mu}_{\alpha[n]}$. The term 'velocity' here has been adapted by analogy with velocity in mechanics in the sense that $\dot{\tilde{\mu}}_\alpha$ denotes first order time-derivative of 'position', namely the bare variable $\tilde{\mu}_\alpha$. Prepared with this extra theoretical construct, the gradient descent scheme is restated in the FEP as

$$\dot{\mu}_{\alpha[n]} - D\mu_{\alpha[n]} = -\kappa \hat{\mu}_{\alpha[n]} \cdot \nabla_{\tilde{\mu}_\alpha} E(\{\tilde{\mu}_\alpha\}, \{\tilde{\varphi}_\alpha\}) \tag{49}$$

where $D\mu_{\alpha[n]} = \mu'_{\alpha[n]}$. According to this formulation, $E$ is minimised with respect to the generalised state $\tilde{\mu}_\alpha$ when the 'path of the mode' (generalised velocity) is equal to the 'mode of the path' (average velocity), in other words the gradient of $E$ vanishes when $\dot{\tilde{\mu}}_\alpha = D\tilde{\mu}_\alpha$. It is worth noting that in 'static' situations where generalised motions are not required (see Section 8.4), the concept of the 'mode of the path' is not needed, i.e. $D\tilde{\mu}_\alpha \equiv 0$ by construction. In such situations we consider the relevant brain variables $\mu_\alpha$ to reach the desired minimum when there is no more temporal change in $\mu_\alpha$ in the usual sense, i.e. when $\dot{\mu}_\alpha = 0$.

In sum, these equations specify sets of first order ordinary differential equations that could be straightforwardly integrated by neuronal processing, e.g., they are very similar equations for firing rate dynamics in neural networks (e.g., see Haykin & Network, 2004). Continuously integrating these equations in the presence of stream of sensory data would make brain states continuously minimise VFE and thus implement approximate inference on environmental states. Furthermore, with some additional assumptions about their implementation (Friston & Kiebel, 2009c) they become strongly analogous to the predictive coding framework (Rao & Ballard, 1999).

## 7. Active inference

A central appeal of the FEP is that it suggests not only an account of perceptual inference but also an account of action within the same framework: active inference. Specifically while perception minimises VFE by changing brain states to better predict sensory data, action instead acts on the environment to alter sensory input to better fit sensory predictions. Thus action minimises VFE indirectly by changing sensations.

In this section we describe a gradient-descent scheme analogous to that in the previous section but for action. To ground this idea for action, and combine it with the framework for perceptual inference discussed in previous sections, we present an implementation of a simple agent-based model.

Under the FEP action does not appear explicitly in the formulation of VFE but minimises VFE by changing sensory data. To evaluate this the brain must have a inverse model (Wolpert, 1997) of how sensory data change with action (Friston et al., 2010). Specifically, for a single brain state variable $\mu$ we write this as $\varphi = \varphi(a)$ where $a$ represents the action and $\varphi$ is a single sensory channel. Action in this context could be moving ones limbs or eyes and thus changing sensory input. Given an inverse model we can write how the Laplace-encoded energy changes with respect to action using the chain rule as,

$$\frac{dE(\mu, \varphi)}{da} \equiv \frac{d\varphi}{da} \frac{\partial E(\mu, \varphi)}{\partial \varphi}. \tag{50}$$

Thus we can write the same gradient decent scheme outlined in the last section to calculate the actions that minimise the Laplace-encoded energy as

$$\dot{a} = -\kappa_a \frac{d\varphi}{da} \frac{dE(\mu, \varphi)}{d\varphi} \tag{51}$$

where $\kappa_a$ is the learning rate associated with action.

It is straightforward to write this gradient descent scheme for a vector of brain states in generalised coordinates as

$$\dot{a} = -\kappa_a \sum_\alpha \frac{d\tilde{\varphi}_\alpha}{da} \cdot \nabla_{\tilde{\varphi}_\alpha} E(\{\tilde{\mu}_\alpha\}, \{\tilde{\varphi}_\alpha\}). \tag{52}$$

The idea that brains innately possess inverse models, at first glance, seems somewhat troublesome. However, under the FEP

the execution of motor control depends only on predictions about proprioceptors (internal sensors) which can be satisfied by classic reflex arcs (Friston, 2011; Friston et al., 2010). On this reading exteroceptive, and perhaps interoceptive (Seth, 2015), sensations are only indirectly minimised by action. While a full assessment of this idea's implications is outside the remit of this work, it provides an interesting alternative to conventional notions of motor control, or behaviour optimisation, that rest on maximising a value function or minimising a cost function (Friston, 2011).

To give a concrete example of how perceptual and active inference work we present an implementation of a simple agent-based model. Specifically we present a model that comprises a mobile agent that must move to achieve some desired local temperature, $T_{desire}$.

**The agent's world.** The agent's environment, or *generative process* (Friston et al., 2010), consists of a 1-dimensional line and a simple temperature source. The agent's position on this plane is denoted by the environmental variable $\vartheta$ and the agent's temperature depends on its position in the following manner,

$$T(\vartheta) = \frac{T_0}{\vartheta^2 + 1}, \tag{53}$$

where $T_0$ is the temperature at the origin, i.e., this equation gives the dynamics of the agents' environment (the environmental causes of its sensory signals). The corresponding temperature gradient is readily given by,

$$\frac{dT}{d\vartheta} = -T_0 \frac{2\vartheta}{(\vartheta^2 + 1)^2} \equiv T_\vartheta. \tag{54}$$

The temperature profile is depicted by the black line in Fig. 3(a). We allow the agent to sense both the local temperature and the temporal derivative of this temperature

$$\varphi = T + z_{gp} \tag{55}$$

$$\varphi' = T_\vartheta \vartheta' + z'_{gp} \tag{56}$$

where $z_{gp}$ and $z'_{gp}$ are normally distributed noise in the sensory readings. Note that the subscript *gp* reminds us that this noise is a part of the agent's environment (rather than its brain model) described by the generative process.

In this model the agent is presumed to sit on a flat frictionless plane and, thus, in the absence of action the agent is stationary. We allow the agent to set its own velocity by setting it equal to the action variable $a$ as,

$$\vartheta' = a. \tag{57}$$

**The agent's brain**. To construct a scheme to minimise VFE we first write down what the agents believe in terms of it brain states. Note: here we will assumed the agent knows the dynamics of the environment and how sensory data is generated, i.e., we provide it with an appropriate generative model *a priori*. In Section 8 we consider how the agent could learn this model but we do not deal with this possibility in the simple simulation presented here.

The agent has brain state $\mu$ which represents the agents estimate of its temperature in the environment. Following Eqs. (35), we write a generative model for the agent, up to third order, as

$$\mu' = f(\mu) + w \quad \text{where} \quad f(\mu) \equiv -\mu + T_{desire} \tag{58}$$

$$\mu'' = -\mu' + w' \tag{59}$$

$$\mu''' = w'', \tag{60}$$

where the third order term is just random fluctuations with large variance and thus is effectively eliminated from the expression for the Laplace-encoded energy, see Section 5.2. Following Eq. (34), we write the agent's belief about its sensory data only to first order as,

$$\varphi = g(\mu) + z \quad \text{where} \quad g(\mu) \equiv \mu$$
$$\varphi' = \mu' + z'$$

which follow Eqs. (55) and (56), i.e., the agent knows how its sensory data is generated. It is important to note that in this dynamic formulation agents do not desire specific states themselves but rather have beliefs about the dynamics of the world. For example the agent we present does not explicitly desire to be at $T_{desire}$ (so there is not an explicit prior on $\mu$, or rather this is a flat prior). However, examining the agent's generative model we easily see that it possesses a stable equilibrium point at $T_{desire}$. In effect the agent believes in a environment where the forces it experiences naturally move it to its desired temperature, see Section 2 and (Friston et al., 2010). However, examining the agent's generative model, see Eqs. (58)–(60) we easily see that it possesses a stable equilibrium point at $T_{desire}$ .

We can write the Laplace-encoded energy, Eq. (45), for this model, as

$$E(\tilde{\mu}, \tilde{\varphi}) = \frac{1}{2}\left[\frac{1}{\sigma_{z[0]}}(\varepsilon_{z[0]})^2 + \frac{1}{\sigma_{z[1]}}(\varepsilon_{z[1]})^2 + \frac{1}{\sigma_{w[0]}}(\varepsilon_{w[0]})^2 + \frac{1}{\sigma_{w[1]}}(\varepsilon_{w[1]})^2\right], \tag{61}$$

where the various error terms are given as

$$\varepsilon_{z[0]} = \varphi - \mu$$
$$\varepsilon_{z[1]} = \varphi' - \mu'$$
$$\varepsilon_{w[0]} = \mu' + \mu - T_{desire}$$
$$\varepsilon_{w[1]} = \mu'' + \mu'.$$

Also, $\sigma_{z[0]}$, $\sigma_{z[1]}$, $\sigma_{w[0]}$, and $\sigma_{w[1]}$ in Eq. (61) are the variances corresponding to the noise terms $z$, $z'$, $w$, and $w'$, respectively. In addition we have dropped logarithm of variance terms, see Eq. (24) because they play no role when we minimise these equations with respect to the brain variable $\mu$. A schematic of the generative model for this system is given in Fig. 2. Note, that the noise terms in the agents internal model are distinct from those in Eqs. (55) and (56) and represent the agent's beliefs about the noise on environmental states and sensory data rather than the actual noise on these variables. As we will see these terms effectively represent the confidence of the agent in its own sensory input.

Using the gradient decent scheme described in Eq. (49) we write the recognition dynamics as

$$\begin{aligned}
\dot{\mu} &= \mu' - \kappa_a\left[-\frac{\varepsilon_{z[0]}}{\sigma_{z[0]}} + \frac{\varepsilon_{w[0]}}{\sigma_{w[0]}}\right] \\
\dot{\mu}' &= \mu'' - \kappa_a\left[-\frac{\varepsilon_{z[1]}}{\sigma_{z[1]}} + \frac{\varepsilon_{w[0]}}{\sigma_{w[0]}} + \frac{\varepsilon_{w[1]}}{\sigma_{w[1]}}\right] \\
\dot{\mu}'' &= -\kappa_a \frac{\varepsilon_{w[1]}}{\sigma_{w[1]}}.
\end{aligned} \tag{62}$$

Here we have considered generalised coordinates up to second order only. To allow the agent to perform action we must provide it with an inverse model, which we assume is hard-wired (Friston et al., 2010). Replacing the agent's velocity with the action variable $a$ in Eq. (56) we specify this as

$$\frac{d\varphi'}{da} = \frac{d}{da}\left(aT_\vartheta + z'_{gp}\right) = T_\vartheta. \tag{63}$$

Effectively the agent believes that action changes the temperature in a way that is consistent with its beliefs about the temperature gradient. Given this inverse model we can write down the minimisation scheme for action as.

$$\dot{a} = -\kappa_a\left[\frac{d\varphi'}{da}\frac{\partial E}{\partial \varphi'}\right] = -\kappa_a T_\vartheta \frac{\varepsilon_{z[1]}}{\sigma_{z[1]}}. \tag{64}$$

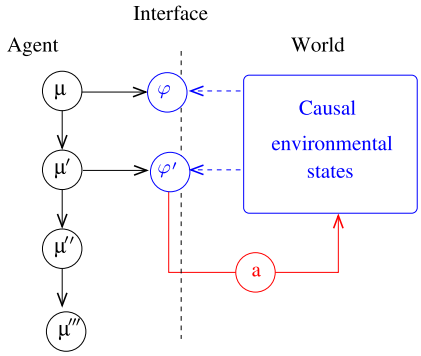


**Fig. 2.** The generative model for an active inference agent: Terms have the same meanings as in Fig. 1. The agent acts on the world, via variable $a$ (red arrow), to change sensory input $\varphi'$ and minimise VFE indirectly. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Thus, Eqs. (62) through (64) describe the complete agent–environment system and can be straightforwardly integrated.

**A simulation of the agent-world system**. Simulating agent-world system involves three steps:

1. We simulate the agent's environment by numerically integrating Eqs. (53)–(56) (i.e the generative process).
2. Perception is implemented by updating the brain states constrained by this sensory data using the gradient descent scheme, this achieved by numerically integrating Eq. (62).
3. Action involves the agent changing the environment, through action $a$, which here involves simply moving along a line, so that $\varphi'$ also minimises the VFE. This is given by Eq. (63) and the gradient descent update for this is given in Eq. (64).

See the code in Appendix B for details.

Fig. 3 shows the behaviour of the agent in the absence of action, i.e., when the agent is unable to move. We examine two conditions. In a first condition the agent's sensory variances $\sigma_{z[0]}$, $\sigma_{z[1]}$ are several orders of magnitude smaller than model variances $\sigma_{w[0]}$ and $\sigma_{w[1]}$. Thus the agent has higher confidence (see Section 5.1) in sensory input than in its internal model. Under this condition the agent successfully infers both the local temperature and its corresponding derivatives, see Fig. 3(b) black lines. In effect the agent ignores its internal model and the gradient descent scheme is equivalent to a least mean square estimation on the sensory data, see supplied code in Appendix B. In a second condition, see Fig. 4 red lines, we equally balance internal model and sensory variances ($\sigma_{z[i]} = \sigma_{w[i]}$, $i = 0, 1$). Now minimisation of VFE cannot satisfy both sensory perception and predictions of the agent's internal model, i.e., what the agent perceives is in conflict with what it desires. Thus the inferred local temperature sits somewhere between its desired and sensed temperature, see Fig. 3(b).

In Fig. 4, after an initial period, the agent is allowed to act according to Eq. (64). It does so by changing the environment to bring it in line with sensory predictions and the desires encoded within its dynamic model, i.e., the agent moves towards the desired temperatures.

The reduction of surprisal can be quantified as the difference between the Laplace-encoded energy (and thus VFE) in presence and absence of action, i.e., the difference between black and red traces in Fig. 4(e), respectively. Specifically, it is the portion of the VFE that must be minimised by acting on the environment rather than through optimisation of the agent's environment model. We

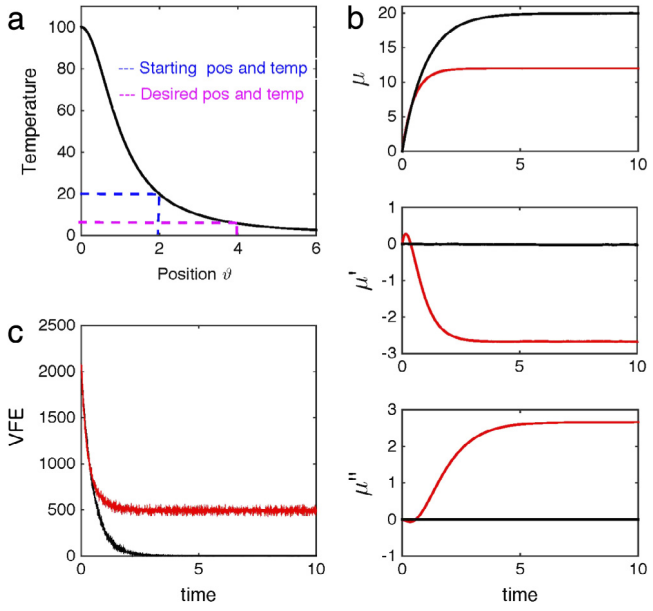


**Fig. 3.** Perceptual inference: The agent's environment comprises a simple temperature gradient (a), the blue and magenta lines give the actual and desired positions of the agent, respectively. The agent performs simple perceptual inference (b), the dynamics of three generalised coordinates, $\mu$, $\mu'$ and $\mu''$, are given in the top, middle and bottom panels, respectively. Two conditions are shown, when the confidence in the sensory input is high (i.e. $\sigma_{z[i]}$ is small in comparison to $\sigma_{w[i]}$), black line, and when confidence is equal between the internal model and sensory input, red line, respectively. VFE in both conditions monotonically decreases (c): black and red traces, respectively. The tension between sensory input and internal model manifests a relatively high value of VFE (c) (red curve), compared to the case where sensation has much higher confidence than the internal model (black curve). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

leave a more explicit quantification of the dynamics of surprisal for future work.

In summary we have presented an example of an agent performing a very simple task under the FEP. The model demonstrates how the minimisation of VFE can underpin both perception and action. Furthermore, it shows how a tension between desires and perception can be reconciled through action. Many other agent based implementations of the FEP have been presented in the literature, see for example (Friston et al., 2010), which can be constructed in a similarly way.

## 8. Hierarchical inference and learning

In the previous sections we developed the FEP for organisms given simple dynamical generative models. We then investigated the emergence of behaviour in a simulated organism (agent) furnished with an appropriate generative model of a simple environment. The assumption here was that organisms possess some knowledge or beliefs of about how the environment works a priori, in the form of a pre-specified generative model. However, another promise of the FEP is the ability to learn and infer arbitrary environmental dynamics (Friston, 2008a). To achieve this it is suggested that the brain starts out with a very general hierarchical generative model of environmental dynamics which is moulded and refined through experience. The advantage of using hierarchical models, as we will see, is that they suggest a way of avoiding specifying an explicit and fixed prior, and thus can implement empirical Bayes (Casella & Berger, 2002). In what follows, we will first consider inference and then turn to learning. We start by providing a description of a hierarchical G-density which is capable of hierarchical inference which is equivalent to empirical Bayes (Casella

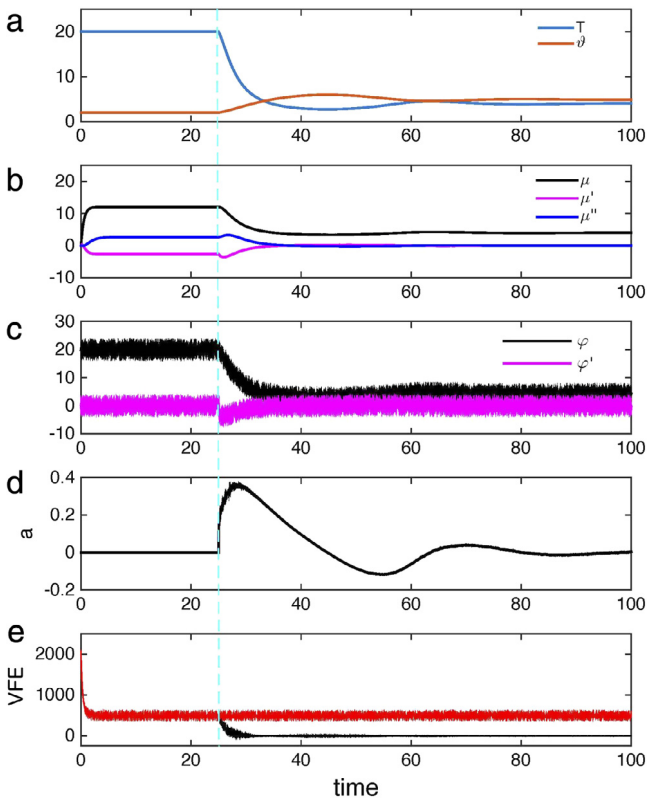


**Fig. 4.** Perceptual and active inference: An agent with equal confidence in its internal model and sensory input $\sigma_{z[i]} = \sigma_{w[i]} = 1$ is allowed to act at $t = 25$. The agent acts, see (d), to alter its position, see (a: orange line), to bring down its initial temperature ($T = 20$) to the desired temperature ($T = T_{desire} = 4$), see (a: blue line). It does this by bringing its sensory data (c) in line with its desire, i.e., $\varphi = T_{desire}$ and thus the brain state becomes equal to its desired state, see (b). VFE was calculated in the presence and absence of the onset of action at $t = 25$, see e, black and red lines, respectively. First VFE is reduced by inference ($t < 25$), then later through action (eg., black line). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

& Berger, 2002). We then combine this with dynamical generative model described in Eq. (45) to define what we shall call the *full construct*. We go on to describe how appropriate parameters (e.g. parameters that define generative functions) hyperparameters (e.g. inverse variances or precisions) of the G-density for given world could be discovered through learning. Note: there is a fundamental distinction between inference and learning. Inference refers to recognising the current causes of sensory input by inferring hidden or latent states of the world that vary with time. This contrasts with inferring the (time invariant) parameters that mediate dependencies among (time varying) states. We finish this section by showing how action can be described in this construct.

### 8.1. Hierarchical generative model

A key challenge for Bayesian inference models is how to specify the priors. Hierarchical models provide a powerful response to this challenge, in which higher levels can provide empirical priors or constraints on lower levels (Kass & Steffey, 1989). In the FEP, hierarchical models are mapped onto the hierarchical organisation of the cortex (Felleman & Van Essen, 1991; Zeki & Shipp, 1988), which requires extension of the simple generative model described above.

We denote $\mu^{(i)}$ as a brain state at hierarchical level $i$ and we assume $M$ cortical levels, with $i = 1$ the *lowest* level and $i = M$ as the *highest*. Then, the hierarchical model may be written explicitly

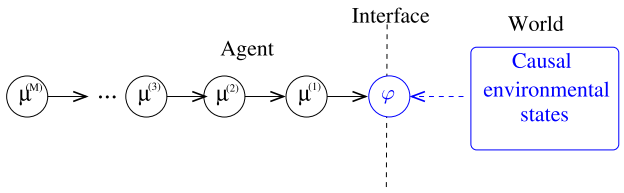


**Fig. 5.** A hierarchical generative model for the univariate case: Interactions between hierarchical brain states, $\{\mu^{(1)}, \mu^{(2)}, \ldots\}$, and sensory data at the lowest hierarchical level, $\varphi$ . The arrows denote where one variable (source) specifies the mean of the other (target).

as (Friston, 2008a)

$$\begin{aligned}\varphi &= g^{(1)}(\mu^{(1)}) + z^{(0)}\\ \mu^{(1)} &= g^{(2)}(\mu^{(2)}) + z^{(1)}\\ \mu^{(2)} &= \cdots\\ &\vdots\\ \mu^{(M)} &= z^{(M)}\end{aligned}$$

which can be written compactly as

$$\mu^{(i)} = g^{(i+1)}(\mu^{(i+1)}) + z^{(i)} \tag{65}$$

where $i$ runs through $1, 2, \ldots, M$. We further assume that the sensory data $\varphi$ reside exclusively at the lowest cortical level $\mu^{(1)}$ and dynamics at the highest level $\mu^{(M)}$ are governed by a random fluctuation $z^{(M)}$, i.e:

$$\mu^{(0)} \equiv \varphi \quad \text{and} \quad g^{(M+1)} \equiv 0. \tag{66}$$

The hierarchy Eq. (65) specifies that a cortical state $\mu^{(i)}$ is connected to higher level $\mu^{(i+1)}$ through the generative function $g^{(i+1)}$. The fluctuations $z^{(i)}$ exist at each level, in particular $z^{(0)}$ designating the observation noise at the sensory interface, and are assumed to be statistically independent. A schematic of the interaction in the hierarchical generative model is given in Fig. 5.

Having defined the hierarchical model, one can write the corresponding G-density as

$$\begin{aligned}p(\varphi, \mu) &= p(\mu^{(0)}|\mu^{(1)}, \mu^{(2)}, \ldots, \mu^{(N)})p(\mu^{(1)}, \mu^{(2)}, \ldots, \mu^{(M)})\\ &\equiv p(\mu^{(0)}|\mu^{(1)})p(\mu^{(1)}|\mu^{(2)}) \cdots p(\mu^{(M-1)}|\mu^{(M)})p(\mu^{(M)}).\end{aligned} \tag{67}$$

The second step in Eq. (67) assumes that the transition probabilities from higher levels to lower levels are Markovian. Consequently, Eq. (67) asserts that the likelihood associated with a given level, for instance $p(\mu^{(i)}|\mu^{(i+1)})$, serves as a prior density for the level immediately below, $i - 1$. The prior at the highest level $p(\mu^{(M)})$ contains information only with respect to its spontaneous noise, which may be given by a Gaussian form

$$p(\mu^{(M)}) = \frac{1}{\sqrt{2\pi\sigma_z^{(M)}}} \exp\left\{-[\mu^{(M)}]^2/\left(2\sigma_z^{(M)}\right)\right\} \tag{68}$$

where the mean has been assumed to be zero and $\sigma_z^{(M)}$ is the variance. We shall further assume that the Gaussian noises are responsible for the (statistically independent) fluctuations at all hierarchical levels. Accordingly, the likelihoods $p(\mu^{(i)}|\mu^{(i+1)})$ are given as

$$\begin{aligned}p(\mu^{(i)}|\mu^{(i+1)}) = &\frac{1}{\sqrt{2\pi\sigma_z^{(i)}}}\\ &\times \exp\left[-\left\{\mu^{(i)} - g^{(i+1)}(\mu^{(i+1)})\right\}^2/\left\{2\sigma_z^{(i)}\right\}\right].\end{aligned} \tag{69}$$

**Table 4**
Mathematical objects relating to the hierarchical generative model.

| Symbol | Name & description |
|---|---|
| Hierarchical model | $p(\varphi,\mu)=p(\mu^{(M)})\prod_{i=0}^{M-1}p(\mu^{(i)}\|\mu^{(i+1)})$ |
| $\mu^{(i)}$ | Brain states at cortical level $i$ ($i=1,2,\ldots,M$); $\mu^{(0)}\equiv\varphi$ denotes the sensory data which reside at the lowest cortical level. |
| $g^{(i)}(\mu^{(i)})$ | Generative map (or function) of the brain state $\mu^{(i)}$ to estimate one-level lower state $\mu^{(i-1)}$ in the cortical hierarchy via $\mu^{(i-1)}=g^{(i)}(\mu^{(i)})+z^{(i-1)}$; where $z^{(i-1)}$ is Gaussian noise. |
| $p(\mu^{(i)}\|\mu^{(i+1)})$ | Likelihood of $\mu^{(i)}$ given a value for $\mu^{(i+1)}$; which acts as a prior for $p(\mu^{(i-1)}\|\mu^{(i)})$ in the cortical hierarchy. |
| $p(\mu^{(M)})$ | Probabilistic representation of brain states at the highest level, which forms the highest prior. |

and the G-density reduces to

$$p(\varphi,\mu)=\left[\prod_{i=0}^{M}\frac{1}{\sqrt{2\pi\sigma_z^{(i)}}}\right]\exp\left(-\sum_{i=0}^{M}\frac{1}{2\sigma_z^{(i)}}[\varepsilon^{(i+1)}]^2\right)\tag{70}$$

where the auxiliary variables $\varepsilon^{(i)}$ have been introduced as

$$\varepsilon^{(i)}\equiv\mu^{(i-1)}-g^{(i)}(\mu^{(i)}).\tag{71}$$

The quantity $\varepsilon^{(i)}$ measures the discrepancy between the prediction (estimation) at a given level $\mu^{(i)}$ via $g^{(i)}$ and $\mu^{(i-1)}$ at a lower-level, which comprises a *prediction error*.

Finally, by substituting the G-density, constructed in Eq. (70), into Eq. (19), after a simple manipulation, the Laplace-encoded energy $E$ is given up to a constant as

$$E(\mu,\varphi)=\sum_{i=0}^{M}\left\{\frac{1}{2\sigma_z^{(i)}}[\varepsilon^{(i+1)}]^2+\frac{1}{2}\ln\sigma_z^{(i)}\right\}.\tag{72}$$

The variance of the noise at the top level of hierarchy is typically assumed to be large and thus the corresponding term in the Laplace-encoded energy Eq. (72) is approximately zero. As with the higher dynamical orders discussed above Section 5.2 this means that the level below is effectively unconstrained (has no prior) and thus this type of inference constitutes an example of empirical Bayes (Casella & Berger, 2002).

Table 4 itemises the mathematical objects associated with the hierarchical generative model.

### 8.2. Combining hierarchical and dynamical models: the full construct

We now combine the dynamical structure and the multivariate brain states in a single expression. First we note that under the FEP brain states representing neuronal activity $\mu_\alpha$ are divided into the *hidden* states $x_\alpha$ and the *causal* states $v_\alpha$,

$$\mu_\alpha=(x_\alpha,v_\alpha).$$

Here causal and hidden states distinguish between states that are directly observable from those that are not. At the lowest level of the hierarchy causal states refer to sensory variables (e.g., the colour of a red hot poker) and hidden states the variables that constitute the generative process (e.g., the temperature of the poker). However this distinction is generalised throughout the hierarchy and each level is thought of as 'observing' (through causal states) the level below. Then, the full FEP implementation can be derived formally by extending Eqs. (43) and (44) (Eq. (65))

$$\tilde{v}_\alpha^{(i)}=\tilde{g}_\alpha^{(i+1)}(\tilde{x}_\alpha^{(i+1)},\tilde{v}_\alpha^{(i+1)})+\tilde{z}_\alpha^{(i)},\quad i=0,1,\ldots M\tag{73}$$

$$D\tilde{x}_\alpha^{(i)}=\tilde{f}_\alpha^{(i)}(\tilde{x}_\alpha^{(i)},\tilde{v}_\alpha^{(i)})+\tilde{w}_\alpha^{(i)},\quad i=1,2\ldots,M\tag{74}$$

where the brain-state index runs through $\alpha=1,2,\ldots,N$ and $\tilde{v}_\alpha^{(0)}$ designates the sensory data at the lowest cortical level, $i=1$. Inter-level hierarchical links are made through the causal states and intra-hierarchical level dynamics through the hidden states. The generalised coordinates of neuronal brain state $\alpha$ in hierarchical level $i$ are given by the infinite-dimensional vectors

$$\tilde{x}_\alpha^{(i)}\equiv(x_{\alpha[0]}^{(i)},x_{\alpha[1]}^{(i)},x_{\alpha[2]}^{(i)},\ldots)\quad\text{and}\quad\tilde{v}_\alpha^{(i)}\equiv(v_{\alpha[0]}^{(i)},v_{\alpha[1]}^{(i)},v_{\alpha[2]}^{(i)},\ldots)$$

where the components are labelled by the subscripts $[n]$, $n=0,1,\ldots,\infty$. Note that we have introduced different notations in the vector components: The subscript $\alpha$ for brain states at a given hierarchical level, the superscript $(i)$ for the hierarchical indices, and the subscript $[n]$ for the dynamical orders. Recall that the $n$th component of the vector $\tilde{x}_\alpha^{(i)}$ and $\tilde{v}_\alpha^{(i)}$ are time-derivatives of order $n$, namely

$$x_{\alpha[n]}^{(i)}\equiv\frac{d^n}{dt^n}x_\alpha^{(i)}\quad\text{and}\quad v_{\alpha[n]}^{(i)}\equiv\frac{d^n}{dt^n}v_\alpha^{(i)}.$$

The other mathematical quantities in Eqs. (73) and (74) are given explicitly as:

$$D\tilde{x}_\alpha^{(i)}=(x_{\alpha[1]}^{(i)},x_{\alpha[2]}^{(i)},x_{\alpha[3]}^{(i)},\ldots),$$

$$\tilde{z}_\alpha^{(i)}\equiv(z_{\alpha[0]}^{(i)},z_{\alpha[1]}^{(i)},z_{\alpha[2]}^{(i)},\ldots),\quad\text{and}\quad\tilde{w}_\alpha^{(i)}\equiv(w_{\alpha[0]}^{(i)},w_{\alpha[1]}^{(i)},w_{\alpha[2]}^{(i)},\ldots).$$

The generative functions appearing in Eqs. (73) and (74) are specified for $n\geqslant 1$, under the local-linearity assumption, as

$$g_{\alpha[n]}(x_{\alpha[n]}^{(i+1)},v_{\alpha[n]}^{(i+1)})\equiv\frac{\partial g}{\partial v_{\alpha[n]}^{(i+1)}}v_{\alpha[n]}^{(i+1)}\equiv g_{\alpha[n]}^{(i+1)}$$

and

$$f_{\alpha[n]}(x_{\alpha[n]}^{(i)},v_{\alpha[n]}^{(i)})\equiv\frac{\partial f}{\partial x_{\alpha[n]}^{(i)}}x_{\alpha[n]}^{(i)}\equiv f_{\alpha[n]}^{(i)}.$$

For the lowest dynamical order of $n=0$,

$$g_{\alpha[0]}^{(i+1)}=g(x_{\alpha[0]}^{(i+1)},v_{\alpha[0]}^{(i+1)})\quad\text{and}\quad f_{\alpha[0]}^{(i)}=f(x_{\alpha[0]}^{(i)},v_{\alpha[0]}^{(i)}).$$

It is evident from Eq. (73) that the causal states $\tilde{v}_\alpha^{(i)}$ at one hierarchical level are predicted from states at one level higher in the hierarchy $\tilde{v}_\alpha^{(i+1)}$ via the map $\tilde{g}_\alpha^{(i+1)}$: $\tilde{z}_\alpha^{(i)}$ specifies the fluctuations associated with these inter-level links. Eq. (74) asserts that the dynamical transitions of the hidden states $\tilde{x}_\alpha^{(i)}$ are induced *within* a given hierarchical level via $\tilde{f}_\alpha^{(i)}$: The corresponding fluctuations are given by $\tilde{w}_\alpha^{(i)}$. In order to describe these transitions more transparently, we spell out Eqs. (73) and (74) explicitly:

$$\begin{aligned}
\tilde{v}_\alpha^{(0)}&=\tilde{g}_\alpha^{(1)}(\tilde{x}_\alpha^{(1)},\tilde{v}_\alpha^{(1)})+\tilde{z}_\alpha^{(0)} & D\tilde{x}_\alpha^{(1)}&=\tilde{f}_\alpha^{(1)}(\tilde{x}_\alpha^{(1)},\tilde{v}_\alpha^{(1)})+\tilde{w}_\alpha^{(1)}\\
\tilde{v}_\alpha^{(1)}&=\tilde{g}_\alpha^{(2)}(\tilde{x}_\alpha^{(2)},\tilde{v}_\alpha^{(2)})+\tilde{z}_\alpha^{(1)} & D\tilde{x}_\alpha^{(2)}&=\tilde{f}_\alpha^{(2)}(\tilde{x}_\alpha^{(2)},\tilde{v}_\alpha^{(2)})+\tilde{w}_\alpha^{(2)}\\
&\vdots & &\vdots\\
\tilde{v}_\alpha^{(M-1)}&=\tilde{g}_\alpha^{(M)}+\tilde{z}_\alpha^{(M-1)} & D\tilde{x}^{(M-1)}&=\tilde{f}^{(M-1)}+\tilde{w}^{(M-1)}\\
\tilde{v}_\alpha^{(M)}&=\tilde{z}_\alpha^{(M)} & D\tilde{x}_\alpha^{(M)}&=\tilde{f}_\alpha^{(M)}+\tilde{w}_\alpha^{(M)}
\end{aligned}$$

where we have set that

$$\tilde{\varphi}_\alpha\equiv\tilde{v}_\alpha^{(0)}\quad\text{and}\quad\tilde{g}_\alpha^{(M+1)}\equiv 0.$$

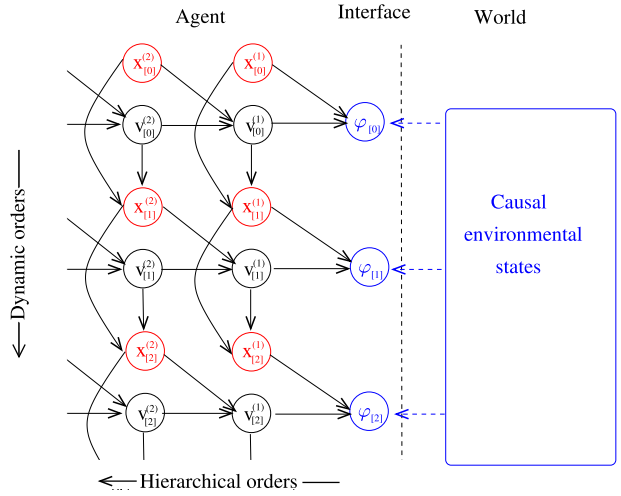


**Fig. 6.** The full construct for the univariate case: See text for description of the variables. The arrows denote where one variable (source) specifies the mean of the other (target).

Note that the sensory data $\tilde{\varphi}_\alpha$ reside at the lowest hierarchical level and are to be inferred by the causal states $\tilde{v}_\alpha^{(1)}$ at the corresponding dynamical orders. At the highest cortical level $M$ the causal states $\tilde{v}_\alpha^{(M)}$ are described by the spontaneous fluctuations $\tilde{z}_\alpha^{(M)}$ around their means (which have been set to be zero without loss of generality). Note that the generalised motions of hidden states are still present at the highest cortical level, in just the same way that they manifest at all the other hierarchical levels: the corresponding spontaneous fluctuations are given by $\tilde{w}_\alpha^{(M)}$. A schematic of the interactions between the variables that comprise the full construct is given in Fig. 6.

Separating brain states into causal and hidden states, we can now express the G-density by generalising Eq. (67) as

$$\begin{aligned} p(\tilde{\varphi}, \tilde{\mu}) &= \prod_{\alpha=1}^{N} p(\tilde{\varphi}_\alpha, \tilde{\mu}_\alpha) = \prod_{\alpha=1}^{N} p(\tilde{\mu}_\alpha^{(M)}) \prod_{i=0}^{M-1} p(\tilde{\mu}_\alpha^{(i)} | \tilde{\mu}_\alpha^{(i+1)}) \\ &\Rightarrow \prod_{\alpha=1}^{N} p(\tilde{x}_\alpha^{(M)}, \tilde{v}_\alpha^{(M)}) \prod_{i=0}^{M-1} p(\tilde{x}_\alpha^{(i)}, \tilde{v}_\alpha^{(i)} | \tilde{x}_\alpha^{(i+1)}, \tilde{v}_\alpha^{(i+1)}) \\ &= \prod_{\alpha=1}^{N} p(\tilde{x}_\alpha^{(M)}, \tilde{v}_\alpha^{(M)}) \prod_{i=0}^{M-1} p(\tilde{x}_\alpha^{(i)} | \tilde{v}_\alpha^{(i)}) p(\tilde{v}_\alpha^{(i)} | \tilde{x}_\alpha^{(i+1)}, \tilde{v}_\alpha^{(i+1)}) \end{aligned} \tag{75}$$

where in the second step we have used $\tilde{\mu}_\alpha^{(i)} = (\tilde{x}_\alpha^{(i)}, \tilde{v}_\alpha^{(i)})$ and only the causal states $\tilde{v}_\alpha^{(i)}$ are involved in the inter-level transitions in the third step. Also, it must be understood that $p(\tilde{x}_\alpha^{(0)} | \tilde{v}_\alpha^{(0)}) \equiv 1$ in Eq. (75), which appears solely for a mathematical compactness. The intra-level conditional probabilities $p(\tilde{x}_\alpha^{(i)} | \tilde{v}_\alpha^{(i)})$ are given as

$$\begin{aligned} p(\tilde{x}_\alpha^{(i)} | \tilde{v}_\alpha^{(i)}) &= p(x_{\alpha[0]}^{(i)}, x_{\alpha[1]}^{(i)}, \dots | v_{\alpha[0]}^{(i)}, v_{\alpha[1]}^{(i)}, \cdots) \\ &= \prod_{n=0}^{\infty} p(x_{\alpha[n]}^{(i)} | v_{\alpha[n]}^{(i)}) \end{aligned} \tag{76}$$

where in the second step we have made use of the assumption of statistical independence among the generalised states at different dynamical orders. The quantity $p(x_{\alpha[n]}^{(i)} | v_{\alpha[n]}^{(i)})$ specifies the conditional density at the dynamical order $n$ within the hierarchical level $i$, where the corresponding fluctuations $w_{\alpha[n]}^{(i)}$ are assumed to take Gaussian form as

$$\begin{aligned} p(x_{\alpha[n]}^{(i)} | v_{\alpha[n]}^{(i)}) \equiv & \frac{1}{\sqrt{2\pi \sigma_{w[n]}^{\alpha(i)}}} \\ & \times \exp\left[ -\left( x_{\alpha[n+1]}^{(i)} - f_{\alpha[n]}^{(i)} \right)^2 / \left( 2\sigma_{w[n]}^{\alpha(i)} \right) \right]. \end{aligned} \tag{77}$$

The conditional densities $p(\tilde{v}_\alpha^{(i)} | \tilde{x}_\alpha^{(i+1)}, \tilde{v}_\alpha^{(i+1)})$ appearing in Eq. (75) link two successive causal states in the cortical hierarchy which are specified by a similar Gaussian fluctuation for $z_{\alpha[n]}^{(i)}$ via Eq. (73) as

$$\begin{aligned} p(\tilde{v}_\alpha^{(i)} | \tilde{x}_\alpha^{(i+1)}, \tilde{v}_\alpha^{(i+1)}) \equiv & \prod_{n=0}^{\infty} \frac{1}{\sqrt{2\pi \sigma_{z[n]}^{\alpha(i)}}} \\ & \times \exp\left[ -\left( v_{\alpha[n]}^{(i)} - g_{\alpha[n]}^{(i+1)} \right)^2 / \left( 2\sigma_{z[n]}^{\alpha(i)} \right) \right]. \end{aligned} \tag{78}$$

What is left unspecified in constructing the G-density fully, i.e. Eq. (75), is the prior density $p(\tilde{x}_\alpha^{(M)}, \tilde{v}_\alpha^{(M)})$ at the highest cortical level. It is given here explicitly as

$$\begin{aligned} p(\tilde{x}_\alpha^{(M)}, \tilde{v}_\alpha^{(M)}) \equiv & \prod_{n=0}^{\infty} \frac{1}{\sqrt{2\pi \sigma_{w[n]}^{\alpha(M)}}} \exp\left\{ -[x_{\alpha[n+1]}^{(M)} - f_{\alpha[n]}^{(M)}]^2 / \left( 2\sigma_{w[n]}^{\alpha(M)} \right) \right\} \\ & \times \prod_{n=0}^{\infty} \frac{1}{\sqrt{2\pi \sigma_{z[n]}^{\alpha(M)}}} \exp\left\{ -[v_{\alpha[n]}^{(M)}]^2 / \left( 2\sigma_{z[n]}^{\alpha(M)} \right) \right\}. \end{aligned} \tag{79}$$

The prior in the highest cortical level, Eq. (79), comprises the lateral generalised motions of the hidden states and the spontaneous, random fluctuations associated with the causal states. It is assumed that both causal and hidden states fluctuate about zero means.

Next, the Laplace-encoded energy $E$ can be written explicitly by substituting Eq. (75) into Eq. (19) and incorporating the likelihood and prior densities, Eq. (77), (78), and (79), at all hierarchical levels and dynamical orders. After a straightforward manipulation, we obtain the Laplace-encoded energy for a specific brain variable $\mu_\alpha$ as

$$\begin{aligned} E_\alpha(\tilde{\mu}_\alpha, \tilde{\varphi}_\alpha) = & \sum_{n=0}^{\infty} \left\{ \frac{1}{2\sigma_{w[n]}^{\alpha(M)}} \left( x_{\alpha[n+1]}^{(M)} - f_{\alpha[n]}^{(M)} \right)^2 + \frac{1}{2} \ln \sigma_{w[n]}^{\alpha(M)} \right\} \\ & + \sum_{n=0}^{\infty} \left\{ \frac{1}{2\sigma_{z[n]}^{\alpha(M)}} \left( v_{\alpha[n]}^{(M)} \right)^2 + \frac{1}{2} \ln \sigma_{z[n]}^{\alpha(M)} \right\} \\ & + \sum_{i=1}^{M-1} \sum_{n=0}^{\infty} \left\{ \frac{1}{2\sigma_{w[n]}^{\alpha(i)}} \left( x_{\alpha[n+1]}^{(i)} - f_{\alpha[n]}^{(i)} \right)^2 + \frac{1}{2} \ln \sigma_{w[n]}^{\alpha(i)} \right\} \\ & + \sum_{i=0}^{M-1} \sum_{n=0}^{\infty} \left\{ \frac{1}{2\sigma_{z[n]}^{\alpha(i)}} \left( v_{\alpha[n]}^{(i)} - g_{\alpha[n]}^{(i+1)} \right)^2 + \frac{1}{2} \ln \sigma_{z[n]}^{\alpha(i)} \right\} \end{aligned}$$

where the first and second terms are from prior-densities at the highest level, Eq. (79), the third term is from Eq. (77), and last term from Eq. (78). A quick inspection reveals that the first and second terms can be absorbed into the third and fourth terms, respectively. Then, the Laplace-encoded energy for multiple brain variables is written compactly as

$$\begin{aligned} E(\tilde{\mu}, \tilde{\varphi}) &= \sum_{\alpha=1}^{N} E_\alpha(\tilde{\mu}_\alpha, \tilde{\varphi}_\alpha) \\ &= \frac{1}{2} \sum_{\alpha=1}^{N} \sum_{n=0}^{\infty} \sum_{i=1}^{M} \left\{ \frac{1}{\sigma_{w[n]}^{\alpha(i)}} \left( \varepsilon_{w[n]}^{\alpha(i)} \right)^2 + \ln \sigma_{w[n]}^{\alpha(i)} \right\} \\ &+ \frac{1}{2} \sum_{\alpha=1}^{N} \sum_{n=0}^{\infty} \sum_{i=0}^{M} \left\{ \frac{1}{\sigma_{z[n]}^{\alpha(i)}} \left( \varepsilon_{z[n]}^{\alpha(i+1)} \right)^2 + \ln \sigma_{z[n]}^{\alpha(i)} \right\} \end{aligned} \tag{80}$$

**Table 5**
Mathematical objects relating to the full generative model.

| Symbol | Name & description |
|---|---|
| Full construct | $p(\tilde{\varphi}_\alpha, \tilde{\mu}_\alpha) = p(\tilde{x}_\alpha^{(M)}, \tilde{v}_\alpha^{(M)}) \prod_{i=0}^{M-1} p(\tilde{x}_\alpha^{(i)} \vert \tilde{v}_\alpha^{(i)}) p(\tilde{v}_\alpha^{(i)} \vert \tilde{x}_\alpha^{(i+1)}, \tilde{v}_\alpha^{(i+1)})$ |
| $\tilde{\mu}_\alpha^{(i)}$ | Brain state $\alpha$ in cortical level $i$ in generalised coordinates, whose $n$th component is denoted as $\mu_{\alpha[n]}^{(i)}$. |
| $\tilde{x}_\alpha^{(i)}, \tilde{v}_\alpha^{(i)}$ | Two distinct neuronal representations, $\tilde{\mu}_\alpha^{(i)} = (\tilde{x}_\alpha^{(i)}, \tilde{v}_\alpha^{(i)})$; designated as hidden and causal states, respectively. |
| $\tilde{g}_\alpha^{(i)}$ | Generative map of the causal state $\tilde{v}_\alpha^{(i)}$ to learn the state one level below, $\tilde{v}_\alpha^{(i-1)} = \tilde{g}_\alpha^{(i)}(\tilde{x}_\alpha^{(i)}, \tilde{v}_\alpha^{(i)}) + \tilde{z}_\alpha^{(i-1)}$. |
| $\tilde{f}_\alpha^{(i)}$ | Generative function which induces the Langevin-type equation of motion of the hidden state $\tilde{x}_\alpha^{(i)}$, $\dot{\tilde{x}}_\alpha^{(i)} = \tilde{f}_\alpha^{(i)}(\tilde{x}_\alpha^{(i)}, \tilde{v}_\alpha^{(i)}) + \tilde{w}_\alpha^{(i)}$. |
| $\tilde{z}_\alpha^{(i)}, \tilde{w}_\alpha^{(i)}$ | Random fluctuations treated as Gaussian noise. |
| $p(\tilde{x}_\alpha^{(M)}, \tilde{v}_\alpha^{(M)})$ | Prior density of the brain state $\tilde{\mu}_\alpha$ at the highest cortical level ($M$). |
| $p(\tilde{x}_\alpha^{(i)} \vert \tilde{v}_\alpha^{(i)})$ | Probabilistic representation of the intra-level dynamics of hidden states $\tilde{x}_\alpha^{(i)}$ conditioned on the causal state $\tilde{v}_\alpha^{(i)}$ via $\tilde{f}_\alpha^{(i)}$; dynamic transition from order $n$ to $n+1$ is hypothesised as the Gaussian fluctuation of $w_{\alpha[n]}^{(i)} = x_{\alpha[n+1]}^{(i)} - f_{\alpha[n]}^{(i)}$. |
| $p(\tilde{v}_\alpha^{(i)} \vert \tilde{x}_\alpha^{(i+1)}, \tilde{v}_\alpha^{(i+1)})$ | Likelihood density of the causal state $\tilde{v}^{(i+1)}$ which serves as a prior for one level lower density, representing statistically the inter-level map between two successive causal states, $z_{\alpha[n]}^{(i)} = v_{\alpha[n]}^{(i)} - g_{\alpha[n]}^{(i+1)}$, by the Gaussian fluctuation. |

where we have defined the prediction errors

$$\varepsilon_{z[n]}^{\alpha(i)} \equiv v_{\alpha[n]}^{(i-1)} - g_{\alpha[n]}^{(i)}\left(x_{\alpha[n]}^{(i)}, v_{\alpha[n]}^{(i)}\right) \tag{81}$$

$$\varepsilon_{w[n]}^{\alpha(i)} \equiv x_{\alpha[n+1]}^{(i)} - f_{\alpha[n]}^{(i)}\left(x_{\alpha[n]}^{(i)}, v_{\alpha[n]}^{(i)}\right). \tag{82}$$

Thus, it turns out that the Laplace-encoded energy is expressed essentially as a sum of the prediction-errors (squared) and their associated variances. It appears in Eq. (80) that the structure of the first term differs from the second term: In the first term the hierarchical index runs from $i = 1$ which indicates the lowest cortical level, while the second term includes additional $i = 0$ in the hierarchical sum which designates the sensory data, $\tilde{\varphi} \equiv \tilde{v}^{(0)}$. Note also in Eq. (81) that $\varepsilon_{z[n]}^{\alpha(M+1)} = v_{\alpha[n]}^{(M)}$ because the highest hierarchical level is at $i = M$, accordingly $g_{\alpha[n]}^{(M+1)} \equiv 0$ by construction.

Table 5 provides the glossary of the mathematical objects involved in the G-density in the full construct for a single brain activity $\mu_\alpha$.

To summarise, the 'full construct' incorporates into the G-density, both multi-level hierarchies corresponding to cortical architecture, and multi-scale dynamics in each level via generalised coordinates. The G-density is expressed as the sequential product of the priors and the likelihoods, cascading down the cortical hierarchy to the lowest level where the sensory data are registered (mediated by causal states), and taking into account the intra-level dynamics, mediated by hidden states. The final form of the Laplace-encoded energy, Eq. (80), has been derived from Eq. (19) which specifies the Laplace-encoded energy as the (negative) logarithm of the generative density constructed for the hidden and causal brain states.

### 8.3. The full-construct recognition dynamics and neuronal activity

We now describe recognition dynamics incorporating the full construct (Section 8.2), given the Laplace-encoded energy $E(\tilde{\mu}, \tilde{\varphi})$, Eq. (80). In the full construct, the brain states $\tilde{\mu}_\alpha$ are decomposed into the causal states $\tilde{v}_\alpha$ which link the cortical hierarchy and the hidden states $\tilde{x}_\alpha$ which implement the dynamical ordering within a cortical level.

Distinguishing the 'path of the modes' from the 'modes of the path', see Section 6, the learning algorithm for the dynamical causal states on the cortical level $i$ can be constructed from

$$\dot{v}_{\alpha[n]}^{(i)} - D v_{\alpha[n]}^{(i)} \equiv -\kappa_z \hat{v}_{\alpha[n]}^{(i)} \cdot \nabla_{\tilde{v}_\alpha} E(\tilde{\mu}, \tilde{\varphi}) \tag{83}$$

where $\kappa_z$ is the learning rate and $\hat{v}_{\alpha[n]}^{(i)}$ is the unit vector along $v_{\alpha[n]}^{(i)}$. As mentioned in Section 6, the crucial assumption here is that when the path of modes becomes identical to the modes of the path, i.e. $\dot{\tilde{v}}_\alpha^{(i)} - D\tilde{v}_\alpha^{(i)} \to 0$, the Laplace-encoded energy $E$ takes its minimum, and *vice versa*. The gradient operation in the RHS of Eq. (83) can be made explicit to give

$$\begin{aligned}
&\hat{v}_{\alpha[n]}^{(i)} \cdot \nabla_{\tilde{v}_\alpha} E(\tilde{\mu}, \tilde{\varphi}) \\
&= \frac{\partial}{\partial v_{\alpha[n]}^{(i)}} \left[ \frac{1}{2\sigma_{z[n]}^{\alpha(i-1)}} \left\{ \varepsilon_{z[n]}^{\alpha(i)} \right\}^2 + \frac{1}{2\sigma_{z[n]}^{\alpha(i)}} \left\{ \varepsilon_{z[n]}^{\alpha(i+1)} \right\}^2 \right. \\
&\quad \left. + \frac{1}{2\sigma_{w[n]}^{\alpha(i)}} \left\{ \varepsilon_{w[n]}^{\alpha(i)} \right\}^2 \right] \\
&= \frac{1}{\sigma_{z[n]}^{\alpha(i-1)}} \varepsilon_{z[n]}^{\alpha(i)} \frac{\partial \varepsilon_{z[n]}^{\alpha(i)}}{\partial v_{\alpha[n]}^{(i)}} + \frac{1}{\sigma_{z[n]}^{\alpha(i)}} \varepsilon_{z[n]}^{\alpha(i+1)} \frac{\partial \varepsilon_{z[n]}^{\alpha(i+1)}}{\partial v_{\alpha[n]}^{(i)}} \\
&\quad + \frac{1}{\sigma_{w[n]}^{\alpha(i)}} \varepsilon_{w[n]}^{\alpha(i)} \frac{\partial \varepsilon_{w[n]}^{\alpha(i)}}{\partial v_{\alpha[n]}^{(i)}}
\end{aligned} \tag{84}$$

where one can further see that

$$\frac{\partial \varepsilon_{z[n]}^{\alpha(i)}}{\partial v_{\alpha[n]}^{(i)}} = -\frac{\partial g_{z[n]}^{\alpha(i)}}{\partial v_{\alpha[n]}^{(i)}}, \quad \frac{\partial \varepsilon_{z[n]}^{\alpha(i+1)}}{\partial v_{\alpha[n]}^{(i)}} = 1, \quad \text{and} \quad \frac{\partial \varepsilon_{w[n]}^{\alpha(i)}}{\partial v_{\alpha[n]}^{(i)}} = -\frac{\partial f_{\alpha[n]}^{(i)}}{\partial v_{\alpha[n]}^{(i)}}.$$

The additional auxiliary variables are introduced:

$$\xi_{z[n]}^{\alpha(i)} \equiv \varepsilon_{z[n]}^{(i)} / \sigma_{z[n]}^{\alpha(i-1)} \equiv \Lambda_{z[n]}^{\alpha(i-1)} \left\{ v_{\alpha[n]}^{(i-1)} - g_{\alpha[n]}^{(i)}\left(x_{\alpha[n]}^{(i)}, v_{\alpha[n]}^{(i)}\right) \right\}, \tag{85}$$

$$\xi_{w[n]}^{\alpha(i)} \equiv \varepsilon_{w[n]}^{(i)} / \sigma_{w[n]}^{(i)} \equiv \Lambda_{w[n]}^{\alpha(i)} \left\{ x_{\alpha[n+1]}^{(i)} - f_{[n]}^{(i)}\left(x_{\alpha[n]}^{(i)}, v_{\alpha[n]}^{(i)}\right) \right\}, \tag{86}$$

where $\Lambda_{z[n]}^{\alpha(i)}$ and $\Lambda_{w[n]}^{\alpha(i)}$ are the inverse of the variances,

$$\Lambda_{z[n]}^{\alpha(i)} \equiv 1/\sigma_{z[n]}^{\alpha(i)} \quad \text{and} \quad \Lambda_{w[n]}^{\alpha(i)} \equiv 1/\sigma_{w[n]}^{\alpha(i)}, \tag{87}$$

which are called the *precisions*. Note that the precisions reflect the magnitude of the prediction errors.

Its is proposed that the auxiliary variables $\xi_{z[n]}^{\alpha(i)}$ and $\xi_{w[n]}^{\alpha(i)}$ represent *error units* and that the brain states, $v_{\alpha[n]}^{(i)}$ and $x_{\alpha[n]}^{(i)}$, similarly represent *state units* or, equivalently, *representation units*, within neuronal populations (Friston, 2010c; Friston & Kiebel, 2009a).

In terms of 'predictive coding' or (more generally) hierarchical message passing in cortical networks (Friston, 2008a), Eq. (85)

implies that the error-units $\xi^{\alpha(i)}_{z[n]}$ receive signals from causal states $v^{(i-1)}_{\alpha[n]}$ lying in immediately lower hierarchical level and also from causal and hidden states in the same level, $v^{(i)}_{\alpha[n]}$ and $x^{(i)}_{\alpha[n]}$, via the generative function $g^{(i)}_{\alpha[n]}$. Similarly, Eq. (86) implies that the error-units $\xi^{\alpha(i)}_{w[n]}$ specify prediction-error in the within-level (lateral) dynamics: $\xi^{\alpha(i)}_{w[n]}$ designates prediction error between the objective hidden-state $x^{(i)}_{\alpha[n+1]}$ and its estimation from one-order lower causal- and hidden-states $v^{(i)}_{\alpha[n]}$ and $x^{(i)}_{\alpha[n]}$, via the different generative function $f^{(i)}_{\alpha[n]}$.

With the help of Eq. (84), one can recast Eq. (83) to give the dynamics of the causal states as

$$\dot{v}^{(i)}_{\alpha[n]} = Dv^{(i)}_{\alpha[n]} + \kappa_z \frac{\partial g^{\alpha(i)}_{[n]}}{\partial v^{(i)}_{\alpha[n]}} \xi^{\alpha(i)}_{z[n]} - \kappa_z \xi^{\alpha(i+1)}_{z[n]} + \kappa_z \frac{\partial f^{(i)}_{\alpha[n]}}{\partial v^{(i)}_{\alpha[n]}} \xi^{\alpha(i)}_{w[n]} \tag{88}$$

which shows clearly how hierarchical links are made among nearest-neighbour cortical levels. Specifically, the representation units of causal states $v^{(i)}_{\alpha[n]}$ are updated by the error units $\xi^{\alpha(i+1)}_{z[n]}$ which reside in the level immediately above, and also by the error-units $\xi^{\alpha(i)}_{z[n]}$ and $\xi^{\alpha(i)}_{w[n]}$ in the same hierarchical level, all at the same dynamical order.

The intra-level dynamics of hidden states are generated similarly as

$$\begin{aligned} \dot{x}^{(i)}_{\alpha[n]} &\equiv Dx^{(i)}_{\alpha[n]} - \kappa_w \hat{x}^{(i)}_{\alpha[n]} \cdot \nabla_{\tilde{x}_\alpha} E(\tilde{\mu}, \tilde{\varphi}) \\ &= Dx^{(i)}_{\alpha[n]} - \kappa_w \xi^{\alpha(i)}_{w[n-1]} + \kappa_w \frac{\partial f^{(i)}_{\alpha[n]}}{\partial x^{(i)}_{\alpha[n]}} \xi^{\alpha(i)}_{w[n]} + \kappa_w \frac{\partial g^{(i)}_{\alpha[n]}}{\partial x^{(i)}_{\alpha[n]}} \xi^{\alpha(i)}_{z[n]} \end{aligned} \tag{89}$$

where $\kappa_w$ is the leaning rate. In passing to the second line in Eq. (89), one needs to evaluate

$$\begin{aligned} &\hat{x}^{(i)}_{\alpha[n]} \cdot \nabla_{\tilde{x}_\alpha} E(\tilde{\mu}, \tilde{\varphi}) \\ \rightarrow\ &\frac{1}{\sigma^{\alpha(i)}_{w[n-1]}} \varepsilon^{\alpha(i)}_{w[n-1]} \frac{\partial \varepsilon^{\alpha(i)}_{w[n-1]}}{\partial x^{(i)}_{\alpha[n]}} + \frac{1}{\sigma^{\alpha(i)}_{w[n]}} \varepsilon^{\alpha(i)}_{w[n]} \frac{\partial \varepsilon^{\alpha(i)}_{w[n]}}{\partial x^{(i)}_{\alpha[n]}} \\ &+ \frac{1}{\sigma^{\alpha(i-1)}_{z[n]}} \varepsilon^{\alpha(i)}_{z[n]} \frac{\partial \varepsilon^{\alpha(i)}_{z[n]}}{\partial x^{(i)}_{\alpha[n]}}, \end{aligned}$$

and an explicit evaluation of the derivatives of the prediction errors, Eqs. (81) and (82). The hidden-state learning algorithm, Eq. (89), specifies how the representation-units $x^{(i)}_{\alpha[n]}$ are driven by the error-units in the current level $i$ at both the immediately lower dynamical order $\xi^{\alpha(i)}_{w[n-1]}$ and the same dynamical order $\xi^{\alpha(i)}_{w[n]}$, and also by the error units $\xi^{\alpha(i)}_{z[n]}$ in the current level at the same dynamical order.

To summarise, the hierarchical, dynamical causal structure of the generative model is fully implemented in the mathematical constructs given by Eqs. (85) and (86) (specifying prediction errors), and Eqs. (88) and (89) (specifying update rules for state-units).

According to these equations, the state units come to encode the conditional expectations of the environmental causes of sensory data, and the error units measure the discrepancy between these expectations and the data. Error units are driven by state units at the same level and from the level below, whereas state units are driven by error units at the same level and the level above. Thus, prediction errors are passed up the hierarchy (bottom-up) and predictions (conditional expectations) are passed down the hierarchy (top-down), fully consistent with predictive coding (Rao & Ballard, 1999).

### 8.4. Parameters and hyperparameters: synaptic efficacy and gain

Thus far we have discussed how environmental variables can be inferred given an appropriate G-density. In this section we discuss how the G-density itself can be learned. It has been proposed that the dynamics of neural systems is captured by three time-scales, $\tau_\mu < \tau_\theta < \tau_\gamma$. The first, $\tau_\mu$, represents the timescale of the dynamics of sufficient statistics of the encoded in the R-density i,.e $\mu \equiv (x, v)$ as described above. In contrast $\tau_\theta$ and $\tau_\gamma$ represent the slow timescale of synaptic efficacies and gains which are parameterised implicitly in Eq. (80) through the generative functions, $f$ and $g$, and the variances $\sigma$ (or the precisions $\Lambda$, Eq. (87)), respectively. Under the FEP slow variables are assumed to be approximately 'static' or 'time-invariant' in contrast to the 'time-varying' neuronal states $\mu$ (Friston & Kiebel, 2009a). Second, changes in $\theta$ and $\gamma$ (with respect to a small $\delta t$) have a much smaller effect on the Laplace-encoded energy (or VFE) than do changes in $\mu$, i.e.

$$\frac{\partial F}{\partial \theta} \frac{\delta \theta}{\delta t} \ll \frac{\partial F}{\partial \mu} \frac{\delta \mu}{\delta t}.$$

The latter point implies that, from the perspective of gradient-descent, what is relevant for $\theta$ and $\gamma$ is not the VFE $F$ but the accumulation, more precisely the integration of $F$ over time (Friston & Stephan, 2007)

$$S[F] \equiv \int dt F(\tilde{\mu}, \tilde{\varphi}; \theta, \gamma) \tag{90}$$

where the time-dependence of $F$ is implicit through the arguments. To distinguish their different roles, $\theta^{(i)}_\alpha$ are called *parameters* and $\gamma^{(i)}_\alpha$ are called *hyperparameters*, corresponding to brain state $\mu_\alpha$, in each hierarchical level $i$. Eqs. (85) and (86) can now be generalised to include these parameters and hyperparameters as

$$\xi^{\alpha(i)}_{z[n]} = \Lambda^{\alpha(i-1)}_{z[n]}(\gamma^{(i-1)}_\alpha) \left\{ v^{(i-1)}_{\alpha[n]} - g^{(i)}_{\alpha[n]} \left( x^{(i)}_{\alpha[n]}, v^{(i)}_{\alpha[n]}; \theta^{(i)}_\alpha \right) \right\}, \tag{91}$$

$$\xi^{\alpha(i)}_{w[n]} = \Lambda^{\alpha(i)}_{w[n]}(\gamma^{(i)}_\alpha) \left\{ x^{(i)}_{\alpha[n+1]} - f^{(i)}_{\alpha[n]} \left( x^{(i)}_{\alpha[n]}, v^{(i)}_{\alpha[n]}; \theta^{(i)}_\alpha \right) \right\}. \tag{92}$$

The Laplace-encoded energy including $\theta$ and $\gamma$ may therefore be written as

$$\begin{aligned} E(\tilde{\mu}, \tilde{\varphi}; \theta, \gamma) = &\frac{1}{2} \sum_{\alpha=1}^{N} \sum_{n=0}^{\infty} \sum_{i=1}^{M} \left\{ \varepsilon^{\alpha(i)}_{w[n]} \xi^{\alpha(i)}_{w[n]} - \ln \Lambda^{\alpha(i)}_{w[n]} \right\} \\ &+ \frac{1}{2} \sum_{\alpha=1}^{N} \sum_{n=0}^{\infty} \sum_{i=0}^{M} \left\{ \varepsilon^{\alpha(i+1)}_{z[n]} \xi^{\alpha(i+1)}_{z[n]} - \ln \Lambda^{\alpha(i)}_{z[n]} \right\}. \end{aligned} \tag{93}$$

We are now in a position to write down the recognition dynamics for the slow synaptic efficacy $\theta$ and for the slower synaptic gain $\gamma$. Specifically, gradient descent for the parameters $\theta^{(i)}_\alpha$ is applied using the time-integral of $F$, given in Eq. (90), assuming a static model (i.e., without dynamical order indices), as

$$\dot{\theta}^{(i)}_\alpha = -\kappa_\theta \hat{\theta}^{(i)}_\alpha \cdot \nabla_\theta S$$

which, when temporal differentiation is repeated on both sides, gives rise to

$$\ddot{\theta}^{(i)}_\alpha = -\kappa_\theta \hat{\theta}^{(i)}_\alpha \cdot \nabla_\theta E(\tilde{\mu}, \tilde{\varphi}; \theta, \gamma). \tag{94}$$

After explicitly carrying out the gradient on the RHS of Eq. (94), one obtains an equation to minimise $\theta^{(i)}_\alpha$ corresponding to brain variable $\mu_\alpha$ at cortical level $i$

$$\ddot{\theta}^{(i)}_\alpha = \sum_{n=0}^{\infty} \left[ \kappa_\theta \frac{\partial g^{(i)}_{\alpha[n]}}{\partial \theta^{(i)}_\alpha} \xi^{\alpha(i)}_{z[n]} + \kappa_\theta \frac{\partial f^{(i)}_{\alpha[n]}}{\partial \theta^{(i)}_\alpha} \xi^{\alpha(i)}_{w[n]} \right] \tag{95}$$

where the summation over the dynamic index $n$ reflects the generalised motion over causal as well as hidden states. According to

**Table 6**
Mathematical objects relating to the recognition dynamics.

| Symbol | Name & description |
|---|---|
| $\nabla_{\tilde{\mu}} E(\tilde{\mu}, \tilde{\varphi})$ | 'Gradient' of the Laplace encoded-energy: Multi-dimensional derivative of the scalar function $E$; which vanishes at an optimum $\tilde{\mu}^*$. |
| Dynamical construct | $\dot{\tilde{\mu}}_\alpha^{(i)} - D\tilde{\mu}_\alpha^{(i)} = -\kappa_\alpha \nabla_{\tilde{\mu}_\alpha^{(i)}} E(\tilde{\mu}, \tilde{\varphi}),\ \tilde{\mu}_\alpha^{(i)} = (\tilde{x}_\alpha^{(i)}, \tilde{v}_\alpha^{(i)})$ |
| $\tilde{\mu}_\alpha^{(i)}$ | Generalised brain states: A point in the generalised state space to represent fast 'time-dependent' neuronal activity $\tilde{\mu}_\alpha$ on each cortical level $i$ [see Eqs. (73) and (74)]. |
| $\dot{\tilde{\mu}}_\alpha^{(i)}, D\tilde{\mu}_\alpha^{(i)}$ | $\dot{\tilde{\mu}}_\alpha^{(i)}$ is the 'path of the mode'; $D\tilde{\mu}_\alpha^{(i)}$ is the 'mode of the path'. $\dot{\tilde{\mu}}_\alpha^{(i)}$ represents the rate of change of a brain state in generalised state space, while $D\tilde{\mu}_\alpha^{(i)}$ represents the encoded motion in the brain; when the two become identical, i.e. $\dot{\tilde{\mu}}_\alpha^{(i)} = D\tilde{\mu}_\alpha^{(i)}$, in the course of recognition dynamics, $E$ reaches its minimum. |
| Static construct | $\ddot{\mu}_\beta^{(i)} = -\kappa_\beta \hat{\mu}_\beta^{(i)} \cdot \nabla_{\mu_\beta} E(\tilde{\mu}, \tilde{\varphi}; \theta, \gamma),\ \mu_\beta = \theta, \gamma$ |
| $\tilde{\Lambda}_z^{\alpha(i)}, \tilde{\Lambda}_w^{\alpha(i)}$ | Precisions: Inverse variances in the generalised coordinates [see Eq. (87)]. |
| $\theta_\alpha^{(i)}, \gamma_\alpha^{(i)}$ | Parameters, hyperparameters: Slow brain states that are treated as 'static' and are associated with $\theta_\alpha^{(i)}$ and $\gamma_\alpha^{(i)}$, respectively, on each cortical level; where $\theta_\alpha^{(i)}$ appear as parameters in the generative functions $g_\alpha^{(i)}$ and $f_\alpha^{(i)}$, and $\gamma_\alpha^{(i)}$ are hyperparameters in the precisions $\Lambda_z^{\alpha(i)}$ and $\Lambda_w^{\alpha(i)}$. |
| $\tilde{\xi}_z^{\alpha(i)}, \tilde{\xi}_w^{\alpha(i)}$ | Prediction errors; measuring the discrepancy between the observation and the evaluation [*e.g.* Eqs. (91), (92)] |

Eq. (95) synaptic efficacy is influenced by error-units only in the same cortical level.

Similarly, the learning algorithm for the hyperparameters $\gamma$, specifically for $\gamma_\alpha^{(i)}$ associated with brain's representation of environmental states $\mu_\alpha$ at cortical level $i$, is given from

$$\dot{\gamma}_\alpha^{(i)} = -\kappa_\gamma \hat{\gamma}_\alpha^{(i)} \cdot \nabla_\gamma S$$

which results in

$$\ddot{\gamma}_\alpha^{(i)} = -\frac{1}{2}\sum_{n=0}^{\infty}\left[\kappa_\gamma \frac{\partial \Lambda_{w[n]}^{\alpha(i)}}{\partial \gamma_\alpha^{(i)}}\left\{\xi_{w[n]}^{\alpha(i)}\right\}^2 - \kappa_\gamma \frac{\partial}{\partial \gamma_\alpha^{(i)}} \ln \Lambda_{w[n]}^{\alpha(i)}\right] - \frac{1}{2}\sum_{n=0}^{\infty}\left[\kappa_\gamma \frac{\partial \Lambda_{z[n]}^{\alpha(i)}}{\partial \gamma_\alpha^{(i)}}\left\{\xi_{z[n]}^{\alpha(i+1)}\right\}^2 - \kappa_\gamma \frac{\partial}{\partial \gamma_\alpha^{(i)}} \ln \Lambda_{z[n]}^{\alpha(i)}\right]. \quad (96)$$

According to this equation, synaptic gains are influenced by error units in the same level $\xi_w^{(i)}$ and also by error units in one-level above $\xi_z^{(i+1)}$.

Note that the equations for $\theta$ and $\gamma$, Eqs. (95) and (96), are *by construction* second-order differential equations, unlike the corresponding equations for state-units $\mu$ [Eqs. (88) and (89)], which are first-order in time (Friston et al., 2008). Table 6 provides the summary of mathematical symbols appearing in the recognition dynamics in the dynamical construct and also in the static construct.

To summarise the FEP prescribes *recognition dynamics* by gradient descent with respect to the sufficient statistics $\tilde{\mu}$, parameters $\theta$, and hyperparameters $\gamma$ on the Laplace-encoded energy $E(\tilde{\mu}, \tilde{\varphi}; \theta, \gamma)$, given the sensory input $\tilde{\varphi}$. At the end of this process, an optimal $\tilde{\mu}^*$ is specified which represents the brain's posterior expectation of the environmental cause of the observed sensory data. In theory the second term in the VFE $F$, Eq. (18), can be fixed according to Eq. (17) thereby completing the minimisation of the VFE, although in practice this is rarely done and the focus is on approximating the means, parameters and hyperparameters.

This whole minimisation process is expressed abstractly as

$$\tilde{\mu}^* = \arg \min_{\tilde{\mu}} F(\tilde{\mu}, \tilde{\varphi}) \quad (97)$$

where $\tilde{\mu}^*$ is the minimising (optimal) solution. The resulting minimised VFE can be calculated by substituting the optimising $\tilde{\mu}^*$ for $\tilde{\mu}$ as

$$F^* = F(\tilde{\mu}^*, \tilde{\varphi}\,).$$

The only remaining task is to specify the generative functions $f$ and $g$, which will depend on the particular system being modelled. We have utilised a concrete model in our calculation in Section 7. Examples of various generating functions have already been provided (Friston, 2008a; Friston et al., 2010, 2016; Pezzulo et al., 2015; Pio-Lopez, Nizard, Friston, & Pezzulo, 2016), to which we refer the reader.

### 8.5. Active inference on the full construct

The VFE also accounts for an active inference by minimising the VFE with respect to action, for which a formal procedure can be written as

$$a^* = \arg \min_a F(\tilde{\mu}, \tilde{\varphi}(a)) \quad (98)$$

where $a^*$ is the minimising solution. Similarly with Eq. (51) we can write down the gradient descent scheme for the minimisation in the full construct for action corresponding to brain's representation $\mu_\alpha$ as

$$\dot{a}_\alpha = -\kappa_a \hat{a}_\alpha \cdot \nabla_{a_\alpha} E(\tilde{\mu}, \tilde{\varphi}(a)) \quad (99)$$

where Eq. (93) is to be used for the Laplace-encoded energy. Then, after the gradient operation is completed, the organism's action is implemented explicitly as,

$$\dot{a}_\alpha = -\kappa_a \sum_{n=0}^{\infty} \frac{d\tilde{\varphi}_{\alpha[n]}}{da_\alpha} \Lambda_{z[n]}^{\alpha(0)} \varepsilon_{z[n]}^{\alpha(1)} \quad (100)$$

where $\varepsilon_{z[n]}^{\alpha(1)} = \varphi_{\alpha[n]} - g_{\alpha[n]}^{(1)}(x_{\alpha[n]}^{(1)}, v_{\alpha[n]}^{(1)}; \theta_\alpha^{(1)})$ is the prediction-error associated with learning of the sensory data on the dynamical order $n$ at the lowest cortical level and $\Lambda_{z[n]}^{\alpha(0)} = \Lambda_{z[n]}^{\alpha(0)}(\gamma_\alpha^{(0)})$ is the precision of the sensory noise.

In summary, we have seen how the principle of variational free energy minimisation can be applied to hidden states of generative models to provide a formal description of perceptual inference. Furthermore we have shown how action can be described as a process that modifies the environment, to change sensory data, and indirectly minimise VFE. Lastly we have shown how learning can be described when the same principles are applicable to the time invariant parameters that govern the dynamics of states.

## 9. Discussion

The FEP framework is an ambitious project, spanning a chain of reasoning from fundamental principles of biological self-maintenance essential for sustainable life, to a mechanistic brain theory that proposes to account for a startling range of properties of perception, cognition, action and learning. It draws conclusions about neurocognitive mechanisms from extremely general statistical considerations regarding the viability of organism's survival in unpredictable environments. Under certain assumptions – which we discuss in more detail below – it entails a hierarchical predictive processing model geared towards the inference and control of the hidden causes of sensory inputs, which both sheds new light on existing data about functional neuroanatomy and motivates a number of specific hypotheses regarding brain function in health and in disease. At the same time, the current status of much of the research under the rubric of the FEP does depend on, to different degrees, a variety of assumptions and approximations, both at the level of the overarching theory and with regard to the specific implementation (or 'process theory') the theory proposes. In this section, we discuss the consequences of some of more important of these assumptions and approximations, with respect to the framework and implementation described in the body of this paper.

A central assumption in this (representative) exposition of the FEP is that the brain utilises properties of Gaussian distributions in order to carry out probabilistic computation. Specifically, the Laplace approximation assumes a Gaussian functional form for the R-density and G-density which are encoded by sufficient statistics, see Sections 4 and 5. Additionally, it is assumed that the R-density is tightly peaked, i.e., the variance and covariance are small, see Section 4. At first glance this assumption may appear troublesome, because it suggests that organisms do not directly represent the uncertainty of environmental variables (hidden causes of sensory signals). However, this worry is misplaced and the organism in fact does represent a distribution over states. Representations of uncertainty are accommodated via precisions on the expectations of brain states that comprise the G-density, see Eq. (32). Intuitively this means that organisms encode uncertainties about their model of how hidden causes relate to each other and to sensory signals.

The main advantage of adopting Gaussian assumptions is that they vastly simplify the implementation of the FEP, and make it formally equivalent to the more widely known predictive coding framework (Clark, 2013; Elias, 1955; Friston & Kiebel, 2009c), see the Introduction. Furthermore, it can be argued this implementation is compatible with a plausible neuronal functional architecture in terms of message passing in cortical hierarchies (Friston, 2005). Specifically, inferred variables (hidden causes) can be represented in terms of neural firing rates; the details of generative models encoded as patterns of synaptic connectivity, and the process of VFE minimisation by the relaxation of neuronal dynamics (Bastos, Usrey, Adams, Mangun, Fries, & Friston, 2012). The concept of hierarchical generative models, see Section 8, also maps neatly onto the hierarchical structure of cortical networks, at least in the most frequently studied perceptual modalities like vision. Here, the simple idea is that top-down cortical signalling conveys predictions while bottom-up activity returns prediction errors (Bastos et al., 2012). However, it remains an open question whether representing the world in terms of Gaussian distributions is sufficient given the complexities of real-world sensorimotor interactions. For example, standard robotics architectures have long utilised practical strategies for representing more complex distributions (Thrun, Burgard, & Fox, 2005) including (for example) multimodal peaks (Otworowska, Kwisthout, & van Rooij, 2014). Other authors have proposed that brains engage in Bayesian sampling rather than the encoding of probability distributions, suggesting that sampling schemes parsimoniously explain classic cognitive reasoning errors (Knill & Pouget, 2004a). Whether these alternate schemes can be used to construct more versatile and behaviourally powerful implementations of the FEP, and whether they remain compatible with neuronally plausible process theories, remains to be seen.

The minimisation of VFE, for both inference and learning, is assumed to be implemented as a gradient descent scheme. While this has the major advantage of transforming difficult or infeasible inference problems into relatively straightforward optimisation problems, it is not clear whether the proposed gradient descent schemes always have good convergence properties. For example, the conditions under which gradient descent will become stuck in local minima, or fail to converge in an appropriate amount of time, are not well understood. Furthermore, parameters such as learning rate will be crucial for the timely inference of the dynamics of variables, as well as central to the dynamics of control, see Figs. 3 and 4. Parameters like these, which play important roles in the estimation of – but not specification of – the VFE, can be incorporated into process theories in many ways, with as yet no clear consensus (though see, for one proposal (Joffily & Coricelli, 2013)).

The implementation described in this paper supports inference in dynamical environments. This is based on the concept of generalised motions, whereby it is assumed that the brain infers not only the current value of environmental variables (e.g., position) but also their higher-order derivatives (i.e., velocity, acceleration, jerk, etc.). This requires that both that the relevant sensory noise is differentiable, and, that interactions between derivatives are linear (Friston et al., 2007). The extent to which these assumptions are justifiable remains unclear, as does the utility of encoding generalised motions in practical applications. It is likely, for example, that signal magnitudes after the second derivative will be small and carry considerable noise, thus practical usefulness of including higher order derivatives is unclear, although this may be justifiable in some cases (Balaji & Friston, 2011).

Under active inference, prediction errors are minimised by acting on the world to change sensory input, rather than by modifying predictions. Active inference therefore depends on the ability to make conditional predictions about the sensory consequences of actions. To achieve this the FEP assumes that agents have a model of the relationship between action and sensation, in the form of an inverse model, in addition to their generative model (Friston et al., 2016; Seth, 2014). In the general case the specification of an inverse model is non-trivial (Wolpert, 1997), thus at first glance this seems like a strong assumption. However, the FEP suggests generation of motor actions are driven through the fulfilment of proprioceptive predictions only, where relations between actions and (proprioceptive) sensations are assumed to be relatively simple such that minimisation of prediction error can be satisfied by simple reflex arcs (Friston, 2011; Friston et al., 2010). On this view, action only indirectly affects exteroceptive or interoceptive sensations, obviating the need for complicated inverse models like those described in the motor control literature (Friston, 2011; Wolpert, 1997). In the implementation of the FEP given in this paper there is no distinction between different types of sensory input.

In Section 7 we showed that behaviour is extremely sensitive to precisions. This is often presented as an advantage of the framework, allowing an agent to balance sensory inputs against internal predictions in an optimal and context sensitive manner, through precision weighting (which is associated with attention) (Clark, 2013). Supposedly the appropriate regulation of precision should also emerge as a consequence of the minimisation of free energy, see Section 8 for a description of this. But how the interplay between brain states and precisions will unfold in an active agent involved in a complex behaviour is far from clear.

Where do the priors come from? This is an intuitive way to put a key challenge for models involving Bayesian inference (Kass & Steffey, 1989). To some extent the FEP circumvents this problem via the concept of hierarchical models, which maps neatly onto the framework of 'empirical Bayes' (Casella & Berger, 2002). In this view, the hierarchical structure allows priors at one level to be supplied by posteriors at a higher level. Sensory data are assumed to reside only at the lowest level in the hierarchy, and the highest level is assumed to generate only spontaneous random fluctuations. While this is a powerful idea within formal frameworks, its practicality for guiding inference in active agents remains to be established.

These discussion points merely scratch the surface of the promises and pitfalls of the FEP formalism, a formalism which is rapidly advancing both in its theoretical aspects and in its various implementations and applications. Nevertheless, research directed towards addressing these issues should further clarify both the explanatory power and the practical utility of this increasingly influential framework. In this paper, we have tried to present an inclusive and self-contained presentation of the essential aspects of the free energy principle. We then unpacked a particular application of the free energy principle to continuous state space models in generalised coordinates of motion. This is a particularly important example because it provides a process theory that has a degree of biological plausibility, in terms of neuronal implementation. In doing so we hope to clarify the scientific contributions of the FEP, facilitate discussions of some of the core issues and assumptions underlying it, and motivate additional research to explore how far the grand ambitions of the FEP can be realised in scientific practice.

## Acknowledgements

C.S.K. is grateful to the hospitality of the School of Engineering and Informatics at the University of Sussex where he spent a sabbatical. Support is also gratefully acknowledged from the Dr. Mortimer and Theresa Sackler Foundation.

## Appendix A. Variational Bayes: ensemble learning

In this section, we present a general treatment of ensemble learning that does not assume a particular form for the marginal recognition density is (e.g., the Laplace approximation). The key approximation we focus on is the mean field approximation; in other words, the factorisation of the posterior into marginal approximate posteriors. Note that when the true posterior factorises in the same way as the approximate posterior, minimising variational free energy renders the approximate posterior exactly equal to the true posterior. This corresponds to exact Bayesian inference – as opposed to approximate Bayesian inference. This approach makes no assumptions about how the R-density is encoded in the brain's state; namely the Laplace approximation for the R-density is dispensed with. Technically the method we describe in Section 4 is known as 'Generalised Filtering' in Friston et al. (2010) while the one we present here is known as 'Variational Filtering' in Friston (2008b).

According to Eq. (7) the VFE is a functional of the R-density $q(\vartheta)$ where the variable $\vartheta$ denotes the environmental states collectively. The environmental sub-states $\vartheta_\alpha, \alpha = 1, 2, \ldots, N$, must vary on distinctive time-scale, $\tau_1 < \tau_2 < \cdots < \tau_N$, where $\tau_\alpha$ is associated with $\vartheta_\alpha$. Then, the sub-densities may be assumed to be statistically-independent to allow the *factorisation approximation* for $q(\vartheta)$ as

$$q(\vartheta) \equiv \prod_{i=1}^{N} q_\alpha(\vartheta_\alpha). \tag{A.1}$$

Eq. (4) gives rise to the individual normalisation condition:

$$\int d\vartheta\ q(\vartheta) = \prod_{\alpha=1}^{N} \int d\vartheta_\alpha\ q_\alpha(\vartheta_\alpha) = 1$$

which asserts that

$$\int d\vartheta_\alpha\ q_\alpha(\vartheta_\alpha) = 1. \tag{A.2}$$

When the factorisation approximation, Eq. (A.1) is substituted into Eq. (7), the VFE is written as

$$\begin{aligned} F &= \int \prod_\alpha [d\vartheta_\alpha q_\alpha(\vartheta_\alpha)] \left\{ E(\vartheta, \varphi) + \sum_\sigma \ln q_\sigma(\vartheta_\sigma) \right\} \\ &\equiv F[q(\vartheta); \varphi] \end{aligned}$$

where the last expression indicates explicitly that the VFE is to be treated as a *functional* of the R-density. We now optimise the VFE functional by taking the variation of $F$ with respect to a particular R-density $q_\beta(\vartheta_\beta)$. We treat the remainder of the ensemble densities as constant and use the normalisation constraint, Eq. (4), in the form

$$\lambda \left( \prod_\alpha \int d\vartheta_\alpha\ q_\alpha(\vartheta_\alpha) - 1 \right) = 0 \tag{A.3}$$

where $\lambda$ is a Lagrange multiplier.

A straightforward manipulation brings about

$$\begin{aligned} \delta_\beta F = \int d\vartheta_\beta \Bigg\{ & \int \prod_{\alpha \neq \beta} d\vartheta_\alpha q_\alpha(\vartheta_\alpha) \left( E(\vartheta, \varphi) + \sum_\sigma \ln q_\sigma(\vartheta_\sigma) \right) \\ & + 1 + \lambda \Bigg\} \delta q_\beta \end{aligned}$$

where $\delta_\beta$ represents a functional derivative with respect to $q_\beta(\vartheta_\beta)$. Next, by imposing $\delta_\beta F \equiv 0$ it follows that the integration must vanish identically for any change in $\delta q_\beta$,

$$\int \prod_{\alpha \neq \beta} d\vartheta_\alpha q_\alpha(\vartheta_\alpha) \left( E(\vartheta, \varphi) + \sum_\sigma \ln q_\sigma(\vartheta_\sigma) \right) + 1 + \lambda = 0$$

which is to be solved for $q_\beta(\vartheta_\beta)$. The result brings out the optimal density for the sub-state $\vartheta_\beta$ as

$$\begin{aligned} q^*_\beta = \exp \Bigg\{ & -(\lambda + 1) - \sum_{\sigma \neq \beta} \int \prod_{\alpha \neq \beta} d\vartheta_\alpha q_\alpha(\vartheta_\alpha) \ln q_\sigma(\vartheta_\sigma) \\ & - \mathcal{E}_\beta(\vartheta_\beta, \varphi) \Bigg\} \end{aligned} \tag{A.4}$$

where use has been made of the definition

$$\mathcal{E}_\beta(\vartheta_\beta, \varphi) \equiv \int \prod_{\alpha \neq \beta} d\vartheta_\alpha q_\alpha(\vartheta_\alpha) E(\vartheta, \varphi) \tag{A.5}$$

which is the partially-averaged energy (Friston et al., 2006, 2007). Here, it is worthwhile to note that the following relation holds

$$\int d\vartheta_\beta\ q_\beta(\vartheta_\beta) \mathcal{E}_\beta(\vartheta_\beta, \varphi) = \int d\vartheta\ q(\vartheta) E(\vartheta, \varphi),$$

which states that the expectation of the partially-averaged energy $\mathcal{E}_\beta(\vartheta_\beta, \varphi)$ under $q_\beta(\vartheta_\beta)$ is the average energy, i.e. the first term in Eq. (9). The undetermined Lagrange multiplier is now fixed by the

normalisation constraint, Eq. (A.2), which results in

$$\left[\int d\vartheta_\beta\, e^{-\mathcal{E}_\beta(\vartheta_\beta,\varphi)}\right]\exp\left\{-(\lambda+1)\right.$$
$$\left.-\sum_{\sigma\neq\beta}\int\prod_{\alpha\neq\beta}d\vartheta_\alpha q_\alpha(\vartheta_\alpha)\ln q_\sigma(\vartheta_\sigma)\right\}=1,$$

which is to be solved for $\lambda$. When the *determined* $\lambda$ is substituted back into Eq. (A.4), the resulting *ensemble-learned* R-density can be expressed formally as[3]

$$q^*_\beta(\vartheta_\beta)=\frac{1}{Z_\beta}e^{-\mathcal{E}_\beta(\vartheta_\beta,\varphi)} \tag{A.6}$$

where $Z_\beta$ has been defined to be

$$Z_\beta\equiv\int d\vartheta_\beta\, e^{-\mathcal{E}_\beta(\vartheta_\beta,\varphi)}. \tag{A.7}$$

The superscript $*$ appearing in $q^*_\beta$ indicates that it is the solution which optimises the VFE. The functional form of Eq. (A.6) is reminiscent of the equilibrium canonical ensemble in statistical physics in which the normalisation factor $Z_\beta$ is called the *partition function* of the subsystem $\{\vartheta_\beta\}$ (Huang, 1987).

Under the factorisation approximation, by substituting Eq. (A.6) into Eq. (A.1), the R-density becomes

$$q^*(\vartheta)=\frac{1}{Z_T}e^{-\mathcal{E}_T(\vartheta,\varphi)} \tag{A.8}$$

where

$$\mathcal{E}_T(\vartheta,\varphi)\equiv\sum_{\alpha=1}^{N}\mathcal{E}_\alpha(\vartheta_\alpha,\varphi)\quad\text{and}\quad Z_T\equiv\prod_{\alpha=1}^{N}Z_\alpha=\int d\vartheta e^{-\mathcal{E}_T(\vartheta,\varphi)}.$$

In Eq. (A.8) $Z_T$ may be called the 'total' partition function of the environmental states and $\mathcal{E}_T$ is the sum of the partially-averaged energies. Note that, as a consequence of the ensemble-learning, the optimising R-density approximates the posterior density $p(\vartheta|\varphi)$ (see Section 3 and below). In principle, the optimising R-density, Eq. (A.8), completes the ensemble-learning of the sensory data. However, it does not provide a functionally fixed-form for the optimal R-density. This is because the partially-averaged energy appearing on the RHS of Eq. (A.8) is a functional of the R-density itself (see Eq. (A.5)). One possible way to obtain a closed form of $q^*(\vartheta,\varphi)$ is to seek a *self-consistent solution*: One starts with an educated guess (an 'ansatz') for the optimal R-density to evaluate the partially-averaged energy, Eq. (A.5) and uses the outcome to update the R-density, Eq. (A.6). This iterative process is to be continued until a convergence reaches between estimation and evaluation of the R-densities.

We now exploit the optimal R-density, $q^*_\beta(\vartheta_\beta)$ given in Eq. (A.6). The partially averaged-energy appearing in $q^*_\beta$ can be manipulated as

$$\begin{aligned}\mathcal{E}_\beta(\vartheta_\beta,\varphi)&=\int\prod_{\alpha\neq\beta}d\vartheta_\alpha q_\alpha(\vartheta_\alpha)E(\vartheta,\varphi)\\&=-\sum_\sigma\int\prod_{\alpha\neq\beta}d\vartheta_\alpha q_\alpha(\vartheta_\alpha)\ln p(\vartheta_\sigma,\varphi_\sigma),\end{aligned} \tag{A.9}$$

where we have used the factorisation approximation for the G-density appearing in the energy $E=-\ln p(\vartheta,\varphi)$ as

$$p(\vartheta,\varphi)=\prod_\sigma p(\vartheta_\sigma,\varphi_\sigma)=\prod_\sigma p(\vartheta_\sigma|\varphi_\sigma)p(\varphi_\sigma). \tag{A.10}$$

Next, one can separate out the environmental sub-state $\vartheta_\beta$ among summation on the RHS of Eq. (A.9) to cast it into

$$\begin{aligned}\mathcal{E}_\beta(\vartheta_\beta,\varphi)&=-\ln p(\vartheta_\beta,\varphi_\beta)\\&\quad-\sum_{\sigma\neq\beta}\int\prod_{\alpha\neq\beta}d\vartheta_\alpha q_\alpha(\vartheta_\alpha)\ln p(\vartheta_\sigma,\varphi_\sigma).\end{aligned} \tag{A.11}$$

Then, it follows from Eq. (A.6) that

$$q^*_\beta(\vartheta_\beta)=\frac{e^{-\mathcal{E}_\beta(\vartheta_\beta,\varphi)}}{\int d\vartheta_\beta\, e^{-\mathcal{E}_\beta(\vartheta_\beta,\varphi)}}\to\frac{p(\vartheta_\beta,\varphi_\beta)}{\int d\vartheta_\beta p(\vartheta_\beta,\varphi_\beta)}=p(\vartheta_\beta|\varphi_\beta),$$

where the last step can be obtained by noticing the identity, $p(\vartheta_\beta,\varphi_\beta)=p(\vartheta_\beta|\varphi_\beta)p(\varphi_\beta)$, and $\int d\vartheta_\beta p(\vartheta_\beta,\varphi_\beta)=p(\varphi_\beta)$. Finally, the ensemble-learned R-density, Eq. (A.8), is given by

$$q^*(\vartheta)=\prod_\alpha q^*_\alpha(\vartheta_\alpha)=\prod_\alpha p(\vartheta_\alpha|\varphi_\alpha)=p(\vartheta|\varphi). \tag{A.12}$$

Note that we have expressed the true posterior as a product of marginal posteriors and thus we have exact Bayesian inference. In other situations, the above equation becomes an approximate equality and we have approximate Bayesian inference.

By substituting the optimal R-density, Eq. (A.12), into expression for VFE given in Eq. (7), we can also obtain the minimised VFE as

$$\begin{aligned}F^*&=\int d\vartheta\; q^*(\vartheta)\ln\frac{q^*(\vartheta)}{p(\vartheta,\varphi)}\\&=\int d\vartheta\; q^*(\vartheta)\ln\frac{p(\vartheta|\varphi)}{p(\vartheta|\varphi)p(\varphi)}\\&=-\ln p(\varphi)\int d\vartheta\; q^*(\vartheta)\\&=-\ln p(\varphi)\end{aligned} \tag{A.13}$$

where we have used Eq. (A.12) in moving to second line and the normalisation condition for $q^*(\vartheta)$ in the last step. Note that we have made it explicit that the sensory density $p(\varphi)$ is conditioned on the biological agent $m$. Thus, we have come to a conclusion that the minimum VFE provides a *tight bound* on surprisal.

In summary, the variation of the VFE functional with respect to the R-density (ensemble-learning) has allowed us to specify an optimal (ensemble-learned) R-density, $q^*(\vartheta,\varphi)$, selected among an ensemble of R-densities. The specified R-density is the brain's solution to statistical inference of the posterior density about the environmental states given sensory inputs. The minimum VFE, fixed in this way, is identical to the surprisal. To achieve this it was assumed that distinctive independent timescales characterise environmental sub-states (the factorisation approximation). The ensemble-learned R-density of each partitioned variable set $\vartheta_\beta$, $q^*_\beta(\vartheta_\beta)$, is specified by the corresponding partially-averaged energy (see Eq. (A.6)). The influence from other environmental variables $\{\vartheta_\sigma\}$ $(\sigma\neq\beta)$ occurs through their mean, i.e, their complicated interactions are averaged out in Eq. (A.5). In this sense, $\vartheta_\beta$ may be regarded as a 'mean-field' of the environmental states and the procedure described as a *mean-field approximation* (Friston, 2008b; Friston et al., 2008).

[3] Note that the minus sign arises in the exponent because we have defined the energy as Eq. (10) differently from other papers on the free energy principle. We have made this choice because our definition resembles the Boltzmann factor in the canonical ensemble in statistical physics.

## Appendix B. Dynamic Bayesian thermostat

```
% A Simple Bayesian Thermostat
% The free energy principle for action and perception: A mathematical review, Journal of Mathematical Psychology
% Christopher L. Buckley,Chang Sub Kim, Simon M. McGregor and Anil K. Seth
clear;
rng(6);
%simulation parameters
simTime=100; dt=0.005; time =0:dt:simTime;
N =length(time);
action =true;
%Generative Model Parameters
Td = 4; %desired temperature

%The time that action onsets
actionTime =simTime/4;

%initialise sensors
rho_0(1) =0;
rho_1(1)=0;

%sensory variances
Omega_z0 =0.1;
Omega_z1 =0.1;
%hidden state variances
Omega_w0 =.1;
Omega_w1 =.1;

%Params for generative process
T0 = 100; %temperature at x=0

%Initialise brain state variables
mu_0(1)=0;
mu_1(1)=0;
mu_2(1)=0;

%Sensory noise in the generative process
zgp_0 = randn(1,N)*.1;
zgp_1 = randn(1,N)*.1;

%Initialise the action variable
a(1) =0;

%Initialise generative process
x_dot(1) = a(1);
x(1) = 2;
T(1) = T0/(x(1)^2+1);
Tx(1)= -2*T0*x(1)*(x(1)^2+1)^-2;
T_dot(1) =  Tx(1)*(x_dot(1));

%Initialise sensory input
rho_0(1) =  T(1);
rho_1(1) =  T_dot(1);

%Initialise error terms
epsilon_z_0 = (rho_0(1)-mu_0(1));
epsilon_z_1 = (rho_1(1)-mu_1(1));

epsilon_w_0 = (mu_1(1)+mu_0(1)-Td);
epsilon_w_1 = (mu_2(1)+mu_1(1));

%Initialise Variational Energy
VFE(1) = 1/Omega_z0*epsilon_z_0^2/2 ...
    + 1/Omega_z1*epsilon_z_1^2/2 ...
    +1/Omega_w0*epsilon_w_0^2/2 ...
    +1/Omega_w1*epsilon_w_1^2/2 ...
    +1/2*log(Omega_w0*Omega_w1*Omega_z0*Omega_z1);


%Gradient descent learning parameters
k=.1; %for inference
ka=.01; %for learning

for i=2:N

    %The generative process (i.e. the real world)
```

```
    x_dot(i) = a(i-1);%action
    x(i) = x(i-1)+dt*(x_dot(i));
    T(i) = T0/(x(i)^2+1);
    Tx(i)= -2*T0*x(i)*(x(i)^2+1)^-2;
    T_dot(i) =  Tx(i)*(x_dot(i));

    rho_0(i) =  T(i) + zgp_0(i); %calclaute sensory input
    rho_1(i) =  T_dot(i) + zgp_1(i);

    %The generative model (i.e. the agents brain)
    epsilon_z_0 = (rho_0(i-1)-mu_0(i-1));%error terms
    epsilon_z_1 = (rho_1(i-1)-mu_1(i-1));

    epsilon_w_0 = (mu_1(i-1)+mu_0(i-1)-Td);
    epsilon_w_1 = (mu_2(i-1)+mu_1(i-1));

    VFE(i) = 1/Omega_z0*epsilon_z_0^2/2 ...
        +1/Omega_z1*epsilon_z_1^2/2 ...
        +1/Omega_w0*epsilon_w_0^2/2 ...
        +1/Omega_w1*epsilon_w_1^2/2 ...
        +1/2*log(Omega_w0*Omega_w1*Omega_z0*Omega_z1);

    mu_0(i)  = mu_0(i-1) ...
        +dt*(mu_1(i-1)-k*(-epsilon_z_0/Omega_z0 ...
        +epsilon_w_0/Omega_w0));

    mu_1(i)  = mu_1(i-1) +dt*(mu_2(i-1)- k*(-epsilon_z_1/Omega_z1 ...
        +epsilon_w_0/Omega_w0+epsilon_w_1/Omega_w1));

    mu_2(i)  = mu_2(i-1)...
        +dt*-k*(epsilon_w_1/Omega_w1);


    if(time(i) >25)
        a(i) = a(i-1) +dt*-ka*Tx(i)*epsilon_z_1/Omega_z1; %active inference
    else
        a(i) = 0;
    end
end
figure(1);clf;

subplot(5,1,1)
plot(time,T); hold on;
plot(time,x); hold on;
legend('T','x')

subplot(5,1,2)
plot(time,mu_0,'k');hold on;
plot(time,mu_1,'m');hold on;
plot(time,mu_2,'b');hold on;

legend('\mu','\mu','\mu');

subplot(5,1,3)
plot(time,rho_0,'k');hold on;
plot(time,rho_1,'m');hold on;

legend('\rho','\rho');

subplot(5,1,4)
plot(time, a,'k');
ylabel('a')

subplot(5,1,5)
plot(time, VFE,'k'); xlabel('time');hold on;
ylabel('VFE')
```

## References


Adams, R. A., Shipp, S., & Friston, K. J. (2013). Predictions not commands: active inference in the motor system. *Brain Structure and Function*, *218*(3), 611–643.

Adkins, A. (1983). *Equilibrium thermodynamics*. (3rd ed.). Cambridge: Cambridge University Press.

Balaji, B., & Friston, K. J. (2011). Bayesian state estimation using generalized coordinates. In *SPIE defense, security, and sensing*. International Society for Optics and Photonics, 80501Y–1.

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, *76*(4), 695–711.

Bogacz, R. (2017). A tutorial on the free-energy framework for modelling perception and learning. *Journal of Mathematical Psychology*, *76*(B), 198–211.

Bubic, A., von Cramon, D. Y., & Schubotz, R. I. (2010). Prediction, Cognition and the Brain. *Frontiers of Human Neuroscience*, *4*, 25–25.

Carhart-Harris, R. L., & Friston, K. J. (2010). The default-mode, ego-functions and free-energy: a neurobiological account of Freudian ideas. *Brain*, *133*(4), 1265–1283.

Casella, G., & Berger, R. (2002). *Statistical inference*. Pacific Grove: Duxbury.

Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science *Behavioral and Brain Sciences*, *36*(03), 181–204.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York: Wiley-Interscience.

Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The Helmholtz machine. *Neural Computation*, *7*, 889–904.

Elias, P. (1955). Predictive coding–I. *IRE Transactions on Information Theory*, *1*(1), 16–24.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, *1*, 1–47.

Friston, K. J. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society, Series B (Biological Sciences)*, *360*, 815–836.

Friston, K. J. (2008a). Hierarchical models in the brain. *PLoS Computational Biology*, *4*, e1000211–e1000211.

Friston, K. J. (2008b). Variational filtering. *NeuroImage*, *41*, 747–766.

Friston, K. J. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Science*, *13*, 293–301.

Friston, K. J. (2010a). Is the free-energy principle neurocentric? *Nature Reviews Neuroscience*, *11*(8), 605–605.

Friston, K. J. (2010b). Some free-energy puzzles resolved: response to thornton. *Trends in Cognitive Science*, *14*, 54.

Friston, K. J. (2010c). The free-energy principle: a unified brain theory? *Nature Reviews. Neuroscience*, *11*, 127–138.

Friston, K. J. (2011). What is optimal about motor control? *Neuron*, *72*(3), 488–498.

Friston, K. J. (2013). Life as we know it. *Journal of the Royal Society Interface*, *10*(86).

Friston, K. J., Adams, R., Perrinet, L., & Breakspear, M. (2012). Perceptions as hypotheses: Saccades as experiments. *Frontiers in Psychology*, *3*, 151.

Friston, K. J., Daunizeau, J., & Kiebel, S. J. (2009). Reinforcement learning or active inference? *PLoS One*, *4*, e6421–e6421.

Friston, K. J., Daunizeau, J., Kilner, J., & Kiebel, S. J. (2010). Action and behavior: A free-energy formulation. *Biological Cybernetics*, *102*(3), 227–260.

Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2016). Active inference: A process theory. *Neural Computation*.

Friston, K. J., & Kiebel, S. (2009a). Cortical circuits for perceptual inference. *Neural Networks*, *22*, 1093–1104.

Friston, K. J., & Kiebel, S. (2009b). Predictive Coding Under the Free-energy Principle. *Philosophical Transactions of the Royal Society, Series B (Biological Sciences)*, *364*, 1211–1221.

Friston, K. J., & Kiebel, S. (2009c). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society, Series B (Biological Sciences)*, *364*(1521), 1211–1221.

Friston, K. J., Kilner, J., & Harrison, L. (2006). A free Energy Principle for the Brain. *Journal de Physiologie (Paris)*, *100*, 70–87.

Friston, K. J., Mattout, J., Trujillo-Barreto, N., Ashburner, J., & Penny, W. (2007). Variational free energy and the Laplace approximation. *Neuroimage*, *34*, 220–234.

Friston, K. J., & Stephan, K. E. (2007). Free-energy and the brain. *Synthese*, *159*, 417–458.

Friston, K. J., Stephan, K., Li, B., & Daunizeau, J. (2010). Generalised filtering, *Mathematical Problems in Engineering*, *2010*.

Friston, K. J., Trujillo-Barreto, N., & Daunizeau, J. (2008). DEM: a variational treatment of dynamic systems. *Neuroimage*, *41*, 849–885.

Haykin, S., & Network, N. (2004). A comprehensive foundation. *Neural Networks*, *2*(2004).

Hinton, G. E. (2007). Learning multiple layers of representation. *Trends in Cognitive Sciences*, *11*(10), 428–434.

Hinton, G. E., & Zemel, R. (1994a). Autoencoders, minimum description length, and helmholtz free energy. In G. T. J. D. Cowan, & J. Alspector (Eds.), *Advances in neural information processing systems 6* (pp. 3–10). San Mateo, CA: Morgan Kaufmann.

Hinton, G. E., & Zemel, R. S. (1994b). Autoencoders, minimum description length, and helmholtz free energy. *Advances in Neural Information Processing Systems* 3–3.

Huang, K. (1987). *Statistical mechanics*. New York: John Wiley and Sons.

Joffily, M., & Coricelli, G. (2013). Emotional valence and the free-energy principle. *PLoS Computational Biology*, *9*(6), e1003094.

Kass, R. E., & Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *Journal of the American Statistical Association*, *407*, 717–726.

Knill, D. C., & Pouget, A. (2004a). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712–719.

Knill, D. C., & Pouget, A. (2004b). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neuroscience*, *27*, 712–719.

Neal, R., & Hinton, G. E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. Jordan (Ed.), *Learning in Graphical Models* (pp. 355–368). Cambridge, MA: MIT Press.

Otworowska, M., Kwisthout, J., & van Rooij, I. (2014). Counter-factual mathematics of counterfactual predictive models. *Frontiers in Psychology*, *5*.

Pezzulo, G., Rigoli, F., & Friston, K. J. (2015). Active inference, homeostatic regulation and adaptive behavioural control. *Progress in Neurobiology*, *134*, 17–35.

Pio-Lopez, L., Nizard, A., Friston, K. J., & Pezzulo, G. (2016). Active inference and robot control: a case study. *Journal of the Royal Society Interface*, *13*(122), 20160616.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a Functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*, 79–87.

Seth, A. K. (2014). A predictive processing theory of sensorimotor contingencies: explaining the puzzle of perceptual presence and its absence in synesthesia. *Cognitive Neuroscience*, *5*(2), 97–118.

Seth, A. K. (2015). *The cybernetic Bayesian brain: From interoceptive inference to sensorimotor contingencies*. Open MIND. Frankfurt a. M: MIND Group.

Thrun, S., Burgard, W., & Fox, D. (2005). *Probabilistic robotics*. MIT press.

Von Helmholtz, H., & Southall, J. P. C. (2005). *Treatise on physiological optics. Vol. 3*. Courier Corporation.

Wolpert, D. M. (1997). Computational approaches to motor control. *Trends in Cognitive Sciences*, *1*(6), 209–216.

Zeki, S., & Shipp, S. (1988). The functional logic of cortical connections. *Nature*, *335*, 311–317.

Zemel, R., & Hinton, G. E. (1995). Learning population codes by minimizing description length. *Neural Computation*, *7*, 549–564.

Zwanzig, R. (2001). *Nonequilibrium statistical mechanics*. Oxford: Oxford University Press.