PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

# The anatomy of choice: dopamine and decision-making

Karl Friston, Philipp Schwartenbeck, Thomas FitzGerald, Michael Moutoussis, Timothy Behrens and Raymond J. Dolan

*Phil. Trans. R. Soc. B* 2014 **369**, 20130481, published 29 September 2014

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

rstb.royalsocietypublishing.org

Opinion piece

**Cite this article:** Friston K, Schwartenbeck P, FitzGerald T, Moutoussis M, Behrens T, Dolan RJ. 2014 The anatomy of choice: dopamine and decision-making. *Phil. Trans. R. Soc. B* **369**: 20130481.
http://dx.doi.org/10.1098/rstb.2013.0481

One contribution of 18 to a Theme Issue 'The principles of goal-directed decision-making: from neural mechanisms to computation and robotics'.

**Subject Areas:**
behaviour

**Keywords:**
active inference, agency, Bayesian inference, bounded rationality, free energy, utility theory

**Author for correspondence:**
Karl Friston
e-mail: k.friston@ucl.ac.uk

# The anatomy of choice: dopamine and decision-making

Karl Friston[1], Philipp Schwartenbeck[1], Thomas FitzGerald[1], Michael Moutoussis[1], Timothy Behrens[1,2] and Raymond J. Dolan[1]

[1]The Wellcome Trust Centre for Neuroimaging, University College London, 12 Queen Square, London WC1N 3BG, UK
[2]Centre for Functional MRI of the Brain, The John Radcliffe Hospital, Headley Way, Oxford OX3 9DU, UK

This paper considers goal-directed decision-making in terms of embodied or active inference. We associate bounded rationality with approximate Bayesian inference that optimizes a free energy bound on model evidence. Several constructs such as expected utility, exploration or novelty bonuses, softmax choice rules and optimism bias emerge as natural consequences of free energy minimization. Previous accounts of active inference have focused on *predictive coding*. In this paper, we consider *variational Bayes* as a scheme that the brain might use for approximate Bayesian inference. This scheme provides formal constraints on the computational anatomy of inference and action, which appear to be remarkably consistent with neuroanatomy. Active inference contextualizes optimal decision theory within embodied inference, where goals become prior beliefs. For example, expected utility theory emerges as a special case of free energy minimization, where the *sensitivity* or inverse temperature (associated with softmax functions and quantal response equilibria) has a unique and Bayes-optimal solution. Crucially, this sensitivity corresponds to the *precision* of beliefs about behaviour. The changes in precision during variational updates are remarkably reminiscent of empirical dopaminergic responses—and they may provide a new perspective on the role of dopamine in assimilating reward prediction errors to optimize decision-making.

## 1. Introduction

This paper considers decision-making and action selection as variational Bayesian inference. It tries to place heuristics in decision theory (in psychology) and expected utility theory (in economics) within the setting of embodied or active inference. In brief, we treat the problem of selecting behavioural sequences or policies as an inference problem. We assume that policies are selected under the prior belief that they minimize the difference (relative entropy) between a probability distribution over states that can be reached and states that agents believe they should occupy. In other words, choices are based upon beliefs about alternative policies, where the most likely policy minimizes the difference between attainable and desired outcomes. By formulating the problem in this way, three important aspects of optimal decision-making emerge.

First, because relative entropy can always be decomposed into entropy and expected utility, the ensuing choices necessarily maximize both expected utility and the entropy over final states. This is closely related to maximizing extrinsic and intrinsic rewards in embodied cognition and artificial intelligence. In this setting, utility or *extrinsic reward* is supplemented with *intrinsic reward* to ensure some efficient information gain, exploratory behaviour or control over outcomes. Important examples here include artificial curiosity [1], empowerment [2], information to go [3], computational complexity [4] and self-organization in non-equilibrium systems [5]. In the current setting, a policy that maximizes the entropy over final states is intrinsically rewarding because it keeps 'options open'.



Royal Society Publishing

Second, because choices are based upon beliefs about policies, these beliefs must be associated with a confidence or precision—that is itself optimized. This furnishes a unique and Bayes-optimal sensitivity or inverse temperature of the sort associated with softmax choice rules and quantal response equilibria (QRE) [6].

Third, because beliefs about policies depend upon beliefs about the current state of the world, and vice versa, there is an inevitable optimism bias [7] in which inferences about ambiguous states are biased towards those that support an optimal policy [8].

We motivate the premises that underlie this formulation and unpack its implications using formal arguments and simulations. These simulations are described in detail in a technical companion paper [8]. The novel contribution of this work is the notion that the brain might use variational Bayes for approximate Bayesian inference—and that this variational scheme provides constraints on the computational anatomy of inference and action. In particular, variational Bayes specifies a unique and optimal precision, where Bayesian updates of expected precision (or confidence about desired outcomes) look very much like dopaminergic responses—providing a new interpretation of dopamine that goes beyond reporting reward prediction errors.

The basic idea behind active inference is that behaviour can be understood in terms of inference: in other words, action and perception are integral parts of the same inferential process and one can only be understood in light of the other. It is fairly straightforward to show that self-organizing systems are necessarily inferential in nature [9]. This notion dates back to Helmholtz and Ashby [10–12] and has been formalized recently as minimizing a variational free energy bound on Bayesian model evidence [13,14]. A corollary of this active inference scheme is that agents must perform some form of *Bayesian inference*. Bayesian inference can be approximate or exact, where exact inference is rendered tractable by making plausible assumptions about the approximate form of probabilistic representations—representations that are used to predict responses to changes in the sensorium. The key question, from this perspective, is how do agents perform approximate Bayesian inference? This contrasts with utilitarian and normative accounts of behaviour, which ask how agents maximize some expected value or utility function of their states [15–17].

Normative approaches assume that *perfectly rational* agents maximize value [18], without considering the cost of optimizing behaviour [19]. By contrast, *bounded rational* agents consider processing costs and do not necessarily choose the most valuable option [20]. Most attempts to formalize bounded rationality focus on the Boltzmann distribution, where optimal behaviour involves choosing states with a high value or low energy [4,21]. For example, QRE models assume that choice probabilities are prescribed by a Boltzmann distribution and that rationality is determined by a *temperature* parameter [6,22]. Related stochastic choice rules have a long history in psychology and economics, particularly in the form of logit choice models [23,24]. These choice rules are known as *softmax rules* and are used to describe stochastic sampling of actions, particularly in the context of the exploration–exploitation dilemma [25,26]. In this setting, the temperature models the *sensitivity* of stochastic choices to value, where perfect rationality corresponds to a very high sensitivity (low temperature). The purpose of this paper is to suggest that sensitivity can itself be optimized and corresponds to the confidence or precision associated with beliefs about the consequences of choices.

In active inference, there is no value function—free energy is the only quantity that is optimized. In this context, bounded rationality is an emergent feature of free energy minimization and the value of a state is a consequence of behaviour producing that state, not its cause. In other words, the consequences of minimizing free energy are that some states are occupied more frequently than others—and these states are valuable. Crucially, in active inference, parameters like sensitivity or inverse temperature must themselves minimize free energy. This means that sensitivity ceases to be a free parameter that is adjusted to describe observed behaviour and becomes diagnostic of the underlying (approximate) Bayesian inference scheme. We will see that sensitivity corresponds to the *precision* of beliefs about the future and behaves in a way that is remarkably similar to the firing of dopaminergic cells in the brain. Furthermore, QRE, logit choice models and softmax rules can be derived as formal consequences of free energy minimization, using variational Bayes.

Variational Bayes or ensemble learning is a general and widely used scheme for approximate Bayesian inference [27]. It rests on a partition of probabilistic representations (approximate posterior probability distributions) that renders Bayesian inference tractable. A simple example would be estimating the mean and precision (inverse variance) of some data, under the approximating assumption that uncertainty about the mean does not depend upon uncertainty about the variance and vice versa. This enables a straightforward computation of descriptive statistics that would otherwise be extremely difficult (see [28] for details). Neurobiologically, a partition into conditionally independent representations is nothing more than functional segregation—in which specialized neuronal systems can be regarded as performing variational Bayesian updates by passing messages to each other. This paper tries to relate variational Bayes to the functional anatomy of inference and action selection in the brain. This provides a functional account of neuronal representations and functional integration (message passing) among different systems. A particularly important example will be the exchange of signals among systems encoding posterior beliefs about precision with systems representing hidden states of the world and action, respectively—an exchange we associate with the convergent control of dopaminergic firing and its divergent influence on Bayesian updates in the prefrontal cortex and striatum.

Although variational Bayes uses discrete updates, variational updates still possess a dynamics that can be compared to neuronal responses, particularly dopaminergic responses. We focus on this comparison because understanding the computational role of dopamine is important for understanding the psychopathology and pathophysiology of conditions such as Parkinson's disease, schizophrenia and autism. Traditionally, dopamine has been associated with the reporting of reward prediction errors [29]. However, this may provide an incomplete account of dopamine, because it fails to account for its putative role in action (e.g. the bradykinesia of Parkinson's disease) and perception (e.g. hallucinations and delusions in schizophrenia). Much of current thinking in computational psychiatry points to dopamine as mediating a representation of uncertainty or precision that can account for both false inference [30–33] and impoverished action [34]. In what follows, we will see how precision relates to value and thereby resolves the

dialectic between the role of dopamine in reporting reward prediction errors and as a neuromodulator of action and attentional selection [35,36].

This paper comprises three sections: §2 introduces active inference and describes a general model of control or agency, in which purposeful behaviour rests on prior beliefs that agents will minimize the (relative) entropy of their final states. This leads naturally to expected utility theory and exploration bonuses. §3 considers the inversion of the generative model using variational Bayes, with a special focus on belief updates and message passing. §4 considers the implications for the functional anatomy of inference and decision-making, namely reciprocal message passing between systems supporting perceptual inference, action selection and the encoding of uncertainty or precision.

## 2. Active inference

This section introduces active inference, in which beliefs about (hidden or fictive) states of the world maximize model evidence or the marginal likelihood of observations. In contrast to classic formulations, active inference makes a distinction between *action* that is a physical state of the real world and beliefs about action that we will refer to as *control* states. This changes the problem fundamentally from selecting an optimal action to making optimal inference about control. In other words, under the assumption that action is sampled from posterior beliefs about control, we can treat decision-making and action selection as a pure inference problem that necessarily entails optimizing beliefs about behaviour and its consequences. Sampling actions from posterior beliefs is known as Thompson sampling [37,38]; see [38] which is especially relevant as it provides a free energy derivation.

The following summarizes the material in ref. [8]. We use bold-italic typeface to indicate true states of the world and italic typeface for hidden or fictive states assumed by an agent. The parameters (expectations) of categorical distributions over discrete states $s \in \{1, \ldots, J\}$ are denoted by $J \times 1$ vectors $\hat{s} \in [0, 1]$, while the $\sim$ notation denotes sequences of variables over time.

**Definition.** Active inference rests on the tuple $(\Omega, \boldsymbol{S}, A, P, P, Q, R, S, U)$

- A finite set of observations $\Omega$.
- A finite set of true states and actions $\boldsymbol{S} \times A$.
- A finite set of fictive or hidden states $S \times U$.
- A *generative process* over observations, states and action $R(\tilde{o}, \tilde{\boldsymbol{s}}, \tilde{a}) = \Pr(\{o_0, \ldots, o_t\} = \tilde{o}, \{\boldsymbol{s}_0, \ldots, \boldsymbol{s}_t\} = \tilde{\boldsymbol{s}}, \{a_0, \ldots, a_t\} = \tilde{a})$.
- A *generative model* over observations and hidden states $P(\tilde{o}, \tilde{s}, \tilde{u}|m) = \Pr(\{o_0, \ldots, o_t\} = \tilde{o}, \{s_0, \ldots, s_t\} = \tilde{s}, \{u_t, \ldots, u_T\} = \tilde{u})$.
- An *approximate posterior probability* over hidden states with expectations $\mu \in \mathbb{R}^d$ such that $Q(\tilde{s}, \tilde{u}|\mu) = \Pr(\{s_0, \ldots, s_t\} = \tilde{s}, \{u_t, \ldots, u_T\} = \tilde{u})$.

**Remarks**. In this set-up, the *generative process* describes transitions among real states of the world that depend upon action and generate outcomes. This process models the environment that the agent samples through action. Actions are sampled from approximate posterior beliefs based on a *model* of the generative process. In the generative model, actions $A$ are replaced by control states $U$. The generative model is embodied by an agent (denoted by $m$) that is coupled to the environment through observations (sampled from the generative process) and actions (sampled from its posterior beliefs). Finally, approximate posterior beliefs about hidden states $S \times U$ are encoded by expectations $\mu \in \mathbb{R}^d$.

As it stands, this definition does not describe a process. This is because the dependencies among real states and expectations are not specified. In other words, the agent's generative model of observations $P(\tilde{o}, \tilde{s}, \tilde{u}|m)$ and its approximate posterior distribution over their causes $Q(\tilde{s}, \tilde{u}|\mu)$ does not refer to the process of eliciting outcomes through action $R(\tilde{o}, \tilde{\boldsymbol{s}}, \tilde{a})$. To couple the agent to its environment, we have to specify how its expectations depend upon observations and how its action depends upon expectations. In active inference, the expectations minimize free energy and the ensuing beliefs about control states prescribe action

$$\left.\begin{aligned} \mu_t &= \arg\min_\mu F(\tilde{o}, \mu) \\ \Pr(a_t = u_t) &= Q(u_t|\mu_t) \\ F(\tilde{o}, \mu) &= D_{\mathrm{KL}}[Q(\tilde{s}, \tilde{u}|\mu)||P(\tilde{s}, \tilde{u}|\tilde{o})] - \ln P(\tilde{o}|m). \end{aligned}\right\} \tag{2.1}$$

In summary, the environment is characterized as a distribution $R(\tilde{o}, \tilde{\boldsymbol{s}}, \tilde{a})$ over observations, true states and action, whereas the agent is characterized by two distributions: a generative model $P(\tilde{o}, \tilde{s}, \tilde{u}|m)$ that connects observations to hidden states and posterior beliefs about those states $Q(\tilde{s}, \tilde{u}|\mu)$ parametrized by its expectations. True states control environmental responses but are never observed directly. Instead, the agent infers hidden states based on its observations. Crucially, these hidden states include control states that prescribe action. Here, the generative model plays a dual role—it is a predictive model over observations and encodes optimal policies (in terms of prior beliefs about control states). The agent and the environment interact in cycles. In each cycle, the agent first figures out which hidden states are most likely by optimizing its expectations with respect to the free energy of observations. After optimizing its posterior beliefs, an action is sampled from the posterior marginal over control states. The environment then picks up this action, generates a new observation and a new cycle starts.

The optimization above is that usually portrayed in terms of *perception* (inference about hidden states) and *action* (a choice model in which action is a function of inferred states). Action and perception couple the agent to the environment; where expectations depend upon observations—through perception, whereas observations depend upon expectations—through action. Usually, expectations are associated with neuronal activity or connection strengths and action is associated with the state of effectors. In brief, expectations about the state of the world minimize free energy, while action is selected from the ensuing posterior beliefs about control states.

The expression for free energy above shows that it upper bounds the negative logarithm of Bayesian model evidence $-\ln P(\tilde{o}|m)$ or *surprise*. This is because the relative entropy or Kullback–Leibler (KL) divergence term cannot be less than zero [27]. This means minimizing free energy corresponds to minimizing the divergence between the approximate and true posterior. This formalizes the notion of unconscious

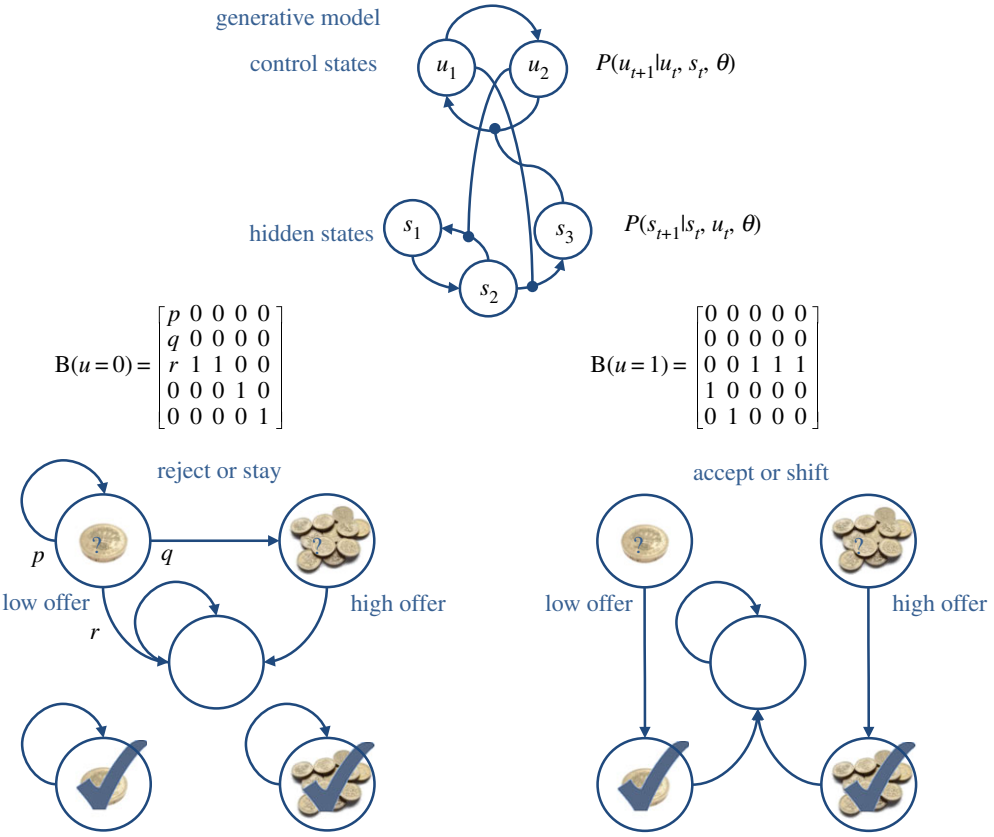


**Figure 1.** *Upper panel*: a schematic of a hierarchical generative model with discrete states. The key feature of this model is that it entertains a subset of hidden states called control states. The transitions among one subset depend upon the state occupied in the other. *Lower panels*: this provides an example of a particular model with two control states; reject (stay) or accept (shift). The control state determines transitions among hidden states that comprise a low offer (first state), a high offer (second state), a no-offer state (third state) and absorbing states that are entered whenever a low (fourth state) or high (fifth state) offer is accepted. The probability of moving from one state to another is unity, unless otherwise specified by the transition probabilities shown in the middle row. The (hazard rate) parameter $r$ controls the rate of offer withdrawal. Note that absorbing states—that re-enter themselves with unit probability—render this Markovian process irreversible. We will use this example in later simulations of choice behaviour.

inference in perception [10,39,13] and, under some simplifying assumptions, reduces to predictive coding [40].

In summary, minimizing free energy corresponds to approximate Bayesian inference and, in active inference, choosing the least surprising outcomes. However, if agents model their environment, they have to entertain posterior beliefs about the control of state transitions producing outcomes. This means that we have moved beyond classical formulations—in which deterministic actions are selected—and have to consider posterior beliefs about putative choices. In §2*a*, we consider the optimization of posterior beliefs and the confidence or precision with which these beliefs are held.

## (a) A generative model of goal-directed agency

Surprise or model evidence is an attribute of a generative model. This model comprises prior beliefs that determine the states an agent frequents. It is these beliefs that specify the attracting states (goals) that action will seek out—to avoid surprise. We now consider how prior beliefs can be understood in terms of expected utility.

The models we consider rest on transitions among hidden states that are coupled to transitions among control states. This coupling is illustrated in the upper panel of figure 1. Here, control states modify the transition probabilities among hidden states, while hidden states modify the transitions among control states (as denoted by the connections ending with circles). This form of model allows context-sensitive-state transitions among states generating outcomes—that themselves can induce changes in the control states providing that context. The lower panels of figure 1 depict a particular example that we will use later.

The generative model used to model these (finite horizon Markovian) processes can be expressed in terms of the following likelihood and prior distributions over observations and hidden states to time $t \in (0, \ldots, T)$ and subsequent control states (omitting normalization constants)

$$\left.\begin{aligned} P(\tilde{o}, \tilde{s}, \tilde{u}, \gamma|\tilde{a}, m) &= P(\tilde{o}|\tilde{s})P(\tilde{s}, \tilde{u}|\gamma, \tilde{a})P(\gamma|m) \\ P(\tilde{o}|\tilde{s}) &= P(o_0|s_0)P(o_1|s_1)\ldots P(o_t|s_t) \\ P(\tilde{s}, \tilde{u}|\gamma, \tilde{a}) &= P(\tilde{u}|s_t, \gamma)P(s_t|s_{t-1}, a_{t-1})\ldots P(s_1|s_0, a_0)P(s_0|m) \\ \ln P(\tilde{u}|s_t, \gamma) &= \gamma \cdot \mathbf{Q} \\ \mathbf{Q}(\tilde{u}|s_t) &= -D_{\mathrm{KL}}[P(s_T|s_t, \tilde{u})||P(s_T|m)]. \end{aligned}\right\} \tag{2.2}$$

The first equality expresses the generative model in terms of the likelihood of observations given the hidden states (first term), while subsequent terms represent *empirical* prior beliefs. Empirical priors are just probability distributions over unknown variables that depend on other unknown variables. The likelihood says that observations depend on, and only on,

concurrent hidden states. The third equality expresses beliefs about state transitions that embody Markovian dependencies among successive hidden states. For simplicity, we have assumed that the agent knows its past actions by observing them. The important part of this generative model lies in the last equalities—describing prior beliefs about control sequences or *policies* that determine which action is selected next.

These beliefs take the form of a Boltzmann distribution, where the policy with the largest prior probability minimizes the relative entropy or divergence between the distribution over final states—given the current state and policy—and the marginal distribution over final states. This marginal distribution defines the agent's *goals* in terms of (desired) states the agent believes it should end up in. One can interpret the negative divergence $\mathbf{Q}(\tilde{u}|s_t)$ as the *value* of policies available from the current state. In other words, a valuable policy minimizes divergence between expected and desired states. We use $\mathbf{Q}(\tilde{u}|s_t)$ in analogy with action-value in Q-learning [41].

Crucially, the precision of beliefs about policies is determined by a hidden variable $\gamma \in \mathbb{R}^+$ that has to be inferred. We will see in §3 that the expected precision minimizes variational free energy in exactly the same way as expectations about hidden states. There is nothing mysterious about this: we estimate the precision of estimators—by minimizing variational free energy or marginal likelihood—in everyday data analysis when estimating their standard error. In our setting, the expected precision reflects the confidence that goals can be reached from current state. In other words, it encodes the confidence placed in beliefs about optimal outcomes, given observations to date. Note that expected precision is context sensitive and, unlike classical sensitivity parameters, changes with each observation. In summary, this model represents past hidden states and future choices, under the belief that controlled transitions from the current state will minimize the divergence between the distribution over final states and desired states.

## (b) Prior beliefs, entropy and expected utility

Basing beliefs about choices on relative entropy is formally related to KL optimization; particularly, risk sensitive control (e.g. [42]). This is also a cornerstone of utility-based free energy treatments of bounded rationality [4,21]. These schemes consider optimal agents to minimize the KL divergence between controlled and desired outcomes. All we have done here is to equip agents with prior beliefs that they are KL optimal. These beliefs are then enacted through active inference. The advantage of doing this is that the precision of beliefs about control (i.e. sensitivity to value) can now be optimized—because we have cast optimal control as an inference problem. These arguments may seem a bit abstract but, happily, familiar notions like exploration, exploitation and expected utility emerge as straightforward consequences.

The KL divergence can be thought of as a prediction error—not between expected and observed outcomes—but between the final outcomes predicted with and without considering the current state. In other words, the difference between what can be attained from the current state and the goals encoded by prior beliefs. Unlike classic reward prediction errors, this prediction error is a divergence between probability distributions over states, as opposed to a scalar function of states. Value is the complement of this divergence, which means that the value of the current state decreases when a previously predicted reward can no longer be reached from the current state.

Mathematically, value can be decomposed into two terms that have an important interpretation

$$\mathbf{Q}(\tilde{u}|s_t) = \underbrace{H[P(s_T|s_t, \tilde{u})]}_{\text{exploration bonus}} + \sum_{s_T} \underbrace{P(s_T|s_t, \tilde{u})c(s_T|m)}_{\text{expected utility}}. \tag{2.3}$$

The first is the entropy (intrinsic reward) of the distribution over final states, given the current state and policy. The second is the expected *utility* of the final state, where utility (extrinsic reward) or negative cost is the log probability of the final state under the priors encoding goals: $c(s_T|m) = \ln P(s_T|m)$.

This decomposition means that agents (believe they) will maximize the entropy of their final states while, at the same time, maximizing expected utility. The relative contribution of entropy and expected utility depends upon the relative utility of different states. If prior goals are very precise (informative), they will dominate and the agent will (believe it will) maximize expected utility. Conversely, with imprecise (flat) priors—that all final states are equally likely—the agent will keep its options open and maximize the entropy over those states: in other words, it will explore, according to the maximum entropy principle [43]. This provides a simple account of *exploration–exploitation* that is consistent with expected utility theory. The entropy term implies that (beliefs about) choices are driven not just to maximize expected value but to explore options in a way that confers an exploratory aspect on behaviour. In the absence of (or change in) beliefs about ultimate states, there will be a bias towards visiting all (low cost) states with equal probability. Similarly, the *novelty bonus* [44] of a new state is, in this formulation, conferred by the opportunity to access states that were previously unavailable—thereby increasing the entropy over final states. This means that the value of a choice comprises an exploration bonus and an expected utility, where the former depends upon the current state and the latter does not.

In summary, if agents occupy a limited set of attracting states, their generative models must be equipped with prior beliefs that controlled state transitions will minimize the divergence between a distribution over attainable states and a distribution that specifies states as attractive. These prior beliefs can be expressed in terms of a KL divergence that defines the value of policies. This value is the same objective function in KL control schemes that grandfather conventional utility-based schemes [4,45]. The value of a policy can be decomposed into its expected utility and an exploration or novelty bonus that corresponds to the entropy over final states. In this setting, notions like value, expected utility and exploration bonus are consequences of the underlying imperative to minimize (relative) entropy. The balance between exploration (entropy) and exploitation (expected value) is uniquely determined by the relative utility of future states—not by inverse temperature: the sensitivity or precision applies to both exploratory and utilitarian behaviour. In other words, explorative behaviour is not just a random version of exploitative behaviour but can itself be very precise, with a clearly defined objective (to maximize the entropy of final outcomes). We will see in §4c that precision plays a different and fundamental role in moderating an *optimism bias* when forming beliefs about hidden states

of the world [7]. First, we need to consider the form of the generative model and its inversion.

## 3. Variational Bayesian inversion

This section illustrates active inference using the generative model in §2 and its variational Bayesian inversion. To simplify notation, we represent allowable policies with $\pi \in \{1, \ldots, K\}$, where each policy prescribes a sequence of control states $(\tilde{u}|\pi) = (u_t, \ldots, u_T|\pi)$. The model considered here is parametrized as follows (omitting constants):

$$\left.\begin{aligned} P(o_t = i|s_t = j, \mathbf{A}) &= \mathbf{A}_{ij} \\ P(s_{t+1} = i|s_t = j, \pi, \mathbf{B}) &= \mathbf{B}(u_t|\pi)_{ij} \\ \ln P(\pi = i|s_t = j, \gamma, \mathbf{Q}) &= \gamma \cdot \mathbf{Q}_{ij} \\ P(s_T = i|\mathbf{c}) &= \mathbf{c}_i \\ P(s_0 = i|\mathbf{d}) &= \mathbf{d}_i \\ P(\gamma|m) &= \Gamma(\alpha, \beta) \\ P(s_T = i|s_t = j, \pi, \mathbf{c}) &= \mathbf{T}(\pi)_{ij} \\ \mathbf{T}(\pi) &= \mathbf{B}(u_t|\pi)\mathbf{B}(u_{t+1}|\pi)\ldots\mathbf{B}(u_T|\pi) \\ \mathbf{Q}_{ij} &= \ln \mathbf{c}^T \cdot \mathbf{T}(\pi = i)_j - \ln \mathbf{T}(\pi = i)_j^T \cdot \mathbf{T}(\pi = i)_j. \end{aligned}\right\} \tag{3.1}$$

The categorical distributions over observations, given the hidden states, are parametrized by the matrix $\mathbf{A}$ that maps from hidden states to outcomes. Similarly, the transition matrices $\mathbf{B}(u_t|\pi)$ encode transition probabilities from one state to the next under the current policy. The vectors $\mathbf{c}$ and $\mathbf{d}$ encode the prior distribution over the last and first states, respectively. The former specify utility $c(s_T|m) = \ln P(s_T|m) = \ln \mathbf{c}$. The prior over precision has a standard $\gamma$-distribution with shape and rate parameters (in this paper) $\alpha = 8$ and $\beta = 1$. The matrix $\mathbf{Q}$ contains the values of the $i$th policy from the $j$th hidden state and $\mathbf{T}(\pi)$ encodes the probability of transition from the current state to a final state, under a particular policy. This is simply the iterated composition of the appropriate transition matrices from the present time until the end of the game.

### (a) Approximate Bayesian inference

Having specified the generative model, we now need to find the expectations that minimize free energy. Variational Bayes provides a generic scheme for approximate Bayesian inference that finesses the combinatoric and analytic intractability of exact inference [27,46]. Variational Bayes rests on a factorization of approximate posterior beliefs that greatly reduces the number of expectations required to encode it. The particular factorization we focus on exploits the Markovian nature of the generative model and has the following form (see [8] for details).

$$\left.\begin{aligned} Q(\tilde{s}, \tilde{u}, \gamma|\mu) &= Q(s_0|\tilde{s}_0)\ldots Q(s_t|\tilde{s}_t)Q(u_t, \ldots, u_T|\hat{\pi})Q(\gamma|\hat{\gamma}) \\ Q(\gamma|\hat{\gamma}) &= \Gamma(\alpha, \hat{\beta}) \\ \hat{\beta} &= \frac{\alpha}{\hat{\gamma}} \end{aligned}\right\}. \tag{3.2}$$

This assumes a factorization over (past) hidden states, (future) control states and precision. The details of the mean field assumption above are not terribly important. The main point is that the formalism of variational Bayes allows one to specify constraints on the form of the approximate posterior that makes prior assumptions or beliefs about choices explicit. For example, in ref. [47], we used a mean field assumption where every choice could be made at every time point. Equation (3.2) assumes that the approximate marginal over precision is, like its conjugate prior, a $\gamma$-distribution—where the rate parameter is optimized. This rate parameter $\hat{\beta} = \alpha/\hat{\gamma}$ corresponds to temperature in classic formulations. However, it is no longer a free parameter but a sufficient statistic of the unknown precision of beliefs about policies.

Given the generative model in equation (3.1) and the mean field assumption in equation (3.2), the expectations can be expressed as functions of themselves [8] to produce the following remarkably simple variational updates, where $\sigma(\cdot)$ is a softmax function

$$\left.\begin{aligned} \hat{s}_t &= \sigma(\ln \mathbf{A}^T \cdot o_t + \ln \mathbf{B}(a_{t-1}) \cdot \hat{s}_{t-1} + \hat{\gamma} \cdot \mathbf{Q}^T \cdot \hat{\pi}) \\ \hat{\pi} &= \sigma(\hat{\gamma} \cdot \mathbf{Q} \cdot \hat{s}_t) \\ \hat{\gamma} &= \frac{\alpha}{\beta - \hat{\pi}T \cdot \mathbf{Q} \cdot \hat{s}_t}. \end{aligned}\right\} \tag{3.3}$$

By iterating these equalities until convergence, one obtains a solution that minimizes free energy and provides Bayesian estimates of the hidden variables. This means the expectations change over two timescales—a fast timescale that updates posterior beliefs given the current observations—and a slow timescale that updates posterior beliefs as new observations arrive and action is taken. We have speculated [47] that these updates may be related to nested electrophysiological oscillations, such as phase coupling between $\gamma$- and $\theta$-oscillations in prefrontal–hippocampal interactions [48]. This speaks to biological implementations of variational Bayes, which we now consider in terms of neuronal and cognitive processing.

## 4. The functional anatomy of decision-making

The computational form of variational Bayes resembles many aspects of neuronal processing in the brain: if we assume that neuronal activity encodes expectations, then the variational update scheme could provide a metaphor for *functional segregation*—the segregation of representations, and *functional integration*—the recursive (reciprocal) exchange of expectations during approximate Bayesian inference. In terms of the updates themselves, the expectations of hidden states and policies are softmax functions of (mixtures of) the other expectations. This is remarkable because these updates are derived from basic variational principles and yet have exactly the form of neural networks that use, integrate and fire neurons. Furthermore, the softmax functions are of linear mixtures of expectations (neuronal activity) with one key exception—the modulation by precision when updating beliefs about the current state and selecting the next action. It is tempting to equate this modulation with the neuromodulation by dopaminergic systems that send projections to (prefrontal) systems involved in working memory [49,50] and striatal systems involved in action selection [51,52]. We now consider the variational updates from a cognitive and neuroanatomical perspective (see figure 2 for a summary):

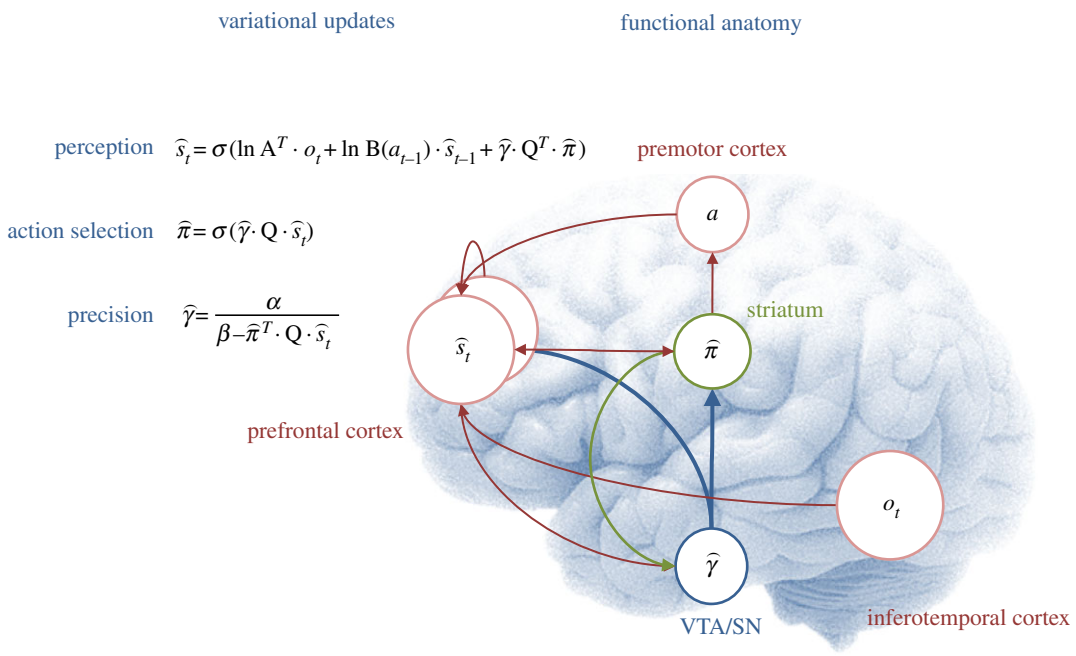


**Figure 2.** This figure illustrates the cognitive and functional anatomy implied by the variational scheme—or more precisely, the mean field assumption implicit in variational updates. Here, we have associated the variational updates of expected states with perception, of future control states (policies) with action selection and, finally, expected precision with evaluation. The updates suggest the expectations from each subset are passed among each other until convergence to an internally consistent (Bayes optimal) solution. In terms of neuronal implementation, this might be likened to the exchange of neuronal signals via extrinsic connections among functionally specialized brain systems. In this (purely iconic) schematic, we have associated perception (inference about the current state of the world) with the prefrontal cortex, while assigning action selection to the basal ganglia. Crucially, precision has been associated with dopaminergic projections from VTA and SN. See main text for a full description of the equations.

## (a) Perception

The first updates beliefs about the state of the world using observations and beliefs about the preceding state and action. However, there is a third term based upon the expected value of each state, averaged over policies. This can be regarded as an optimism bias in the sense that it biases perception towards high value states—much like dopamine [7]. Figure 2 ascribes these updates to the frontal cortex—assuming neuronal populations here encode the current state. Figure 2 should not be taken too seriously: representations of the current state could have been placed in working memory circuits in the dorsolateral prefrontal cortex [53], ventromedial prefrontal cortex or the anterior cingulate cortex, depending upon the task at hand (e.g. [54]).

## (b) Action selection

The second variational update is a softmax function of the expected value of competing choices under the current state. Figure 2 places this update in the striatum, where the expected value of a policy requires posterior beliefs about the current state from prefrontal cortex and expected precision from the ventral tegmental area (VTA). Crucially, this is the softmax choice rule that predominates in QRE and other normative models [22]. Again, it is remarkable that this utilitarian rule is mandated by the form of variational updates. However, utilitarian theories overlook the symmetry between the expected value over states—that provides the value of a choice, and the expected value over choices—that provides the value of a state. In other words, there are two expected values, one for action $\mathbf{Q} \cdot \hat{s}$ and one for perception $\mathbf{Q}^T \cdot \pi$. Finally, the expected value under choices *and* states $\hat{\pi}^T \cdot \mathbf{Q} \cdot \hat{s}_t$ specifies the optimal precision or inverse temperature. Neurobiologically, the softmax policy updates would correspond to biased competition among choices, where precision modulates the selection of competing policies (c.f. [35,36,55]).

## (c) Evaluating confidence

The final variational step estimates the precision of beliefs about policies, using expectations about hidden states and choices. We have associated expected precision with dopaminergic projections from the VTA (and substantia nigra (SN)), which receive messages from the prefrontal cortex and striatum.

The basic tenet of this scheme is that precision must be optimized. So what would happen if (estimated) precision was too high or low? If precision was zero, then perception would be unbiased and represent a veridical representation of worldly states. However, there would be a failure of action selection, in that the value of all choices would be identical. One might heuristically associate this with the pathophysiology of Parkinson's disease—that involves a loss of dopaminergic cells and a poverty of action selection. Conversely, if precision was too high, there would be a predisposition to false perceptual inference—through an augmented optimism bias. This might be a metaphor for the positive symptoms of schizophrenia, putatively associated with hyper-dopaminergic states [31]. In short, there is an optimal precision for any context and the expected precision has to be evaluated carefully on the basis of current beliefs about the state of the world.

In summary, increasing precision biases perceptual inference towards those states that are consistent with prior beliefs about future (choice-dependent) outcomes and increases the precision of action selection. Crucially, the

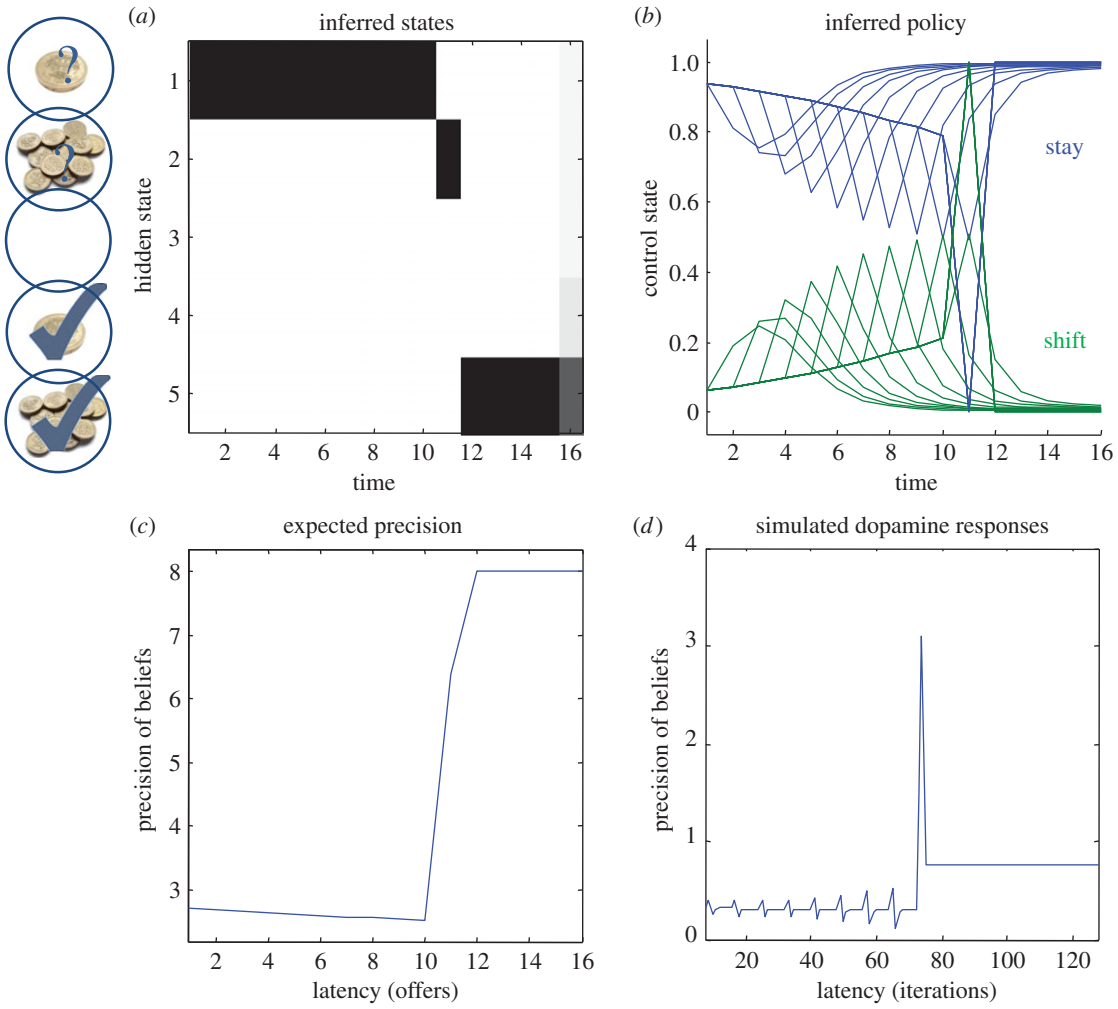


**Figure 3.** This figure shows the results of a simulation of 16 trials, where a low offer was replaced by high offer on the 11th trial, which was accepted on the subsequent trial. Panel (*a*) shows the expected states as a function of trials or time, where the states are defined in figure 1. Panel (*b*) shows the corresponding expectations about control in the future, where the dotted lines are expectations during earlier trials and the full lines correspond to expectations during the final trial. Black corresponds to reject (stay) and grey to accept (shift). Panels (*c*,*d*) show the time-dependent changes in expected precision, after convergence on each trial (*c*) and deconvolved updates after each iteration of the variational updates (*d*).

update for expected precision is an increasing function of value, expected under current beliefs about states and choices. This means that the optimal precision depends upon the attainability of goals: if a goal cannot be obtained from the current state, then precision will be low—reducing confidence in predictions about behaviour. Conversely, if there is a clear and precise path from the current state to a goal, then precision will be high. In short, precision encodes the confidence that a goal can be attained and reports the expected value—it plays a dual role in biasing perceptual inference and action selection. We will now look more closely at the neurobiology of precision and can consider not just the role of precision but also how it is controlled by the representations (posterior expectations) it optimizes.

## (d) Precision, dopamine and decision-making under uncertainty

Figure 3 shows a simulation based on the transition probabilities in figure 1 (see [8] for details). In this 'limited offer' game, the agent has to choose between a low offer—that might be withdrawn at any time—and a high offer—that may replace the low offer with some fixed probability. The problem the agent has to solve is how long to wait. If it waits too long, the low offer may be withdrawn and it will end up with nothing. Conversely, if it chooses too soon, it may miss the opportunity to accept a high offer. In this example, the low offer was replaced with a high offer on the eleventh trial, which the agent accepted. It accepts because this is most probable choice, under its prior belief that it will have accepted the higher offer by the end of the game. The expected probabilities of staying or shifting are shown in the upper right panel (in blue and green, respectively), as a function of time for each trial (thin lines) and the final beliefs (thick lines). The interesting thing here is that before the high offer, the agent believes that it will accept the low offer three or four trials in the future. Furthermore, the propensity to accept (in the future) increases with time (see dotted lines). This means that it waits, patiently, because it thinks it is more likely to accept an offer in the future than to accept the current offer.

The expected precision of these posterior beliefs is shown in the lower left panel and declines gently until the high offer is made. At this point, the expected precision increases markedly, and then remains high. This reflects the fact that the final outcome is assured with a high degree of confidence. These precisions are the expected precisions after convergence of the variational iterations. The equivalent dynamics in the lower right panel show the expected precision over all updates

in terms of simulated dopamine responses. These are a least-squares deconvolution of the variational updates, using an exponentially decaying kernel. In other words, these (simulated) dopamine responses reproduce the fluctuations in expected precision when convolved with an exponential kernel (with a time constant of eight iterations). This accounts for the postsynaptic effects of dopamine that, we imagine, decay after its release. The resulting updates show phasic responses to the arrival of new sensory information that converge to tonic values, which minimize free energy.

Many readers will have noted a similarity between the dynamics of precision and the firing of dopaminergic cells. In fact, nearly every anatomical and physiological feature of dopaminergic neurotransmission can be found in these precision updates:

— expected precision modulates the contribution of expected value during the optimization of posterior beliefs about the state of the world and action selection. This fits comfortably with the broadcasting of dopaminergic signals from the VTA and SN to the cortex (for perception) by the mesocortical system—and to the ventral striatum (for action) via nigrostriatal projections. Crucially, the mean field effects implicit in variational Bayes mandate this bilateral (perception and action) dissemination of precision (dopaminergic) signals;
— precision is updated by posterior expectations from the representations it modulates. This is consistent with the projections that control dopaminergic firing in the VTA/SN that are reciprocated by the targets of the ascending dopaminergic system. In other words, nearly every system receiving projections from the VTA projects back to it [56]. Similarly, the main input to the pars reticulata of the SN derives from medium spiny cells in the striatum via the *direct* and *indirect* pathways. These pathways originate in tightly intermingled striatal cells that express different dopamine receptors [57];
— the effect of precision is to modulate the effect of posterior expectations about the current state on control and vice versa. This modulation is exactly congruent with the postsynaptic effects of dopamine: at a synaptic level, dopamine activates G-protein-coupled receptors to modulate the cAMP second messenger system and modify the sensitivity of postsynaptic responses to presynaptic inputs;
— the modulatory role of (expected) precision effectively increases signal to noise during the competition among posterior beliefs about the state of the world (implicit in the softmax function), while doing the same for posterior beliefs about policies. Similarly, dopamine has a dual role in modulating prefrontal cortical responses in working memory circuits [58,53], while at the same time playing a key role in action selection [35,36]. This dialectic may also be reflected by the role of dopamine in schizophrenia and Parkinson's disease [33];
— precision increases monotonically with expected value, where value is composed of an exploration bonus and expected value. Similarly, dopamine is traditionally thought to report novelty, particularly in relation to action [59] and expected value in the same setting [60];
— precision shows phasic (variational update) dynamics in response to new sensory information, which converge to the expected precision. Similarly, dopamine shows characteristic phasic responses to sensory cues that predict rewards, which return to tonic firing levels that may encode uncertainty or predictability [60,61];
— precision increases whenever a predictable path to a goal is signified by sensory input. For example, the appearance of a high offer in figure 3 elicits a greater increase in precision than receipt of the offer *per se*—or its subsequent retention. Similarly, dopamine responses are elicited by sensory cues that, in higher order operant conditioning paradigms, lead to reward but thereafter 'remain uninfluenced by events that are as good as predicted' [62]. Indeed, it was the transfer of dopamine responses—from early to late conditioned stimuli—that motivated normative theories of reinforcement learning based upon temporal difference models [29]; and
— precision decreases with the withdrawal of an opportunity to fulfil prior beliefs (shown in [8]). Similarly, dopamine firing decreases in the absence of an expected reward [62].

For people familiar with discussions of dopamine in the context of active inference, the correspondence between precision and dopaminergic neurotransmission will come as no surprise—exactly the same conclusions have been reached when examining predictive coding schemes [34] and hierarchical inference using volatility models [63]. 'In brief, the emergent role of dopamine is to report the precision or salience of perceptual cues that portend a predictable sequence of sensorimotor events. In this sense, it mediates the affordance of cues that elicit motor behaviour; in much the same way that attention mediates the salience of cues in the perceptual domain.' [34, p. 1].

## 5. Conclusion

The arguments in this paper can be summarized as follows:

— optimal behaviour can be cast as a pure inference problem, in which valuable outcomes are defined in terms of prior beliefs about future states;
— exact Bayesian inference (perfect rationality) cannot be realized physically, which means that optimal behaviour rests on approximate Bayesian inference (bounded rationality);
— variational free energy provides a bound on Bayesian model evidence (marginal likelihood) that is optimized by bounded rational behaviour;
— bounded rational behaviour requires (approximate Bayesian) inference on both hidden states of the world and (future) control states. This mandates beliefs about action (control) that are distinct from action *per se*—beliefs that entail a precision;
— these beliefs can be cast in terms of minimizing the relative entropy or divergence between prior beliefs about goals and posterior beliefs, given the current state of the world and future choices;
— value can be equated with negative divergence and comprises entropy (exploration or novelty bonus) and expected utility (utilitarian) terms that account for exploratory and exploitative behaviour, respectively;
— variational Bayes provides a formal account of how posterior expectations about hidden states of the world, control states and precision depend upon each other; and may provide a metaphor for message passing in the brain;

— beliefs about the state of the world depend upon expected value over choices, whereas beliefs about choices depend upon expected value over states. Beliefs about precision depend upon expected value under both states and choices;
— precision has to be optimized to balance prior beliefs about choices and sensory evidence for hidden states. In other words, precision nuances an inherent optimism bias when inferring the current state of the world;
— variational Bayes induces distinct probabilistic representations (functional segregation) of hidden states, control states and precision—and highlights the role of reciprocal message passing. This may be particularly important for expected precision that is required for optimal inference about hidden states (perception) and control states (action selection); and
— the dynamics of precision updates, and their computational architecture, are consistent with the physiology and anatomy of the dopaminergic system—providing an account of (mesocortical) projections that encode the precision of valuable states—and (nigrostriatal) projections that encode the precision of valuable actions.

One might ask why these conclusions do not follow from normative accounts of optimal behaviour. One reason is that normative accounts do not distinguish between action and beliefs about action (control). These beliefs entail both content (expectations) and confidence (precision). This means that both expectations about behaviour and the precision of these beliefs have to be optimized. It is this optimization of precision that provides a complete account of bounded rationality (approximate Bayesian inference) and a plausible account of the control of dopaminergic firing (c.f. [64]).

Clearly, this account of dopamine does not address many important issues in the neurobiology of dopamine and its modelling. As with most free energy formulations, the objective is not to replace existing accounts but to contextualize them—usually by appealing to simpler and more fundamental imperatives. For example, we have seen that minimizing surprise (or its free energy bound) provides a principled account of goal-directed behaviour that is not biologically implausible. Furthermore, this account is consistent with many established formulations, converging on softmax choice rules, reconciling the contribution of intrinsic and extrinsic rewards and accounting for a range of anatomical and physiological properties of dopaminergic projections. Having said this, it remains an outstanding challenge to understand more detailed models of dopaminergic function in terms of approximate Bayesian inference. For example, several models consider tonic dopamine firing to influence action *per se* [65]. Others consider its effects on performance and learning [66]. Many studies suggest that the performance effects of dopamine can be explained in terms of costs and benefits—such that high dopamine levels allow an animal to ignore the costs of actions if the benefit is sufficiently high.

Action costs in variational (Bayesian) formulations are normally treated in terms of prior beliefs about control [47]—such that a costly action is unlikely *a priori*. A differential effect of dopamine on costs (prior beliefs about control states) and benefits (prior beliefs about hidden states) speaks to the possibility that these beliefs are equipped with their own precision—and the fact that dopaminergic systems have a multitude of neuromodulatory mechanisms (e.g. D1 versus D2 receptor targets, direct versus indirect pathways, nigrostriatal versus mesocortical projections, etc.). Perhaps the deeper question here is not about whether dopamine mediates expected precision but which beliefs or probabilistic representations are contextualized in terms of their precision. For example, pathophysiology involving the nigrostriatal system is likely to produce very different deficits when compared with abnormal mesocortical dopaminergic function. In short, the challenge may be to map the physiological and anatomical diversity of dopaminergic projections to the plurality of functions in which dopamine has been implicated [61,34,67]. Much progress has been made along these lines—and the encoding of precision may provide a common computational role for dopamine that is consistent with its (neuromodulatory) mechanism of action.

We have previously asserted that the values of *states* are the consequence of behaviour, not its cause [68]. The current formulation finesses this assertion because the value of an *action* is both cause and consequence of behaviour. This is self-evidently true by the circular causality implicit in the action perception cycle [69] of embodied (active) inference. This fits comfortably with the finding of action-value coding in the brain prior to overt choice—for both positive and negative action values [70,71]. Furthermore, optogenetic studies show that stimulating distinct populations of striatal neurons during choice can effectively add or subtract action-value and bias behaviour to select or avoid an associated action [72]. Stimulating these populations during outcome induces subsequent approach or avoidance [73]. These results again point to the plurality of postsynaptic dopaminergic effects. In this instance, converse effects on action selection depending upon whether (facilitatory) D1 receptors or (inhibitory) D2 receptors are activated. These complementary effects of dopaminergic innervation are potentially important in the context of encoding precision in hierarchical models that may underlie action selection [35]: in computational terms, a key determinant of posterior expectations is the relative precision at different levels of hierarchical representations. It may be the case that dopaminergic projections mediate the relative precision or confidence in representations [34]—in a way that relies upon the balanced opposition of distinct pathways or receptor subtypes [67].

In conclusion, the account on offer considers dopamine to report the precision of divergence or prediction errors (in their nuanced sense) and partly resolves the dialectic between dopamine as reporting reward prediction errors [29] and the predictability of rewards [59,60,74]. The notion that dopamine encodes precision is now receiving support from several lines of evidence from theoretical treatments of hierarchical Bayesian inference [63], theoretical neurobiology [31,35,32,34] and empirical studies [75–78]. Having said this, a proper validation of the active inference will require careful model comparison using empirical choice behaviours and a detailed mapping between putative model variables and their neuronal correlates. The approach adopted in this paper highlights the intimate relationship between inferring states of the world and optimal behaviour [79,80], the confidence or precision of that inference [81] and the functional plurality of dopaminergic neuromodulation [61].

Acknowledgements. We would like to thank the Editors and two anonymous reviewers for helpful guidance in presenting this work.

Funding statement. The Wellcome Trust funded this work to K.J.F. (Ref: 088130/Z/09/Z).

## References


1. Schmidhuber J. 1991 Curious model-building control systems. In *Proc. IEEE Int. Joint Conf. on Neural Networks*, *Singapore*, *18–21 Nov. 1991*. Neural Networks, vol. 2, pp. 1458–1463. (doi:10.1109/IJCNN.1991.170605)
2. Klyubin AS, Polani D, Nehaniv CL. 2005 Empowerment: a universal agent-centric measure of control. In *Proc. IEEE Congress on Evolutionary Computation, Edinburgh, UK, 5 September 2005*, vol. 1, pp. 128–135. (doi:10.1109/CEC.2005.1554676)
3. Tishby N, Polani D. 2010 Information theory of decisions and actions. In *Perception–reason–action cycle: models, algorithms and systems* (eds V Cutsuridis, A Hussain, J Taylor), pp. 1–37. Berlin, Germany: Springer.
4. Ortega PA, Braun DA. 2011 Information, utility and bounded rationality. In *Artificial general intelligence* (eds J Schmidhuber, KR Thorisson, M Looks). Lecture Notes in Computer Science, vol. 6830, pp. 269–274. Berlin, Germany: Springer.
5. Wissner-Gross AD, Freer CE. 2013 Causal entropic forces. *Phys. Rev. Lett.* **110**, 168702. (doi:10.1103/PhysRevLett.110.168702)
6. McKelvey R, Palfrey T. 1995 Quantal response equilibria for normal form games. *Games Econ. Behav.* **10**, 6–38. (doi:10.1006/game.1995.1023)
7. Sharot T, Guitart-Masip M, Korn CW, Chowdhury R, Dolan RJ. 2012 How dopamine enhances an optimism bias in humans. *Curr. Biol.* **22**, 1477–1481. (doi:10.1016/j.cub.2012.05.053)
8. Friston K, Schwartenbeck P, FitzGerald T, Moutoussis M, Behrens T, Raymond J, Dolan RJ. 2013 The anatomy of choice: active inference and agency. *Front. Hum. Neurosci.* **7**, 598. (doi:10.3389/fnhum.2013.00598)
9. Friston K. 2012 A free energy principle for biological systems. *Entropy* **14**, 2100–2121. (doi:10.3390/e14112100)
10. Helmholtz H. 1866/1962 Concerning the perceptions in general. In *Treatise on physiological optics*, 3rd edn. New York, NY: Dover.
11. Ashby WR. 1947 Principles of the self-organizing dynamic system. *J. Gen. Psychol.* **37**, 125–128. (doi:10.1080/00221309.1947.9918144)
12. Conant RC, Ashby RW. 1970 Every good regulator of a system must be a model of that system. *Int. J. Syst. Sci.* **1**, 89–97. (doi:10.1080/00207727008920220)
13. Dayan P, Hinton GE, Neal R. 1995 The Helmholtz machine. *Neural Comput.* **7**, 889–904. (doi:10.1162/neco.1995.7.5.889)
14. Friston K. 2010 The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* **11**, 127–138. (doi:10.1038/nrn2787)
15. Camerer CF. 2003 Behavioural studies of strategic thinking in games. *Trends Cogn. Sci.* **7**, 225–231. (doi:10.1016/S1364-6613(03)00094-9)
16. Daw ND, Doya K. 2006 The computational neurobiology of learning and reward. *Curr. Opin. Neurobiol.* **16**, 199–204. (doi:10.1016/j.conb.2006.03.006)
17. Dayan P, Daw ND. 2008 Decision theory, reinforcement learning, and the brain. *Cogn. Affect Behav. Neurosci.* **8**, 429–453. (doi:10.3758/CABN.8.4.429)
18. Savage LJ. 1954 *The foundations of statistics*. New York, NY: Wiley.
19. Von Neumann J, Morgenstern O. 1944 *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.
20. Simon HA. 1956 Rational choice and the structure of the environment. *Psychol. Rev.* **63**, 129–138. (doi:10.1037/h0042769)
21. Ortega PA, Braun DA. 2013 Thermodynamics as a theory of decision-making with information-processing costs. *Proc. R. Soc. A* **469**, 20120683. (doi:10.1098/rspa.2012.0683)
22. Haile PA, Hortaçsu A, Kosenok G. 2008 On the empirical content of quantal response equilibrium. *Am. Econ. Rev.* **98**, 180–200. (doi:10.1257/aer.98.1.180)
23. Luce RD. 1959 *Individual choice behavior*. Oxford, UK: Wiley.
24. Fudenberg D, Kreps D. 1993 Learning mixed equilibria. *Games Econ. Behav.* **5**, 320–367. (doi:10.1006/game.1993.1021)
25. Sutton RS, Barto AG. 1998 *Reinforcement learning: an introduction*. Cambridge, MA: MIT Press.
26. Cohen JD, McClure SM, Yu AJ. 2007 Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Phil. Trans. R. Soc. B* **362**, 933–942. (doi:10.1098/rstb.2007.2098)
27. Beal MJ. 2003 Variational algorithms for approximate bayesian inference. PhD thesis, University College London, London, UK.
28. MacKay DJC. 2003 *Information theory, inference and learning algorithms*. Cambridge, UK: Cambridge University Press.
29. Schultz W, Dayan P, Montague PR. 1997 A neural substrate of prediction and reward. *Science* **275**, 1593–1599. (doi:10.1126/science.275.5306.1593)
30. Kapur S. 2003 Psychosis as a state of aberrant salience: a framework linking biology, phenomenology, and pharmacology in schizophrenia. *Am. J. Psychiatry* **160**, 13–23. (doi:10.1176/appi.ajp.160.1.13)
31. Fletcher PC, Frith CD. 2009 Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nat. Rev. Neurosci.* **10**, 48–58. (doi:10.1038/nrn2536)
32. Pellicano E, Burr D. 2012 When the world becomes 'too real': a Bayesian explanation of autistic perception. *Trends Cogn. Sci.* **16**, 504–510. (doi:10.1016/j.tics.2012.08.009)
33. Adams RA, Stephan KE, Brown HR, Frith CD, Friston KJ. 2013 The computational anatomy of psychosis. *Front. Psychiatry* **4**, 47. (doi:10.3389/fpsyt.2013.00047)
34. Friston KJST *et al*. 2012 Dopamine, affordance and active inference. *PLoS Comput. Biol.* **8**, e1002327. (doi:10.1371/journal.pcbi.1002327)
35. Frank MJ, Scheres A, Sherman SJ. 2007 Understanding decision-making deficits in neurological conditions: insights from models of natural action selection. *Phil. Trans. R. Soc. B* **362**, 1641–1654. (doi:10.1098/rstb.2007.2058)
36. Cisek P. 2007 Cortical mechanisms of action selection: the affordance competition hypothesis. *Phil. Trans. R. Soc. B* **362**, 1585–1599. (doi:10.1098/rstb.2007.2054)
37. Thompson WR. 1933 On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* **25**, 285–294. (doi:10.1093/biomet/25.3-4.285)
38. Ortega PADA. 2010 A minimum relative entropy principle for learning and acting. **38**, 475–511.
39. Dayan P, Hinton GE. 1997 Using expectation maximization for reinforcement learning. *Neural Comput.* **9**, 271–278. (doi:10.1162/neco.1997.9.2.271)
40. Rao RP, Ballard DH. 1999 Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2**, 79–87. (doi:10.1038/4580)
41. Watkins CJ, Dayan P. 1992 Q-learning. *Mach. Learn.* **8**, 279–292. (doi:10.1007/BF00992698)
42. van den Broek JL, Wiegerinck WAJJ, Kappen HJ. 2010 Risk-sensitive path integral control. *UAI* **6**, 1–8.
43. Jaynes ET. 1957 Information theory and statistical mechanics. *Phys. Rev. Ser. II* **106**, 620–630.
44. Kakade S, Dayan P. 2002 Dopamine: generalization and bonuses. *Neural Netw.* **15**, 549–559. (doi:10.1016/S0893-6080(02)00048-5)
45. Kappen HJ, Gomez Y, Opper M. 2012 Optimal control as a graphical model inference problem. *Mach. Learn.* **87**, 159–182. (doi:10.1007/s10994-012-5278-7)
46. Fox C, Roberts S. 2011 A tutorial on variational Bayes. In *Artificial intelligence review*. New York, NY: Spinger.
47. Friston K, Samothrakis S, Montague R. 2012 Active inference and agency: optimal control without cost functions. *Biol. Cybern.* **106**, 523–541. (doi:10.1007/s00422-012-0512-8)
48. Canolty RT *et al*. 2006 High gamma power is phase-locked to theta oscillations in human neocortex. *Science* **313**, 1626–1628. (doi:10.1126/science.1128115)
49. Goldman-Rakic PS. 1997 The cortical dopamine system: role in memory and cognition. *Adv. Pharmacol.* **42**, 707–711. (doi:10.1016/S1054-3589(08)60846-7)
50. Moran RJ, Symmonds M, Stephan KE, Friston KJ, Dolan RJ. 2011 An *in vivo* assay of synaptic function mediating human cognition. *Curr. Biol.* **21**, 1320–1325. (doi:10.1016/j.cub.2011.06.053)

51. O'Doherty J, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ. 2004 Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* **304**, 452–454. (doi:10.1126/science.1094285)
52. Surmeier D, Plotkin J, Shen W. 2009 Dopamine and synaptic plasticity in dorsal striatal circuits controlling action selection. *Curr. Opin. Neurobiol.* **19**, 621–628. (doi:10.1016/j.conb.2009.10.003)
53. Goldman-Rakic PS, Lidow MS, Smiley JF, Williams MS. 1992 The anatomy of dopamine in monkey and human prefrontal cortex. *J. Neural Transm. Suppl.* **36**, 163–177.
54. Solway A, Botvinick M. 2012 Goal-directed decision making as probabilistic inference: a computational framework and potential neural correlates. *Psychol. Rev*. **119**, 120–154. (doi:10.1037/a0026435)
55. Jocham G, Hunt LT, Near J, Behrens TE. 2012 A mechanism for value-guided choice based on the excitation–inhibition balance in prefrontal cortex. *Nat. Neurosci.* **15**, 960–961. (doi:10.1038/nn.3140)
56. Geisler S, Derst C, Veh RW, Zahm DS. 2007 Glutamatergic afferents of the ventral tegmental area in the rat. *J. Neurosci.* **27**, 5730–5743. (doi:10.1523/JNEUROSCI.0012-07.2007)
57. Smith Y, Bevan MD, Shink E, Bolam JP. 1998 Microcircuitry of the direct and indirect pathways of the basal ganglia. *Neuroscience* **86**, 353–387. (doi:10.1016/S0306-4522(97)00608-8)
58. Lidow MS, Goldman-Rakic PS, Gallager DW, Rakic P. 1991 Distribution of dopaminergic receptors in the primate cerebral cortex: quantitative autoradiographic analysis using [$^3$H]raclopride, [$^3$H]spiperone and [$^3$H]SCH23390. *Neuroscience* **40**, 657–671. (doi:10.1016/0306-4522(91)90003-7)
59. Redgrave P, Gurney K. 2006 The short-latency dopamine signal: a role in discovering novel actions? *Nat. Rev. Neurosci.* **7**, 967–975. (doi:10.1038/nrn2022)
60. Fiorillo CD, Tobler PN, Schultz W. 2003 Discrete coding of reward probability and uncertainty by dopamine neurons. *Science* **299**, 1898–1902. (doi:10.1126/science.1077349)
61. Schultz W. 2007 Multiple dopamine functions at different time courses. *Annu. Rev. Neurosci.* **30**, 259–288. (doi:10.1146/annurev.neuro.28.061604.135722)
62. Schultz W. 1998 Predictive reward signal of dopamine neurons. *J. Neurophysiol.* **80**, 1–27.
63. Mathys C, Daunizeau J, Friston KJ, Stephan KE. 2011 A Bayesian foundation for individual learning under uncertainty. *Front. Hum. Neurosci.* **5**, 39. (doi:10.3389/fnhum.2011.00039)
64. Gurney K, Prescott TJ, Redgrave P. 2001 A computational model of action selection in the basal ganglia. I. A new functional anatomy. *Biol. Cybern.* **84**, 401–410. (doi:10.1007/PL00007984)
65. Niv Y. 2007 Cost, benefit, tonic, phasic: what do response rates tell us about dopamine and motivation? *Ann. NY Acad. Sci.* **1104**, 357–376. (doi:10.1196/annals.1390.018)
66. Beeler JA, Frank MJ, McDaid J, Alexander E, Turkson S, Bernandez MS, McGehee DS, Zhuang X. 2012 A role for dopamine-mediated learning in the pathophysiology and treatment of Parkinson's disease. *Cell Rep.* **2**, 1747–1761. (doi:10.1016/j.celrep.2012.11.014)
67. Dayan P. 2012 Twenty-five lessons from computational neuromodulation. *Neuron* **76**, 240–256. (doi:10.1016/j.neuron.2012.09.027)
68. Friston K. 2011 What is optimal about motor control? *Neuron* **72**, 488–498. (doi:10.1016/j.neuron.2011.10.018)
69. Fuster JM. 2004 Upper processing stages of the perception–action cycle. *Trends Cogn. Sci.* **8**, 143–145. (doi:10.1016/j.tics.2004.02.004)
70. Samejima K, Ueda Y, Doya K, Kimura M. 2005 Representation of action-specific reward values in the striatum. *Science* **310**, 1337–1340. (doi:10.1126/science.1115270)
71. Lau B, Glimcher PW. 2007 Action and outcome encoding in the primate caudate nucleus. *J. Neurosci.* **27**, 14 502–14 514. (doi:10.1523/JNEUROSCI.3060-07.2007)
72. Tai LH, Lee AM, Benavidez N, Bonci A, Wilbrecht L. 2012 Transient stimulation of distinct subpopulations of striatal neurons mimics changes in action value. *Nat. Neurosci.* **15**, 1281–1289. (doi:10.1038/nn.3188)
73. Kravitz AV, Tye LD, Kreitzer AC. 2012 Distinct roles for direct and indirect pathway striatal neurons in reinforcement. *Nat. Neurosci.* **15**, 816–818. (doi:10.1038/nn.3100)
74. Schultz W *et al*. 2008 Explicit neural signals reflecting reward uncertainty. *Phil. Trans. R. Soc. B* **363**, 3801–3811. (doi:10.1098/rstb.2008.0152)
75. Fiorillo CD, Newsome WT, Schultz W. 2008 The temporal precision of reward prediction in dopamine neurons. *Nat. Neurosci.* **11**, 966–973. (doi:10.1038/nn.2159)
76. Zokaei N, Gorgoraptis N, Husain M. 2012 Dopamine modulates visual working memory precision. *J. Vis.* **12**, 350. (doi:10.1167/12.9.350)
77. Galea JM, Bestmann S, Beigi M, Jahanshahi M, Rothwell JC. 2012 Action reprogramming in Parkinson's disease: response to prediction error is modulated by levels of dopamine. *J. Neurosci.* **32**, 542–550. (doi:10.1523/JNEUROSCI.3621-11.2012)
78. Coull JT, Cheng RK, Meck W. 2011 Neuroanatomical and neurochemical substrates of timing. *Neuropsychopharmacology* **36**, 3–25. (doi:10.1038/npp.2010.113)
79. Toussaint M, Storkey A. 2006 Probabilistic inference for solving discrete and continuous state Markov Decision Processes. In *Proc. 23nd Int. Conf. on Machine Learning*, pp. 945–952. New York, NY: Association for Computing Machinery. (doi:10.1145/1143844.1143963)
80. Gläscher J, Daw N, Dayan P, O'Doherty JP. 2010 States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* **66**, 585–595. (doi:10.1016/j.neuron.2010.04.016)
81. De Martino B, Fleming SM, Garrett N, Dolan RJ. 2012 Confidence in value-based choice. *Nat. Neurosci.* **16**, 105–110. (doi:10.1038/nn.3279)