City Research Online

City, University of London Institutional Repository

Citation: Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P. & Pezzulo, G. (2017). Active Inference: A Process Theory. Neural Computation, 29(1), pp. 1-49. doi: 10.1162/neco_a_00912

This is the published version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/16683/

Link to published version: https://doi.org/10.1162/neco_a_00912

Copyright: City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

Reuse: Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

City Research Online: http://openaccess.city.ac.uk/ publications@city.ac.uk

ARTICLE **Communicated by Ollie Hulme**

Active Inference: A Process Theory

Karl Friston
k.friston@ucl.ac.uk
Wellcome Trust Centre for Neuroimaging, UCL, London WC1N 3BG, U.K.

Thomas FitzGerald
thomas.fitzgerald@ucl.ac.uk
Wellcome Trust Centre for Neuroimaging, UCL, London WC1N 3BG, U.K., and Max Planck–UCL Centre for Computational Psychiatry and Ageing Research, London WC1B 5BE, U.K.

Francesco Rigoli
f.rigoli@ucl.ac.uk
Wellcome Trust Centre for Neuroimaging, UCL, London WC1N 3BG, U.K.

Philipp Schwartenbeck
philipp.schwartenbeck.12@alumni.ucl.ac.uk
Wellcome Trust Centre for Neuroimaging, UCL, London WC1N 3BG, U.K.; Max Planck–UCL Centre for Computational Psychiatry and Ageing Research, London, WC1B 5BE, U.K.; Centre for Neurocognitive Research, University of Salzburg, 5020 Salzburg, Austria; and Neuroscience Institute, Christian-Doppler-Klinik, Paracelsus Medical University Salzburg, A-5020 Salzburg, Austria

Giovanni Pezzulo
giovanni.pezzulo@gmail.com
Institute of Cognitive Sciences and Technologies, National Research Council, 00185 Rome, Italy

This article describes a process theory based on active inference and belief propagation. Starting from the premise that all neuronal processing (and action selection) can be explained by maximizing Bayesian model evidence—or minimizing variational free energy—we ask whether neuronal responses can be described as a gradient descent on variational free energy. Using a standard (Markov decision process) generative model, we derive the neuronal dynamics implicit in this description and reproduce a remarkable range of well-characterized neuronal phenomena. These include repetition suppression, mismatch negativity, violation responses, place-cell activity, phase precession, theta sequences, theta-gamma coupling, evidence accumulation, race-to-bound dynamics, and transfer of dopamine responses. Furthermore, the (approximately Bayes' optimal)

Neural Computation **29, 1–49** (2017)
doi:10.1162/NECO_a_00912
© 2016 Massachusetts Institute of Technology

behavior prescribed by these dynamics has a degree of face validity, providing a formal explanation for reward seeking, context learning, and epistemic foraging. Technically, the fact that a gradient descent appears to be a valid description of neuronal activity means that variational free energy is a Lyapunov function for neuronal dynamics, which therefore conform to Hamilton's principle of least action.

1 Introduction

There has been a paradigm shift in the cognitive neurosciences over the past decade toward the Bayesian brain and predictive coding (Ballard, Hinton, & Sejnowski, 1983; Rao & Ballard, 1999; Knill & Pouget, 2004; Yuille & Kersten, 2006; De Bruin & Michael, 2015). At the same time, there has been a resurgence of enactivism; emphasizing the embodied aspect of perception (O'Regan & Noë, 2001; Friston, Mattout, & Kilner, 2011; Ballard, Kit, Rothkopf, & Sullivan, 2013; Clark, 2013; Seth, 2013; Barrett & Simmons, 2015; Pezzulo, Rigoli, & Friston, 2015). Even in consciousness research and philosophy, related ideas are finding traction (Clark, 2013; Hohwy, 2013, 2014). Many of these developments have informed (and have been informed by) a variational principle of least free energy (Friston, Kilner, & Harrison, 2006; Friston, 2012), namely, active (Bayesian) inference.

However, the enthusiasm for Bayesian theories of brain function is accompanied by an understandable skepticism about their usefulness, particularly in furnishing testable process theories (Bowers & Davis, 2012). Indeed, one could argue that many current normative theories fail to provide detailed and physiologically plausible predictions about the processes that might implement them. And when they do, their connection with a normative or variational principle is often obscure. In this work, we show that process theories can be derived in a relatively straightforward way from variational principles. The level of detail we consider is fairly coarse; however, the explanatory scope of the resulting process theory is remarkable—and provides an integrative (and simplifying) perspective on many phenomena that are studied in systems neuroscience. The aim of this article is to describe the basic ideas and illustrate the emergent processes using simulations of neuronal responses. We anticipate revisiting some issues in depth: in particular, a companion paper focuses on learning and the emergence of habits as a natural consequence of observing one's own behavior (Friston et al., 2016).

This article has three sections. The first describes active inference, combining earlier formulations of planning as inference (Botvinick & Toussaint, 2012; Friston et al., 2014) with Bayesian model averaging (FitzGerald, Dolan, & Friston, 2014) and learning (FitzGerald, Dolan, & Friston, 2015). Importantly, action (i.e., policy selection), perception (i.e., state estimation), and learning (i.e., reinforcement learning) all minimize the same quantity: variational free energy. This refinement of previous schemes considers an explicit representation of past and future states, conditioned on competing

policies. This leads to Bayesian belief updates that are informed by beliefs about the future (prediction) and context learning that is informed by beliefs about the past (postdiction). Technically, these updates implement a form of Bayesian smoothing, with explicit representations of states over time, which include future (i.e., counterfactual) states. Furthermore, the implicit variational updates have some biological plausibility in the sense that they eschew neuronally implausible computations. For example, expectations about future states are sigmoid functions of linear mixtures of the preceding and subsequent states. An alternative parameterization, which did not appeal to explicit representations over time, would require recursive matrix multiplication, for which no neuronally plausible implementation has been proposed. Under this belief parameterization, learning is mediated by classical associative (synaptic) plasticity. The remaining sections use simulations of foraging in a radial maze to illustrate some key aspects of inference and learning, respectively.

The inference section describes the behavioral and neuronal correlates of belief updating during inference or planning, with an emphasis on electrophysiological correlates and the encoding of precision by dopamine. It illustrates a number of phenomena that are ubiquitous in empirical studies. These include repetition suppression (de Gardelle, Waszczuk, Egner, & Summerfield, 2013), violation and omission responses (Bendixen, SanMiguel, & Schroger, 2012), and neuronal responses that are characteristic of the hippocampus, namely, place cell activity (Moser, Rowland, & Moser, 2015), theta-gamma coupling, theta sequences and phase precession (Burgess, Barry, & O'Keefe, 2007; Lisman & Redish, 2009). We also touch on dynamics seen in parietal and prefrontal cortex, such as evidence accumulation and race-to-bound or threshold (Huk & Shadlen, 2005, Gold & Shadlen, 2007; Hunt et al., 2012; Solway & Botvinick, 2012; de Lafuente, Jazayeri, & Shadlen, 2015; FitzGerald, Moran, Friston, & Dolan, 2015; Latimer, Yates, Meister, Huk, & Pillow, 2015).

The final section considers context learning and illustrates the transfer of dopamine responses to conditioned stimuli, as agents become familiar with experimental contingencies (Fiorillo, Tobler, & Schultz, 2003). We conclude with a brief demonstration of epistemic foraging. The aim of these simulations is to illustrate how all of the phenomena emerge from a single imperative (to minimize free energy) and how they contextualize each other.

2 Active Inference and Learning

This section provides a brief overview of active inference that builds on our previous treatments of Markov decision processes. Specifically, it introduces a parameterization of posterior beliefs about the past and future that makes state estimation (i.e., belief updating) biologically plausible. (A slightly fuller version of this material can be found in Friston et al., 2016.) Active inference is based on the premise that everything minimizes

variational free energy (Friston, 2013). This leads to some surprisingly simple update rules for action, perception, policy selection, learning, and the encoding of uncertainty or its complement, precision. Although some of the intervening formalism looks complicated, what comes out at the end are update rules that will be familiar to many readers (e.g., integrate-and-fire dynamics with sigmoid activation functions and plasticity with associative and decay terms). This means that the underlying theory can be tied to neuronal processes in a fairly straightforward way. Furthermore, the formalism accommodates a number of established normative approaches, thereby providing an integrative framework.

In principle, the scheme described in this section can be applied to any paradigm or choice behavior. Indeed, earlier versions have been used to model waiting games (Friston et al., 2013), the urn task and evidence accumulation (FitzGerald, Schwartenbeck, Moutoussis, Dolan, & Friston, 2015), trust games from behavioral economics (Moutoussis, Trujillo-Barreto, El-Deredy, Dolan, & Friston, 2014; Schwartenbeck, FitzGerald, Mathys, Dolan, Kronbichler et al., 2015), addictive behavior (Schwartenbeck, FitzGerald, Mathys, Dolan, Wurst et al., 2015), two-step maze tasks (Friston, Rigoli et al., 2015), and engineering benchmarks such as the mountain car problem (Friston, Adams, & Montague, 2012). It has also been used in the setting of computational fMRI (Schwartenbeck, FitzGerald, Mathys, Dolan, & Friston, 2015).

In brief, active inference separates the problems of optimizing action and perception by assuming that action fulfills predictions based on inferred states of the world. Optimal predictions are therefore based on (sensory) evidence that is evaluated using a generative model of (observed) outcomes. This allows one to frame behavior as fulfilling optimistic predictions, where the optimism is prescribed by prior preferences or goals (Friston et al., 2014). In other words, action realizes predictions that are biased toward preferred outcomes. More specifically, the generative model entails beliefs about future states and policies, where policies that lead to preferred outcomes are more likely. This enables action to realize the next (proximal) outcome predicted by the policy that leads to (distal) goals. This behavior emerges when action and inference maximize the evidence or marginal likelihood of the model generating predictions. Note that action is prescribed by predictions of the next outcome and is not itself part of the inference process. This separation of action and perceptual inference or state estimation can be understood by associating action with peripheral reflexes in the motor system that fulfill top-down motor predictions about how we move (Feldman, 2009; Adams, Shipp, & Friston, 2013).

The models considered in this article include states of the world in the past and the future. This enables agents to select policies that will maximize model evidence in the future by minimizing expected free energy. Furthermore, it enables learning about contingencies based on state transitions that are inferred retrospectively. We will see that this leads to a Bayes-optimal arbitration between epistemic (explorative) and pragmatic

(exploitative) behavior that is formally related to several established ideas (e.g., the infomax principle, Bayesian surprise, the value of information, artificial curiosity, and expected utility theory).

We start by describing the generative model on which predictions and actions are based. We then describe how action is specified by beliefs about states of the world under different policies. The section concludes by considering the optimization of these beliefs through Bayesian belief updating and implicit neuronal processing.

The parameters of categorical distributions over discrete states $s \in \{0, 1\}$ are denoted by column vectors of expectations $\mathbf{s} \in [0, 1]$, where the $\sim$ notation denotes sequences of variables over time, for example, $\tilde{s} = (s_1, \ldots, s_T)$. The entropy of a probability distribution $P(s) = \Pr(S = s)$ is denoted by $H(S) = H[P(s)] = E_P[-\ln P(s)]$, while the relative entropy or Kullback-Leibler (KL) divergence is denoted by $D[Q(s)||P(s)] = E_Q[\ln Q(s) - \ln P(s)]$. Inner and outer products are indicated by $A \cdot B = A^T B$, and $A \otimes B = AB^T$, respectively. We use a hat notation $\widehat{s} = \ln s$ to denote (natural) logarithms. Finally, $P(o|s) = Cat(\mathbf{A})$ implies $\Pr(o = i|s = j) = Cat(\mathbf{A}_{ij})$. Definitions of the variables referred to are in Table 1.

Definition. *Active inference rests on the tuple* $(O, P, Q, R, S, T, \Upsilon)$*:*

- *A finite set of outcomes* O
- *A finite set of control states or actions* Υ
- *A finite set of hidden states* S
- *A finite set of time-sensitive policies* T
- *A generative process* $R(\tilde{o}, \tilde{s}, \tilde{u})$ *that generates probabilistic outcomes* $o \in O$ *from (hidden) states* $s \in S$ *and action* $u \in \Upsilon$
- *A generative model* $P(\tilde{o}, \tilde{s}, \pi, \eta)$ *with parameters* η*, over outcomes, states, and policies* $\pi \in T$*, where* $\pi \in \{0, \ldots, K\}$ *returns a sequence of actions* $u_t = \pi(t)$
- *An approximate posterior* $Q(\tilde{s}, \pi, \eta) = Q(s_0|\pi) \ldots Q(s_T|\pi)Q(\pi)Q(\eta)$ *over states, policies and parameters with expectations* $(\boldsymbol{s}_0^\pi, \ldots, \boldsymbol{s}_T^\pi, \boldsymbol{\pi}, \boldsymbol{\eta})$

Remark. The generative process describes transitions among states in the world that generate observed outcomes. These states are referred to as hidden because they cannot be observed directly. Their transitions depend on action, which depends on posterior beliefs about the next state. In turn, these beliefs are formed using a generative model of how observations are generated. The generative model describes what the agent believes about the world, where beliefs about hidden states and policies are encoded by expectations. Note the distinction between actions (that are part of the generative process in the world) and policies (that are part of the generative model of an agent). This distinction allows actions to be specified by beliefs about policies, effectively converting an optimal control problem into an optimal inference problem (Attias, 2003; Botvinick & Toussaint, 2012).

Table 1: Glossary of Expressions.

Expression		Description
o_τ	$\in \{0,1\}$	Outcomes, their posterior expectations and logarithms
$\mathbf{o}_\tau$	$\in [0,1]$	
$\widehat{\mathbf{o}}_\tau$	$= \ln \mathbf{o}_\tau$	
$\tilde{o}$	$= (o_1, \ldots, o_t)$	Sequences of outcomes until the current time point
s_τ	$\in \{0,1\}$	Hidden states and their posterior expectations and logarithms, conditioned on each policy
$\mathbf{s}_\tau^\pi$	$\in [0,1]$	
$\widehat{\mathbf{s}}_\tau^\pi$	$= \ln \mathbf{s}_\tau^\pi$	
$\tilde{s}$	$= (s_1, \ldots, s_T)$	Sequences of hidden states until the end of the current trial
π	$= (\pi_1, \ldots, \pi_K) : \pi \in \{0,1\}$	Policies specifying action sequences, their posterior expectations, and logarithms
$\boldsymbol{\pi}$	$= (\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_K) : \boldsymbol{\pi} \in [0,1]$	
$\widehat{\pi}$	$= \ln \boldsymbol{\pi}$	
u	$= \pi(t)$	Action or control variables
$\mathbf{A}$	$\in [0,1]$	Likelihood matrix mapping from hidden states to outcomes and its expected logarithm
$\widehat{\mathbf{A}}$	$= \psi(\mathbf{a}) - \psi(\mathbf{a}_0)$	
$\mathbf{B}_\tau^\pi$	$= \mathbf{B}(\mathrm{u} = \pi(\tau)) \in [0,1]$	Transition probability for hidden states under each action prescribed by a policy at a particular time and its logarithm
$\widehat{\mathbf{B}}_\tau^\pi$	$= \ln \mathbf{B}_\tau^\pi$	
$\mathbf{D}$	$\in [0,1]$	Prior expectation of the hidden state at the beginning of each trial
$\mathbf{U}_\tau$	$= \ln P(o_\tau) \Leftrightarrow P(o_\tau) = \sigma(\mathbf{U}_\tau)$	Logarithm of prior preference over outcomes or utility
$\mathbf{F} : \mathbf{F}_\pi$	$= F(\pi) = \sum_\tau F(\pi, \tau) \in \mathbb{R}$	Variational free energy for each policy
$\mathbf{G} : \mathbf{G}_\pi$	$= G(\pi) = \sum_\tau G(\pi, \tau) \in \mathbb{R}$	Expected free energy for each policy
$\mathbf{s}_t$	$= \sum_\pi \boldsymbol{\pi}_\pi \cdot \mathbf{s}_t^\pi$	Bayesian model average of hidden states over policies
$\mathbf{H}$	$= -diag(\check{\mathbf{A}} \cdot \widehat{\mathbf{A}})$	The vector encoding the entropy or ambiguity over outcomes for each hidden state
$\widehat{\mathbf{A}}$	$= E_Q[\ln \mathbf{A}] = \psi(\mathbf{a}) - \psi(\mathbf{a}_0)$	Expected outcome probabilities for each hidden states and their expected logarithms
$\check{\mathbf{A}}$	$= E_Q[\mathbf{A}_{ij}] = \mathbf{a} \times \mathbf{a}_0^{-1}$	
$\mathbf{a}_{0ij}$	$= \sum_i \mathbf{a}_{ij}$	

2.1 The Generative Model. The generative model is at the heart of (active) Bayesian inference. In simple terms, the generative model is just a way of formalizing beliefs about the way outcomes are caused. Usually a generative model is specified in terms of the likelihood of each outcome, given their causes and the prior probability of those causes. Inference then corresponds to inverting the model, which means computing the posterior probability of (unknown or hidden) causes, given observed outcomes. In approximate Bayesian inference, this entails optimizing an approximate posterior so that it minimizes variational free energy. In other words, the difficult problem of exact Bayesian inference is converted into an easy optimization problem, where the approximate posterior minimizes a (variational free energy) functional of observed outcomes, under a given generative model. We will see later that when variational free energy is minimized, it approximates the (negative) log evidence or marginal likelihood of the outcomes, namely, the probability of the outcomes under the generative model.

In our case, the generative model can be parameterized in a general way as follows, where the model parameters are $\eta = \{a, b, d, \beta\}$:

$$
\begin{aligned}
P\left(\tilde{o}, \tilde{s}, \pi, \eta\right) &= P(\pi)P(\eta)\prod_{t=1}^{T} P(o_t|s_t)P(s_t|s_{t-1}, \pi) \\
P\left(o_t|s_t\right) &= Cat(\mathbf{A}) \\
P\left(s_{t+1}|s_t, \pi\right) &= Cat(\mathbf{B}(u = \pi(t))) \\
P\left(s_1|s_0\right) &= Cat(\mathbf{D}) \\
P(\pi) &= \sigma(-\gamma \cdot \mathbf{G}(\pi)) \\
P(\mathbf{A}) &= Dir(a) \\
P(\mathbf{B}) &= Dir(b) \\
P(\mathbf{D}) &= Dir(d) \\
P(\gamma) &= \Gamma(1, \beta).
\end{aligned}
\tag{2.1}
$$

An approximate posterior over hidden states and parameters $x = (\tilde{s}, \pi, \eta)$ can be expressed in terms of its sufficient statistics, which are expectations $\mathbf{x} = (\mathbf{s}_0^\pi, \ldots, \mathbf{s}_T^\pi, \boldsymbol{\pi}, \boldsymbol{\eta})$ and $\boldsymbol{\eta} = (\mathbf{a}, \mathbf{b}, \mathbf{d}, \boldsymbol{\beta})$:

$$
\begin{aligned}
Q(x) &= Q(s_1|\pi)\ldots Q(s_T|\pi)Q(\pi)Q(\mathbf{A})Q(\mathbf{B})Q(\mathbf{D})Q(\gamma) \\
Q\left(s_t|\pi\right) &= Cat(\mathbf{s}_t^\pi) \\
Q(\pi) &= Cat(\boldsymbol{\pi}) \\
Q(\mathbf{A}) &= Dir(\mathbf{a}) \\
Q(\mathbf{B}) &= Dir(\mathbf{b})
\end{aligned}
\tag{2.2}
$$

$$Q(\mathbf{D}) = Dir(\mathbf{d})$$
$$Q(\gamma) = \Gamma(1, \boldsymbol{\beta})$$

In this model, observations depend only on the current state, while state transitions depend on a policy or sequence of actions. This sequential policy is sampled from a Gibbs distribution or softmax function of expected free energy, with inverse temperature or precision γ. Here $\mathbf{G}(\pi)$ is the free energy expected under each policy (see below). The role of the model parameters will be unpacked later, when we consider model inversion.

Note that the policy is a random variable that has to be inferred. In other words, the agent entertains competing hypotheses or models of its behavior in terms of policies. This contrasts with standard formulations in which a single state-action policy returns an action as a function of each state $u = \pi(s)$, as opposed to time, $u = \pi(t)$. Furthermore, the approximate posterior is parameterized in terms of expected states under each policy. In other words, we assume that the agent keeps a separate record of expected states—in the past and future—for each allowable policy.

The predictions that guide action are based on a Bayesian model average of these policy-specific states. This means that expectations about policies (and their precision) also have to be optimized. All the posterior probabilities over model parameters, including the initial state, are Dirichlet distributions (FitzGerald, Dolan et al., 2015). The sufficient statistics of these distributions are concentration parameters that can be regarded as the number of occurrences encountered in the past. In what follows, we first describe how actions are selected, given beliefs about the hidden state of the world and the policies currently being pursued. We then turn to the more difficult problem of optimizing the beliefs on which action is based.

2.2 Behavior Action and Reflexes. We associate action with reflexes that minimize the expected difference between the outcomes predicted at the next time step and the outcome following an action. Mathematically, this can be expressed in terms of (outcome) prediction errors as follows:

$$\begin{aligned} u_t &= \min_u E_Q[D[P(o_{t+1}|s_{t+1})||R(o_{t+1}|s_t, u)]] \\ &= \min_u \mathbf{o}_{t+1} \cdot \varepsilon^u_{t+1} \end{aligned}$$

$$\begin{aligned} \varepsilon^u_{t+1} &= \widehat{\mathbf{o}}_{t+1} - \widehat{\mathbf{o}}^u_{t+1} \\ \mathbf{o}_{t+1} &= \mathbf{A}\mathbf{s}_{t+1} \\ \mathbf{o}^u_{t+1} &= \mathbf{A}\mathbf{B}(u)\mathbf{s}_t \\ \mathbf{s}_t &= \sum_\pi \boldsymbol{\pi}_\pi \cdot \mathbf{s}^\pi_t . \end{aligned} \tag{2.3}$$

This specification of action is considered reflexive by analogy to motor reflexes that minimize the discrepancy between proprioceptive signals (i.e., primary afferents) and descending motor commands or predictions. Heuristically, action realizes expected outcomes by minimizing the expected outcome prediction error (Adams et al., 2013). Expectations about the next outcome therefore enslave behavior. If we regard competing policies as models of behavior, the predicted outcome is formally equivalent to a Bayesian model average of outcomes, under posterior beliefs about policies (last equality in equation 2.3).

For simplicity, we assume the agent has learned the consequences of action. More complete schemes would incorporate learning the consequences of action by analogy with learning transitions among hidden states.

Having specified action selection in terms of expected outcomes, we now consider how these expectations are optimized. In active inference, there are no stimulus-response links found in conventional formulations: choices or actions are separated from inference in the same way that peripheral reflexes are separated from processing in the central nervous system. This means all behavior rests on optimizing beliefs or expectations about the next state of the world. These expectations furnish predictions of the next outcome that action simply fulfills. Following action, a new observation becomes available, and the perception-action cycle starts again.

2.3 Free Energy and Expected Free Energy. In active inference, all the heavy lifting is done by minimizing free energy with respect to expectations about hidden states, policies, and parameters. Variational free energy can be expressed as a function of these posterior beliefs in a number of ways:

$$
\begin{aligned}
Q(x) &= \arg\min_{Q(x)} F \\
&\approx P(x|\tilde{o})
\end{aligned}
$$

$$
\begin{aligned}
F &= E_Q[\ln Q(x) - \ln P(x, \tilde{o})] \\
&= E_Q[\ln Q(x) - \ln P(x|\tilde{o}) - \ln P(\tilde{o})] \\
&= E_Q[\ln Q(x) - \ln P(\tilde{o}|x) - \ln P(x)] \\
&= \underbrace{D[Q(x)||P(x|\tilde{o})]}_{\textit{relative entropy}} - \underbrace{\ln P(\tilde{o})}_{\textit{log evidence}} \\
&= \underbrace{D[Q(x)||P(x)]}_{\textit{complexity}} - \underbrace{E_Q[\ln P(\tilde{o}|x)]}_{\textit{accuracy}},
\end{aligned}
\tag{2.4}
$$

where $\tilde{o} = (o_1, \ldots, o_t)$ denotes observed outcomes up until the current time.

Because KL divergences cannot be less than zero, the penultimate equality in equation 2.4 means that free energy is minimized when the approximate posterior becomes the true posterior. At this point, the free energy becomes the negative log evidence for the generative model (Beal, 2003). This means that minimizing free energy is equivalent to maximizing model evidence, which is equivalent to minimizing the complexity of accurate explanations for observed outcomes (the last equality in equation 2.4).

With this equivalence in mind, we now turn to the prior beliefs about policies that shape posterior beliefs—and the Bayesian model averaging that determines action. Minimizing free energy with respect to expectations of hidden states and parameters ensures that they encode posterior beliefs, given observed outcomes. However, beliefs about policies rest on outcomes in the future, because these beliefs determine action and action determines subsequent outcomes. This means that policies should, a priori, minimize the free energy of beliefs about the future. Equation 2.1 expresses this formally by making the log probability of a policy proportional to the expected free energy if that policy was pursued. The expected free energy of a policy follows from equation 2.4 (Friston, Rigoli et al., 2015):

$$G(\pi) = \sum_{\tau} G(\pi, \tau),$$

$$\begin{aligned} G(\pi, \tau) &= E_{\tilde{Q}}[\ln Q(s_\tau|\pi) - \ln P(s_\tau, o_\tau|\tilde{o}, \pi)] \\ &= E_{\tilde{Q}}[\ln Q(s_\tau|\pi) - \ln P(s_\tau|o_\tau, \tilde{o}, \pi) - \ln P(o_\tau)], \end{aligned} \tag{2.5}$$

$$\begin{aligned} &\approx \underbrace{E_{\tilde{Q}}[\ln Q(s_\tau|\pi) - \ln Q(s_\tau|o_\tau, \pi)]}_{(-ve)\ mutual\ information} - \underbrace{E_{\tilde{Q}}[\ln P(o_\tau)]}_{expected\ log\ evidence} \\ &= \underbrace{E_{\tilde{Q}}[\ln Q(o_\tau|\pi) - \ln Q(o_\tau|s_\tau, \pi)]}_{(-ve)\ epistemic\ value} - \underbrace{E_{\tilde{Q}}[\ln P(o_\tau)]}_{extrinsic\ value} \\ &= \underbrace{D[Q(o_\tau|\pi)||P(o_\tau)]}_{expected\ cost} + \underbrace{E_{\tilde{Q}}[H[P(o_\tau|s_\tau)]]}_{expected\ ambiguity}, \end{aligned}$$

where $\tilde{Q} = Q(o_\tau, s_\tau|\pi) = P(o_\tau|s_\tau)Q(s_\tau|\pi) \approx P(o_\tau, s_\tau|\tilde{o}, \pi)$ and $Q(o_\tau|s_\tau, \pi) = P(o_\tau|s_\tau)$.

In the expected free energy, the relative entropy becomes the mutual information between hidden states and the outcomes they cause (and vice versa), while the log evidence becomes the log evidence expected under predicted outcomes. By associating the log-prior over outcomes with utility

or prior preferences, $U(o_\tau) = \ln P(o_\tau)$, the expected free energy can also be expressed in terms of epistemic and extrinsic value (the penultimate equality in equation 2.5). This means that extrinsic value is the (log) evidence for a generative model expected under a particular policy. In other words, because our model of the world entails prior preferences, any outcomes that provide evidence for our model (and implicit preferences) have pragmatic or extrinsic value. In practice, utilities are defined only to within an additive constant, such that the prior probability of an outcome is a softmax function of utility: $P(o_\tau) = \sigma(U(o_\tau))$. This means prior preferences depend only on utility differences and are inherently context sensitive (Rigoli, Friston, & Dolan, 2016).

Epistemic value is the expected information gain (i.e., mutual information) afforded to hidden states by future outcomes and vice-versa.[1] We will see below that epistemic value can be thought of as driving curiosity and novelty-seeking behavior, by which we resolve uncertainty and ignorance. A final rearrangement shows that complexity becomes expected cost—namely, the KL divergence between the posterior predictions and prior preferences—while accuracy becomes the accuracy expected under predicted outcomes (i.e., negative ambiguity). This last equality in equation 2.5 shows how expected free energy can be evaluated relatively easily; it is just the divergence between the predicted and preferred outcomes plus the ambiguity (i.e., entropy) expected under predicted states.

In summary, expected free energy is defined in relation to prior beliefs about future outcomes. These define the expected cost or complexity and complete the generative model. It is these priors that lend inference and action a purposeful or goal-directed aspect because they represent preferences or goals. These preferences define agents in terms of characteristic states they expect to occupy and, through action, tend to frequent.

There are several interpretations of expected free energy that appeal to and contextualize/established constructs. For example, maximizing epistemic value is equivalent to maximizing (expected) Bayesian surprise (Schmidhuber, 1991; Itti & Baldi, 2009), where Bayesian surprise is the KL divergence between posterior and prior beliefs. This can also be interpreted in terms of the principle of maximum mutual information or minimum redundancy (Barlow, 1961; Linsker, 1990; Olshausen & Field, 1996; Laughlin, 2001). This is because epistemic value is the mutual information between hidden states and observations: $I(S_\tau, O_\tau|\pi) = H[Q(s_\tau|\pi)] - H[Q(s_\tau|o_\tau, \pi)]$. In other words, it reports the reduction in uncertainty about hidden states afforded by observations. Because the KL divergence or information gain

[1] Note that the negative mutual information (which is never positive) is not an expected KL divergence (which is never negative). This is because the expectation is under the joint distribution over outcomes and hidden states. Furthermore, epistemic value is never positive, which means that the best one can do is to have an epistemic value of zero; in other words, a preferred outcome is expected with probability one.

cannot be less than zero, it disappears when the (predictive) posterior beliefs are not informed by new observations. Heuristically, this means that epistemic policies will search out observations that resolve uncertainty about the state of the world (e.g., foraging to locate a prey or fixating on informative part of a face, such as the eyes or mouth). However, when there is no posterior uncertainty and the agent is confident about the state of the world, there can be no further information gain, and epistemic value will be the same for all policies, allowing preferences to dictate action.

Conversely, with no preferences (i.e., all outcomes are deemed equally likely), the most likely policies maximize uncertainty over outcomes (i.e., keeping all options open), in accord with the maximum entropy principle (Jaynes, 1957), while minimizing the entropy of outcomes, given the state. Heuristically, this means agents will try to avoid uninformative (low entropy) outcomes (e.g., closing one's eyes) while avoiding states that produce ambiguous (high-entropy) outcomes (e.g., a noisy discotheque) (Schwartenbeck, Fitzgerald, Dolan, & Friston, 2013). This resolution of uncertainty is closely related to satisfying artificial curiosity (Schmidhuber, 1991; Still & Precup, 2012) and speaks to the value of information (Howard, 1966). It is also referred to as intrinsic value (see Barto, Singh, & Chentanez, 2004) for a discussion of intrinsically motivated learning). In one sense, epistemic value can be regarded as the drive for novelty-seeking behavior (Wittmann, Daw, Seymour, & Dolan, 2008; Krebs, Schott, Schütze, & Düzel, 2009; Schwartenbeck et al., 2013), in which we anticipate uncertainty that can be resolved (e.g., opening a birthday present: see also Barto, Mirolli, & Baldassarre, 2013).

The expected complexity or cost is exactly the same quantity minimized in risk-sensitive or KL control (Klyubin, Polani, & Nehaniv, 2005; van den Broek, Wiegerinck, & Kappen, 2010), and underpins related variational formulations of bounded rationality based on complexity costs (Braun, Ortega, Theodorou, & Schaal, 2011; Ortega & Braun, 2013). In other words, minimizing expected complexity renders behavior risk-sensitive, while maximizing expected accuracy induces ambiguity-sensitive behavior. In short, expected free energy covers nearly all measures that have been proposed to explain adaptive behavior, and has each as a special case.

Although the expressions above may appear complicated, expected free energy can be expressed in a simple form in terms of the generative model:

$$
\begin{aligned}
G(\pi, \tau) &= \underbrace{D[Q(o_\tau|\pi)||P(o_\tau)]}_{\textit{expected cost}} + \underbrace{E_{\tilde{Q}}[H[P(o_\tau|s_\tau)]]}_{\textit{expected ambiguity}} \\
&= \underbrace{\mathbf{o}_\tau^\pi \cdot (\widehat{\mathbf{o}}_\tau^\pi - \mathbf{U}_\tau)}_{\textit{risk}} + \underbrace{\mathbf{s}_\tau^\pi \cdot \mathbf{H}}_{\textit{ambiguity}},
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{o}_\tau^\pi &= \breve{\mathbf{A}} \cdot \mathbf{s}_\tau^\pi \\
\widehat{\mathbf{o}}_\tau^\pi &= \ln \mathbf{o}_\tau^\pi \\
\mathbf{U}_\tau &= U(o_\tau) = \ln P(o_\tau) \\
\mathbf{H} &= -diag(\breve{\mathbf{A}} \cdot \widehat{\mathbf{A}}),
\end{aligned}
\tag{2.6}
$$

$$
\begin{aligned}
\widehat{\mathbf{A}} &= E_Q[\ln \mathbf{A}] = \psi(\mathbf{a}) - \psi(\mathbf{a}_0) \\
\breve{\mathbf{A}} &= E_Q[\mathbf{A}_{ij}] = \mathbf{a} \times \mathbf{a}_0^{-1} : \mathbf{a}_{0ij} = \sum_i \mathbf{a}_{ij}.
\end{aligned}
$$

The two terms in the first expression for expected free energy represent risk- and ambiguity-sensitive contributions, respectively, where utility is a vector of preferences over outcomes. This decomposition lends a formal meaning to risk and ambiguity: risk is the relative entropy or uncertainty about outcomes, in relation to preferences, while ambiguity is the uncertainty about outcomes given the state of the world. This is largely consistent with the use of risk and ambiguity in economics (Kahneman & Tversky, 1979; Zak, 2004; Knutson & Bossaerts, 2007; Preuschoff, Quartz, & Bossaerts, 2008), where ambiguity reflects uncertainty about the context (e.g., which lottery is currently in play).

In summary, the above formalism suggests that expected free energy can be carved in two complementary ways. First, it can be decomposed into a mixture of epistemic and extrinsic value, promoting explorative, novelty seeking, and exploitative, reward-seeking behavior, respectively (Friston, Rigoli et al., 2015). Equivalently, minimizing expected free energy can be formulated as minimizing a mixture of expected cost or risk and ambiguity. This completes our description of free energy. We now turn to belief updating that is based on minimizing free energy under the generative model we have described.

2.4 Belief Updating and Belief Propagation. Belief updating mediates inference and learning, where *inference* means optimizing expectations about hidden states (policies and precision), while *learning* refers to optimizing model parameters. This optimization entails finding the sufficient statistics of posterior beliefs that minimize variational free energy. These solutions are (see appendix A):

$$
\left.
\begin{aligned}
\mathbf{s}_\tau^\pi &= \sigma(\widehat{\mathbf{A}} \cdot \mathrm{o}_\tau + \widehat{\mathbf{B}}_{\tau-1}^\pi \mathbf{s}_{\tau-1}^\pi + \widehat{\mathbf{B}}_\tau^\pi \cdot \mathbf{s}_{\tau+1}^\pi) \\
\boldsymbol{\pi} &= \sigma(-\mathbf{F} - \boldsymbol{\gamma} \cdot \mathbf{G}) \\
\boldsymbol{\beta} &= \beta + (\boldsymbol{\pi} - \boldsymbol{\pi}_0) \cdot \mathbf{G}
\end{aligned}
\right\} \text{Inference},
$$

$$\left.\begin{array}{ll} \widehat{\mathbf{A}} = \psi(\mathbf{a}) - \psi(\mathbf{a}_0) & \mathbf{a} = a + \sum_\tau o_\tau \otimes \mathbf{s}_\tau \\ \widehat{\mathbf{B}} = \psi(\mathbf{b}) - \psi(\mathbf{b}_0) & \mathbf{b}(u) = b(u) + \sum_{\pi(\tau)=u} \boldsymbol{\pi}_\pi \cdot \mathbf{s}_\tau^\pi \otimes \mathbf{s}_{\tau-1}^\pi \\ \widehat{\mathbf{D}} = \psi(\mathbf{d}) - \psi(\mathbf{d}_0) & \mathbf{d} = d + \mathbf{s}_1 \end{array}\right\} \text{Learning.} \tag{2.7}$$

For notational simplicity, we have used $\widehat{\mathbf{B}}_\tau^\pi = \widehat{\mathbf{B}}(\pi(\tau))$, $\widehat{\mathbf{D}} = \widehat{\mathbf{B}}_0^\pi \mathbf{s}_0^\pi$, $\boldsymbol{\gamma} = 1/\beta$, and $\boldsymbol{\pi}_0 = \sigma(-\boldsymbol{\gamma} \cdot \mathbf{G})$. Usually one would iterate the equalities in equation 2.7 until convergence. However, we can also obtain the solution in a robust and biologically more plausible fashion using a gradient descent on free energy (see appendixes B and C):

$$\begin{aligned} \dot{\widehat{\mathbf{s}}}_\tau^\pi &= \partial_{\widehat{\mathbf{s}}} \mathbf{s}_\tau^\pi \cdot \varepsilon_\tau^\pi \\ \mathbf{s}_\tau^\pi &= \sigma(\widehat{\mathbf{s}}_\tau^\pi) \\ \dot{\boldsymbol{\beta}} &= \boldsymbol{\gamma}^2 \varepsilon^\gamma \end{aligned}$$

$$\begin{aligned} \varepsilon_\tau^\pi &= (\widehat{\mathbf{A}} \cdot o_\tau + \widehat{\mathbf{B}}_{\tau-1}^\pi \mathbf{s}_{\tau-1}^\pi + \widehat{\mathbf{B}}_\tau^\pi \cdot \mathbf{s}_{\tau+1}^\pi) - \widehat{\mathbf{s}}_\tau^\pi \\ \varepsilon^\gamma &= (\beta - \boldsymbol{\beta}) + (\boldsymbol{\pi} - \boldsymbol{\pi}_0) \cdot \mathbf{G}. \end{aligned} \tag{2.8}$$

This converts the discrete updates above into dynamics for inference that minimize state and precision prediction errors $\varepsilon_\tau^\pi = -\partial_{\mathbf{s}} F$ and $\varepsilon^\gamma = \partial_{\boldsymbol{\gamma}} F$, where these prediction errors are free energy gradients.

Solving these equations produces posterior expectations that minimize free energy to provide Bayesian estimates of hidden variables. This means that expectations change over several timescales: a fast timescale that updates posterior beliefs about hidden states after each observation (to minimize free energy over peristimulus time) and a slower timescale that updates posterior beliefs as new observations are sampled (to mediate evidence accumulation over observations): (see also Penny, Zeidman, & Burgess, 2013). Finally, at the end of each sequence of observations (i.e., trial of observation epochs), the expected (concentration) parameters are updated to mediate learning over trials (FitzGerald, Dolan, & Friston, 2015). These updates are remarkably simple and have intuitive (neurobiological) interpretations:

2.5 Belief Updating and Neuronal Dynamics. Updating hidden states corresponds to state estimation, under each policy. Because beliefs about the current state are informed by expectations about past and future states, this scheme has the form of a Bayesian smoother that combines (empirical)

prior expectations about hidden states with the likelihood of the current observation (Kass & Steffey, 1989). However, the scheme does not use conventional forward and backward sweeps (Penny et al., 2013; Pezzulo, Rigoli, & Chersi, 2013), because all future and past states are encoded explicitly. In other words, representations always refer to the same hidden state at the same time in relation to the start of the trial, not in relation to the current time. This may seem counterintuitive, but this form of spatiotemporal (place and time) encoding finesses belief updating considerably and, as we will see later, has a degree of plausibility in relation to empirical findings.

The formulation in equation 2.8 is important because it describes dynamics that can be related to neuronal processes. In other words, we move a variational Bayesian scheme toward a process theory that can predict neuronal responses during state estimation and action selection (e.g., Solway & Botvinick, 2012). This process theory associates the expected probability of a state with the probability of a neuron (or population) firing and the logarithm of this probability with postsynaptic membrane potential. This fits comfortably with theoretical proposals and empirical work on the accumulation of evidence (Kira, Yang, & Shadlen, 2015) and the neuronal encoding of probabilities (Deneve, 2008), while rendering the softmax function a (sigmoid) activation function that converts membrane potentials to firing rates. The postsynaptic depolarization caused by afferent input can now be interpreted in terms of free energy gradients (i.e., state prediction errors) that are linear mixtures of firing rates in other neurons (or populations). These prediction errors play the role of postsynaptic currents, which drive changes in membrane potential and subsequent firing rates. This means that when there are no prediction errors, postsynaptic currents disappear and depolarizations (and firing rates) converge to the free energy minimum. Note that the above expressions imply a self-inhibition because prediction errors decrease when log expectations increase.

Technically, replacing the explicit solutions, equation 2.7, with a gradient ascent, equation 2.8, is exactly the same generalization of variational Bayes found in variational Laplace (Friston et al., 2007), namely, a generalized coordinate descent. This is nice, because it means one can think about process theories for variational treatments of Markov decision processes as formally similar to equivalent process theories for state-space models, such as predictive coding (Rao & Ballard, 1999; Bastos et al., 2012). There are some finer, neurobiologically plausible details of the dynamics of expectations about hidden states that we will consider elsewhere. For example, the modulation by $\partial_{\tilde{\mathbf{s}}} \mathbf{s}_\tau^\pi$ implies activity-dependent (e.g., NMDA-R dependent) depolarization that enforces an excitation-inhibition balance (see appendix B).

2.6 Action Selection, Precision, and Dopamine. The policy updates are just a softmax function of their log probability, which has two components:

the free energy based on past outcomes and the expected free energy based on preferences about future outcomes. In other words, prior beliefs about policies in the generative model are supplemented or informed by the free energy based on outcomes. Policy selection also entails the optimization of expected uncertainty or precision. This is expressed above in terms of the temperature (inverse precision), which encodes posterior beliefs about precision: $\boldsymbol{\beta} = 1/\gamma$.

Interestingly, the updates for temperature are determined by the difference between the expected free energy under posterior and prior beliefs about policies, that is, the prediction error based on expected free energy. This endorses the notion of reward prediction errors as an update signal that the brain might use, in the sense that if posterior beliefs based on current observations reduce the expected free energy, relative to prior beliefs, then precision will increase (FitzGerald, Dolan et al., 2015). This can be related to dopamine discharges that have been interpreted in terms of changes in expected reward (Schultz & Dickinson, 2000; Fiorillo et al., 2003) and marginal utility (Stauffer, Lak, & Schultz, 2014). We have previously considered the intimate (monotonic) relationship between expected precision and expected utility in this context (see Friston et al., 2014, for a fuller discussion). The role of the neuromodulator dopamine in encoding precision is also consistent with its multiplicative effect in equation 2.7, to nuance the selection among competing policies (Fiorillo et al., 2003; Frank, Scheres, & Sherman, 2007; Humphries, Wood, & Gurney, 2009; Humphries, Khamassi, & Gurney, 2012; Solway & Botvinick, 2012; Mannella & Baldassarre, 2015). We will return to this later.

2.7 Learning and Associative Plasticity. Finally, the updates for the parameters bear a marked resemblance to classical Hebbian plasticity (Abbott & Nelson, 2000). The parameter updates for state transitions comprise two terms: an associative term that is a digamma function of the accumulated coincidence of past (postsynaptic) and current (presynaptic) states (or observations under hidden causes) and a decay term that reduces each connection as the total afferent connectivity increases. The associative and decay terms are strictly increasing but saturating functions of the concentration parameters. Note that the updates for the connectivity parameters accumulate coincidences over time, because parameters are time invariant (in contrast to states that change over time). Furthermore, the parameters encoding state transitions have associative terms that are modulated by policy expectations.

In addition to learning contingencies through the parameters of the transition matrices, the vectors encoding beliefs about initial states accumulate evidence by simply counting the number of times an initial state occurs. In other words, if a particular state is encountered frequently, it will come to dominate posterior expectations. This mediates context learning in terms of the initial state. In practice, the parameters are updated at the end of each

trial or sequence of observations. This ensures that learning benefits from postdicted states, after ambiguity has been resolved through epistemic behavior. For example, the agent can learn about the initial state even if the initial cues were completely ambiguous.

Collectively, the updates above constitute a formal description of perception and learning. In what follows, we will associate electrophysiological responses with depolarization (i.e., state prediction error) driving changes in neuronal activity. For simplicity, we recover this from the rate of change of the associated expectation (see equation 2.8).

2.8 Summary. By assuming a generic (Markovian) form for the generative model, it is fairly easy to derive Bayesian updates that clarify the relationships among perception, policy selection, precision, and action and how these quantities shape beliefs about hidden states of the world and subsequent behavior. In brief, the agent first infers the hidden states under each model or policy that it entertains. It then evaluates the evidence for each policy based on prior beliefs or preferences about future outcomes. Having optimized the precision or confidence in beliefs about policies, they are used to form a Bayesian model average of the next outcome, which is realized through action. The anatomy of the implicit message passing is not inconsistent with functional anatomy in the brain (see Friston et al., 2014, and Figures 1 and 2). Figure 1 reproduces the (solutions to) belief updating and assigns them to plausible brain structures. Figure 2 rehearses the belief updating in terms of the implicit computations. This functional anatomy rests on reciprocal message passing among expected policies (e.g., in the striatum) and expected precision (e.g., in the substantia nigra). Expectations about policies depend on expected outcomes and states of the world for example, in the prefrontal cortex (Mushiake, Saito, Sakamoto, Itoyama, & Tanji, 2006) and hippocampus (Pezzulo, van der Meer, Lansink, & Pennartz, 2014). Crucially, this scheme entails reciprocal interactions between the prefrontal cortex and basal ganglia (Botvinick & An, 2009), in particular, selection of expected motor outcomes by the basal ganglia (Mannella & Baldassarre, 2015).

In this scheme, the scope and depth of the policy search is exhaustive, in the sense that all policies entertained by an agent are encoded explicitly and all hidden states over the sequence of actions entailed by policy are continuously updated. This may sound like an overcomplete representation of policies; however, this sort of architecture is implicit in salience maps in the brain (Santangelo, 2015; Zelinsky & Bisley, 2015). This is because a salience map represents the value (e.g., epistemic value or Bayesian surprise) of all possible actions (e.g., saccadic eye movements), from which the best action is selected: see Mirza, Adams, Mathys, and Friston (2016) for a simulation of saccadic searches and scene construction using the current scheme. In the simulations below, each policy comprises two actions, whereas in Mirza et al. (2016), we used just a single action: each policy specified where to

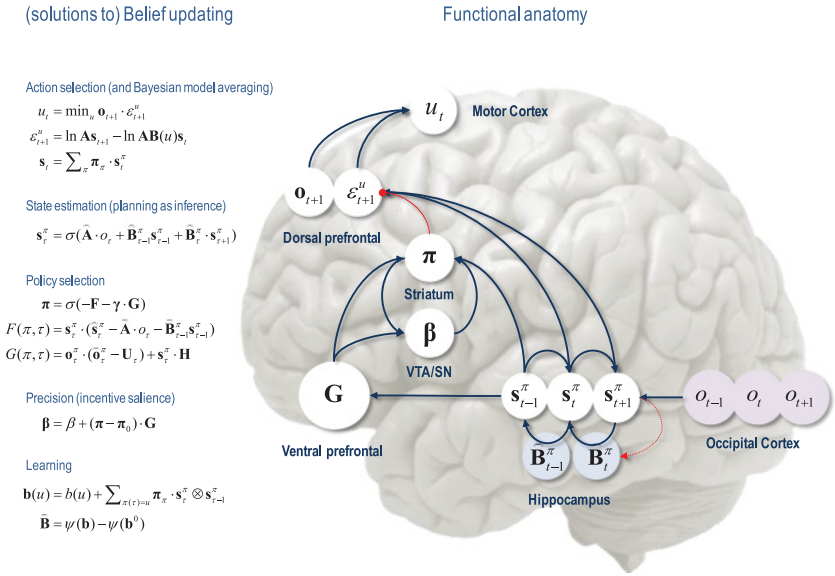


Figure 1: Schematic overview of belief updates for active inference under discrete Markovian models. The left panel lists the solutions in the main text, associating various updates with action, perception, policy selection, precision, and learning. It assigns the variables (sufficient statistics or expectations) that are updated to various brain areas. This attribution should not be taken too seriously but serves to illustrate a rough functional anatomy, implied by the form of the belief updates. In this simplified scheme, we have assigned observed outcomes to visual representations in the occipital cortex and state estimation to the hippocampal formation. The evaluation of policies, in terms of their (expected) free energy, has been placed in the ventral prefrontal cortex. Expectations about policies per se and the precision of these beliefs have been attributed to striatal and ventral tegmental areas to indicate a putative role for dopamine in encoding precision. Finally, beliefs about policies are used to create Bayesian model averages over future states that are fulfilled by action. The blue arrows denote message passing, and the solid red line indicates a modulatory weighting that implements Bayesian model averaging. The broken red lines indicate the updates for parameters or connectivity (in blue circles) that depend on expectations about hidden states. This scheme is described heuristically in Figure 2. See the appendixes and Table 1 for an explanation of the equations and variables.

look next. In the next section, we use equation 2.8 to simulate neuronal responses and show that many familiar electrophysiological phenomena emerge.

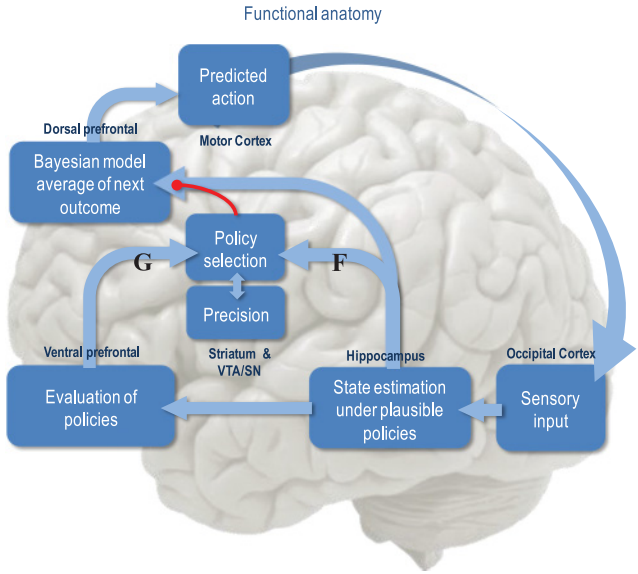


Figure 2: Summary of belief updates in terms of functional anatomy. Sensory evidence is accumulated to optimize expectations about the current state of the world, which are constrained by expectations of past and future states. This corresponds to state estimation under each policy the agent entertains. The quality of each policy is evaluated in the ventral prefrontal cortex, possibly in combination with ventral striatum (van der Meer, Kurth-Nelson, & Redish, 2012), in terms of its expected free energy. This evaluation and the ensuing policy selection rest on expectations about future states. Note that the explicit encoding of future states lends this scheme the ability to plan and explore. After the free energy of each policy has been evaluated, it is used to predict the subsequent hidden state through Bayesian model averaging (over policies). This enables an action to be selected that is most likely to realize the predicted outcome. Once an action has been selected, it generates a new observation, and the cycle begins again.

3 Simulations of Inference

This section considers inference using simulations of foraging in a maze. Its aim is to illustrate belief updating as a process theory for commonly observed electrophysiological and behavioral responses. We first describe the simulation setup and then establish the construct validity of the scheme in terms of simulated electrophysiological responses. The simulations involve searching for rewards in a T-maze. This T-maze contains primary rewards such as food and cues that are not rewarding per se but disclose

the location of rewards. The basic structure of this problem can be translated to any number of scenarios (e.g., saccadic eye movements to visual targets). The simulations use the same setup as in Friston et al. (2015) and is as simple as possible while illustrating some fairly complicated behaviors. This example can also be interpreted in terms of responses elicited in reinforcement learning paradigms by unconditioned (US) and conditioned (CS) stimuli. Strictly speaking, our paradigm is instrumental, and the cue is a discriminative stimulus; however, we retain the Pavlovian nomenclature when relating precision updates to dopaminergic discharges.

3.1 The Setup. An agent, such as a rat, starts in the center of a T-maze, where either the right or left arms are baited with a reward (US). The lower arm contains a discriminative cue (CS) that tells the animal whether the reward is in the upper right or left arm. Crucially, the agent can make only two moves. Furthermore, the agent cannot leave the baited arms after they are entered. This means that the optimal behavior is to first go to the lower arm to find where the reward is located and then retrieve the reward at the cued location.

In terms of a Markov decision process, there are four control states that correspond to visiting, or sampling, the four locations (the center and three arms). For simplicity, we assume that each control state takes the agent to the associated location, as opposed to moving in a particular direction from the current location. This is analogous to place-based navigation strategies mediated by the hippocampus (e.g., Moser, Kropff, & Moser, 2008). There are eight hidden states (four locations by two contexts) and seven possible outcomes. The outcomes correspond to being in the center of the maze plus the (two) outcomes at each of the (three) arms that are determined by the context (the right or left arm is more rewarding).

Having specified the state-space, it is now necessary to specify the $(\mathbf{A},\mathbf{B})$ matrices encoding contingencies. These are shown in Figure 3, where the $\mathbf{A}$ matrix maps from hidden states to outcomes, delivering an ambiguous cue at the center (first) location and a definitive cue at the lower (fourth) location. The remaining locations provide a reward with probability $p = 98\%$ depending on the context. The $\mathbf{B}(u)$ matrices encode action-specific transitions, with the exception of the baited (second and third) locations, which are absorbing hidden states that the agent cannot leave.

In general treatments, we would consider learning contingencies by updating the prior concentration parameters (a,b) of the transition matrices, but we will assume the agent knows (i.e., has very precise beliefs about) the contingencies. This corresponds to making the prior concentration parameters very large. Conversely, we will use small values of d to enable context learning. Preferences in the vector $\mathbf{U}_\tau = \ln P(o_\tau) \leq 0$ encode the utility of outcomes. Here, the (relative) utilities of a rewarding and unrewarding outcome were 3 and -3, respectively (and zero otherwise). This means, that the agent expects to be rewarded $\exp(3) \approx 20$ times more than experiencing

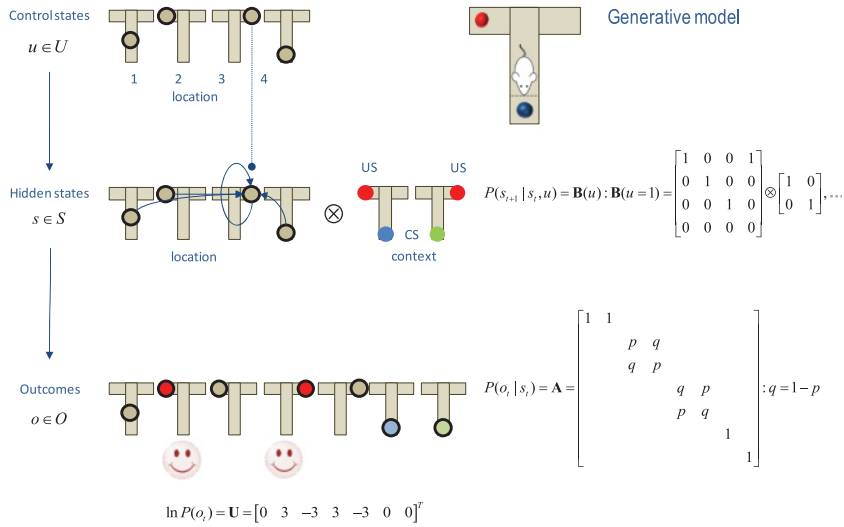


Figure 3: The generative model used to simulate foraging in a three-arm maze (insert on the upper right). This model contains four control states that encode movement to one of four locations (three arms and a central location). These control the transition probabilities among hidden states that have a tensor product form with two factors: the first is place (one of four locations), and the second is one of two contexts. These correspond to the location of rewarding (red) outcomes and the associated cues (blue or green circles). Each of the eight hidden states generates an observable outcome, where the first two hidden states generate the same outcome that just tells the agent that it is at the center. Some selected transitions are shown as arrows, indicating that control states attract the agent to different locations, where outcomes are sampled. The equations define the generative model in terms of its parameters (**A**,**B**), which encode mappings from hidden states to outcomes and state transitions, respectively. The lower vector corresponds to prior preferences—namely, the agent expects to find a reward. Here, ⊗ denotes a Kronecker tensor product.

a neutral outcome. Note that utility is always relative because the probabilities over outcomes must sum to one. As noted above, this means the prior preferences are a softmax function of utility $P(o_\tau) = \sigma(\mathbf{U}_\tau)$. Associating utility with log probabilities is important because it endows utility with the same measure as information, namely, nats (i.e., units of information or entropy based on natural logarithms). This highlights the close connection between value and information (Howard, 1966).

Having specified the state-space and contingencies, one can solve the belief updating equations in equation 2.8 to simulate behavior. Prior beliefs about the initial state were initialized to $d = 8$ for the central location for

each context and zero otherwise. These concentration parameters can be regarded as the number of times each state, transition, or policy has been encountered in previous trials.

Figure 4 summarizes simulated behavioral and physiological responses over 32 successive trials using a format that will be used in subsequent figures. Each trial comprises two actions following an initial outcome. The first panel shows the initial states on each trial (as colored circles) and subsequent policy selection (in image format) over the 10 policies considered. These correspond to staying at the center and then moving to each of the four possible locations (policies 1–4; ending in the center, left, right, or lower arm), moving to the left or right arm and staying there (policies 5 and 6), or moving to the lower arm and then to each of the four locations (policies 7–10). The second panel reports the final outcomes (encoded by colored circles) and performance. Performance is reported in terms of preferred (i.e., utility of) outcomes, summed over time (black bars) and reaction times (cyan dots). Note that because utilities are log probabilities, they are always negative, and the best outcome is zero. The reaction times here are based on the actual processing time in the simulations (using the Matlab *tic-toc* facility) and are shown after normalization to a mean of zero and standard deviation of one.

In this example, the first couple of trials alternate between the two contexts with rewards on the right and left. After this, the context (indicated by the cue) remained unchanged. For the first 20 trials, the agent selects epistemic policies—first going to the lower arm and then proceeding to the reward location (i.e., left for policy 8 and right for policy 9). After this, the agent becomes increasingly confident about the context and starts to visit the reward location directly. The differences in performance—between these epistemic and pragmatic behaviors—are revealed in the second panel as a decrease in reaction time and an increase in the average utility. This increase follows because the average is over trials and the agent spends two trials enjoying its preferred outcome when seeking reward directly, as opposed to one trial when behaving epistemically. Note that on trial 12, the agent received an unexpected (null) outcome that induces a degree of posterior uncertainty about which policy it was pursuing, indicated by the red dot. This is seen as a nontrivial posterior probability for three policies: the correct (context-sensitive) epistemic policy and the best alternatives that involve staying in the lower arm or returning to the center. This loss of certainty is accompanied by a low-utility outcome and a suppression of phasic dopamine responses reporting the confidence in behavior.

The marked reduction in reaction times, with the emergence of pragmatic behavior, reflects the fact that the estimation of hidden states under policies that have a small posterior probability is omitted. This is a common device in Bayesian model averaging, where the evidence for implausible models that fall outside Occam's window are not evaluated. Here, we removed policies with a relative posterior probability of 1/128 or less.

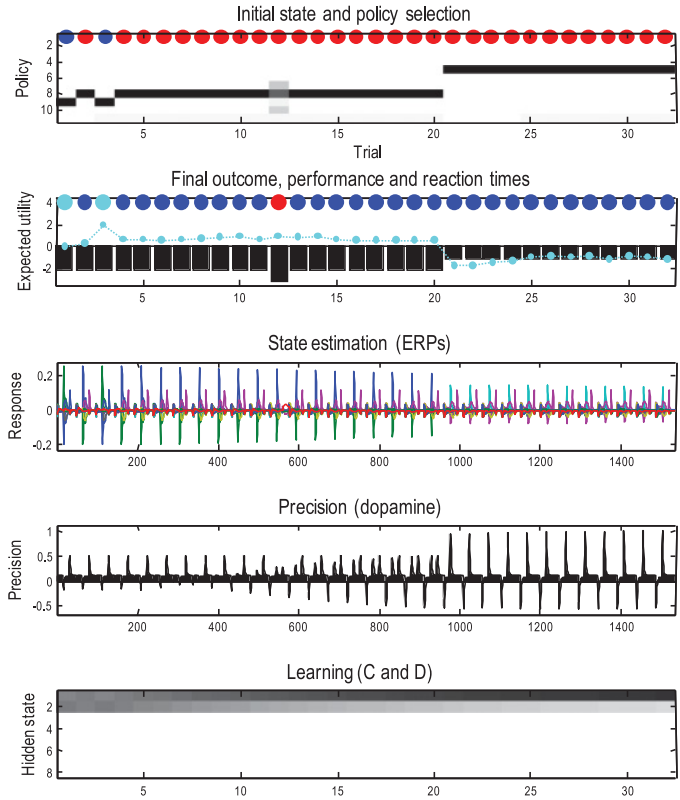


Figure 4: Simulated responses over 32 trials. The behavioral and (simulated) physiological responses during successive trials. The first panel shows, for each trial, the initial state (as blue and red circles indicating the context) and the selected policy (in image format) over the policies considered. The policies selected in the first two trials correspond to epistemic policies (8 and 9), which involve examining the cue in the lower arm and then going to the left or right arm to secure the reward (depending on the context). After the agent becomes sufficiently confident that the context does not change (after trial 21), it indulges in pragmatic behavior, accessing the reward directly. The second panel reports the final outcomes (encoded by colored circles: cyan and blue for rewarding outcomes in the left and right arms) and performance measures in terms of preferred outcomes, summed over time (black bars) and reaction times (cyan dots). The third panel shows a succession of simulated event-related potentials following each outcome. The different colors correspond to expectations about different hidden states. These are the rate of change of neuronal activity, encoding the expected probability of hidden states. The fourth panel shows phasic fluctuations in posterior precision that can be interpreted in terms of dopamine responses. The final panel shows the accumulated posterior beliefs about the initial state, where black denotes a posterior expectation of one and white a posterior expectation of zero.

Neurobiologically, this would entail a selective suspension of belief updating, mediated by neuromodulatory projections (omitted from Figure 1). When the agent becomes increasingly confident about the context, the precision of competing policies increases, enabling it to focus on a smaller number and select one quickly and efficiently.

The third panel shows a succession of simulated event-related potentials following each outcome. These are the rates of change of neuronal activity, encoding expectations about hidden states. The fourth panel shows phasic fluctuations in posterior precision that can be interpreted in terms of dopamine responses. Here, the phasic component of simulated dopamine responses corresponds to the rate of change of precision (multiplied by eight) and the tonic component to the precision per se (divided by eight; see appendix 5). The phasic part reflects the precision prediction error (cf. reward prediction error: see equation 2.8). These simulated responses reveal a phasic response to the cue (CS) during epistemic trials that emerges with context learning over repeated trials. This reflects an implicit transfer of dopamine responses from the US to the CS. When the reward (US) is accessed directly, there is a profound increase in the phasic response relative to the response elicited after it has been predicted by the CS.

The final panel illustrates learning in terms of the accumulated posterior expectations about the initial state. The implicit learning reflects an accumulation of evidence that the reward will be found in the same location. In other words, initially ambiguous priors over the first two hidden states come to reflect the agent's experience that it always starts in the first hidden state. It is this context learning that underlies the pragmatic behavior in later trials. We talk about context learning (as opposed to inference) because, strictly speaking, Bayesian updates to model parameters (between trials) are referred to as learning, while updates to hidden states (within trial) correspond to inference.

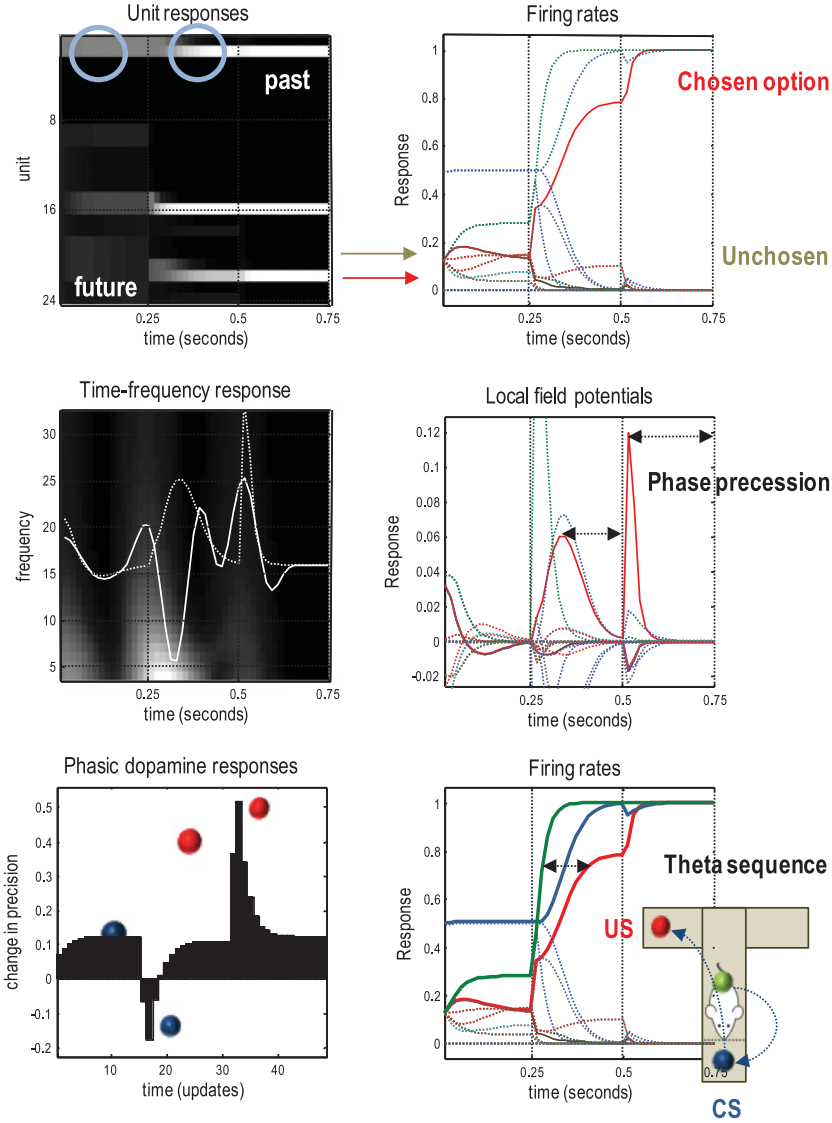
3.2 Electrophysiological Correlates of Variational Belief Updating. Figure 5 shows responses during the first trial in a way that speaks to empirical responses in studies of spatial navigation and decision making. The upper left panel shows simulated neuronal activity (firing rate) for units encoding hidden states using an image (or raster) format. There are eight hidden states for each of the three epochs or moves. These responses are organized such that the first eight rows show the probability of the eight states in the first observation epoch (i.e., period before moving), while subsequent epochs are shown in the middle and lower rows. This format illustrates the encoding of states over time, where the past lies in the upper diagonal blocks and the future in the lower diagonal blocks. To interpret these responses in relation to empirical results, we assume that outcomes are sampled every 250 ms. Although this is a little fast for overt exploratory movements in a maze, it corresponds to the intervals between saccadic eye

movements in visual exploration (Srihasam, Bullock, & Grossberg, 2009) and the rate at which syllables are articulated in normal speech (Gross et al., 2013). Furthermore, it corresponds to the timescale of neuronal dynamics in the hippocampus (e.g., the duty cycle of theta activity).

Note the changes in activity after each new outcome is observed. For example, the two units encoding the first two hidden states in the first epoch (circled) maintain their firing rate at equivalent levels, reflecting uncertainty about which of the two hidden states are occupied. However, after observing the cue, their activity diverges to properly infer that the first state was the central location under the second context. In other words, representations of the past are informed by current outcomes. The implicit postdiction enables the agent to update its representation (i.e., memory) of the initial state (i.e., past), which it can call on for context learning (see below).

The upper right panel plots the same information, highlighting two units (in solid lines), encoding the upper left and right location on the third epoch. These are the chosen and unchosen states, respectively. Initially, both units encode the same uncertain beliefs about the state that will be occupied, which are resolved in the second epoch and confirmed in the third. The ensuing pattern of firing reflects a saltatory or stepwise evidence accumulation in which expectations about occupying the chosen and unchosen states diverge as the trial progresses. This belief updating is formally identical to evidence accumulation described by drift diffusion or race-to-bound models (Solway & Botvinick, 2012; Zhang & Maloney, 2012; de Lafuente et al., 2015; Kira et al., 2015) and nicely recapitulates the emergence of a choice as evaluation of options proceeds (Hunt et al., 2012). Furthermore, the separation of timescales implicit in variational updating reproduces the stepping dynamics seen in parietal responses during decision making (Latimer et al., 2015).

The right middle panel shows the associated local field potentials, which are simply the rate of change of neuronal firing shown on the upper right. These simulated responses show that units encoding locations later in the trial peak earlier, as successive outcomes are observed. This necessarily results in a phase precession (Burgess et al., 2007; Lisman & Buzsaki, 2008; Lisman & Redish, 2009). In other words, units (e.g., place cells) encoding the same location at the same point in the trial reach their maximum activity more quickly with each successive (theta cycle) of evidence accumulation (see the arrows in the middle right panel of Figure 5). This phenomenon reflects the fact that locations visited toward the end of a trial only receive sensory evidence when they are encountered, at which point they quickly converge to their posterior expectations. The implicit encoding of trajectories through (state) space has many similarities with the notion of a to-do list that has been invoked to explain phase precession (Jensen, Gips, Bergmann, & Bonnefond, 2014).



The lower left panel illustrates simulated dopamine responses. Here, we see a phasic suppression when the cue (conditioned stimulus—CS) is located, followed by a phasic burst when the reward (unconditioned stimulus—US) is secured. The suppressive responses to the CS shown here are during the first trial. As noted above, these reductions quickly reverse and come to resemble the responses to the US after a few trials. We will

return to this; however, we first consider the place coding responses of units representing hidden states.

3.3 Theta-Gamma Coupling and Place Cell Activity. The lower right panel of Figure 5 shows the same firing rate responses above but highlights units encoding the three locations visited (the thick green blue and red lines). These responses reflect increases in activity (during the second theta epoch) in the same sequence that the locations are visited. Empirically, this phenomenon is called a theta sequence: short (3–5) sequences of place cells that fire sequentially within each theta cycle, as if they were encoding time-compressed trajectories (Lisman & Redish, 2009).

In our setting, theta-gamma coupling is a straightforward consequence of belief updating every 250 ms (i.e., theta), where each observation induces phasic updates that necessarily possess high-frequency (i.e., gamma) components. This is illustrated in the middle left panel of Figure 5, which shows

Figure 5: Simulated electrophysiological responses for the first trial. This figure reports the belief updating described in the text. It presents responses in several formats that emulate empirical characterizations of spatial navigation and decision-making responses. The upper left panel shows the activity (firing rate) of all units encoding hidden states in image (raster) format. There are eight hidden states for each of the three epochs in this trial, where each (250 ms or theta) epoch starts with an observation and ends with an action. These responses are organized such that the upper rows encode the probability of the eight states in the first epoch, with subsequent epochs in the middle and lower rows. Note the fluctuations in activity after each new outcome is observed. The upper right panel plots the same information highlighting two units (in solid lines), encoding the upper left (rewarded and chosen state) and upper right location on the third epoch (unrewarded and unchosen state). The simulated local field potentials for these units (i.e., their rate of change of neuronal firing) are shown in the middle right panel. This pattern of firing reflects a saltatory evidence accumulation (stepping dynamics), in which expectations about occupying the chosen and unchosen states diverge as the trial progresses. The simulated local field potentials also show that responses in units encoding locations later in the trial peak earlier, as successive outcomes are observed. This necessarily results in a phase precession that is also illustrated in the middle left panel. This panel shows the response of the rewarded hidden state unit before (dotted line) and after (solid line) filtering at 4 Hz, superimposed on a time-frequency decomposition of the local field potential (averaged over all units). The key observation here is that depolarization in the 4 Hz range coincides with induced responses, including gamma activity. The lower left panel illustrates simulated dopamine responses in terms of a mixture of precision and its rate of change. Finally, the lower right panel reproduces the upper right panel but highlights responses in units encoding the states visited (green, – first; blue, second; and red, final state).

the response of the second (rewarded hidden state) unit before (dotted line) and after (solid line) filtering at 4 Hz. These responses are superimposed on a time frequency decomposition of the local field potential averaged over all units. The key observation here is that depolarization in the theta range coincides with induced responses, including gamma activity. The implicit theta-gamma coupling during navigation can be seen more clearly in Figure 6. This figure reports simulated electrophysiological responses over the first eight trials, with the top panel showing the responses of units encoding hidden states and the second panel showing the associated time frequency response (and depolarization of the first unit, after filtering at 4 Hz). The final two panels show the simulated local field potentials and dopamine responses using the same format as the previous figure. The key observation in this here is that fluctuations in gamma power (averaged over all units) are tightly coupled to the depolarization in the theta range (of single units).

Phase precession and theta-gamma coupling are typically observed in the context of place cell activity, in which units respond selectively when an animal passes through particular locations. This sort of response is easy to demonstrate under the current scheme. Figure 7 (upper right panel) plots the activity of two units encoding the rewarded locations at the right (green dots) and left (red dots) arms as a function of the location in the maze over the first eight trials. The trajectories (dotted lines) were constructed by adding random displacements (with a standard deviation of an eighth) to the trajectory prescribed by action. The dots indicate times at which the unit approached its maximal firing rate (i.e., greater than 80%) and illustrate place cell activity that is specific to the locations they encode. However, this response profile is unique to the units encoding the final location: units encoding the location in the second epoch fire maximally at both the target location and the preceding (cue) location (lower right panel).

We present these results to address an interesting question. Hitherto, we have assumed that units encode states (location) in a frame of reference that is locked to the beginning of a trial or trajectory. The alternative is that each unit encodes the state in relation to the current time, in a moving time frame. This distinction is shown schematically in the lower left panel of Figure 7. If we use a fixed frame of reference, the successive activities of the two units are described by rows of the raster, indicated with white numbers. Conversely, if the encoding uses a moving frame of reference, these units would show the activity along the leading diagonal of the raster, indicated by the red numbers. Crucially, in a moving frame of reference, all units would show classical place cell responses, whereas in a fixed frame of reference, some units will encode the location of states that will be visited in the future. This would lead to a more complicated relationship between neuronal firing and the location of the animal.

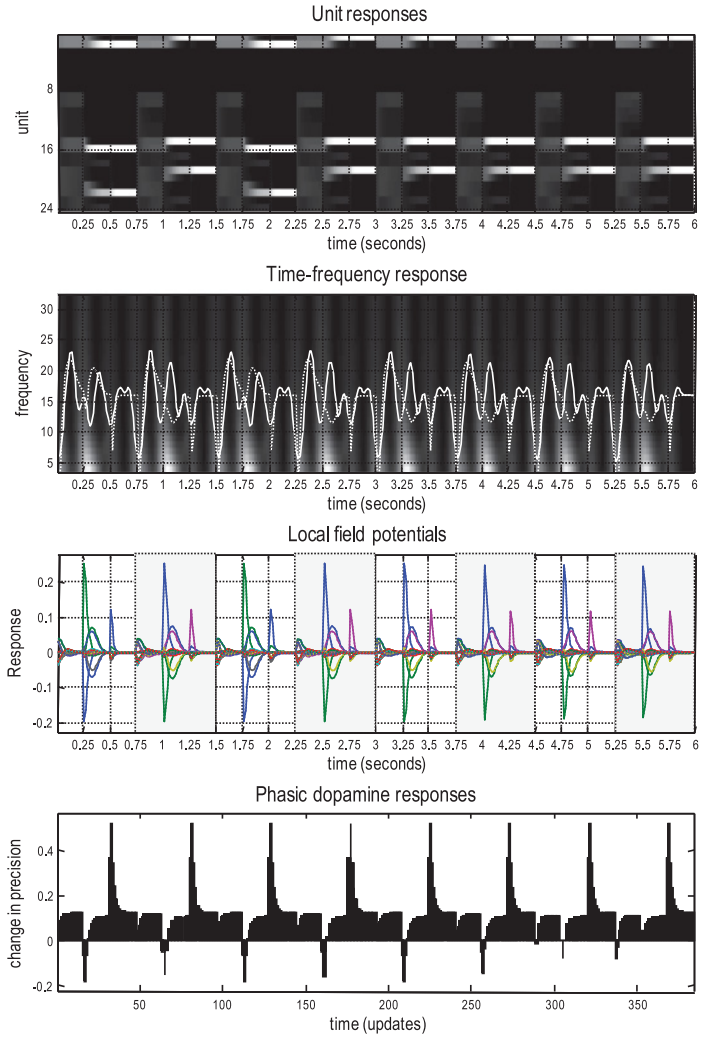


Figure 6: Theta-gamma phase coupling during spatial navigation. Simulated electrophysiological responses as in the previous figure. Here, the first eight trials are shown, with the top panel reporting the responses of units encoding hidden states and the second panel showing the associated time frequency response (and depolarization of a single unit after bandpass filtering at 4 Hz). The final two panels show the simulated local field potentials and dopamine responses using the same format as the previous figure. Every other trial is highlighted with a gray background, where each trial comprises three epochs (following the first and subsequent outcomes after two movements). The key observation here is that fluctuations in gamma power (averaged over all units) are tightly coupled to the depolarization in the theta range (of single units).

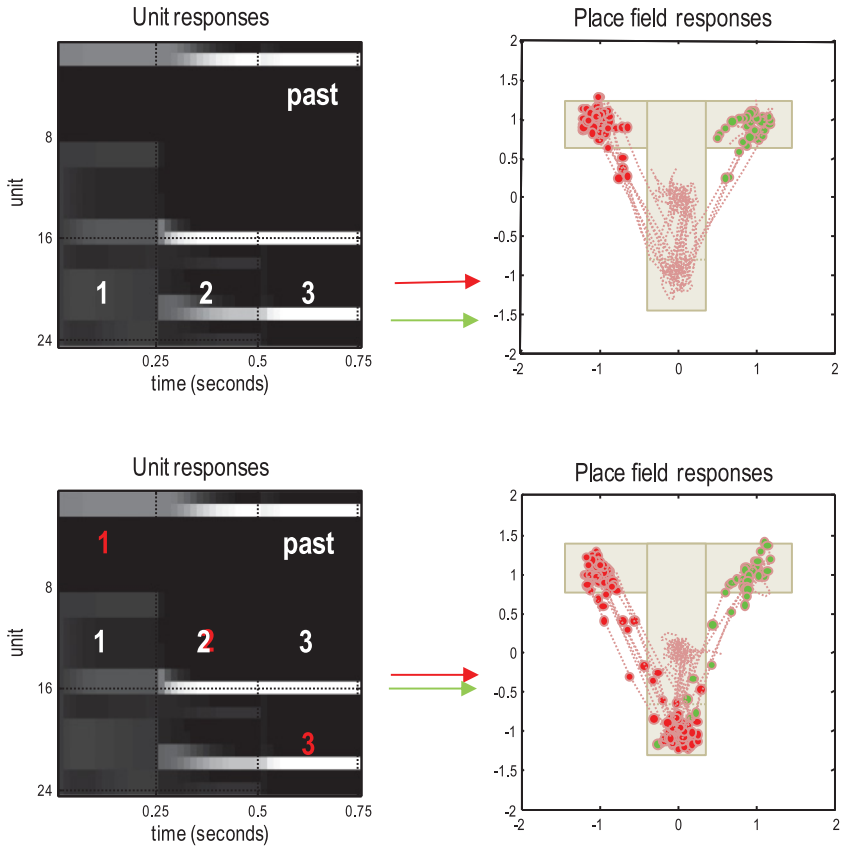


Figure 7: Place cell responses. The upper right panel plots the activity of two units encoding the rewarded locations at the right (green dots) and left (red dots) arms, as a function of the location in the maze, over the first eight trials. The trajectories (dotted lines) were constructed by adding (smooth) random displacements (with a standard deviation of an eighth) to the trajectory prescribed by action. The dots indicate times at which the unit exceeded 80% of its maximum activity and illustrate place cell activity that is specific to the locations encoded. However, this response profile is unique to the units encoding the final location: units encoding the location in the second epoch fire maximally at both the target location and the preceding (cue) location (lower right panel). The left panel reproduces the neural activity in raster format for two trials to indicate expectations about hidden states that are plotted.

Mathematically, both encoding schemes are viable and can be expressed following equation 2.8 (where $\lfloor t \rfloor$ is the floor function that returns epoch as a function of time):

$$\mathbf{s}_\tau^\pi(t+\Delta t) = \sigma(\widehat{\mathbf{s}}_\tau^\pi(t) - \Delta t \cdot (\widehat{\mathbf{s}}_\tau^\pi(t) - \ldots - \widehat{\mathbf{B}}(\pi(\tau)) \cdot \mathbf{s}_{\tau+1}^\pi(t))),$$

$$\mathbf{s}_\tau^\pi(t+\Delta t) = \sigma(\widehat{\mathbf{s}}_\tau^\pi(t) - \Delta t \cdot (\widehat{\mathbf{s}}_\tau^\pi(t) - \ldots - \widehat{\mathbf{B}}(\pi(\lfloor t \rfloor + \tau)) \cdot \mathbf{s}_{\tau+1}^\pi(t))). \tag{3.1}$$

The key difference between these formulations is that in the moving frame of reference, the connectivity changes from epoch to epoch, whereas in a fixed frame of reference, the connectivity remains the same. In light of this, we have elected to simulate responses assuming a fixed frame of reference, which suggests that a subset of hippocampal (or parietal) units should show extraclassical place cell activity, encoding trajectories over multiple locations (Grosmark & Buzsaki, 2016).

4 Context Learning

Having established that the Bayesian updates of expected hidden states and parameters have a degree of biological plausibility, we now turn to the correlates of parameter learning. In this article, the only parameters that are updated are those encoding prior beliefs about the initial state or context. These are the concentration parameters d. In what follows, we look at the effects of context learning on electrophysiological responses and what would happen if we removed prior preferences to reveal purely epistemic behavior.

4.1 Repetition Suppression and Dopamine Transfer. Figure 8 uses the same format as Figure 6; however, here we compare two identical trials that differ only in terms of the agent's prior beliefs about context. These trials are indicated by the arrows on the insert from Figure 4 (upper right in Figure 8) and have been associated with oddball and standard trials, respectively. The only difference is that the agent has become familiar with the context in which it enacts its epistemic policy. The increased efficiency and confidence afforded by context learning are expressed in terms of a faster encoding of hidden states and the emergence of a phasic dopamine (precision) response to the CS. In other words, the familiarity effects of repetitions of standard trials suppress evoked responses in units encoding the first state in the second epoch (blue circles). This can be seen clearly if we subtract the evoked response during the standard trial from the equivalent response during the oddball trial at the point of anticipation, in the second epoch. The result is shown in the right panel as a negative difference waveform that peaks at around 80 ms (or 180 ms allowing 100 ms conduction delays to occipital cortex). This is exactly the form of difference elicited in empirical oddball studies using sequences of repeating stimuli, where it is known as the mismatch negativity (Bendixen et al., 2012).

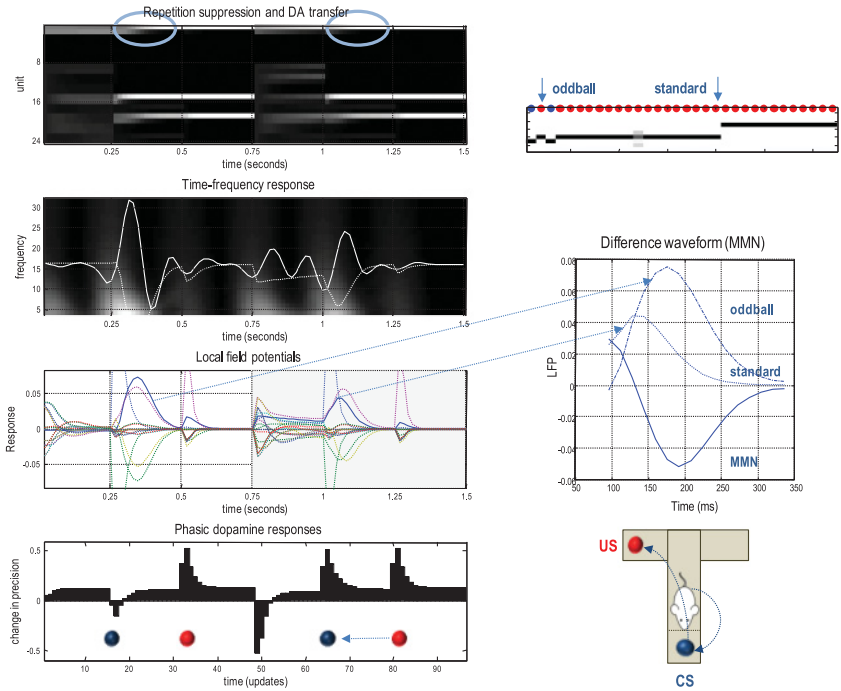


Figure 8: Repetition suppression and transfer of dopamine responses. This figure uses the same format as Figure 6; however, here we compare two (oddball and standard) trials that are indicated by the arrows on the insert from Figure 4 (upper right). The only difference between these trials is that the agent has become familiar with the context. This means it is more efficient and confident in its inference. This is expressed in terms of a slightly faster and lower-amplitude belief updating about hidden states and increases in expected precision when sampling the cue. The familiarity effects due to repetitions of the standard trials suppress evoked responses in units encoding the first state (cyan circles). This can be seen clearly in the right panel, when we subtract the responses during the standard trial from the equivalent updates during the oddball trial (at the point of anticipation, in the second epoch). The result is a negative difference wave that peaks at around 80 ms (or 180 ms, allowing 400 ms conduction delays). Inspection of the (simulated) phasic dopamine responses shows that the large-amplitude responses to the reward (US) in the first trial are transferred to the cue (CS) after the context has been learned. This pattern corresponds to the transfer of dopamine responses observed in reinforcement learning paradigms.

This repetition suppression is accompanied by profound changes in simulated dopamine responses that effectively reproduce the transfer of phasic dopamine responses from unconditioned to conditioned stimuli

during learning (Schultz, Apicella, & Ljungberg, 1993; Bromberg-Martin & Hikosaka, 2009). In this instance, the learning corresponds to increasing confidence about the context in which choices are made (Fiorillo et al., 2003). This translates into a higher precision of beliefs about competing policies once the CS has resolved residual uncertainty. Note that this transfer from the US to the CS is direct and does not require any representation of intervening states (see (FitzGerald, Dolan et al., 2015) for a fuller discussion). The differences in responses in these two trials can be explained only by differences in prior beliefs about context, because the actions and outcomes were identical. But what about responses when outcomes are unpredicted?

4.2 Violation Responses and Simulated P300 Waveforms. Figure 9 uses the same format as Figure 6 but focuses on consecutive trials after a degree of context learning (the trials indicated by the arrows above the insert from Figure 4). The first trial is a standard one in which the agent interrogates the cue location and then acquires the reward from the appropriate arm. In the subsequent trial, we forced the agent to stay at the cue location (by preventing it from moving), thereby inducing protracted belief updating about hidden states. This is most evident in the hidden state encoding the true location in the third (final) epoch (blue circles). These violation responses reach peak amplitude at about 100 ms—or 200 ms in peristimulus time (allowing for 100 ms conduction delays). Although earlier than classical P300 and N400 responses, this protracted and late response is reminiscent of violation responses in event-related potential (ERP) studies when the outcome is inconsistent with the preceding succession of states, such as semantic violations in sentence processing and action observation (Friederici, 2005; Maffongelli et al., 2015). These late violation responses contrast with the early mismatch responses in the previous figure. Finally, note that the phasic dopamine response to the unexpected outcome is attenuated although not abolished. This may reflect the fact that the agent finds it difficult to believe it has not secured its reward. In other words, the agent partly believes it has pursued the epistemic policy despite evidence to the contrary (see upper panel).

4.3 Foraging for Information. One might ask what would happen if rewards were devalued by setting their (relative) utility to zero. Figure 10 shows the results of a simulation, using the same setup as in Figure 4. The only difference here was that there were no explicit preferences or utilities. However, the resulting behavior is still structured and purposeful because it is driven by epistemic value. In every trial, the agent moves to the cue location to resolve ambiguity about the context (see lower panels). After the cue is sampled, uncertainty cannot be reduced further, and the agent either stays where it is or returns to the central location, avoiding the baited arms. It avoids the baited arms because they are mildly ambiguous (given our partial reinforcement schedule). This sort of simulation can, in

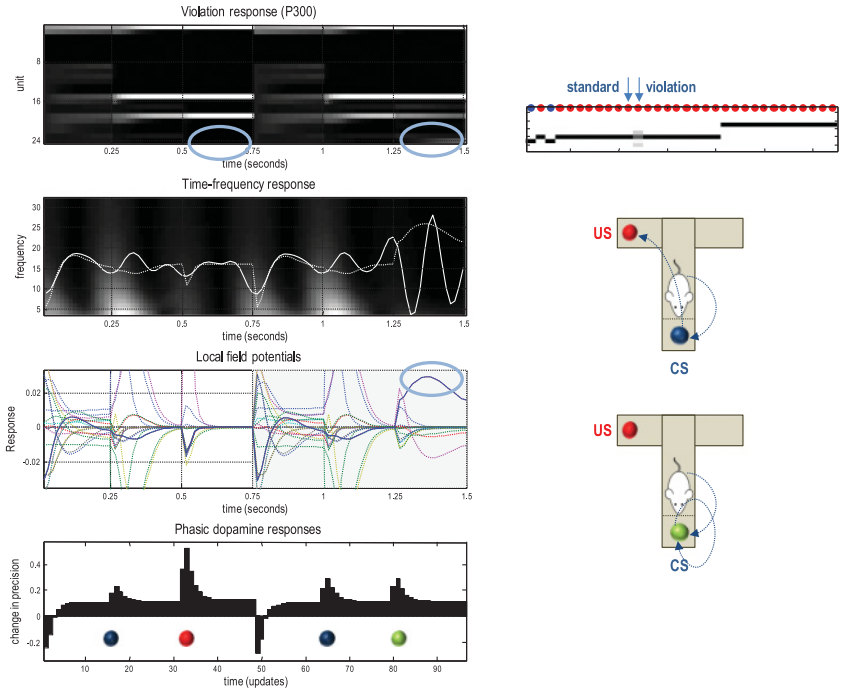


Figure 9: Violation responses and simulated P300 waveforms. This figure uses the same format as the previous figure but focuses on consecutive trials indicated by the arrows above the insert. The first trial is an epistemic trial in which the agent interrogates the cue location and then acquires the reward. In the subsequent trial, we forced the agent to stay where it was, thereby inducing protracted and high-amplitude belief updating about hidden states. This is most evident in the hidden states encoding the (cue) location in the third (final) epoch (cyan circles). Assuming each epoch lasts 250 ms, these responses reach peak amplitude at about 150 ms—or 250 ms in peristimulus time (allowing for 100 ms conduction delays).

principle, be used to simulate foraging for information using saccadic eye movements.

This simulation illustrates the fact that behavior can still be purposeful even in the absence of extrinsic value or prior preferences about outcomes. In other words, epistemic value can, on its own, specify behavior even if there are no explicit or extrinsic goals. The implication here is that the balance between purely exploratory and exploitative behavior rests on the precision of prior preferences. In the simulations, the removal of preferences corresponds to making every outcome equally plausible, thereby setting the precision of prior preferences to zero. Having said this, outcomes are

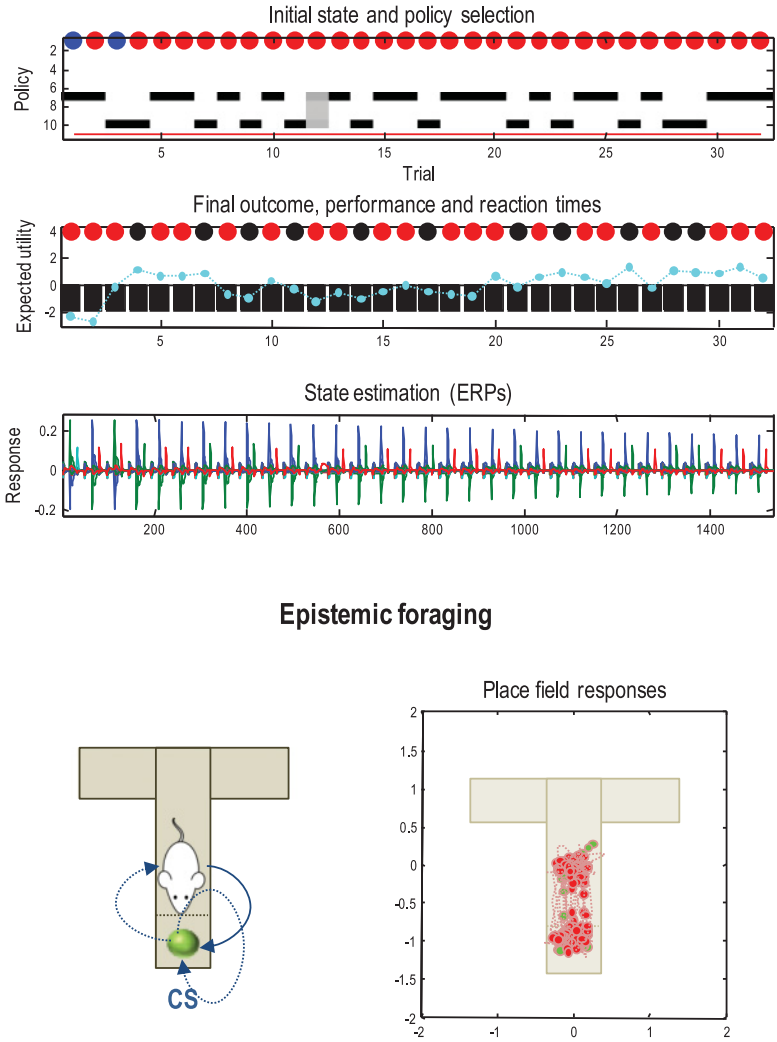


Figure 10: Epistemic foraging. This figure reports the (behavioral and physiological) responses over the 32 trials as in Figure 4. However, in this simulation, all outcomes were assigned the same utility. This means there is no extrinsic value, and the agent maximizes epistemic value by first resolving its uncertainty about the context (by going to the cue location) and then avoiding (the mildly ambiguous) upper arms. This behavior is shown schematically, and in terms of place cell firing, in the lower panels.

still limited to those entertained by an agent's beliefs about the world. The set of outcomes entailed by a particular generative model could be construed as preferred outcomes, where all other possible outcomes have been eliminated and, effectively, have a large negative utility. This means that in one sense, explorative or epistemic behavior is always restricted to outcomes that, a priori, an agent prefers or, equivalently, outcomes that characterise an agent.

4.4 Summary. In summary, we have reviewed several simulated responses that bear a remarkable resemblance to empirical electrophysiological responses in spatial navigation and classical ERP paradigms. We have also seen responses characteristic of dopaminergic activity during instrumental learning, when examining the encoding of precision. Although the similarity between simulated and empirical responses is at best metaphorical, it is interesting to note that all of these behaviors emerged from a standard variational scheme that was applied to a generic state-space model. In other words, there was no attempt to reproduce empirical findings by hand-tuning the generative model or the inversion scheme. The only thing we assumed was that outcomes are sampled every 250 ms. More specifically, the neuronal dynamics in equation 2.8 follow from a gradient descent on variational free energy, where variational free energy is defined completely by the generative model, and the generative model is based on a generic Markovian process. This is important because it provides a putative variational principle for neural dynamics that can be described in terms of a Lyapunov function (variational free energy) from a dynamical systems perspective (Stam, 2005). Alternatively, we can think of neuronal activity as conforming to Hamilton's principle of least action, where action is the path integral of free energy (Friston, 2013). In short, the simulations above constitute a construct validation of the ensuing process theory in relation to empirical electrophysiology and the numerous normative models inspired by these empirical phenomena. Clearly, variational principles do not, in and of themselves, prescribe the aspects of neurobiology we have considered, in the same sense that natural selection does not prescribe a particular phenotype. However, they may offer a relatively straightforward and teleological perspective on neuroanatomy and physiology (Friston & Buzsaki, 2016).

5 Conclusion

We have described an active inference scheme for discrete state-space models of choice behavior that is suitable for modeling a variety of paradigms and phenomena. This generic scheme offers a process theory that is based on a standard (gradient descent) minimization of variational free energy—or approximate Bayesian inference. The ensuing process theory provides a simple (perhaps oversimplified) account of many empirical phenomena

that include repetition suppression, omission responses, violation responses, place cell activity, phase precession, theta sequences, theta-gamma coupling, evidence accumulation, race-to-bound dynamics, and transfer of dopamine responses. It is worth reiterating that these emergent properties follow from, and only from, the form of the underlying generative model.

In this sense, the challenge is to identify the generative models that best explain empirical responses. We have focused on a simple and generic form, but there are clearly many alternatives and extensions. Key among these are hierarchical models with deep temporal structure (George & Hawkins, 2009; Specht, 2014), and models in which prior preferences are absorbed into beliefs about state transitions or contingencies. Appendix F touches on further extensions that consider not the path integral of expected free energy but the expected path integral of free energy and the distinction between naive and sophisticated schemes. This distinction may be particularly important for understanding planning and metacognition and their physiological correlates (Lisman & Redish, 2009; Penny et al., 2013; Pezzulo et al., 2014).

In closing, one should acknowledge that good process theories "should explain what is already known more parsimoniously than any other theory of comparable explanatory scope, but they should also stick their neck out to specify what is forbidden, and what new phenomena have not been observed yet but should be or could be" (personal communication from an anonymous reviewer). We will not meet this challenge here; however, it is interesting to note that the epistemic imperatives implied by minimizing variational free energy lead to a parsimonious (minimally complex) yet accurate description of observable outcomes or facts (see equation 2.4). In this sense, active inference may offer a formal (metatheoretical) description for the scientific process itself.

Appendix A: Belief Updating

Variational updates are a self-consistent set of equalities that minimize variational free energy, which can be expressed as the (time-dependent) free energy under each policy plus the complexity incurred by posterior beliefs about (time-invariant) policies and parameters, where (ignoring constants):

$$\begin{aligned} F &= D[Q(x)||P(x)] - E_Q[\ln P(\tilde{o}|x)] \\ &= \sum_{\tau} E_Q[F(\pi,\tau)] + D[Q(\pi)||P(\pi)] + D[Q(\gamma)||P(\gamma)] \\ &\quad + D[Q(\mathbf{A})||P(\mathbf{A})] + \dots \end{aligned}$$

$$= \boldsymbol{\pi} \cdot (\widehat{\pi} + \mathbf{F} + \boldsymbol{\gamma} \cdot \mathbf{G}) + \ln Z + \beta\boldsymbol{\gamma} - \ln \boldsymbol{\gamma} + \sum_i (\mathbf{a}_i - a_i) \cdot \widehat{\mathbf{A}}_i$$
$$- \ln B(\mathbf{a}_i) + \ldots$$

The free energy of hidden states and the expected free energy are given by

$$\begin{aligned}
\mathbf{F}_\pi &= F(\pi) \\
F(\pi) &= \sum_\tau F(\pi, \tau) \\
F(\pi, \tau) &= \underbrace{E_{\tilde{Q}}[D[Q(s_\tau|\pi)||P(s_\tau|s_{\tau-1}, \pi)]]}_{complexity} - \underbrace{E_{\tilde{Q}}[\ln P(o_\tau|s_\tau)]}_{accuracy} \\
&= \mathbf{s}^\pi_\tau \cdot (\widehat{\mathbf{s}}^\pi_\tau - \widehat{\mathbf{B}}^\pi_{\tau-1}\mathbf{s}^\pi_{\tau-1} - \widehat{\mathbf{A}} \cdot o_\tau)
\end{aligned}$$

$$\begin{aligned}
\mathbf{G}_\pi &= G(\pi) \\
G(\pi) &= \sum_\tau G(\pi, \tau) \\
G(\pi, \tau) &= \underbrace{D[Q(o_\tau|\pi)||P(o_\tau)]}_{expected\ cost} + \underbrace{E_{\tilde{Q}}[H[P(o_\tau|s_\tau)]]}_{expected\ ambiguity} \\
&= \mathbf{o}^\pi_\tau \cdot (\widehat{\mathbf{o}}^\pi_\tau - \mathbf{U}_\tau) + \mathbf{s}^\pi_\tau \cdot \mathbf{H}.
\end{aligned}$$

Here, $\widehat{\mathbf{B}}^\pi_\tau = \widehat{\mathbf{B}}(\pi(\tau))$, $\widehat{\mathbf{B}}^\pi_0 \mathbf{s}^\pi_0 = \widehat{\mathbf{D}}$. $Z = \sum_\pi \exp(-\boldsymbol{\gamma} \cdot \mathbf{G}_\pi)$ and $\widehat{\mathbf{A}} = \psi(\mathbf{a}) - \psi(\mathbf{a}_0)$. The beta function of the column vector $\mathbf{a}_i$is denoted by $B(\mathbf{a}_i)$. Using the standard result, $\partial_{\mathbf{a}} \mathrm{B}(\mathbf{a}) = \mathrm{B}(\mathbf{a})\widehat{\mathbf{A}}$, we can differentiate the variational free energy with respect to the sufficient statistics (with a slight abuse of notation and using $\partial_{\mathbf{s}} F := \partial F(\pi, \tau)/\partial \mathbf{s}^\tau_\pi$):

$$\partial_{\mathbf{s}} F = \widehat{\mathbf{s}}^\pi_\tau - \widehat{\mathbf{A}} \cdot o_\tau - \widehat{\mathbf{B}}^\pi_{\tau-1}\mathbf{s}^\pi_{\tau-1} - \widehat{\mathbf{B}}^\pi_\tau \cdot \mathbf{s}^\pi_{\tau+1}$$

$$\begin{aligned}
\partial_{\boldsymbol{\pi}} F &= \widehat{\pi} + \mathbf{F} + \boldsymbol{\gamma} \cdot \mathbf{G} \\
\partial_{\boldsymbol{\gamma}} F &= \beta + \boldsymbol{\pi} \cdot \mathbf{G} + \frac{1}{Z}\partial_{\boldsymbol{\gamma}} Z - \boldsymbol{\beta} \\
&= \beta + (\boldsymbol{\pi} - \boldsymbol{\pi}_0) \cdot \mathbf{G} - \boldsymbol{\beta} \\
\partial_{\boldsymbol{\gamma}} Z &= -\exp(-\boldsymbol{\gamma} \cdot \mathbf{G}) \cdot \mathbf{G} \\
\boldsymbol{\pi}_0 &= \sigma(-\boldsymbol{\gamma} \cdot \mathbf{G})
\end{aligned}$$

$$\partial_{\mathbf{a}} F = \partial_{\mathbf{a}} \widehat{\mathbf{A}} \cdot (\mathbf{a} - a - \sum_{\tau} o_{\tau} \otimes \mathbf{s}_{\tau})$$

$$\partial_{\mathbf{b}} F = \partial_{\mathbf{b}} \widehat{\mathbf{B}} \cdot (\mathbf{b}(u) - b(u) - \sum_{\pi(\tau)=u} \boldsymbol{\pi}_{\pi} \cdot \mathbf{s}_{\tau}^{\pi} \otimes \mathbf{s}_{\tau-1}^{\pi})$$

$$\partial_{\mathbf{d}} F = \partial_{\mathbf{d}} \widehat{\mathbf{D}} \cdot (\mathbf{d} - d - \mathbf{s}_1)$$

$$\mathbf{s}_{\tau} = \sum_{\pi} \boldsymbol{\pi}_{\pi} \cdot \mathbf{s}_{\tau}^{\pi}$$

Finally, the solutions to these equations give the variational updates in the main text (see equation 2.7).

Appendix B: Generalized Coordinate Descent

Equation 2.8 follows in a straightforward fashion from a gradient ascent on variational free energy:

$$\dot{\widehat{\mathbf{s}}}_{\tau}^{\pi} = -\partial_{\widehat{\mathbf{s}}} F = \partial_{\widehat{\mathbf{s}}} \mathbf{s}_{\tau}^{\pi} \cdot \varepsilon_{\tau}^{\pi}$$

$$\dot{\boldsymbol{\beta}} = -\partial_{\boldsymbol{\beta}} F = \boldsymbol{\gamma}^2 \varepsilon^{\gamma}$$

$$\partial_{\widehat{\mathbf{s}}} F = \partial_{\widehat{\mathbf{s}}} \mathbf{s}_{\tau}^{\pi} \cdot \partial_{\mathbf{s}} F = -\partial_{\widehat{\mathbf{s}}} \mathbf{s}_{\tau}^{\pi} \cdot \varepsilon_{\tau}^{\pi}$$

$$\partial_{\boldsymbol{\beta}} F = \partial_{\boldsymbol{\beta}} \boldsymbol{\gamma} \cdot \partial_{\boldsymbol{\gamma}} F = -\boldsymbol{\gamma}^2 \cdot \varepsilon_{\tau}^{\pi}$$

$$\mathbf{s}_{\tau}^{\pi} = \sigma(\widehat{\mathbf{s}}_{\tau}^{\pi}) = Z^{-1} \exp(\widehat{\mathbf{s}}_{\tau}^{\pi})$$

$$\partial_{\widehat{\mathbf{s}}} Z = Z \cdot \mathbf{s}_{\tau}^{\pi}$$

$$\partial_{\widehat{\mathbf{s}}} \mathbf{s}_{\tau}^{\pi} = diag(\mathbf{s}_{\tau}^{\pi}) - \mathbf{s}_{\tau}^{\pi} \otimes \mathbf{s}_{\tau}^{\pi}$$

where the gradients (prediction errors) are derived in appendix A:

$$-\partial_{\mathbf{s}} F = \varepsilon_{\tau}^{\pi} = (\widehat{\mathbf{A}} \cdot o_{\tau} + \widehat{\mathbf{B}}_{\tau-1}^{\pi} \mathbf{s}_{\tau-1}^{\pi} + \widehat{\mathbf{B}}_{\tau}^{\pi} \cdot \mathbf{s}_{\tau+1}^{\pi}) - \widehat{\mathbf{s}}_{\tau}^{\pi},$$

$$\partial_{\boldsymbol{\gamma}} F = \varepsilon^{\gamma} = (\beta - \boldsymbol{\beta}) + (\boldsymbol{\pi} - \boldsymbol{\pi}_0) \cdot \mathbf{G}.$$

Practically, one can solve these equations using the discrete updates:

$$\mathbf{s}_{\tau}^{\pi}(t + \Delta t) \approx \sigma(\widehat{\mathbf{s}}_{\tau}^{\pi}(t) + \Delta t \cdot \partial_{\widehat{\mathbf{s}}} \mathbf{s}_{\tau}^{\pi} \cdot \varepsilon_{\tau}^{\pi})$$

$$\boldsymbol{\beta}(t+\Delta t) \approx \boldsymbol{\beta}(t) + \Delta t \cdot \varepsilon^{\gamma},$$

In the simulations, we used $\Delta t = 1/4$ but continued iterating for 16 (250 ms) iterations.

Appendix C: Belief Propagation

The mean field assumption approximates the posterior with the product of marginals over the current state, lending free energy the following form:

$$\begin{aligned} F(\pi) = & \underbrace{D[Q(s_1|\pi)\ldots Q(s_T|\pi)||P(s_1,\ldots,s_T|\pi)]}_{\textit{Relative complexity}} \\ & - \underbrace{E_{Q(s_t|\pi)\ldots Q(s_T|\pi)}[P(\tilde{o}|s_1,\ldots,s_T)]}_{\textit{Accuracy}}. \end{aligned}$$

In practice, this leads to an overconfidence that can be finessed by explicitly optimizing the marginal posterior for each time point. This corresponds to belief propagation, which minimizes the following free energy (Yedidia, Freeman, & Weiss, 2005):

$$\begin{aligned} F(\pi) &= \sum_{\tau} F(\pi,\tau) \\ F(\pi,\tau) &= \tfrac{1}{2}\underbrace{D[Q(s_\tau|\pi)||P_F(s_\tau|\pi)]}_{\textit{Forward complexity}} + \tfrac{1}{2}\underbrace{D[Q(s_\tau|\pi)||P_F(s_\tau|\pi)]}_{\textit{Backward complexity}} \\ &\quad - \underbrace{E_{Q(s_t|\pi)}[P(o_\tau|s_\tau)]}_{\textit{Accuracy}} \\ &= \mathbf{s}_\tau^\pi \cdot \left(\widehat{\mathbf{s}}_\tau^\pi - \tfrac{1}{2}\ln(\bar{\mathbf{B}}_{\tau-1}^\pi \mathbf{s}_{\tau-1}^\pi) - \tfrac{1}{2}\ln(\bar{\mathbf{B}}_\tau^{\pi\dagger}\mathbf{s}_{\tau+1}^\pi) - \widehat{\mathbf{A}}\cdot o_\tau\right). \end{aligned}$$

In this formulation, complexity is defined in relation to empirical priors based on the approximate posterior expectations of the preceding (forward) and subsequent (backward) states:

$$\begin{aligned} P_F(s_\tau|\pi) &= E_{Q(s_{\tau-1}|\pi)}[P(s_\tau|s_{\tau-1},\pi)] \approx P(s_\tau|o_{\tau-1},o_{\tau-2},\ldots,\pi), \\ P_B(s_\tau|\pi) &= E_{Q(s_{\tau+1}|\pi)}[P(s_\tau|s_{\tau+1},\pi)] \approx P(s_\tau|o_{\tau+1},o_{\tau+2},\ldots,\pi). \end{aligned}$$

Free energy is minimized when $Q(s_\tau|\pi) = P(s_\tau|o_\tau,o_{\tau-1},\ldots,\pi) = P(s_\tau|o_\tau, o_{\tau+1},\ldots,\pi)$, is the marginal posterior distribution, given past and future

observations. In this case, free energy reduces to (negative) log evidence:

$$\begin{aligned} F(\pi,\tau) &= \tfrac{1}{2}\underbrace{D[Q(s_\tau|\pi)||P(s_\tau|o_\tau,o_{\tau-1},\ldots,\pi)]}_{\textit{forward divergence}} \\ &\quad + \tfrac{1}{2}\underbrace{D[Q(s_\tau|\pi)||P(s_\tau|o_\tau,o_{\tau+1},\ldots,\pi)]}_{\textit{backward divergence}} - \underbrace{\ln P(o_\tau)}_{\textit{evidence}} \\ &= -\underbrace{\ln P(o_\tau)}_{\textit{evidence}}. \end{aligned}$$

Here, we have omitted (uniform) priors over hidden states $P(s_\tau)$. Note that this marginal free energy retains the same form but uses the log of expectations, as opposed to expectations of logs. Furthermore, it uses backward transitions, $\mathbf{B}(u)^\dagger = Dir(\mathbf{b}(u)^T)$, such that the free energy gradients become

$$\partial_{\mathbf{s}} F(\pi,\tau) = \widehat{\mathbf{s}}_\tau^\pi - \widehat{\mathbf{A}}\cdot o_\tau - \tfrac{1}{2}\ln(\mathbf{B}_{\tau-1}^\pi \mathbf{s}_{\tau-1}^\pi) - \tfrac{1}{2}\ln(\mathbf{B}_\tau^{\pi\dagger}\mathbf{s}_{\tau+1}^\pi).$$

Although this formulation is slightly more complicated, it retains a biological plausibility in the sense that the computations are local and are mediated by connections that do not change with time. This computational simplicity of the scheme should be contrasted with the exact inference scheme described in appendix D.

Appendix D: Exact Bayesian Inference

An exact inference over sequences rests on using posterior distributions over the current states, conditioned on the previous states. This leads to a more complicated but exact scheme in which transitions (i.e., sequences) are retained in the posterior. In this instance,

$$\begin{aligned} Q(x) &= Q(s_T|s_{T-1},\pi)\ldots Q(s_1|\pi)Q(\pi)\ldots \\ Q\left(s_t|s_{t-1},\pi\right) &= Cat(\mathbf{s}_t^\pi). \end{aligned}$$

The free energy under any policy now becomes a tight or exact bound on log evidence,

$$\begin{aligned} F(\pi) &= \underbrace{D[Q(s_T|s_{T-1},\pi)\ldots Q(s_1|\pi)||P(s_T|s_{T-1},\pi)\ldots P(s_1|\pi)]}_{\textit{complexity}} \\ &\quad - \underbrace{E_Q[\ln P(o_1,\ldots,o_T|s_1,\ldots,s_T)]}_{\textit{accuracy}} \end{aligned}$$

$$= \sum_{\tau=1}^{T} diag(\mathbf{s}_\tau^\pi \cdot \Delta_\tau^\pi)\mathbf{s}_{\tau-1}^\pi \ldots \mathbf{s}_1^\pi - o_\tau \cdot \widehat{\mathbf{A}}\mathbf{s}_\tau^\pi \ldots \mathbf{s}_1^\pi,$$

$$\Delta_\tau^\pi = \widehat{\mathbf{s}}_\tau^\pi - \widehat{\mathbf{B}}_{\tau-1}^\pi,$$

with the following gradients:

$$\partial_{\mathbf{s}} F = \Delta_\tau^\pi diag(\mathbf{s}_{\tau-1}^\pi \ldots \mathbf{s}_1^\pi) + \sum_{\nu=1}^{T} (diag(\mathbf{s}_\nu^\pi \cdot \Delta_\nu^\pi) - o_\nu \cdot \widehat{\mathbf{A}}\mathbf{s}_\nu^{\pi})(\mathbf{s}_{\nu-1}^\pi \ldots \mathbf{s}_{\tau+1}^\pi) \otimes (\mathbf{s}_{\tau-1}^\pi \ldots \mathbf{s}_1^\pi).$$

In the absence of observations, the solution requires the posterior transitions to be equal to the prior transitions: $\Delta_\tau^\pi = 0 \Rightarrow \widehat{\mathbf{s}}_\tau^\pi = \widehat{\mathbf{B}}_\tau^\pi \Rightarrow \partial_{\mathbf{s}} F = 0$. Otherwise, the hidden states are updated on the basis of outcomes in the past and the future (when they are available).

Appendix E: Simulating Dopamine Responses

To simulate dopamine discharges, we assume that the encoding of expected precision $\boldsymbol{\gamma}$ is the postsynaptic response to dopaminergic input δ, modeled with a first-order Taylor approximation:

$$\dot{\boldsymbol{\gamma}} = \kappa_1 \cdot \delta - \kappa_2 \cdot \boldsymbol{\gamma} \Rightarrow$$

$$\delta = \frac{1}{\kappa_1} \cdot \dot{\boldsymbol{\gamma}} + \frac{\kappa_1}{\kappa_2} \cdot \delta.$$

In this article, $\kappa_1/\kappa_1 = 1/64$, which corresponds to a postsynaptic time constant of about 1 s (Bengtson, Tozzi, Bernardi, & Mercuri, 2004), assuming each iteration corresponds to 16 ms.

Appendix F: Sophisticated Schemes

The scheme described in this article is naive or unsophisticated in the sense that policies are selected that minimize the path integral of expected free energy, as opposed to the expected path integral of free energy. This means that policies are selected to minimize uncertainty about hidden states as opposed to minimizing uncertainty about policies. Note that by construction, the uncertainty about (or entropy of) beliefs about policies corresponds to the expected (path integral of) free energy. In other words, minimizing the expected path integral corresponds to resolving uncertainty about behavior. One could consider more sophisticated agents whose prior beliefs are based on the expected path integral—for example (omitting precision for

simplicity):

$$\ln P(\pi) = -G(\pi|o_t) - G(\pi|o_{t+1}) - \dots$$

$$G(\pi|o_{t+1}) = E_{Q(o_{t+1}|\pi)P(\pi_{t+1})}[G(\pi_{t+1}|o_{t+1})]$$
$$Q(o_{t+1}|\pi) = E_{Q(s_{t+1}|\pi)}Q(o_{t+1}|s_{t+1}) = \mathbf{A}\mathbf{s}^{\pi}_{t+1}$$
$$P(\pi_{t+1}) = \sigma(G(\pi_{t+1}|o_{t+1}))$$

In this case, the expected free energy after the next outcome $G(\pi_{t+1}|o_{t+1})$ is evaluated in the same way as the expected free energy at the current time $G(\pi_t|o_t) \triangleq G(\pi)$ for each (fictive) outcome o_{t+1} by using the posterior over current hidden states as the prior $\mathbf{D} = \mathbf{s}^{\pi}_{t+1}$. Clearly, this scheme is computationally more involved than the naive scheme and calls on recursive variational updating. This means that sophisticated agents are metacognitive in some sense because they perform belief updating (based on fictive outcomes) to optimize their belief updating.

Heuristically, the difference between naive and sophisticated schemes can be seen in terms of the first choice in current paradigm. For the naive agent, the best policy is to sample the cue location and stay there, because moving to a baited arm has, on average, no extrinsic value (and provides ambiguous outcomes). Conversely, the expected free energy of retrieving a reward after observing the cue is low for both (fictive) outcomes. This means the best policies are to behave epistemically on the first move and then pragmatically on the second move. Note that the sophisticated agent, unlike the naive agent, can entertain future switches between policies.

Acknowledgments

K.J.F. is funded by the Wellcome trust (088130/Z/09/Z). P.S. is a recipient of a DOC fellowship of the Austrian Academy of Sciences at the Centre for Cognitive Neuroscience, University of Salzburg. G.P. gratefully acknowledges support of HFSP (Young Investigator Grant RGY0088/2014). We thank our reviewers for detailed help in formulating these ideas.

We have no disclosures or conflict of interest.

References

Abbott, L. F., & Nelson, S. B. (2000). Synaptic plasticity: Taming the beast. *Nat. Neurosci.*, *3*(Suppl.), 1178–1183.

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

Attias, H. (2003). Planning by probabilistic inference. In *Proc. of the 9th Int. Workshop on Artificial Intelligence and Statistics*.

Ballard, D. H., Hinton, G. E., & Sejnowski, T. J. (1983). Parallel visual computation. *Nature*, *306*, 21–26.

Ballard, D. H., Kit, D., Rothkopf, C. A., & Sullivan, B. (2013). A hierarchical modular architecture for embodied cognition. *Multisensory Research*, *26*, 177.

Barlow, H. (1961). Possible principles underlying the transformations of sensory messages. In W. Rosenblith (Ed.), *Sensory communication* (pp. 217–234). Cambridge, MA: MIT Press.

Barrett, L. F., & Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nat. Rev. Neurosci.*, *16*(7), 419–429.

Barto, A., Mirolli, M., & Baldassarre, G. (2013). Novelty or surprise? *Frontiers in Psychology*, *4*.

Barto, A., Singh, S., & Chentanez, N. (2004). Intrinsically motivated learning of hierarchical collections of skills. In *Proceedings of the 3rd International Conference on Development and Learning*. Cambridge, MA: MIT Press.

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.

Beal, M. J. (2003). Variational algorithms for approximate Bayesian inference. Doctoral dissertation, University College London.

Bendixen, A., SanMiguel, I., & Schroger, E. (2012). Early electrophysiological indicators for predictive processing in audition: A review. *Int. J. Psychophysiol*, *83*(2), 120–131.

Bengtson, C. P., Tozzi, A., Bernardi, G., & Mercuri, N. B. (2004). Transient receptor potential-like channels mediate metabotropic glutamate receptor EPSCs in rat dopamine neurones. *J. Physiol.*, *555*(Pt. 2), 323–330.

Botvinick, M., & An, J. (2009). Goal-directed decision making in prefrontal cortex: A computational framework. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems, 22*. Cambridge, MA: MIT Press.

Botvinick, M., & Toussaint, M. (2012). Planning as inference. *Trends Cogn. Sci.*, *16*(10), 485–488.

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychol. Bull.*, *138*(3), 389–414.

Braun, D. A., Ortega, P. A., Theodorou, E., & Schaal, S. (2011). Path integral control and bounded rationality. In *IEEE symposium on adaptive dynamic programming and reinforcement learning*. Piscataway, NJ: IEEE.

Bromberg-Martin, E. S., & Hikosaka, O. (2009). Midbrain dopamine neurons signal preference for advance information about upcoming rewards. *Neuron*, *63*(1), 119–126.

Burgess, N., Barry, C., & O'Keefe, J. (2007). An oscillatory interference model of grid cell firing. *Hippocampus*, *17*(9), 801–812.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain. Sci.*, *36*(3), 181–204.

De Bruin, L., & Michael, J. (2015). Bayesian predictive coding and social cognition. *Consciousness and Cognition*, *36*, 373–375.

de Gardelle, V., Waszczuk, M., Egner, T., & Summerfield, C. (2013). Concurrent repetition enhancement and suppression responses in extrastriate visual cortex. *Cereb. Cortex*, *23*(9), 2235–2244.

de Lafuente, V., Jazayeri, M., & Shadlen, M. N. (2015). Representation of accumulating evidence for a decision in two parietal areas. *J. Neurosci.*, *35*(10), 4306–4318.

Deneve, S. (2008). Bayesian spiking neurons I: Inference. *Neural Comput.*, *20*(1), 91–117.

Feldman, A. G. (2009). New insights into action-perception coupling. *Exp. Brain. Res.*, *194*(1), 39–58.

Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, *299*(5614), 1898–1902.

FitzGerald, T., Dolan, R., & Friston, K. (2014). Model averaging, optimal inference, and habit formation. *Front. Hum. Neurosci.*, *8*, 457. doi:10.3389/fnhum.2014.00457

FitzGerald, T. H., Dolan, R. J., & Friston, K. (2015). Dopamine, reward learning, and active inference. *Front. Comput. Neurosci.*, *9*, 136.

FitzGerald, T. H., Moran, R. J., Friston, K. J., & Dolan, R. J. (2015). Precision and neuronal dynamics in the human posterior parietal cortex during evidence accumulation. *Neuroimage*, *107*, 219–228.

FitzGerald, T. H., Schwartenbeck, P., Moutoussis, M., Dolan, R. J., & Friston, K. (2015). Active inference, evidence accumulation, and the urn task. *Neural Comput.*, *27*(2), 306–328.

Frank, M. J., Scheres, A., & Sherman, S. J. (2007). Understanding decision-making deficits in neurological conditions: Insights from models of natural action selection. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, *362*(1485), 1641–1654.

Friederici, A. D. (2005). Neurophysiological markers of early language acquisition: From syllables to sentences. *Trends Cogn. Sci.*, *9*(10), 481–488.

Friston, K. (2012). A free energy principle for biological systems. *Entropy*, *14*, 2100–2121.

Friston, K. (2013). Life as we know it. *J. R. Soc. Interface.*, *10*(86), 20130475.

Friston, K., Adams, R., & Montague, R. (2012). What is value—accumulated reward or evidence? *Frontiers in Neurorobotics*, *6*, 11.

Friston, K., & Buzsaki, G. (2016). The functional anatomy of time: What and when in the brain. *Trends Cogn. Sci.*, *20*, 500–511. doi:10.1016/j.tics.2016.05.001

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O'Doherty, J., & Pezzulo, G. (2016). Active inference and learning. *Neurosci. Biobehav. Rev.*, *68*, 862–879. doi:10.1016/j.neubiorev.2016.06.022

Friston, K., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *J. Physiol. Paris.*, *100*(1–3), 70–87.

Friston, K., Mattout, J., & Kilner, J. (2011). Action understanding and active inference. *Biol. Cybern.*, *104*, 137–160.

Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., & Penny, W. (2007). Variational free energy and the Laplace approximation. *NeuroImage*, *34*(1), 220–234.

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015). Active inference and epistemic value. *Cogn. Neurosci.*, *6*, 187–214.

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., & Dolan, R. J. (2014). The anatomy of choice: Dopamine and decision-making. *Philos. Trans. R. Soc. Lond.—B. Biol. Sci.*, *369*, 20130481.

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., Raymond, R. J., & Dolan, J. (2013). The anatomy of choice: Active inference and agency. *Front. Hum. Neurosci.*, *7*, 598.

George, D., & Hawkins, J. (2009). Towards a mathematical theory of cortical micro-circuits. *PLoS Comput. Biol.*, *5*(10), e1000532.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annu. Rev. Neurosci.*, *30*, 535–574.

Grosmark, A. D., & Buzsaki, G. (2016). Diversity in neural firing dynamics supports both rigid and learned hippocampal sequences. *Science*, *351*(6280), 1440–1443.

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., & Garrod, S. (2013). Speech rhythms and multiplexed oscillatory sensory coding in the human brain. *PLoS Biol*. *11*(12), e1001750. doi:10.1371/journal.pbio.1001752

Hohwy, J. (2013). *The predictive mind*. Oxford: Oxford University Press.

Hohwy, J. (2014). The self-evidencing brain. *Noûs*, *50*(2). doi:10.1111/nous.1262

Howard, R. (1966). Information value theory. *IEEE Transactions on Systems, Science and Cybernetics*, *SSC-2*(1), 22–26.

Huk, A. C., & Shadlen, M. N. (2005). Neural activity in macaque parietal cortex reflects temporal integration of visual motion signals during perceptual decision making. *J. Neurosci.*, *25*(45), 10420–10436.

Humphries, M. D., Khamassi, M., & Gurney, K. (2012). Dopaminergic control of the exploration-exploitation trade-off via the basal ganglia. *Front. Neurosci.*, *6*, 9.

Humphries, M. D., Wood, R., & Gurney, K. (2009). Dopamine-modulated dynamic cell assemblies generated by the GABAergic striatal microcircuit. *Neural Netw.*, *22*(8), 1174–1188.

Hunt, L. T., Kolling, N., Soltani, A., Woolrich, M. W., Rushworth, M. F., & Behrens, T. E. (2012). Mechanisms underlying cortical activity during value-guided choice. *Nat. Neurosci.*, *15*(3), 470–476, s471–s473.

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Res.*, *49*(10), 1295–1306.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review Series II*, *106*(4), 620–630.

Jensen, O., Gips, B., Bergmann, T. O., & Bonnefond, M. (2014). Temporal coding organized by coupled alpha and gamma oscillations prioritize visual processing. *Trends Neurosci.*, *37*(7), 357–369.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291.

Kass, R. E., & Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *J. Am. Stat. Assoc.*, *407*, 717–726.

Kira, S., Yang, T., & Shadlen, M. N. (2015). A neural implementation of Wald's sequential probability ratio test. *Neuron*, *85*(4), 861–873.

Klyubin, A. S., Polani, D., & Nehaniv, C. I. (2005). Empowerment: A universal agent-centric measure of control. In *Proc. CEC 2005. IEEE* (vol. 1, pp. 128–135). Piscataway, NJ: IEEE.

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

Knutson, B., & Bossaerts, P. (2007). Neural antecedents of financial decisions. *Journal of Neuroscience*, *27*(31), 8174–8177.

Krebs, R. M., Schott, B. H., Schütze, H., & Düzel, E. (2009). The novelty exploration bonus and its attentional modulation. *Neuropsychologia*, *47*, 2272–2281.

Latimer, K. W., Yates, J. L., Meister, M. L., Huk, A. C., & Pillow, J. W. (2015). Neuronal Modeling: Single-trial spike trains in parietal cortex reveal discrete steps during decision-making. *Science*, *349*(6244), 184–187.

Laughlin, S. B. (2001). Efficiency and complexity in neural coding. *Novartis Found. Symp.*, *239*, 177–187.

Linsker, R. (1990). Perceptual neural organization: Some approaches based on network models and information theory. *Annu. Rev. Neurosci.*, *13*, 257–281.

Lisman, J., & Buzsaki, G. (2008). A neural coding scheme formed by the combined function of gamma and theta oscillations. *Schizophr. Bull.*, *34*(5), 974–980.

Lisman, J., & Redish, A. D. (2009). Prediction, sequences and the hippocampus. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, *364*(1521), 1193–1201.

Maffongelli, L., Bartoli, E., Sammler, D., Kolsch, S., Campus, C., Olivier, . . . D'Ausilio, A., (2015). Distinct brain signatures of content and structure violation during action observation. *Neuropsychologia*, *75*, 30–39.

Mannella, F., & Baldassarre, G. (2015). Selection of cortical dynamics for motor behaviour by the basal ganglia. *Biological Cybernetics*, *109*(6), 575–595.

Mirza, M. B., Adams, R. A., Mathys, C. D., & Friston, K. J. (2016). Scene construction, visual foraging and active inference. *Frontiers in Computational Neuroscience*, *10*, 56. doi:10.3388/fncom.2016.00056

Moser, E. I., Kropff, E., & Moser, M. B. (2008). Place cells, grid cells, and the brain's spatial representation system. *Annu. Rev. Neurosci.*, *31*, 69–89.

Moser, M. B., Rowland, D. C., & Moser, E. I. (2015). Place cells, grid cells, and memory. *Cold Spring Harb. Perspect. Biol.*, *7*(2), a021808.

Moutoussis, M., Trujillo-Barreto, N. J., El-Deredy, W., Dolan, R. J., & Friston, K. J. (2014). A formal model of interpersonal inference. *Front. Hum. Neurosci.*, *8*, 160.

Mushiake, H., Saito, N., Sakamoto, K., Itoyama, Y., & Tanji, J. (2006). Activity in the lateral prefrontal cortex reflects multiple steps of future events in action plans. *Neuron*, *50*, 631–641.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*, 607–609.

O'Regan, J., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behav. Brain. Sci.*, *24*, 939–973.

Ortega, P. A., & Braun, D. A. (2013). Thermodynamics as a theory of decision-making with information-processing costs. *Proc. R. Soc. A*, *469*, 2153.

Penny, W., Zeidman, P., & Burgess, N. (2013). Forward and backward inference in spatial cognition. *PLoS Comput. Biol.*, *9*(12), e1003383.

Pezzulo, G., Rigoli, F., & Chersi, F. (2013). The mixed instrumental controller: Using value of information to combine habitual choice and mental simulation. *Front. Psychol.*, *4*, 92.

Pezzulo, G., Rigoli, F., & Friston, K. (2015). Active Inference, homeostatic regulation and adaptive behavioural control. *Prog. Neurobiol.*, *134*, 17–35.

Pezzulo, G., van der Meer, M. A., Lansink, C. S., & Pennartz, C. M. (2014). Internally generated sequences in learning and executing goal-directed behavior. *Trends Cogn. Sci.*, 647–657.

Preuschoff, K., Quartz, S. R., & Bossaerts, P. (2008). Human insula activation reflects risk prediction errors as well as risk. *Journal of Neuroscience*, *28*(11), 2745–2752.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.*, *2*(1), 79–87.

Rigoli, F., Friston, K. J., & Dolan, R. J. (2016). Neural processes mediating contextual influences on human choice behaviour. *Nat. Commun.*, *7*, 12416.

Santangelo, V. (2015). Forced to remember: When memory is biased by salient information. *Behav. Brain Res.*, *283*, 1–10.

Schmidhuber, J. (1991). Curious model-building control systems. In *Proc. International Joint Conference on Neural Networks* (vol. 2, pp. 1458–1463). Piscataway, NJ: IEEE.

Schultz, W., Apicella, P., & Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *Journal of Neuroscience*, *13*, 900–913.

Schultz, W., & Dickinson, A. (2000). Neuronal coding of prediction errors. *Annu. Rev. Neurosci.*, *23*, 473–500.

Schwartenbeck, P., FitzGerald, T., Dolan, R. J., & Friston, K. (2013). Exploration, novelty, surprise, and free energy minimization. *Front. Psychol.*, *4*, 710.

Schwartenbeck, P., FitzGerald, T. H., Mathys, C., Dolan, R., & Friston, K. (2015). The dopaminergic midbrain encodes the expected certainty about desired outcomes. *Cereb. Cortex*, *25*(10), 3434–3445.

Schwartenbeck, P., FitzGerald, T. H., Mathys, C., Dolan, R., Kronbichler, M., & Friston, K. (2015). Evidence for surprise minimization over value maximization in choice behavior. *Sci. Rep.*, *5*, 16575.

Schwartenbeck, P., FitzGerald, T. H., Mathys, C., Dolan, R., Wurst, F., Kronbichler, M., & Friston, K. (2015). Optimal inference with suboptimal models: Addiction and active Bayesian inference. *Med. Hypotheses*, *84*(2), 109–117.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends Cogn. Sci.*, *17*(11), 565–573.

Solway, A., & Botvinick, M. (2012). Goal-directed decision making as probabilistic inference: A computational framework and potential neural correlates. *Psychol. Rev.*, *119*, 120–154.

Specht, K. (2014). Neuronal basis of speech comprehension. *Hear. Res.*, *307*, 121–135.

Srihasam, K., Bullock, D., & Grossberg, S. (2009). Target selection by the frontal cortex during coordinated saccadic and smooth pursuit eye movements. *J. Cogn. Neurosci.*, *21*(8), 1611–1627.

Stam, C. J. (2005). Nonlinear dynamical analysis of EEG and MEG: Review of an emerging field. *Clin. Neurophysiol.*, *116*(10), 2266–2301.

Stauffer, W. R., Lak, A., & Schultz, W. (2014). Dopamine reward prediction error responses reflect marginal utility. *Curr. Biol.*, *24*(21), 2491–2500.

Still, S., & Precup, D. (2012). An information-theoretic approach to curiosity-driven reinforcement learning. *Theory Biosci.*, *131*(3), 139–148.

van den Broek, J. L., Wiegerinck, W. A. J. J., & Kappen, H. J. (2010). Risk-sensitive path integral control. *UAI*, *6*, 1–8.

van der Meer, M., Kurth-Nelson, Z., & Redish, A. D. (2012). Information processing in decision-making systems. *Neuroscientist*, *18*(4), 342–359.

Wittmann, B. C., Daw, N. D., Seymour, B., & Dolan, R. J. (2008). Striatal activity underlies novelty-based choice in humans. *Neuron*, *58*(6), 967–973.

Yedidia, J. S., Freeman, W. T., & Weiss, Y. (2005). Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, *51*(7), 2282–2312.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends Cogn. Sci.*, *10*(7), 301–308.

Zak, P. J. (2004). Neuroeconomics. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, *359*(1451), 1737–1748.

Zelinsky, G. J., & Bisley, J. W. (2015). The what, where, and why of priority maps and their interactions with visual working memory. *Ann. N. Y. Acad. Sci.*, *1339*, 154–164.

Zhang, H., & Maloney, L. T. (2012). Ubiquitous log odds: A common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in Neuroscience*, *6*, 1.

Received May 15, 2016; accepted August 29, 2016.

This article has been cited by:

1. Francesco Donnarumma, Marcello Costantini, Ettore Ambrosini, Karl Friston, Giovanni Pezzulo. 2017. Action perception as hypothesis testing. *Cortex* . [CrossRef]