COGNITIVE NEUROSCIENCE, 2015
Vol. 6, No. 4, 187–224, http://dx.doi.org/10.1080/17588928.2015.1020053
(includes Discussion Paper and Commentaries)

Discussion Paper

Active inference and epistemic value

Karl Friston[1], Francesco Rigoli[1], Dimitri Ognibene[2], Christoph Mathys[1,3,4], Thomas Fitzgerald[1], and Giovanni Pezzulo[5]

[1]The Wellcome Trust Centre for Neuroimaging, Institute of Neurology, London, UK
[2]Centre for Robotics Research, Department of Informatics, King's College London, London, UK
[3]Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zürich and ETH Zürich, Zürich, Switzerland
[4]Laboratory for Social and Neural Systems Research (SNS Lab), Department of Economics, University of Zürich, Zürich, Switzerland
[5]Institute of Cognitive Sciences and Technologies, National Research Council, Rome, Italy

We offer a formal treatment of choice behavior based on the premise that agents minimize the expected free energy of future outcomes. Crucially, the negative free energy or *quality* of a policy can be decomposed into extrinsic and epistemic (or intrinsic) value. Minimizing expected free energy is therefore equivalent to maximizing extrinsic value or expected utility (defined in terms of prior preferences or goals), while maximizing information gain or intrinsic value (or reducing uncertainty about the causes of valuable outcomes). The resulting scheme resolves the exploration-exploitation dilemma: Epistemic value is maximized until there is no further information gain, after which exploitation is assured through maximization of extrinsic value. This is formally consistent with the Infomax principle, generalizing formulations of active vision based upon salience (Bayesian surprise) and optimal decisions based on expected utility and risk-sensitive (Kullback-Leibler) control. Furthermore, as with previous active inference formulations of discrete (Markovian) problems, ad hoc softmax parameters become the expected (Bayes-optimal) precision of beliefs about, or confidence in, policies. This article focuses on the basic theory, illustrating the ideas with simulations. A key aspect of these simulations is the similarity between precision updates and dopaminergic discharges observed in conditioning paradigms.

Keywords: Active inference; Agency; Bayesian inference; Bounded rationality; Free energy; Utility theory; Information gain; Bayesian surprise; Epistemic value; Exploration; Exploitation.

This article introduces a variational (free energy) formulation of explorative behavior and the (epistemic) value of knowing one's environment. This formulation tries to unite a number of perspectives on behavioral imperatives; namely, the exploration-exploitation dilemma and the distinction between the explicit (extrinsic) value of controlled outcomes and their epistemic (intrinsic) value in reducing uncertainty about environmental contingencies (Bialek, Nemenman, & Tishby, 2001; Botvinick & An, 2008; Braun, Ortega, Theodorou, & Schaal, 2011; Bromberg-Martin & Hikosaka 2009; Cohen, McClure,

Correspondence should be addressed to: Karl Friston, The Wellcome Trust Centre for Neuroimaging, Institute of Neurology, 12 Queen Square, London, UK WC1N 3BG. E-mail: k.friston@ucl.ac.uk
The authors have no disclosures or conflict of interest.
The Wellcome Trust funded this work. KJF is funded by the Wellcome Trust [088130/Z/09/Z]. DO is funded by the European Community's Seventh Framework Programme [FP7/2007-2013] project DARWIN [Grant No: FP7-270138]. GP is funded by the European Community's Seventh Framework Programme (FP7/2007-2013) project Goal-Leaders (Grant No: FP7-ICT-270108) and the HFSP (Grant No: RGY0088/2014).

& Yu, 2007; Daw, Niv, & Dayan, 2005; Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Friston et al., 2014; Pezzulo & Castelfranchi, 2009; Schmidhuber, 1991; Solway & Botvinick, 2012; Still, 2009; Tishby & Polani, 2010). In particular, it addresses how resolving uncertainty "makes the world interesting and exploitable" (Still & Precup, 2012, p. 139). It is the resolution of uncertainty that we associate with the intrinsic value of behavior, which we assume is synonymous with epistemic value. Our basic approach is to cast optimal behavior in terms of inference, where actions are selected from posterior beliefs about behavior. This allows one to frame goals and preferences in terms of prior beliefs, such that goals are subsequently fulfilled by action (Botvinick & Toussaint, 2012; Kappen, Gomez, & Opper, 2012; Toussaint & Storkey, 2006). This furnishes an explanation of behavior in terms of one straightforward imperative—to minimize surprise or, equivalently, to maximize Bayesian model evidence.

The resulting *active inference* scheme unifies conventional treatments of normative behavior under uncertainty. Classical treatments generally consider belief updates and action selection separately, calling on Bayesian inference to optimize beliefs and other schemes (such as dynamic programming) to select actions (Bonet & Geffner, 2014; Hauskrecht, 2000; Kaelbling, Littman, & Cassandra, 1998). Treating action selection as (active) inference means that both state estimation and the ensuing behavior can be described as a minimization of variational free energy or surprise.[1] In this setting, action reduces the difference between the current and (unsurprising) goal states that are defined by prior expectations, much like cybernetic formulations (Miller, Galanter, & Pribram, 1960). This difference can be reduced in two ways. First, by executing a *pragmatic action*, that fulfills goals directly (i.e., exploitation); for example, by visiting a known reward site in the context of foraging. Second, by performing an *epistemic action* (i.e., exploration) to disclose information that enables pragmatic action in the long run; for example, exploring a maze to discover unknown reward sites (Kirsh & Maglio, 1994). Clearly, most behavior has both pragmatic and epistemic aspects. Epistemic actions are the focus of much current research (Andreopoulos & Tsotsos, 2013; Ferro, Ognibene, Pezzulo, & Pirrelli, 2010; Kamar & Horvitz, 2013; Lepora, Martinez-Hernandez, & Prescott, 2013; Lungarella & Sporns, 2006; Ognibene, Chinellato, Sarabia, & Demiris, 2013; Ognibene, Volpi, Pezzulo, & Baldassare, 2013; Pezzementi, Plaku, Reyda, & Hager, 2011; Schneider et al., 2009; Singh, Krause, Guestrin, & Kaiser, 2009). For example, the coastal navigation algorithm (Roy, Burgard, Fox, & Thrun, 1999) shows that epistemic actions can sometimes increase the distance from a goal. In this example, agents move toward a familiar location (and away from the goal) to plan a path to the goal with greater confidence. This example illustrates the exploration-exploitation dilemma encountered at every decision point—and the implicit choice between epistemic and pragmatic actions. This choice is usually addressed in the setting of reinforcement learning (Dayan, 2009; Humphries, Khamassi, & Gurney, 2012; Still & Precup, 2012) but here we place more emphasis on planning or inference (Attias, 2003; Botvinick & Toussaint, 2012). In short, we offer a solution to exploration-exploitation dilemma that rests solely on the minimization of expected free energy.

In active inference, constructs like reward, utility, epistemic value, etc. are described in terms of prior beliefs or preferences. In other words, preferred outcomes are simply outcomes one expects, a priori, to be realized through behavior (e.g., arriving at one's destination or maintaining physiological states within some homoeostatic range). This formulation of utility has a number of advantages. First, it eliminates any ad hoc parameters in classical schemes (such as softmax parameters, temporal discounting, etc.). Second, it reveals the formal relationships among classical constructs, enabling their interpretation in terms of beliefs or expectations. For example, we have previously shown that the softmax parameter in classical (utilitarian) choice models corresponds to the precision or confidence in posterior beliefs about policies—and that this precision increases with expected utility (Friston et al., 2014). Third, this formulation is equipped with a relatively simple and biologically plausible process theory based upon variational message passing (Friston et al., 2013). This can be potentially useful when looking for the neuronal correlates of message passing in decision-making paradigms. Finally, casting rewards and value as probabilistic beliefs means that intrinsic and extrinsic values share a common currency. This means one can express (extrinsic) reward in terms of (epistemic) information gain and quantify their relative contributions to behavior.

Formally speaking, we resolve the exploration-exploitation dilemma by endowing agents with prior beliefs that they will minimize the expected free energy

[1]Variational free energy was introduced by Richard Feynman to solve inference problems in quantum mechanics and can be regarded as a generalization of thermodynamic free energy. In this paper, free energy refers to variational free energy. We will see later that minimizing free energy (or maximizing negative free energy) corresponds to maximizing expected value.

of future outcomes. In other words, the agent will be surprised if it behaves in a way that is not Bayes optimal. Because expected free energy determines action selection, the resulting behavior is necessarily Bayes optimal. Crucially, expected free energy is minimized over an extended timescale, making exploration a necessary and emergent aspect of optimal behavior. Expected free energy can be expressed as the Kullback-Leibler (KL) divergence between the posterior (predictive) and prior (preferred) distributions over future outcomes, plus the expected entropy of those observations, given their causes. In brief, minimizing this divergence ensures preferred outcomes are actively sampled from the environment, while minimizing the expected entropy resolves uncertainty about the (hidden) states causing those outcomes.[2] Intuitively, these two aspects of emergent behavior (sampling preferred outcomes and minimizing expected uncertainty) correspond to exploitation and exploration, respectively. Interestingly, the negative expected free energy can also be expressed as the expected divergence between the posterior (predictive) distribution over hidden states with and without future observations, plus the expected utility (defined as the log of the prior probability of future states). We will associate these terms with epistemic and extrinsic value respectively.

We have shown previously that minimizing the divergence between the posterior predictive distribution and prior preferences produces behavior that is risk-sensitive or KL optimal (Friston et al., 2013). Here, we show that this risk-sensitive control is a special case of minimizing expected free energy, which effectively supplements expected utility with a KL divergence that reflects epistemic value, mutual information, information gain, or Bayesian surprise, depending upon one's point of view. The KL divergence is also known as *relative entropy* or *information gain*. This means that minimizing expected free energy maximizes information gain (Ognibene & Demiris, 2013; Sornkarn, Nanayakkara, & Howard, 2014; Tishby & Polani, 2010) or, heuristically, satisfies curiosity by reducing uncertainty about the world (Schmidhuber, 1991). An alternative perspective on this epistemic quantity is afforded by *Bayesian surprise*; namely, the KL divergence between prior and posterior beliefs (Bruce & Tsotsos, 2009; Itti & Baldi, 2009). However, in this case, the Bayesian surprise pertains to future states that have yet to be observed.

For readers who are familiar with our previous work on active inference, this paper introduces a generic formulation that combines earlier work on optimal choice behavior (Friston et al., 2014) with formulations of salience based on sampling the world to resolve uncertainty (Friston, Adams, Perrinet, & Breakspear, 2012). These two formulations can be regarded as special cases of minimizing expected free energy, when sensory cues are unambiguous and when outcomes have only epistemic value, respectively. In this article, we show that minimizing expected free energy provides an inclusive perspective on several other established formulations of behavior.

In what follows, we introduce the basic formalism behind active inference, with a special focus on epistemic value and how this emerges under active (Bayesian) inference. The second section considers (biologically plausible) variational message passing schemes that can be used to simulate active inference in the context of partially observed Markov decision processes (Kaelbling et al., 1998) or to model empirical choice behavior. The final sections present simulations of exploration and exploitation, using a simple foraging game to illustrate the fundamental role of epistemic value in actively resolving uncertainty about goal-directed behavior. These sections consider planning and learning as inference, respectively. In future work, we will consider the optimization of models per se in terms of Bayesian model selection (structure learning) and the role of Bayesian model averaging in contextualizing shallow (model-free) and deep (model-based) models.

ACTIVE INFERENCE

This section describes active inference, in which inference and behavior are seen as consequences of minimizing variational free energy or, equivalently, maximizing Bayesian model evidence (Friston, 2010). We have previously considered epistemic value and salience using continuous time predictive coding schemes and saccadic searches (Friston et al., 2012). Here, we will use a discrete time and state space formulation of Bayes optimal behavior to show that information gain is a necessary consequence of minimizing expected free energy.

This formulation rests upon two key distinctions. First, we distinguish between a real world process that generates observations and an agent's internal model

[2]Note the dialectic between minimizing the entropy expected in the future and maximizing the entropy of current beliefs—implicit in minimizing free energy Friston et al. (2012). "Perceptions as hypotheses: Saccades as experiments." *Front Psychol*. **3**: 151.

of that process. These are referred to as the *generative process* and *generative model* respectively. The process and model are coupled in two directions: (sensory) observations generated by the generative process are observed by the agent, while the agent acts on the world to change that process. We will see that action serves to minimize the same quantity (variational free energy) used to make inferences about the hidden causes of observations. Crucially, *action* is a real variable that acts on the generative process, while the corresponding hidden cause in the generative model is a *control state*. This means the agent has to infer its behavior by forming beliefs about control states, based upon the observed consequences of its action.

We will adopt the formalism of partially observed Markov decision processes (POMDP). This is just a way of describing transitions among (discrete) states, under the assumption that the probability of the next state depends on, and only on, the current state. The partially observed aspect of the ensuing Markovian process means that the states of the generative process are hidden and have to be inferred through a limited set of (possibly noisy) observations.[3]

Notation: We use conventional notation, where the parameters of categorical distributions over discrete states $s \in S \in \{1, \ldots, J\}$ are denoted by $J \times 1$ vectors of expectations $\widehat{S} \in [0, 1]$, while the ~ notation denotes sequences of variables over time. The entropy of a probability distribution $P(a) = \Pr(A = a)$ will be denoted by $H(A) = H[P(a)] = E_{P(a)}[-\ln P(a)]$ and the relative entropy by the Kullback-Leibler (KL) divergence $D[Q(a)||P(a)] = E_{Q(a)}[\ln Q(a) - \ln P(a)]$. The dot notation means $A \cdot B = A^T B$ and $A \times B$ denotes the Hadamard (or element by element) product of two matrices. Similarly, $\ln A$ denotes the logarithm of the elements of a matrix.

Definition: Active inference rests on the tuple $(P, Q, R, S, A, U, \Omega)$:

- A finite set of observations Ω
- A finite set of actions A
- A finite set of hidden states S
- A finite set of control states U
- A *generative process* over observations $\tilde{o} \in \Omega$, hidden states $\tilde{s} \in S$, and action $\tilde{a} \in A$

$$\begin{aligned} R(\tilde{o}, \tilde{s}, \tilde{a}) = \Pr(\{o_0, \ldots, o_t\} = \tilde{o}, \{s_0, \ldots, s_t\} \\ = \tilde{s}, \{a_0, \ldots, a_{t-1}\} = \tilde{a}) \end{aligned}$$

- A *generative model* over observations $\tilde{o} \in \Omega$, hidden $\tilde{s} \in S$, and control $\tilde{u} \in U$ states $P(\tilde{o}, \tilde{s}, \tilde{u}|m) = \Pr(\{o_0, \ldots, o_T\} = \tilde{o}, \{s_0, \ldots, s_T\} = \tilde{s}, \{u_t, \ldots, u_T\} = \tilde{u})$, with parameters θ
- An *approximate posterior* over hidden and control states such that $Q(\tilde{s}, \tilde{u}) = \Pr(\{s_0, \ldots, s_T\} = \tilde{s}, \{u_t, \ldots, u_T\} = \tilde{u})$ with parameters or expectations $(\widehat{s}, \widehat{\pi})$, where $\pi \in \{1, \ldots, K\}$ is a policy that indexes a sequence of control states $(\tilde{u}|\pi) = (u_t, \ldots, u_T|\pi)$

Remark: The *generative process* describes the environment in terms of transitions among hidden states that generate observed outcomes. These transitions depend upon actions, which are sampled from approximate posterior beliefs about control states. In turn, these beliefs are formed using a *generative model* (denoted by m) of how observations are generated. The generative model describes what the agent believes about the world, where (approximate posterior) beliefs about hidden states and control states are encoded by expectations. This is a slightly unusual setup because there is a distinction between actions (that are part of a generative process) and control states (that are part of the generative model). This distinction allows actions to be sampled from posterior beliefs about control, effectively converting an optimal control problem into an optimal inference problem (Attias, 2003; Botvinick & Toussaint, 2012). Furthermore, note that (unlike the generative process) the generative model includes beliefs about future states.

So far, we have just described the agent's world and its model of that world. To describe the agent's exchange with its environment, we have to specify how its expectations depend upon observations and how its action depends upon expectations. In other words, we have to close the perception-action cycle (Fuster, 2004). In brief, we will make one sassumption; namely, that both actions and expectations minimize the free energy of observations. More precisely, we will assume that expectations minimize free energy and the expectations of control states prescribe action at the current time t:

[3]For readers interested in technical details, the simulations (and figures) reported in this paper can be reproduced by downloading the academic freeware SPM. Annotated Matlab scripts can then be accessed through a graphical user interface (invoked by typing DEM and selecting "epistemic value"). Please visit http://www.fil.ion.ucl.ac.uk/spm/software/

$$(\hat{s}^*, \hat{\pi}^*) = \arg\min F(\tilde{o}, \hat{s}, \hat{\pi})$$
$$\Pr(a_t = u_t) = Q(u_t|\hat{\pi}^*)$$

$$\begin{aligned} F(\tilde{o}, \hat{s}, \hat{\pi}) &= E_Q[-\ln P(\tilde{o}, \tilde{s}, \tilde{u}|m)] - H[Q(\tilde{s}, \tilde{u})] \\ &= -\ln P(\tilde{o}|m) + D[Q(\tilde{s}, \tilde{u})||P(\tilde{s}, \tilde{u}|\tilde{o})] \end{aligned} \quad (1)$$

Heuristically, at each decision point or cycle the agent first figures out which states are most likely by optimizing its expectations with respect to free energy (using the generative model). After optimizing its posterior beliefs, an action is sampled from the posterior probability distribution over control states. Given this action, the environment generates a new observation (using the generative process) and a new cycle begins.

The first expression for free energy in Equation 1 shows that it is an expected energy, under the generative model, minus the entropy of the approximate posterior. This expression can be rearranged to give the second expression, which shows that free energy is an upper bound on the negative logarithm of Bayesian model evidence $-\ln P(\tilde{o}|m)$, which is also known as *surprise* or *surprisal*. The free energy is an upper bound on surprise because the divergence term cannot be less than zero (Beal, 2003). Therefore, minimizing free energy corresponds to minimizing the divergence between the approximate and true posterior. This formalizes the notion of approximate Bayesian inference in psychology and machine learning (Dayan & Hinton, 1997; Dayan, Hinton, & Neal, 1995; Helmholtz, 1866/1962). Minimizing surprise provides a nice perspective on perception which, in this setting, corresponds to updating expectations about hidden states of the world in a Bayes optimal fashion. But what about action? If action is sampled from beliefs about control states, then the agent must believe its actions will minimize free energy. We now look at this more closely.

In active inference, agents do not just infer hidden states but actively sample outcomes that minimize free energy. The aim here is to explain how agents restrict themselves to a small number of preferred outcomes (i.e., goals). This is fairly straightforward to explain if agents minimize surprise, while a priori expecting to attain their goals. More formally, if actions depend upon posterior beliefs, then actions depend on prior beliefs. This means prior beliefs entail goals because they specify action. In turn, the generative model entails prior beliefs because it comprises the likelihood over observations, an empirical prior over state transitions and a prior over control states. These correspond to the three marginal distributions of the generative model: $P(\tilde{o}, \tilde{s}, \tilde{u}|m) = P(\tilde{o}|\tilde{s})P(\tilde{s}|\tilde{u})P(\tilde{u}|m)$. Crucially, the only self-consistent prior beliefs an agent can entertain about control states is that they will minimize free energy.[4] One can express this formally by associating the prior probability of a policy with the path integral (from the current to the final state) of free energy expected under that policy (c.f., Hamilton's principle of least action and Feynman's path integral formulation of quantum mechanics). We will call this the quality, value, or the expected (negative) free energy of a policy, denoted by $\mathbf{Q}(\tilde{u}|\pi) := \mathbf{Q}(\pi)$:

$$\begin{aligned} \ln P(\tilde{u}|\gamma) &= \gamma \cdot \mathbf{Q}(\pi) = \gamma \cdot (\mathbf{Q}_{t+1}(\pi) + \ldots + \mathbf{Q}_T(\pi)) \\ \mathbf{Q}_\tau(\pi) &= E_{Q(o_\tau, s_\tau|\pi)}[\ln P(o_\tau, s_\tau|\pi)] + H[Q(s_\tau|\pi)] \end{aligned} \quad (2)$$

This expression says that a policy is a priori more likely if the policy has a high quality or its expected free energy is small. Heuristically, this means that agents believe they will pursue policies that minimize the expected free energy of outcomes and implicitly minimize their surprise about those outcomes. Equivalently, policies that do not minimize expected free energy are a priori surprising and will be avoided. Put simply, not only do agents minimize free energy or surprise (Equation 1) but they also believe they will minimize free energy or surprise (Equation 2). These beliefs (Equation 2) are realized through active inference because agents minimize surprise (Equation 1). This self-consistent recursion leads to behavior that is apparently purposeful, in the sense that it appears to avoid surprising states.

The expected free energy is the free energy of beliefs about the future (not the free energy of future beliefs). More formally, the expected free energy is the energy of counterfactual outcomes and their causes expected under their posterior predictive distribution, minus the entropy of the posterior

[4]This is a fairly subtle assertion that lies at the heart of active inference. Put simply, agents will adjust their expectations to minimize the free energy associated with any given observations. However, when the agent actively samples observations, it has the opportunity to choose observations that minimize free energy—an opportunity that is only realized when the agent believes this is how it behaves. A more formal proof by *reductio ad absurdum*—that appeals to random dynamical systems—can be found in Friston and Mathys (2015). I think therefore I am. *Cognitive Dynamic Systems*. S. Haykin, IEEE press: in press. In brief, to exist, an ergodic system must place an upper bound on the entropy of its states, where entropy is the long-term average of surprise. Therefore, any system that does not (believe it will) minimize the long-term average of surprise does not (believe it will) exist.

predictive distribution over hidden states. The posterior predictive distributions are distributions over future states at $\tau > t$ expected under current beliefs: $Q(o_\tau, s_\tau|\pi) = E_{Q(s_t)}[P(o_\tau, s_\tau|s_t, \pi)]$. Notice that this predictive posterior includes beliefs about future outcomes and hidden states, while the current posterior $Q(s_t)$ just covers hidden states. In this setup, $\gamma \in \theta$ plays the role of a sensitivity or inverse temperature parameter that corresponds to the precision of, or confidence in, prior beliefs about policies.

Note that we have introduced a circular causality by specifying prior beliefs in this way: Prior beliefs about control states depend upon (approximate posterior predictive) beliefs about hidden states, which depend on observations. This means that prior beliefs about policies depend upon past observations. Indeed, we will see later that if the precision parameter γ was known, the prior and posterior beliefs would be identical. However, when the precision is a free (hyper) parameter, posterior beliefs become the prior beliefs expected under posterior precision. This may sound rather complicated but the important role of posterior precision or confidence will become increasingly evident. In brief, by making precision a free parameter, it can be optimized with respect to free energy or model evidence (unlike the inverse temperature parameter of conventional models). We now try to unpack these beliefs about policies in terms of established formulations of goal-directed behavior.

Although Equation 2 has a relatively simple form, it is not easy to see the behaviors it produces. However, with some straightforward rearrangement, two intuitive terms reveal themselves; namely, extrinsic and epistemic value (see also Appendix A).

$$\begin{aligned}\mathbf{Q}_\tau(\pi) &= E_{Q(o_\tau, s_\tau|\pi)}[\ln P(o_\tau, s_\tau|\pi) - \ln Q(s_\tau|\pi)] \\ &= E_{Q(o_\tau, s_\tau|\pi)}[\ln Q(s_\tau|o_\tau, \pi) + \ln P(o_\tau|m) \\ &\quad - \ln Q(s_\tau|\pi)] = \underbrace{E_{Q(o_\tau|\pi)}[\ln P(o_\tau|m)]}_{\text{Extrinsic value}} \\ &\quad + \underbrace{E_{Q(o_\tau|\pi)}[D[Q(s_\tau|o_\tau, \pi)||Q(s_\tau|\pi)]]}_{\text{Epistemic value}}\end{aligned} \tag{3}$$

Here, the generative model of future states $P(o_\tau, s_\tau|\pi) = Q(s_\tau|o_\tau, \pi)P(o_\tau|m)$ comprises the predictive posterior and prior beliefs about future outcomes. Note, the generative model of future states is not the generative model of states in the future, when the *predictive* posterior becomes the future posterior and the generative model of the future becomes the future generative model $P(o_\tau, s_\tau|\pi) = P(s_\tau|o_\tau, \pi)P(o_\tau|m)$. Equation 3 shows that under the generative model of the future, the quality of a policy can be expressed in terms of extrinsic and epistemic value:

Extrinsic value: Extrinsic value is the utility $C(o_\tau|m) = \ln P(o_\tau|m)$ of an outcome expected under the posterior predictive distribution. It is this utility that encodes the preferred outcomes that lend behavior its goal-directed nature. In other words, agents consider outcomes with low utility surprising, irrespective of the policy. This means that agents (believe they) will maximize expected utility to ensure preferred outcomes. Note that, by definition, the utility of an outcome is not a function of the policy. This means the agent believes all (unsurprising) policies lead to the same preferred outcomes or goals. The degree to which expected utility dominates prior beliefs about policies rests on the precision of prior preferences. In the absence of precise goals, epistemic or intrinsic value will come to dominate policy selection.

Epistemic value: Epistemic value is the expected information gain under predicted outcomes. In other words, it reports the reduction in uncertainty about hidden states afforded by observations. Because the KL divergence (or information gain) cannot be less than zero, the information gain is smallest when the posterior predictive distribution is not informed by new observations. Heuristically, this means valuable policies will search out observations, cues or "signs" that resolve uncertainty about the state of the world (e.g., foraging to resolve uncertainty about the hidden location of food or fixating on informative part of a face to identify someone). 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. In this case, extrinsic value will dominate policy selection.

Relationship to established formalisms

The Infomax principle: Epistemic or intrinsic value fits comfortably with a number of formulations from the visual sciences and information theory. As discussed (using continuous time formulations) in Friston et al. (2012), minimizing uncertainty about hidden states necessarily entails an increase in the mutual information between (sensory) outcomes and their (hidden) causes. Formally, this can be seen with a simple rearrangement of epistemic value to show that it is equivalent to the mutual information between hidden states and outcomes, under the posterior predictive distribution:

$$\underbrace{E_{Q(o_\tau|\pi)}[D[Q(s_\tau|o_\tau,\pi)||Q(s_\tau|\pi)]]}_{\text{Epistemic value}}$$
$$= \underbrace{D[Q(s_\tau,o_\tau|\pi)||Q(s_\tau|\pi)Q(o_\tau|\pi)]}_{\text{Predictive mutual information}} \quad (4)$$

This means that policies with epistemic value render observations more informative about their causes. This is one instance of the Infomax principle (Linsker, 1990), which is closely related to the principle of maximum mutual information, or minimum redundancy (Barlow, 1961, 1974; Bialek et al., 2001; Najemnik & Geisler, 2005; Oja, 1989; Olshausen & Field, 1996; Optican & Richmond, 1987).

Bayesian surprise: Epistemic value is also the Bayesian surprise expected under counterfactual outcomes. Bayesian surprise is a measure of salience and is the KL divergence between a posterior and prior distribution (Itti & Baldi, 2009). Empirically, people tend to direct their gaze toward salient visual features with high Bayesian surprise (Itti & Baldi, 2009). In the current setup, the expected Bayesian surprise, or salience, is the epistemic value of a particular policy that samples (sensory) outcomes. Although the value of a policy includes Bayesian surprise, it also comprises expected utility, which contextualizes the influence of salience. In other words, salience will only drive epistemic sampling of salient information if the epistemic value of that sampling is greater than the extrinsic value of an alternative behavior. We will see examples of this later.

Value of information: The value information is the amount an agent would pay to obtain information pertaining to a decision (Howard, 1966; Krause & Guestrin, 2005; Kamar & Horvitz, 2013). In this formulation, information has no epistemic value per se but only relative to choices or policy selection; in other words, information that does not affect a choice has no value. The value of information is generally intractable to compute for complex (e.g., nonstationary) environments. Here, we offer a formulation that contextualizes the value of information (epistemic value) in relation to extrinsic value and provides a tractable (approximate Bayesian inference) scheme for its evaluation.

KL control: Optimal control problems can generally be expressed as minimizing the KL divergence between the preferred and predictive distribution over outcomes. The general idea behind KL control is to select control states that minimize the difference between predicted and desired outcomes, where the difference is measured in terms of the KL divergence between the respective probability distributions. Minimizing this divergence is a cornerstone of risk-sensitive control (Van Den Broek, Wiegerinck, & Kappen, 2010) and utility-based free energy treatments of bounded rationality (Ortega & Braun, 2011, 2013). In the current context, risk-sensitive (KL) control can be seen as a special case of minimizing expected free energy, when outcomes unambiguously specify hidden states. In other words, when the generative process is completely observable, we can associate each outcome with a hidden state such that $o_\tau = s_\tau$ and:

$$\begin{aligned}\mathbf{Q}_\tau(\pi) &= E_{Q(s_\tau|\pi)}[\ln P(s_\tau|\pi) - \ln Q(s_\tau|\pi)] \\ &= -\underbrace{D[Q(s_\tau|\pi)||P(s_\tau|\pi)]}_{\text{KL divergence}} = \underbrace{E_{Q(s_\tau|\pi)}[\ln P(s_\tau|m)]}_{\text{Extrinsic value}} \\ &\quad + \underbrace{H[Q(s_\tau|\pi)]}_{\text{Epistemic value}}\end{aligned} \quad (5)$$

In this special case, minimizing free energy minimizes the divergence between the posterior predictive distribution over states and the prior predictive distribution encoding goals. Here, the extrinsic value now becomes an expected utility over *states* and the epistemic value becomes the novelty or (posterior predictive) entropy over future states. The difference between maximizing the entropy (novelty) and relative entropy (information gain) distinguishes risk-sensitive (KL) control from free energy minimization. Only minimizing free energy allows epistemic value to guide explorative behavior in a way that fully accommodates uncertainty about a partially observed world. This can be seen clearly with a final rearrangement of the expression for the quality of a policy (see Appendix A):

$$\begin{aligned}\mathbf{Q}_\tau(\pi) &= E_{Q(o_\tau,s_\tau|\pi)}[\ln Q(o_\tau|s_\tau,\pi) \\ &\quad + \ln P(o_\tau|m) - \ln Q(o_\tau|\pi)] \\ &= -\underbrace{E_{Q(s_\tau|\pi)}[H[P(o_\tau|s_\tau)]]}_{\text{Predicted uncertainty}} - \underbrace{D[Q(o_\tau|\pi)||P(o_\tau|m)]}_{\text{Predicted divergence}}\end{aligned} \quad (6)$$

This equality expresses the value of a policy in terms of the posterior predictive distribution over outcomes, as opposed to hidden states. In this formulation, expected free energy corresponds to the expected entropy or uncertainty over outcomes, given their causes, plus the KL divergence between the posterior predictive and preferred distributions. In other words, minimizing expected free energy minimizes the divergence between predicted and preferred outcomes (i.e., predicted divergence) and any uncertainty afforded by observations (i.e., predicted uncertainty). Heuristically, this ensures observations are informative. For example, an agent who wants to avoid bright light will move to the shade, as opposed to closing its eyes. If outcomes are always informative, we revert to risk-sensitive (KL) control, expressed in terms of preferences over outcomes, as opposed to states.

In our previous formulations of active inference and risk-sensitive (KL) control, we only considered scenarios in which hidden states could be observed directly. In this paper, we will illustrate the difference between risk-sensitive (KL) control and expected free energy minimization in a more realistic setting, in which hidden states can only be inferred from particular observations. In this context, we will see that risk-sensitive (KL) control is not sufficient to explain purposeful or exploratory responses to salient cues that resolve uncertainty about the environment.

Dopamine and reward prediction errors: In the next section, we will see how approximate Bayesian inference, implicit in active inference, can be implemented using a relatively simple variational message passing scheme. We have previously discussed the biological plausibility of this scheme in terms of recursive neuronal message passing (Friston et al., 2013) and have associated dopamine with the posterior precision of beliefs about control states (Friston et al., 2014). We will see later that changes in the expected (inverse) precision are identical to changes in (negative) expected value. This is potentially important because it may explain why changes in dopamine firing have been associated with reward prediction error (Schultz, 1998). However, it has a deeper implication here: If expected precision changes with expected value, then the current formulation explains why dopamine has a multilateral sensitivity to novelty (Kakade & Dayan, 2002; Krebs, Schott, Schütze, & Düzel, 2009; Wittmann, Daw, Seymour, & Dolan, 2008), salience (Berridge, 2007), expected reward (Bunzeck & Düzel, 2006; D'Ardenne, McClure, Nystrom, & Cohen, 2008; Daw & Doya, 2006; Dayan, 2009; McClure, Daw, & Montague, 2003; O'Doherty et al., 2004; Pessiglione, Seymour, Flandin, Dolan, & Frith, 2006), epistemic value (Fiorillo, Tobler, & Schultz, 2003; Redgrave & Gurney, 2006; Bromberg-Martin & Hikosaka, 2009), and affordance (Cisek, 2007; Gurney, Prescott, & Redgrave, 2001; see also Nepora & Gurney, 2012). The study of Bromberg-Martin and Hikosaka (2009) is particularly interesting in this context because it provides direct evidence linking dopamine responses and epistemic value. The emerging perspective also fits comfortably with recent attempts to reconcile dopamine's role in the exploration-exploitation trade-off with the role of the basal ganglia in action selection, "by testing the hypothesis that tonic dopamine in the striatum, the basal ganglia's input nucleus, sets the current exploration-exploitation trade-off" (Humphries et al., 2012, p. 1).

The close relationship between expected precision and value provides an interesting perspective on the transfer of dopaminergic responses to conditioned stimuli in operant conditioning paradigms (Schultz, 1998). From the perspective of active inference, conditioned stimuli have epistemic value because they resolve uncertainty about future outcomes (unconditioned stimuli). This perspective may also provide an inferential account of blocking and latent inhibition, in the sense that if epistemic uncertainty has already been resolved by one conditioned stimulus, then no further information gain is afforded by another. We will pursue these arguments with simulations of goal-directed behavior below. The important issue here is a dual role for dopaminergic responses in reporting precision in terms of extrinsic and epistemic value. However, functionally, there is only one precision (sensitivity) parameter that applies to, and reconciles, both aspects of value. This eliminates the need for ad hoc parameters to finesse the exploration-exploitation dilemma. We will illustrate these and other points using simulations in the last two sections.

Summary

Although minimizing expected free energy corresponds to maximizing extrinsic and epistemic value, this dual maximization is a particular perspective on the underlying imperative to minimize surprise. This means that both extrinsic and epistemic value work synergistically to increase the likelihood of preferred outcomes with the minimum of uncertainty. For example, extrinsic value depends on the posterior predictive distribution over outcomes, which is only informative when the agent can be confident about the current state (c.f., the coastal navigation example above). This means epistemic uncertainty must first be resolved (by increasing epistemic value) before expected utility comes into play. At the same time, an agent should not indulge in epistemic actions, if it is sufficiently confident it can pursue a successful plan.

These considerations are especially interesting in relation to exploration and exploitation.

In summary, minimizing free energy corresponds to approximate Bayesian inference and, in active inference, choosing the least surprising outcomes. If agents model their environments, they have to entertain posterior beliefs about the control of state transitions producing outcomes. This means we have to consider posterior beliefs about control states, which rest on prior beliefs about controlled outcomes. Using the self-consistent prior that control states minimize expected free energy ("I expect to avoid surprises"), we arrive at a process theory that offers a formal definition of extrinsic and epistemic value. Furthermore, it emphasizes the fact that purposeful behavior rests upon generative models that entertain future outcomes. This formulation accommodates a number of established perspectives; namely, the Infomax principle, the notion of Bayesian surprise in reporting the salience of cues, and KL control, which generalizes risk-sensitive control and expected utility theory. In the next section we will see how this theory prescribes a computational anatomy for Bayesian belief updating that has many similarities with message passing in the brain.

GENERATIVE MODELS AND VARIATIONAL MESSAGE PASSING

The generative model

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

$$\begin{aligned}
P(\tilde{o}, \tilde{s}, \tilde{u}, \gamma | \tilde{a}, m) &= P(\tilde{o}|\tilde{s})P(\tilde{s}|\tilde{a})P(\tilde{u}|\gamma)P(\gamma|m) \\
P(\tilde{o}|\tilde{s}) &= P(o_0|s_0)P(o_1|s_1)\ldots P(o_t|s_t) \\
P(\tilde{s}|\tilde{a}) &= P(s_t|s_{t-1}, a_t)\ldots P(s_1|s_0, a_1)P(s_0|m) \\
P(\tilde{u}|\gamma) &= \sigma(\gamma \cdot \mathbf{Q})
\end{aligned} \tag{7}$$

Here, $\sigma(\cdot)$ is a softmax function. The first equality expresses the generative model in terms of the likelihood of observations given the hidden states (first term) and subsequent *empirical* prior beliefs. Empirical priors are probability distributions over unknown variables that depend on other unknown variables. Empirical priors are a universal aspect of hierarchical Bayesian models; for example, parametric empirical Bayes (Kass & Steffey, 1989). In effect, empirical priors are informed by observations under hierarchical constraints. The likelihood in the second equality implies that observations depend on, and only on, the current hidden state. The third equality expresses (empirical) prior beliefs about state transitions. For simplicity, we assume that agents know their past actions. The final equality expresses beliefs about policies in terms of their quality or value. In short, this model represents past hidden states and future choices, under the belief that controlled transitions from the current state will minimize the expected free energy of future states.

This model can be parameterized in a fairly straightforward way, using the notation $P(o_t = i|s_t = j, \mathbf{A}) = \mathbf{A}_{ij} \Leftarrow P(o_t|s_t) = \mathbf{A}$

$$\begin{aligned}
P(o_t|s_t) &= \mathbf{A} \\
P(s_{t+1}|s_t, u_t) &= \mathbf{B}(u_t) \\
P(o_\tau|m) &= \mathbf{C}_\tau \\
P(s_0|m) &= \mathbf{D} \\
P(\gamma|m) &= \Gamma(\alpha, \beta)
\end{aligned} \tag{8}$$

These equalities mean that the categorical distributions over observations, given the hidden states, are encoded by the matrix $\mathbf{A} \in \theta$ that maps from hidden states to outcomes. Similarly, the transition matrices $\mathbf{B}(u_t) \in \theta$ encode transition probabilities from one state to the next, under the current control state. The vectors $\mathbf{C} \in \theta$ and $\mathbf{D} \in \theta$ encode prior distributions over future outcomes and initial states, respectively. The priors over future outcomes specify their utility $C(o_\tau|m) = \ln P(o_\tau|m) = \ln \mathbf{C}_\tau$. Finally, the prior over precision has a standard gamma distribution, with shape and rate parameters (in this paper) $\alpha = 64$ and $\beta = 4$.

The vector $\mathbf{Q}$ contains the values of each policy at the current time. These values can be expressed in terms of the parameters above using the expression for expected free energy in Equation (6) and Appendix A:

$$\begin{aligned}
\mathbf{Q}(\pi) &= \mathbf{Q}_{t+1}(\pi) + \ldots + \mathbf{Q}_T(\pi) \\
\mathbf{Q}_\tau(\pi) &= \underbrace{\mathbf{1} \cdot (\mathbf{A} \times \ln \mathbf{A})\hat{s}_\tau(\pi)}_{\text{Predicted uncertainty}} \\
&\quad - \underbrace{(\ln \hat{o}_\tau(\pi) - \ln \mathbf{C}_\tau) \cdot \hat{o}_\tau(\pi)}_{\text{Predicted divergence}} \\
\hat{s}_\tau(\pi) &= \mathbf{B}(u_\tau|\pi)\ldots\mathbf{B}(u_t|\pi)\hat{s}_t \\
\hat{o}_\tau(\pi) &= \mathbf{A}\hat{s}_\tau(\pi)
\end{aligned} \tag{9}$$

Where $\widehat{s}_\tau(\pi)$ are the expected states at time τ under policy π and $\mathbf{1}$ is a column vector of ones. Note that when there is no uncertainty about future states, we have $\mathbf{1} \cdot (\mathbf{A} \times \ln \mathbf{A})\widehat{s}_\tau(\pi) = \ln(\mathbf{A}\widehat{s}_\tau(\pi)) \cdot \mathbf{A}\widehat{s}_\tau(\pi)$ and the value of a policy depends only on expected utility $\mathbf{Q}_\tau(\pi) = \widehat{o}_\tau(\pi) \cdot \ln \mathbf{C}_\tau$. In other words, policies have no epistemic value when they lead to no further information gain.

Approximate Bayesian inference

Having specified the generative model, variational Bayes now offers a generic scheme for approximate Bayesian inference that finesses the combinatoric and analytic intractability of exact inference (Beal, 2003; Fox & Roberts, 2011). Variational Bayes rests on a factorization of approximate posterior beliefs that greatly reduces the number of expectations required to encode it. The factorization we focus on exploits the Markovian nature of the generative model and has the following form (see Friston et al., 2013 for details):

$$\begin{aligned} Q(\tilde{s}, \tilde{u}, \gamma|\mu) &= Q(s_0|\widehat{s}_0) \ldots Q(s_T|\widehat{s}_T) \\ &\quad Q(u_t, \ldots, u_T|\widehat{\pi})Q(\gamma|\widehat{\gamma}) \\ &\quad Q(\gamma|\widehat{\gamma}) = \Gamma(\alpha, \widehat{\beta} = \alpha/\widehat{\gamma}) \end{aligned} \tag{10}$$

This assumes a factorization over hidden states, (future) control states, and precision. It is this factorization that renders the inference approximate and resolves many of the intractable problems of exact inference. For example, the factorization does not consider sequences of hidden states, which means we only have to evaluate sequences of control states (as opposed to all possible sequences of controlled state transitions). We have assumed here that the posterior marginal over precision is, like its conjugate prior, a gamma distribution. The rate parameter of this posterior belief $\widehat{\beta} = \alpha \big/ \widehat{\gamma}$ corresponds to temperature in classic formulations. However, it is no longer a fixed parameter but a sufficient statistic of beliefs about policies.

Given the generative model (Equation 7) and the mean field assumption (Equation 10), it is straightforward to solve for the expectations that minimize variational free energy (see Appendix B).

$$\begin{aligned} \widehat{s}_t &= \sigma(\ln \mathbf{A} \cdot o_t + \ln(\mathbf{B}(a_{t-1})\widehat{s}_{t-1})) \\ \widehat{\pi} &= \sigma(\widehat{\gamma} \cdot \mathbf{Q}) \\ \widehat{\gamma} &= \frac{\alpha}{\beta - \mathbf{Q} \cdot \widehat{\pi}} \end{aligned} \tag{11}$$

Iterating these self-consistent equations until convergence produces the posterior expectations that minimize free energy and provides Bayesian estimates of the unknown variables. This means that expectations change over two timescales: A fast timescale that updates posterior beliefs between observations and a slow timescale that updates posterior beliefs as new observations are sampled. We have speculated (Friston, Samothrakis, & Montague, 2012) that these updates may be related to nested electrophysiological oscillations, such as phase coupling between gamma and theta oscillations in prefrontal–hippocampal interactions (Canolty et al., 2006). See also (Penny, Zeidman, & Burgess, 2013). The forms of these updates are remarkably simple and we now consider each in turn.

The first equation updates expectations about hidden states and corresponds to *perceptual inference* or *state estimation*. This is essentially a Bayesian filter that combines predictions based upon expectations about the previous state with the likelihood of the current observation. For simplicity, we have ignored the dependency of value on expected states that would introduce a third (optimism bias) term (see Appendix B).

The second update is just a softmax function of the value of each policy, where the sensitivity parameter or expected precision is an increasing function of expected value. This last point is quite important: It means that the sensitivity or inverse temperature, that determines the precision with which a policy is selected, increases with the expected value of those policies.

The third update optimizes expected precision. If we express these updates in terms of the posterior rate parameter, we see that changes in (inverse) precision are changes in (negative) expected value: $\widehat{\beta} = \beta - \mathbf{Q} \cdot \widehat{\pi}$. In other words, if an observation increases the expected value of the policies entertained by an agent, then expected precision increases (i.e., temperature decreases) and the agent is implicitly more confident in selecting the next action. As noted above, this may explain why dopamine discharges have been interpreted in terms of changes in expected value (e.g., reward prediction errors). The role of the neuromodulator dopamine in encoding precision is further substantiated by noting that precision enters the variational updates in a multiplicative or modulatory fashion. We will pursue this in the next section.

Summary

In summary, by assuming a generic (Markovian) form for the generative model, it is fairly easy to derive Bayesian updates that clarify the interrelationships between expected value and precision—and how these quantities shape beliefs about hidden states of the world and subsequent behavior. Furthermore, the anatomy of this message passing is not inconsistent with functional anatomy in the brain (see Friston et al., 2014, and Figure 1 in this paper). The implicit computational anatomy rests on reciprocal message passing between expected policies (e.g., in the striatum) and expected precision (e.g., in the substantia nigra). Expectations about policies depend upon value that, in turn, depends upon expected states of the world that are iterated forward in time—to evaluate free energy in the future (e.g., in the prefrontal cortex; Mushiake, Saito, Sakamoto, Itoyama, & Tanji, 2006) and possibly hippocampus (Pezzulo, Van der Meer, Lansink, & Pennartz, 2014). In the next section, we illustrate the basic behavior of this scheme using simulations.

INFERENCE AND PLANNING

This section considers inference using simulations of foraging for information in a relatively simple environment. Its focus is on the comparative performance when minimizing expected free energy, relative to the special cases of risk-sensitive control and maximizing expected utility or reward. In particular, we will look at the neuronal correlates of the scheme in terms of simulated dopaminergic responses. The problem we consider can be construed as searching for rewards in a T-maze. This T-maze offers primary rewards (or, in Pavlovian terms, *unconditioned stimuli*; US) such as food and cues (or *conditioned stimuli*; CS) that are not rewarding per se but disclose rewards that can be secured subsequently. The basic principles of this problem can be applied to any number of scenarios (e.g., saccadic eye movements to visual targets). This example was chosen to be as simple as possible, while illustrating a number of key points that follow from the theoretical considerations above. Furthermore, this example can also be interpreted in terms of responses elicited in reinforcement learning paradigms by unconditioned (US) and conditioned (CS) stimuli. We will call on this interpretation when relating precision updates to dopaminergic discharges.

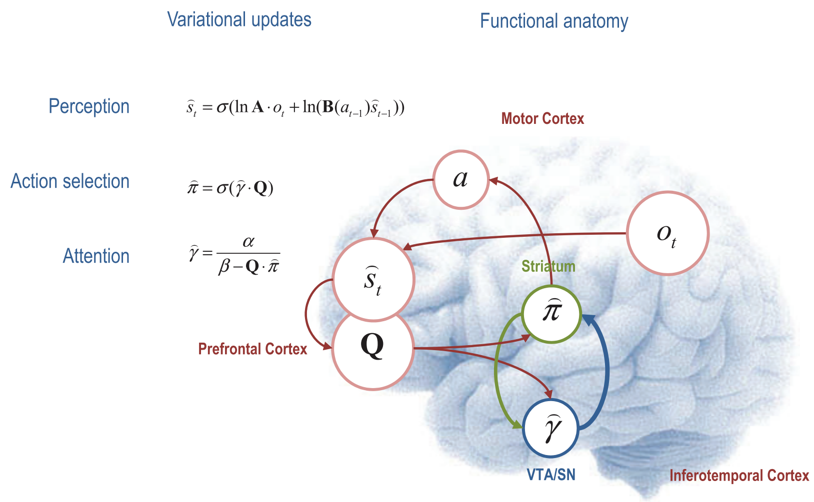

Figure 1. This figure illustrates the cognitive and functional anatomy implied by the variational scheme in the main text. Here, we have associated the variational updates of expected hidden states with perception, of control states (policies) with action selection and, finally, expected precision with attention or salience. In terms of neuronal implementation, the requisite exchange of expectations can 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 (which plausibly interacts with the hippocampus in this context), while assigning action selection to the basal ganglia. Precision has been associated with dopaminergic projections from ventral tegmental area and substantia nigra. See main text for a full description of the equations.

The setup

The agent (e.g., 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 cue (CS), which tells the animal whether the reward is in the upper right or left arm. Crucially, the agent can only make two moves from any location to another (for simplicity, we do not require the agent to visit intermediate locations). 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 secure the reward at the cued location in the appropriate upper arm (i.e., the agent has to move away from the goal so that it can be secured later, as in the coastal navigation example). It is this epistemic behavior we hoped would emerge as a natural consequence of minimizing expected free energy. This may seem a remarkably simple problem but it has all the ingredients necessary to illustrate the basic aspects of behavior under active inference.

Formally, 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 thought to be subserved by the hippocampus (e.g., Moser, Kropff, & Moser, 2008). There are four (locations) times two (right and left reward) hidden states and 16 outcomes. The 16 outcomes correspond to the four locations times four stimuli (cue right, cue left, reward, and no reward). Having specified the state space, it is now only necessary to specify the $(\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D})$ matrices encoding transition probabilities and preferences. These are shown in Figure 2, where the **A** matrix maps from hidden states to outcomes, delivering an uninformative cue at the center (first) location[5] and a definitive cue at the lower (fourth) location. The remaining locations provide a reward (or not) with probability $a = 90\%$ depending upon the hidden context (right versus left reward).

The $\mathbf{B}(u)$ matrices encode control-dependent transitions to the corresponding location, with the exception of the baited (second and third) locations, which are hidden states that the agent cannot leave. The vector **C** determines prior preferences about outcomes. These are expressed in terms of a softmax function of utility, which determines the relative log probability of each outcome. Here, the utility of the rewarding stimulus is c and its absence $-c$. This means, the agent expects a rewarding outcome $\exp(2c)$ times more than a null outcome. For example, if $c = 1$ it would expect a reward about $\exp(2) \approx 8$ times more than no reward. Note that utility is always relative and has a quantitative meaning in terms of relative (log) probabilities of preferred outcomes. This is important because it endows utility with the same measure as information; namely, bits or nats (i.e., units of information or entropy, the former assuming base 2 logarithms and the latter based on natural logarithms). This highlights the close connection between value and information (see below). Finally, the vector **D** specifies the agent's beliefs about the initial conditions; namely, that it starts at the center location with equiprobable baiting of the right or left arm.

Having specified the state space and contingencies, one can iterate the variational updates (in Equation 11 and Figure 1) to simulate behavior. In these simulations the outcomes were generated using the contingencies of the generative model. In other words, we assume the agent has already learned or optimized its model of the generative process (in terms of the model structure and its parameters). We will revisit this assumption in the last section.

Figure 3 shows the results of simulations in terms of performance (upper panel) and the dynamics of Bayesian updating in terms of precision or simulated dopaminergic responses (lower panels). The upper panel shows performance as the percentage of successful (rewarded) trials with increasing levels of utility, using six equally spaced levels from $c = 0$ to $c = 2$. Performance was assessed using 128 trials, under three different schemes: Minimizing expected free energy, risk-sensitive (KL) control, and maximizing expected utility. For completeness, we also provide the results for the free energy minimization when suppressing precision updates. The three schemes can be considered as special cases that result when successively removing terms from the expected free energy (to give reduced forms indicated by the brackets).

$$\mathbf{Q}_\tau(\pi) = E_{Q(o_\tau,s_\tau|\pi)}[\underbrace{\ln P(o_\tau|s_\tau) \underbrace{- \ln Q(o_\tau|\tilde{u}) + \underbrace{\ln P(o_\tau|m)}_{\text{Expected utility}}}_{\text{KL control}}}_{\text{Expected Free energy}}] \tag{12}$$

This expression shows that risk-sensitive control is the same as minimizing expected free energy when ignoring the (predictive) entropy of outcomes given

[5]The values of one half in the first block of the **A** matrix (Figure 2) mean that the agent cannot predict the cue from that location. In other words, there is no precise sensory information and the agent is "in the dark."

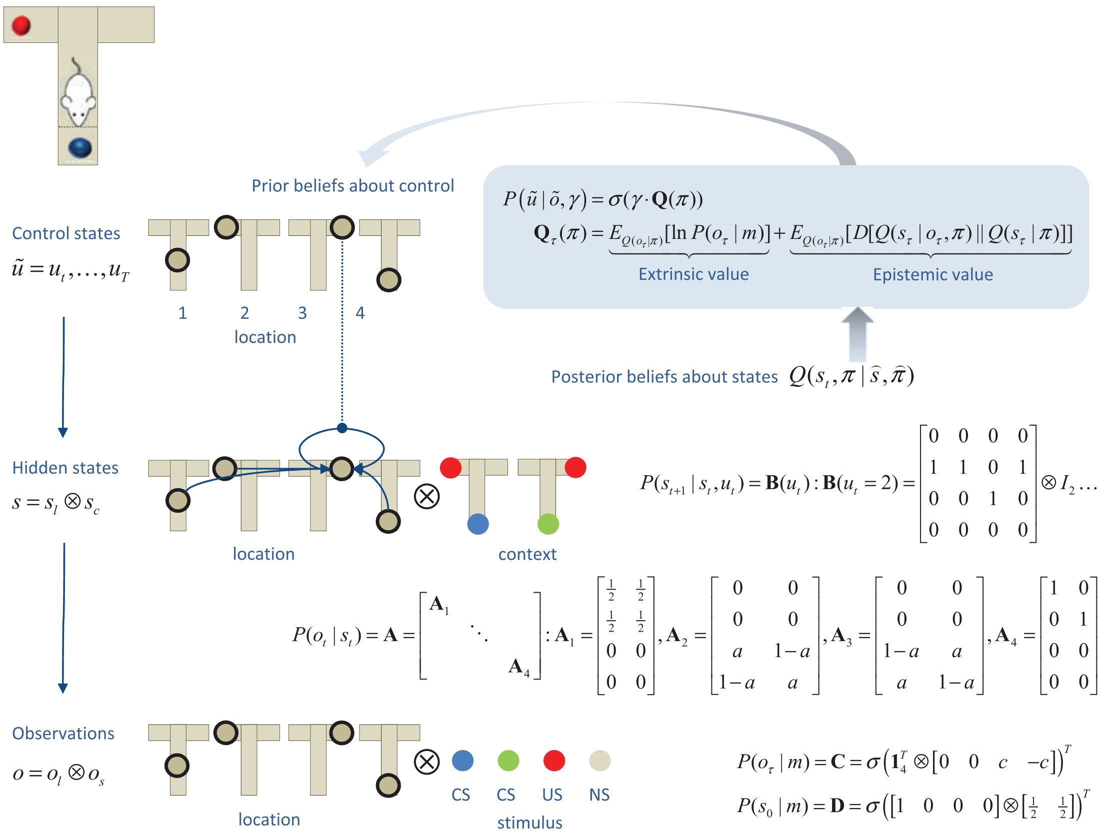

Figure 2. A schematic of the hierarchical generative model used to simulate foraging in a three-arm maze (insert on the upper left). This model contains four control states that encode movement to one of four locations (three peripheral locations and a central location). These control the transition probabilities among hidden states that have a factorial or tensor product form with two factors. The first is the location (one of four locations), while the second is one of two hidden states of the world, corresponding to a combination of cues (blue or green circles) and rewarding (red) outcomes. Each of the ensuing eight hidden states generates an observation. 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 $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}) \subset \theta$ as described in the main text. In this figure, $\sigma(\cdot)$ is a softmax function and $\otimes$ denotes a Kronecker tensor product. Although the graphics are arranged in rows, the vectors of states are actually row vectors.

hidden states. In other words, if every hidden state generates a unique outcome, KL control and expected free energy minimization would be the same. Similarly, maximizing expected utility is the same as minimizing KL divergence, if every outcome is generated by a unique hidden state and we do not have to maximize the entropy of outcomes. In short, expected utility and classical reinforcement schemes (Sutton & Barto, 1998) are special cases of risk-sensitive control that are optimal when (and only when) different hidden states generate different outcomes. Similarly, risk-sensitive control is a special case of free energy minimization that is optimal when (and only when) different outcomes are generated by different hidden states. These special cases are important because they highlight the epistemic value of informative observations, of the sort that are precluded by noisy or context-sensitive observations. This nesting within free energy minimization may also explain the prevalence of classical schemes in the literature, given that they generally assume hidden states are known to the agent.

The performance of the different schemes (see Figure 4, upper panel) speaks to several intuitive and useful points. First, all the schemes show an increased success rate as utility or prior preference increases; however, only expected free energy minimization attains near optimal performance (90%). One might ask why risk-sensitive control performs so poorly, given it is also sensitive to uncertainty. However, KL schemes only consider uncertainty or risk induced by many hidden states causing a single outcome, as opposed to many outcomes caused by a single state. If we had used more locations (say, with a radial maze), the benefits

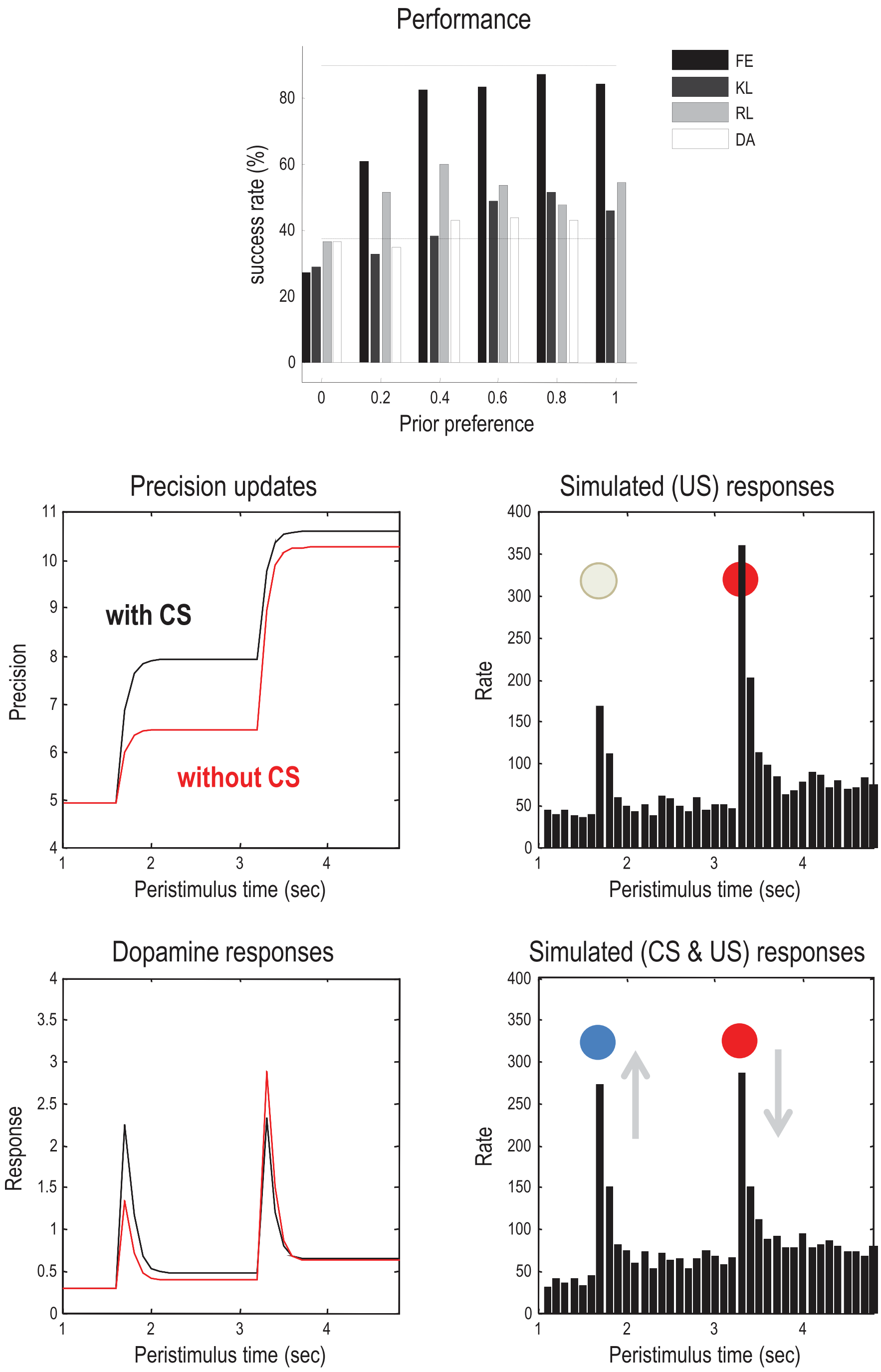

Figure 3. Upper panel: The results of 128 simulated trials assessed in terms of the probability of obtaining a reward. This performance is shown as a function of prior preference over six equally spaced levels. The four profiles correspond to active inference (FE), risk-sensitive control (KL), expected utility (RL), and active inference under fixed levels of precision (DA). See main text for a description of these schemes and how they relate to each other. The two horizontal lines show chance (bottom line) and optimal (top line) performance, respectively. Lower left panels: These report expected precision as a function of time within a trial (comprising three movements). The black lines correspond to a trial in which the cue (CS) was first accessed in the lower arm of the maze in the previous figure, after which the reward (US) was secured. The equivalent results, when staying at the center location and accessing the reward directly, are shown as red lines. The upper panel shows the expected precision and the lower panel shows simulated dopamine responses (that produce an increase in precision, which subsequently decays). Lower right panels: These show the equivalent results in terms of simulated dopamine discharges. The key thing to note here is that the responses to the cue (CS) are increased when it is informative (i.e., accessed in the lower arm), while subsequent responses to the reward (US) are decreased. See main text for details of these simulated responses.

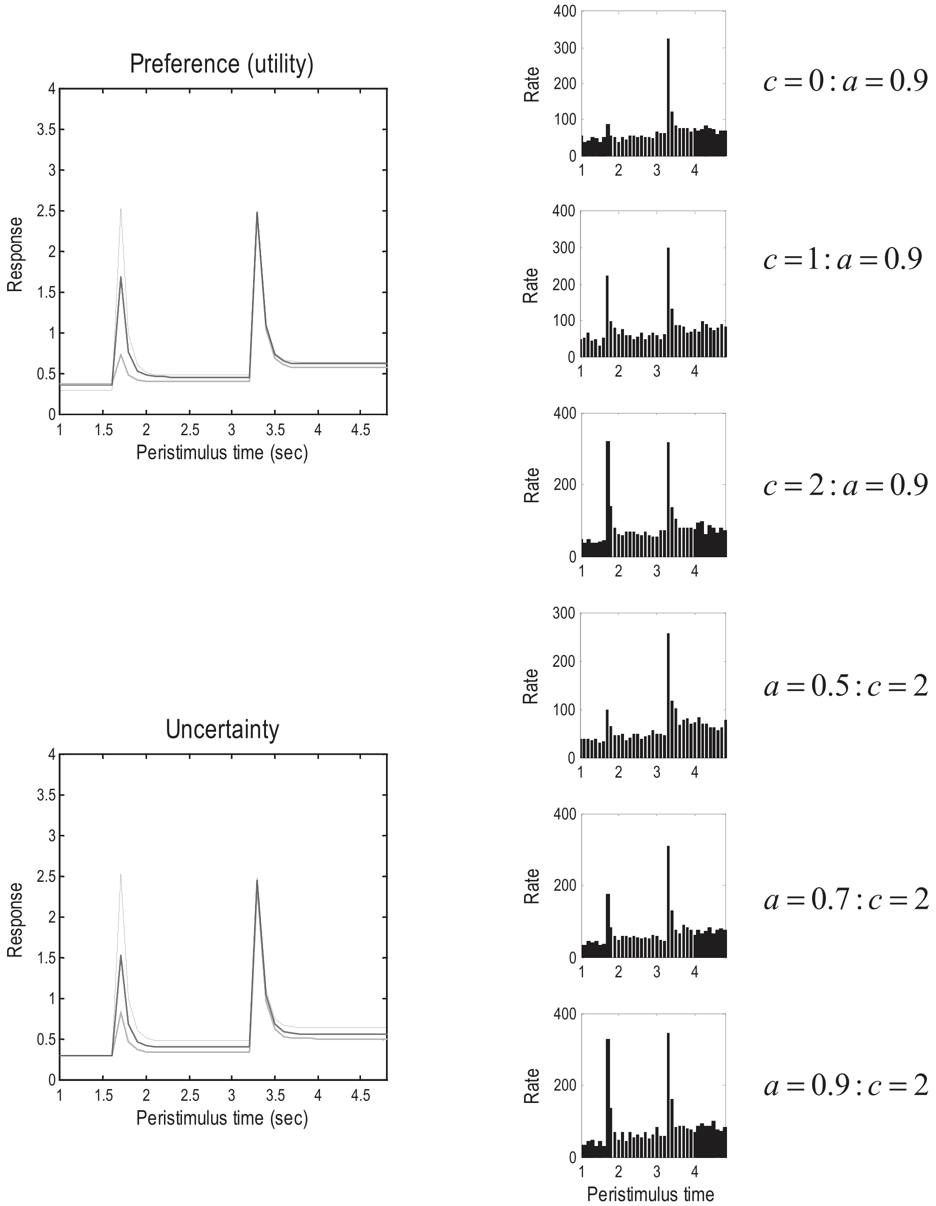

Figure 4. Simulated dopamine responses as a function of preference (upper panels) and uncertainty (lower panels). The left panels show the expected dopaminergic responses using the same format as Figure 3, for three levels of preference (utility) and uncertainty in the upper and lower panels, respectively. The right-hand panels show simulated dopaminergic firing in response to the cue (CS) and reward (US) based upon these expectations. Note that the response to the cue (CS) increases with preference and a reduction in uncertainty.

of risk-sensitive control would have been more apparent (results not shown). This follows because more locations induce more hidden states and a greater degree of uncertainty that would call for risk-sensitive control. The current setup illustrates the importance of considering both sorts of ambiguity in the mapping between causes and consequences (one-to-many and many-to-one) that calls for a minimization of expected free energy. In this example, the most informative location is the lower arm. Visiting this location and sampling the informative cue reduces uncertainty about hidden states and enables expected utility to dominate in the second (and generally successful) move.

This optimal behavior is only apparent when utility is greater than about c > 0.6 nats. This brings us to our second point: If expected utility or preferences are to supervene over epistemic value, then they have to have greater value than the information gain associated with informative outcomes. In this example, the cue resolves uncertainty about the hidden context (which arm is rewarding), thereby providing one bit or $\ln(2) = 0.6931$ nats of information. Intuitively, this means the utility must be greater than the information gain to persuade an agent to leave locations that provide unambiguous (informative) outcomes.

The third point of note is that expected utility (and risk-sensitive control) perform below chance levels when utility is zero (note that chance performance is $\frac{3}{8}$ because the agent can make two moves and is trapped by two locations). This reflects the fact that the upper arms no longer hold any utilitarian attraction and become unattractive because of the slightly ambiguous outcomes implicit in the probabilistic reward schedule. In other words, the most valuable policy is to stay in the epistemically valuable (lower) location for as long as possible.

The last set of simulations under a fixed level of precision $\hat{\gamma} = 1$ show that optimal choice behavior rests on updating expected precision, which we now look at more carefully. The lower panels of Figure 3 show how expected precision is updated during a trial of two movements and three outcomes under a high level of utility $c = 2$. These simulations are presented to highlight the similarity between precision updating and empirical dopamine responses during the presentation of conditioned and unconditioned stimuli. The upper left panel shows the expected precision over variational updates, with 16 updates between observations. The black lines correspond to a trial in which the agent accessed the conditioned stimulus (CS or cue) in the lower arm and then secured the unconditional stimulus (US or reward) on moving to an upper arm. The red lines show the equivalent updates in a second trial, when the agent stayed at the central location for the first move and was then presented with the US. In both situations, the precision increases with each successive outcome; however, the expected precision is higher in the first trial, when the CS reduces uncertainty about the hidden states or context in which the agent is operating. This reflects the greater epistemic value of accessing the cue. Crucially, the precision of the final state is roughly the same for both trials. The implication of this is that expected precision increases on presentation of the CS and, necessarily, increases less on presentation of the subsequent US, relative to presentation of the US alone.

This difference can be highlighted by plotting the expected precision in terms of simulated dopaminergic discharges, which are thought to reflect changes in expected precision or value. More exactly, the lower left panel shows simulated dopamine discharges that would, when convolved with a decaying exponential response function (with a time constant of 16 iterations) reproduce the expected precision in the upper panel. In other words, we are assuming that dopamine mediates increases in precision that subsequently decay with a fixed time constant. In this format, one can clearly see the phasic responses of expected precision (simulated dopaminergic discharges) where, crucially, the presentation of the CS reduces the response to the US. This reproduces the well-known transfer of dopamine responses from a US to a CS in operant paradigms (Schultz, Apicella, & Ljungberg, 1993).

The right panels of Figure 3 shows simulated dopamine discharges assuming that an expected precision of one is encoded by 128 spikes per bin (and firing rates are sampled from a Poisson distribution). These are remarkably similar to empirical results, often interpreted in terms of reward prediction error and temporal difference models of value learning. However, the current framework offers a nuanced perspective; namely, the CS has epistemic value that reduces uncertainty about what will happen next. This uncertainty is already resolved when the US is presented, thereby attenuating the precision-dependent responses it elicits. Put simply, the transfer of dopaminergic responses to conditioned stimuli, in higher-order operant paradigms, can be thought of as reporting the confidence (precision) that policies will bring about predicted outcomes.

The composition of extrinsic and epistemic value implicit in expected free energy can also be used to reproduce the empirical responses of dopaminergic cells to CS under different levels of reward and uncertainty (Fiorillo et al., 2003). Figure 4 shows simulated dopamine responses under increasing utility $c = \{0, 1, 2\} : a = 0.5$ and different levels of uncertainty about the reward probability $a = \{0.5, 0.7, 0.9\} : c = 2$. In both cases, the response to the CS increases in a way that is remarkably reminiscent of empirical results (Fiorillo et al., 2003). Interestingly, the tonic responses appear to be more sensitive to uncertainty (lower panels) than utility (upper panels). This is also seen empirically, although the tonic responses reported in Fiorillo et al. (2003) increased in a ramp-like fashion under higher levels of uncertainty (i.e., $a = 0.5$). This phenomenon is not reproduced in Figure 5, however. Generally, precision increases as the trial progresses because agents become increasingly confident about their policies. One can see this general trend in Figure 4.

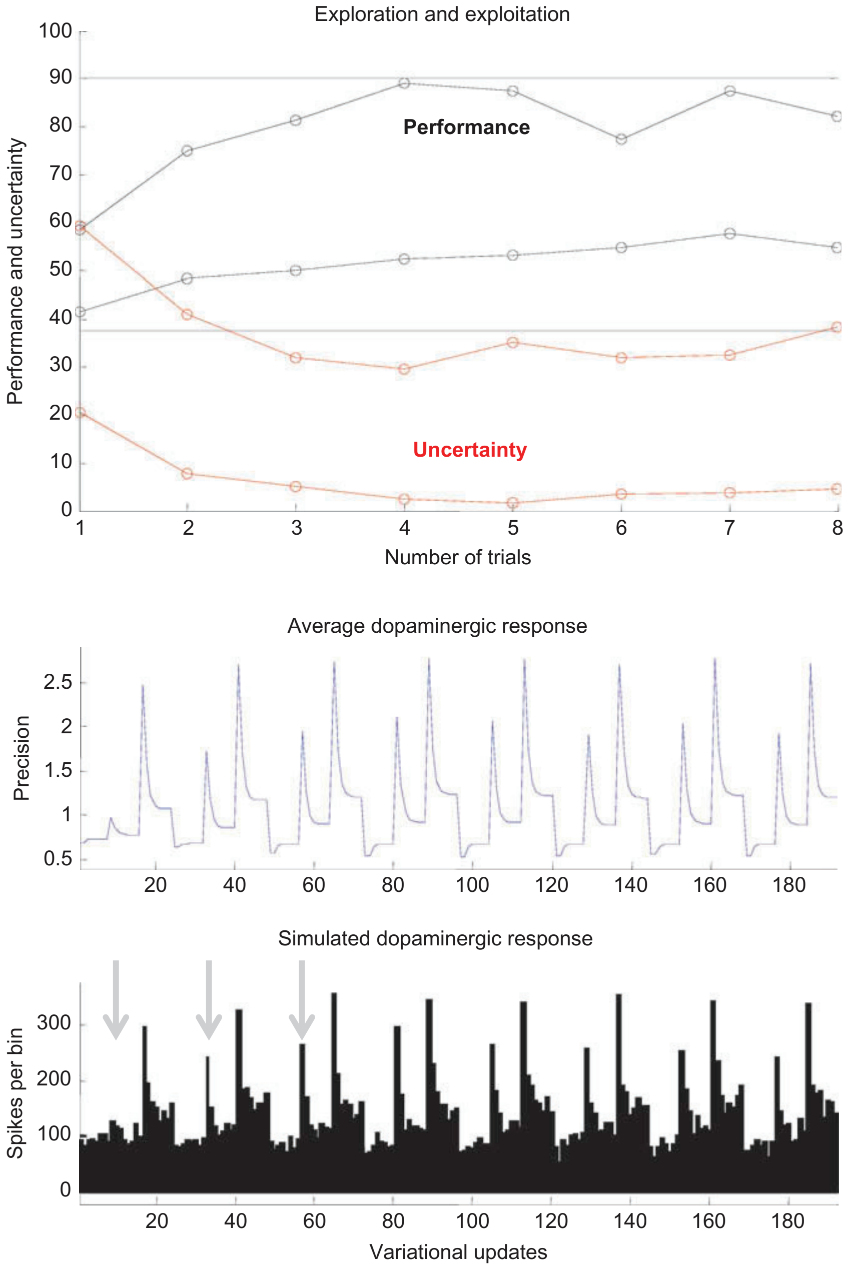

Figure 5. Upper panel: Learning in terms of success rate as a function of trials, for eight successive trials in an initially unknown maze. The results are averaged over 128 realizations. Performance (gray lines) shows a progressive improvement as uncertainty about the hidden states falls (pink lines). The equivalent performance for a conventional expected utility scheme is shown with broken lines. Lower panels: Simulated dopamine responses over all iterations and trials shown in terms of average precision (middle panel) and simulated dopaminergic spike rates (lower panel). These results demonstrate the transfer of simulated dopamine responses to the cue (CS) with learning (gray arrows).

Summary

In summary, we have seen that optimal choice behavior, in a very simple paradigm, rests on resolving uncertainty about future choices implicit in minimizing expected free energy. This aspect of optimal behavior is clearly disclosed when decisions under uncertainty are confounded, not only by a many-to-one mapping between hidden states and outcomes, but also between outcomes and hidden states (c.f., Littman, Sutton, & Singh, 2002). In this general setting, the role of epistemic value becomes paramount in resolving uncertainty about what to do next—a resolution that can be construed in terms of exploration or foraging for information. Furthermore, the integrative framework provided by free energy minimization enforces a dialogue between utility and information by casting both as log probabilities. This means every utility or reward can be quantified in terms of information and every bit of information has utility. We have considered the encoding of this information in terms of precision, showing that biologically plausible variational updates of expected precision are remarkably consistent with empirical dopaminergic responses. A key aspect of this formulation is that the precision of beliefs about the value of policies is itself an increasing function of expected value. This means that if dopamine reports (changes in) precision, it also reports (changes in) expected value and, implicitly, reward prediction error. So far, we have limited our discussion to planning as inference and memory. In the final section, we turn to the role of epistemic value in learning and memory, touching on some important issues that attend hierarchical inference and contextualizing behavior.

LEARNING AND MEMORY AS INFERENCE

This section uses the same setup but considers multiple trials during which the agent has to learn which locations deliver rewards and cues. In other words, we introduce an extra hierarchical level to the problem, where the hidden context now includes the mapping between locations and actions (i.e., moving to the lower arm could take it to a rewarding location). This means the agent has to learn which locations offer cues and which offer rewards. The motivation here is to illustrate a form of learning that rests on exploring the environment and to show that there is a Bayes-optimal transition from exploration to exploitation. Crucially, this solution rests upon exactly the same scheme as above—the only thing that changes is the generative model.

There are many ways of modeling learning in this context. These range from Bayesian model selection and averaging, aka structure learning (FitzGerald, Dolan, & Friston, 2014), through optimization of the model parameters $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}) \subset \theta$ with respect to expected free energy, to casting the problem as a hierarchical inference problem (c.f., Ballard, Kit, Rothkopf, & Sullivan, 2013). We will choose the latter because it requires no extra theory[6] and illustrates how hierarchical inference contextualizes lower-level (habitual) action selection. In brief, we will use Bayesian belief updating that embodies the prior that the mapping between locations and control states does not change from trial to trial, but the location of the reward changes between trials. The agent therefore has to learn (infer) time-invariant (contextual) aspects of its environment through exploration, before it can engage in pragmatic goal-directed behavior. We will see that this learning is an emergent property of minimizing expected free energy at each move.

The setup

Our aim was to illustrate learning as inference by introducing hierarchical uncertainty into the setup. In other words, we wanted to see if the agent could learn about its environment by introducing uncertainty about which locations offered rewards and cues. In discrete state-space formulations, hierarchical extensions involve creating product spaces, such that each lower-level state is reproduced under each level of a higher-level state. Here, we considered four higher-level hidden contexts $S^{(2)}$ corresponding to four mappings between each of the three arms of the T-maze. More specifically, we introduced four mappings between the three control states and the associated hidden location states that determine outcomes. This just involved changing the following matrices, where we denote the hierarchical level of parameters and states with superscripts (such that $\mathbf{A}, \mathbf{B}, \ldots$ above become $\mathbf{A}^{(1)}, \mathbf{B}^{(1)}, \ldots$):

$$\begin{aligned}
\mathbf{A} &= [\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(1)}] \\
\mathbf{B}(i) &= \begin{bmatrix} \mathbf{B}^{(1)}(j_{1i}) & & \\ & \ddots & \\ & & \mathbf{B}^{(1)}(j_{4i}) \end{bmatrix} \\
\mathbf{C} &= \mathbf{C}^{(1)} \\
\mathbf{D} &= \tfrac{1}{4} \begin{bmatrix} \mathbf{D}^{(1)} \\ \vdots \\ \mathbf{D}^{(1)} \end{bmatrix}
\end{aligned} \tag{13}$$

Here, j_{ki} returns the index of the i-th control state under the k-th context; e.g., $j_{k\cdot} = [1, 2, 4, 3]$. This means that there are now 32 hidden states $S = S^{(2)} \otimes S^{(1)}$ (four

[6]For example, we do not have to worry about how the agent learns all possible configurations of the maze.

mazes, times four locations, times two reward contexts). The agent now has two levels of uncertainty to resolve. The first is induced by not knowing which maze it is in and the second is resolved by the cue, if it can be found. Neurobiologically, uncertainty about spatial context may be reflected in hippocampal processing (e.g., the "flickering" reported in Jezek, Henriksen, Treves, Moser, and Moser 2011).

Variational updating at the second level corresponds to replacing prior beliefs about hidden states with posterior beliefs after the previous trial. This corresponds to minimizing variational free energy because beliefs about the initial state $Q(s_0) = \mathbf{D} = \hat{s}_0$ become empirical priors that are informed by previous trials:

$$\begin{aligned}\hat{s}'_0 &= \mathbf{E}\hat{s}'_T \\ \mathbf{E} &= P(s'_0|s'_T, m) \\ &= ((1-e)I_4 + e) \otimes (\mathbf{D}^{(1)} \otimes \mathbf{1}_8^T)\end{aligned} \tag{14}$$

Here, (s_0, s_T) and are the posterior expectations at the end of the previous trial and the beginning of the current trial respectively, and where $\mathbf{E} = P(s_0|s_T, m)$ encodes beliefs that the maze will change with a small probability $e = \frac{1}{8}$. This Bayesian belief updating is a formal way of saying that agents remember what they have learned from previous experience.

Figure 5, shows the results of this memory in terms of success rate as a function of trials, for eight successive trials in the same (randomly selected) mazes. The results are averaged over 128 realizations. Performance (gray lines) shows a progressive improvement as uncertainty about the hidden states falls (pink lines). This uncertainty is the entropy of posterior beliefs at the end of each trial $H[Q(s_T)]$ (multiplied by 100 for visual display). The equivalent performance for a conventional expected utility scheme is shown with dotted lines. The key thing to take from these results is that performance becomes near optimal after about four trials, at which point uncertainty falls to (nearly) zero. This means that, on average, the agent has learned which maze it is in after four trials and can then invoke the exploitative strategy of the previous section, first searching for the cue and then claiming the reward.

Crucially, despite the fact there is no explicit epistemic value involved in inference about the environment (maze) at the between-trial level, a failure to consider epistemic value at the within-trial level has deleterious consequences for learning, in that the expected utility agent fails to learn which maze it is in (and is content to perform at the levels it would even if it knew). Note that the performance in Figure 5 never exceeds the performance shown in Figure 3.

The lower panels of Figure 5 show simulated dopamine responses (using the format of previous figures) over all iterations and trials. These results demonstrate the transfer of simulated dopamine responses to the cue or conditioned stimulus as learning progresses (gray arrows). These trial by trial changes are accompanied by elevated tonic responses after the CS—that reflect increasing confidence or precision about the outcomes of policies as the agent becomes familiar with its new environment (Hollerman & Schultz, 1996; Niv, 2007).

Summary

This section has shown it is straightforward to create hierarchical generative models, in which higher levels provide a context for lower levels, by equipping the model with a joint state-space $S = S^{(2)} \otimes S^{(1)}$ and associated transition matrices. This enables one to consider contingencies that are conserved (or not) over trials. In a multi-trial setting, priors over the initial state of each successive trial become empirical priors that minimize variational free energy (in exactly the same way as beliefs are updated within trials). This is simple to implement using Bayesian belief updating and allows a natural separation of temporal scales across hierarchical levels. It is relatively easy to see how one could generalize this to hierarchically deep models of the sort that real agents have to deal with, e.g., $S = S^{(3)} \otimes S^{(2)} \otimes S^{(1)}$.

This hierarchical augmentation reveals the role of integrating extrinsic and epistemic value in enabling the agent to learn which context it is operating in and then exploit that knowledge. It is tempting to associate this (inevitable and emergent) progression from exploration to exploitation with the transformation of goal-directed behavior into habits (Balleine & Dickinson, 1998; Dolan & Dayan, 2013; Pezzulo, Rigoli, & Chersi, 2013). Here, this Bayes optimal progression rests upon a contextualization of (first level) choice behavior by (second level) Bayesian updating that effectively accumulates evidence to resolve uncertainty about the consequences of behavior. This resolution restricts the repertoire of controlled state transitions that have to be considered in selecting the optimal policy and effectively increases the precision of action selection.

One might think that much of the delicate balance between exploration and exploitation could rest upon hierarchical active inference of this sort.

DISCUSSION

Formal approaches to decision-making under uncertainty generally rest on partially observable Markov decision processes, in which states are not directly observable but have to be inferred from observations. This formalism raises two fundamental issues that can be cast in terms of the exploration-exploitation dilemma. First, in relation to inference, in some circumstances an agent might obtain a larger reward by performing an epistemic (explorative) action rather than a more greedy (pragmatic) action. Second, in relation to learning, an epistemic action may be more appropriate to resolve uncertainty about aspects of its generative model. In classical formulations, the exploration-exploitation dilemma is usually solved with ad hoc solutions (like changing the precision of softmax decision rules). Here, we introduce a theoretical framework within which a solution to the exploration-exploitation dilemma emerges normatively from the minimization of expected free energy. For example, the precision or temperature parameter of softmax response rules becomes a parameter of the generative model and thereby acquires a Bayes optimal value.

More specifically, we have introduced a modeling framework for choice behavior that can be framed in terms of discrete states or (partially observed) Markov decision processes. There are two perspectives on this framework. People familiar with active inference could consider this work to show that the minimization of expected free energy furnishes a sufficient account of choice behavior under uncertainty. This necessarily entails epistemic action, providing a formal account of risk-sensitive or KL control and expected utility theory. The ensuing scheme also has construct validity in relation to Bayesian surprise and information theoretic formulations of search behavior. Crucially, the minimization of expected free energy eschews ad hoc parameters associated with conventional treatments (e.g., softmax parameters). Furthermore, active inference under hierarchical models may provide a useful framework within which to consider the contextualization of low-level behaviors that involves a natural (Bayes-optimal) progression from exploration to exploitation. Finally, it enables one to finesse the combinatorics of difficult or deep Markovian problems using approximate Bayesian inference—and a message passing scheme that is not biologically implausible. In particular, the variational updates for expected precision show many similarities to empirical dopaminergic responses.

Our simulations suggest that it is difficult to completely suppress precision updates (dopaminergic responses), even when outcomes are very predictable (because every event portends something in our finite horizon setup). This contrasts with the classic results of Schultz and colleagues (Schultz et al., 1993), who found negligible responses to conditioned stimuli after learning. On the other hand, we were able to reproduce the empirical findings under conditions of uncertainty and predictive reward (Fiorillo et al., 2003; Schultz, 1998). Furthermore, the simulations reproduce the empirical observation that dopaminergic responses are transferred directly from the unconditioned stimuli to the conditioned stimuli, in the absence of any responses during the intervening period. Detailed response characteristics of this sort may provide important clues that may disambiguate, or further refine, theoretical accounts of dopaminergic function.

The second perspective on this work could be taken by people familiar with reinforcement learning, a branch of machine learning inspired by behavioral psychology (Sutton & Barto, 1998). From this perspective one can trace the steps that lead from normative descriptions based upon expected reward or utility to active inference and variational free energy minimization:

- The first step is to reformulate reinforcement learning or game theory problems as pure inference problems, i.e., planning as inference (Botvinick & Toussaint, 2012; Still, 2009; Still & Precup, 2012; Vijayakumar, Toussaint, Petkos, & Howard, 2009). This means that reward or utility functions become log probabilities defining prior beliefs or preferences about future outcomes. This induces probability distributions over policies that produce outcomes—and the precision of those distributions. This is important because it defines a Bayes-optimal precision for selecting among policies (Friston et al., 2014). Furthermore, casting reward or utility in terms of log probabilities means that they have the same currency as information (nats or bits), thereby providing a natural way to combine the value of an outcome and the value of information.
- The second step rests on accommodating uncertainty or risk over outcomes. When the expected utility of two choices is the same but one leads to several outcomes and the other a single outcome, then optimal behavior is not uniquely defined by expected utility. The simplest way to

accommodate uncertainty (risk) of this sort is to maximize both expected utility and the entropy of outcomes. Maximizing the expected entropy of outcomes effectively keeps one's options open (Klyubin, Polani, & Nehaniv, 2008). However, the sum of an entropy and expected utility can always be expressed as a (negative) KL divergence, leading to risk-sensitive or KL control. Formally, maximizing the expected utility and entropy of outcomes is equivalent to minimizing the KL divergence between the expected (predictive) distribution over outcomes and the distribution specified by the utility function. In behavioral terms, maximizing the entropy of controlled outcomes can be understood in terms of novelty bonuses and related concepts (Bach & Dolan, 2012; Daw et al., 2005; De Martino, Fleming, Garrett, & Dolan, 2012; Kakade & Dayan, 2002; Wittmann et al., 2008). In economics, there is a conceptual link with Shackles' formulation of *potential surprise* and the crucial role of money in facilitating risk-sensitive control (Shackle, 1972): If I am not sure what I want to buy, then I will save my money (liquid assets) and buy something later (maximize the entropy over future purchases—a fiscal Ockham's Razor).

- Risk-sensitive or KL control works fine if there is no uncertainty or ambiguity about hidden states given observed outcomes. However, when the same state can lead to several outcomes (e.g., noisy or ambiguous cues), we have to augment the KL divergence with the expected entropy over outcomes given the hidden states that cause them. Minimizing this entropy ensures that hidden states generating ambiguous (high entropy) outcomes are avoided. In other words, observations that resolve uncertainty about hidden states become intrinsically valuable. However, the sum of the expected conditional entropy and the KL divergence is the expected free energy that scores the quality or value of a policy. This brings us to active inference and the minimization of expected free energy that is sensitive to both risk and ambiguity.

In what follows, we consider some of the theoretical implications of these arguments, in relation to established approaches in psychology and artificial intelligence.

Curiosity and Bayesian surprise

Epistemic value and implicit exploratory behavior are related to curiosity in psychology (Harlow, 1950; Ryan & Deci, 1985) and intrinsic motivation in reinforcement learning (Baldassarre & Mirolli, 2013; Barto, Singh, & Chentanez, 2004; Oudeyer & Kaplan, 2007; Schembri, Mirolli, & Baldassare, 2007; Schmidhuber, 1991). Here *intrinsic* stands in opposition to *extrinsic* (e.g., drive or goal) value. While we have focused on reducing uncertainty during inference, most reinforcement learning research uses curiosity or novelty-based mechanisms to learn a policy or model efficiently. The general idea here is that an agent should select actions that improve learning or prediction, thus avoiding behaviors that preclude learning (either because these behaviors are already learned or because they are unlearnable). It has often been emphasized that adaptive agents should seek out *surprising* stimuli, not *unsurprising* stimuli as assumed in active inference. This apparent discrepancy can be reconciled if one considers that surprising events, in the setting of curiosity and Bayesian surprise, are simply outcomes that are *salient* and minimize uncertainty. In active inference, agents are surprised when they do not minimize uncertainty. It is salient (counterfactual) outcomes that optimize exploration (and model selection) and salience-seeking behavior stems nicely from the more general objective of minimizing expected free energy (or surprise proper).

There is, however, an important difference between active inference and the concepts of curiosity and Bayesian surprise, at least as they are usually used. Salience is typically framed in "bottom-up" terms, in that the agents are not assumed to have a particular goal or task. This is also a characteristic of curiosity (and similar) algorithms that try to learn all possible models, without knowing in advance which will be useful for achieving a specific goal. The active inference scheme considered here contextualizes the utilitarian value of competing policies in terms of their epistemic value, where the implicit reduction in uncertainty is (or can be) tailored for the goals or preferred outcomes in mind.

Active inference and the exploitation-exploration dilemma

The active inference formulation effectively combines belief state updates, action selection, and learning under a single imperative. In principle, this results in the efficient learning of both the structure of the environment and the selection of the suitable policies, thereby avoiding the problems of model-free reinforcement learning algorithms (Sutton & Barto, 1998). Model-free schemes need to relearn a policy every time the environment changes (Ognibene, Pezzulo, & Baldassarre, 2010). Active inference offers a principled solution to the exploration-exploitation dilemma and, in contrast with model-based learning, will not waste time modeling irrelevant aspects of the environment (Atkeson & Santamarıa, 1997). This may enhance learning through generalization, by predominantly sampling features that are conserved when the environmental context changes (Ognibene & Baldassarre, 2014; Walther, Rutishauser, Koch, & Perona, 2005).

Furthermore, active inference extends established metaphors for purely perceptual processing, in particular, hierarchical Bayesian filtering and predictive coding (Clark, 2013; Friston, 2010; Lee & Mumford, 2003; Rao & Ballard, 1999). These perspectives can explain several aspects of cortical hierarchies (Dayan et al., 1995) and provide a nice perspective on the brain as an organ that adapts to model and predict its sensory inputs. This is particularly important because the resulting hierarchical representation (deep generative model) can account for sensorimotor regularities produced by action (Lungarella & Sporns, 2006; O'Regan & Noë, 2001). In turn, this can improve learning and inference, which depend sensitively on an efficient and sparse (hierarchical) representation of active sampling and sensorimotor learning (Ballard et al., 2013; Barto et al., 2004; Tani & Nolfi, 1999). From a modeling perspective, the integration of learning, belief updating, and action selection may allow one to study, in a principled manner, how perception supports learning and how learning can result in different internal representations (Little & Sommer, 2013; Lungarella & Sporns, 2006; Ognibene & Baldassarre, 2014; Verschure, Voegtlin, & Douglas, 2003).

This may be particularly important when modeling inference and behavior in tasks where the agent has no detailed knowledge of the environment, e.g., foraging in open and changing environments, possibly with other agents. These difficult problems have limited the application of the MDP framework to tasks with definitive and detailed representations, such as navigation in grid-based mazes. In open environments, epistemic behaviors have been largely described with heuristic (Brooks, 1991; Itti & Koch, 2001) or stochastic processes such as Lévy flight (Beer, 1995; Viswanathan et al., 1999). However, modeling these problems within the active inference framework may reveal the formal nature of these processes and their neuronal correlates.

Bayesian Reinforcement Learning (e.g., Cao & Ray, 2012) also provides a principled approach to the exploration-exploitation trade-off and explicitly models uncertainty about the quality of alternative policies. Because active inference tackles both the problems of learning and of exploration under partial observations in a coherent manner, it would be interesting to see if Bayesian reinforcement learning could be formulated in terms of active inference. This may be useful, because the current scheme offers computational efficiency, by exploiting variational Bayesian techniques (c.f., Furmston & Barber, 2010), accommodates formal constraints on the structure of policies, and comes with a biologically plausible process theory.

Applications

While these are clearly interesting theoretical issues, the purpose of this paper is also pragmatic. The simulations presented in this paper all use one (Matlab) routine that only requires the specification of the generative model in terms of its $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}) \subset \theta$ parameters. Crucially, integrating this scheme, for any given set of choices and outcomes, provides a generative model of empirical choice behavior. This means, one can estimate the parameters that are unique to a particular subject (human or animal) using standard (meta-Bayesian) schemes (Daunizeau et al., 2010). These parameters include the sensitivity to particular outcomes, beliefs about experimental contingencies, and the overall confidence (and confidence in confidence) encoded by a subject's hyperpriors over precision, e.g., $(c, a, \alpha, \beta) \subset \theta$. This enables a cognitive and possibly physiological phenotyping of subjects using behavioral and physiological responses respectively. Furthermore, one could use Bayesian model comparison to assess whether subjects use expected utility, risk-sensitive control, or full active inference. Indeed, we have shown that the choice behavior and fMRI responses in the dopaminergic midbrain area are better explained in terms of KL control, relative to expected utility using this approach (Schwartenbeck, FitzGerald, Mathys, Dolan, & Friston, 2014). We hope to pursue a similar approach to exploration and decision-making under uncertainty in future work.

Original manuscript received 23 October 2014
Revised manuscript received 25 January 2015
First published online 13 March 2015

REFERENCES

Andreopoulos, A., & Tsotsos, J. (2013). A computational learning theory of active object recognition under uncertainty. *International Journal of Computer Vision*, *101*(1), 95–142. doi:10.1007/s11263-012-0551-6

Atkeson, C., & Santamar´ıa, J. (1997). A comparison of direct and model-based reinforcement learning. In *International Conference on Robotics and Automation (ICRA)*. IEEE Press.

Attias, H. (2003). Planning by probabilistic inference. In C. M. Bishop and B. J. Frey (Eds.), *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*. Key West, FL.

Bach, D. R., & Dolan, R. J. (2012). Knowing how much you don't know: A neural organization of uncertainty estimates. *Nature Reviews. Neuroscience*, *13*(8), 572–586.

Baldassarre, G., & Mirolli, M. (2013). *Intrinsically motivated learning in natural and artificial systems*. Springer: Berlin.

Ballard, D. H., Kit, D., Rothkopf, C. A., & Sullivan, B. (2013). A hierarchical modular architecture for embodied cognition. *Multisensory Research*, *26*, 177. doi:10.1163/22134808-00002414

Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: Contingency and incentive learning and their cortical substrates. *Neuropharmacology*, *37* (4–5), 407–419. doi:10.1016/S0028-3908(98)00033-1

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

Barlow, H. B. (1974). Inductive inference, coding, perception, and language. *Perception*, *3*, 123–134. doi:10.1068/p030123

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 (ICDL 2004)*, San Diego: Salk Institute.

Beal, M. J. (2003). *Variational algorithms for approximate bayesian inference* (PhD. Thesis). University College London.

Beer, R. D. (1995). A dynamical systems perspective on agent-environment interaction. *Artificial Intelligence*, *72* (1–2), 173–215. doi:10.1016/0004-3702(94)00005-L

Berridge, K. C. (2007). The debate over dopamine's role in reward: The case for incentive salience. *Psychopharmacology (Berl.)*, *191*(3), 391–431. doi: 10.1007/s00213-006-0578-x

Bialek, W., Nemenman, I., & Tishby, N. (2001). Predictability, complexity, and learning. *Neural Computation*, *13*(11), 2409–2463. doi:10.1080/01621459.1989.10478825

Bonet, B., & Geffner, H. (2014). Belief tracking for planning with sensing: Width, complexity and approximations. *Journal of Artificial Intelligence Research*, *50*, 923–970.

Botvinick, M., & An, J. (2008). Goal-directed decision making in prefrontal cortex: A computational framework. Advances in Neural Information Processing Systems (NIPS).

Botvinick, M., & Toussaint, M. (2012). Planning as inference. *Trends in Cognitive Sciences*, *16*(10), 485–488. doi:10.1016/j.tics.2012.08.006

Braun, D. A., Ortega, P. A., Theodorou, E., & Schaal, S. (2011). Path integral control and bounded rationality. Adaptive Dynamic Programming And Reinforcement Learning (ADPRL), 2011 IEEE Symposium on, Paris, IEEE.

Bromberg-Martin, E. S., & Hikosaka, O. (2009). Midbrain dopamine neurons signal preference for advance information about upcoming rewards. *Neuron*, *63*(1), 119–126. doi:10.1016/j.neuron.2009.06.009

Brooks, R. (1991). Intelligence without representation. *Artificial Intelligence*, *47*(1–3), 139–159. doi:10.1016/0004-3702(91)90053-M

Bruce, N., & Tsotsos, J. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, *9* (3):5, 1–24. doi:10.1167/9.3.5

Bunzeck, N., & Düzel, E. (2006). Absolute coding of stimulus novelty in the human substantia nigra/VTA. *Neuron*, *51*(3), 369–379. doi:10.1016/j.neuron.2006.06.021

Canolty, R. T., Edwards, E., Dalal, S. S., Soltani, M., Nagarajan, S. S., Kirsch, H. E. … Knight, R. T. (2006). High gamma power is phase-locked to theta oscillations in human neocortex. *Science*, *313*(5793), 1626–1628. doi:10.1126/science.1128115

Cao, F., & Ray, S. (2012). Bayesian hierarchical reinforcement learning. Advances in Neural Information Processing Systems.

Cisek, P. (2007). Cortical mechanisms of action selection: The affordance competition hypothesis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *362*(1485), 1585–1599.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *The Behavioral and Brain Sciences*, *36*(3), 181–204. doi:10.1017/S0140525X12000477

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 933–942. doi:10.1098/rstb.2007.2098

D'Ardenne, K., McClure, S. M., Nystrom, L. E., & Cohen, J. D. (2008). BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. *Science*, *319*(5867), 1264–1267. doi:10.1126/science.1150605

Daunizeau, J., Den Ouden, H. E., Pessiglione, M., Kiebel, S. J., Stephan, K. E., & Friston, K. J. (2010). Observing the observer (I): Meta-bayesian models of learning and decision-making. *PLoS One*, *5*(12), e15554. doi:10.1371/journal.pone.0015554

Daw, N. D., & Doya, K. (2006). The computational neurobiology of learning and reward. *Current Opinion in Neurobiology*, *16*(2), 199–204. doi:10.1016/j.conb.2006.03.006

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8* (12), 1704–1711. doi:10.1038/nn1560

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879. doi:10.1038/nature04766

Dayan, P. (2009). Dopamine, reinforcement learning, and addiction. *Pharmacopsychiatry*, *42*(1), S56–65. doi:10.1055/s-0028-1124107

Dayan, P., & Hinton, G. E. (1997). Using expectation maximization for reinforcement learning. *Neural*

Computation, *9*, 271–278. doi:10.1162/neco.1997.9.2.271

Dayan, P., Hinton, G. E., & Neal, R. (1995). The Helmholtz machine. *Neural Computation*, *7*, 889–904. doi:10.1162/neco.1995.7.5.889

De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2012). Confidence in value-based choice. *Nature Neuroscience*, *16*, 105–110. doi:10.1038/nn.3279

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*(2), 312–325. doi:10.1016/j.neuron.2013.09.007

Ferro, M., Ognibene, D., Pezzulo, G., & Pirrelli, V. (2010). Reading as active sensing: A computational model of gaze planning during word recognition. *Frontiers in Neurorobotics*, *4*, 1.

Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, *299*(5614), 1898–1902. doi:10.1126/science.1077349

FitzGerald, T., Dolan, R., & Friston, K. (2014). Model averaging, optimal inference, and habit formation. *Frontiers in Human Neuroscience*. doi:10.3389/fnhum.2014.00457

Fox, C., & Roberts, S. (2011). A tutorial on variational Bayes. *Artificial Intelligence Review*, Springer. doi:10.1007/s10462-10011-19236-10468

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews. Neuroscience*, *11*(2), 127–138. doi:10.1038/nrn2787

Friston, K., Adams, R. A., Perrinet, L., & Breakspear, M. (2012). Perceptions as hypotheses: Saccades as experiments. *Frontiers in Psychology*, *3*, 151. doi:10.3389/fpsyg.2012.00151

Friston, K., & Mathys, C. (2015). I think therefore I am. Cognitive Dynamic Systems. S. Haykin, IEEE press: in press.

Friston, K., Samothrakis, S., & Montague, R. (2012). Active inference and agency: Optimal control without cost functions. *Biological Cybernetics*, *106*, 523–541. [Epub ahead of print]. doi:10.1007/s00422-012-0512-8

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., & Dolan, R. J. (2014). The anatomy of choice: Dopamine and decision-making. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1655). doi:10.1098/rstb.2013.0481

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., Raymond, R. J., & Dolan, J. (2013). The anatomy of choice: Active inference and agency. *Frontiers in Human Neuroscience*, *7*, 598. doi:10.3389/fnhum.2013.00598

Furmston, T., & Barber, D. (2010). Variational methods for reinforcement learning. In *International Conference on Artificial Intelligence and Statistics* (AISTATS). JMLR: W&CP, vol. 9, (pp. 241–248).

Fuster, J. M. (2004). Upper processing stages of the perception-action cycle. *Trends in Cognitive Sciences*, *8*(4), 143–145. doi:10.1016/j.tics.2004.02.004

Gurney, K., Prescott, T. J., & Redgrave, P. (2001). A computational model of action selection in the basal ganglia. I. A new functional anatomy. *Biological Cybernetics*, *84*(6), 401–410. doi:10.1007/PL00007984

Harlow, H. (1950). Learning and satiation of response to intrinisically motivated complex puzzle performance by monkeys." *Journal of Comparative Physiological Psychology*, *40*, 289–294. doi:10.1037/h0058114

Hauskrecht, M. (2000). Value-function approximations for partially observable Markov decision processes. *Journal of Artificial Intelligence Research*, *13*, 33–94.

Helmholtz, H. (1866/1962). *Concerning the perceptions in general. Treatise on physiological optics*. New York, NY: Dover. **III**.

Hollerman, J. R., & Schultz, W. (1996). Activity of dopamine neurons during learning in a familiar task context. *Social Neuroscience Abstracts*, *22*, 1388.

Howard, R. (1966). Information value theory. *IEEE Transactions on Systems, Science and Cybernetics, SSC-2* (1),22–26. doi:10.1109/TSSC.1966.300074

Humphries, M. D., Khamassi, M., & Gurney, K. (2012). Dopaminergic control of the exploration-exploitation trade-off via the basal ganglia. *Frontiers in Neuroscience*, *6*, 9. doi:10.3389/fnins.2012.00009

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, *49*(10), 1295–1306. doi:10.1016/j.visres.2008.09.007

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews. Neuroscience*, *2*(3), 194–203. doi:10.1038/35058500

Jezek, K., Henriksen, E., Treves, A., Moser, E., & Moser, M.-B. (2011). Theta-paced flickering between place-cell maps in the hippocampus. *Nature*, *478*, 246–249. doi:10.1038/nature10439

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, *101*(1–2), 99–134. doi:10.1016/S0004-3702(98)00023-X

Kakade, S., & Dayan, P. (2002). Dopamine: Generalization and bonuses. *Neural Networks*, *15*(4–6), 549–559. doi:10.1016/S0893-6080(02)00048-5

Kamar., E., & Horvitz, E. (2013). Light at the end of the tunnel: A Monte Carlo approach to computing value of information. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*.

Kappen, H. J., Gomez, Y., & Opper, M. (2012). Optimal control as a graphical model inference problem. *Machine Learning*, *87*(2), 159–182. doi:10.1007/s10994-012-5278-7

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

Kirsh, D., & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive Science*, *18*(4), 513–549. doi:10.1207/s15516709cog1804_1

Klyubin, A. S., Polani, D., & Nehaniv, C. L. (2008). Keep your options open: An information-based driving principle for sensorimotor systems. *PLoS One*, *3*(12), e4018. doi:10.1371/journal.pone.0004018

Krause., A., & Guestrin, C. (2005). Optimal nonmyopic value of information in graphical models: Efficient algorithms and theoretical limits. In *International Joint Conference on Artificial Intelligence (IJCAI)*. Carnegie Mellon University.

Krebs, R. M., Schott, B. H., Schütze, H., & Düzel, E. (2009). The novelty exploration bonus and its attentional modulation. *Neuropsychologia*, *47*, 2272–2281. doi:10.1016/j.neuropsychologia.2009.01.015

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A*, *20*, 1434–1448. doi:10.1364/JOSAA.20.001434

Lepora, N., Martinez-Hernandez, U., & Prescott, T. (2013). Active touch for robust perception under position uncertainty. In *IEEE proceedings of ICRA*.

Linsker, R. (1990). Perceptual neural organization: Some approaches based on network models and information theory. *Annual Review of Neuroscience*, *13*, 257–281. doi:10.1146/annurev.ne.13.030190.001353

Little, D. Y., & Sommer, F. T. (2013). Learning and exploration in action-perception loops. *Front Neural Circuits*, *7*, 37. doi:10.3389/fncir.2013.00037

Littman, M., Sutton, R., & Singh, S. (2002). Predictive Representations of State. Advances in Neural Information Processing Systems 14 (NIPS).

Lungarella, M., & Sporns, O. (2006). Mapping information flow in sensorimotor networks. *PLoS Computational Biology*, *2*, e144. doi:10.1371/journal.pcbi.0020144

McClure, S. M., Daw, N. D., & Montague, P. R. (2003). A computational substrate for incentive salience. *Trends in Neurosciences*, *26*(8), 423–428. doi:10.1016/S0166-2236(03)00177-2

Miller, G., Galanter, E., & Pribram, K. (1960). *Plans and the structure of behavio*. New York, NY: Henry Holt.

Moser, E. I., Kropff, E., & Moser, M.-B. (2008). Place cells, grid cells, and the brain's spatial representation system. *Annual Review of Neuroscience*, *31*, 69–89. doi:10.1146/annurev.neuro.31.061307.090723

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. doi:10.1016/j.neuron.2006.03.045

Najemnik, J., & Geisler, W. (2005). Optimal eye movement strategies in visual search. *Nature*, *434*, 387–391. doi:10.1038/nature03390

Nepora, N., & Gurney, K. (2012). The basal ganglia optimize decision making over general perceptual hypotheses. *Neural Computation*, *24*(11), 2924–2945. doi:10.1162/NECO_a_00360

Niv, Y. (2007). Cost, benefit, tonic, phasic: What do response rates tell us about dopamine and motivation? *Annals of the New York Academy of Sciences*, *1104*, 357–376.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, *304*(5669), 452–454. doi:10.1126/science.1094285

O'Regan, J., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *The Behavioral and Brain Sciences*, *24*, 939–973. doi:10.1017/S0140525X01000115

Ognibene, D., & Baldassarre, G. (2014). Ecological active vision: Four Bio-inspired principles to integrate bottom-up and adaptive top-down attention tested with a simple camera-arm robot. *Autonomous Mental Development, IEEE Transactions on*. IEEE.

Ognibene, D., Chinellato, E., Sarabia, M., & Demiris, Y. (2013). Contextual action recognition and target localization with an active allocation of attention on a humanoid robot. *Bioinspiration & Biomimetics*, *8*, 3. doi:10.1088/1748-3182/8/3/035002

Ognibene, D., & Demiris, Y. (2013). Toward active event perception. In I*nternational Joint conference of Artificial Intelligence*. IJCIA: Beijing.

Ognibene, D.,Pezzulo, G., & Baldassarre, G. (2010). Learning to look in different environments: An active-vision model which learns and readapts visual routines. In S. Doncieux (ed.), *Proceedings of the 11th International Conference on Simulation of Adaptive Behaviour*, SAB 2010, Paris - Clos Lucé, France, August 25–28, 2010. Proceedings. Springer Berlin Heidelberg.

Ognibene, D., Catenacci Volpi, N., Pezzulo, G., & Baldassarre, G. (2013). Learning epistemic actions in model-free memory-free reinforcement learning: Experiments with a neuro-robotic model. In *Second International Conference on Biomimetic and Biohybrid Systems*. Proceedings. pp. 191–203. doi:10.1007/978-3-642-39802-5_17

Oja, E. (1989). Neural networks, principal components, and subspaces. *International Journal of Neural Systems*, *1*, 61–68. doi:10.1142/S0129065789000475

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. doi:10.1038/381607a0

Optican, L., & Richmond, B. J. (1987). Temporal encoding of two-dimensional patterns by single units in primate inferior cortex. II Information Theoretic Analysis. *Journal of Neurophysiology*, *57*, 132–146.

Ortega, P. A., & Braun, D. A. (2011). Information, utility and bounded rationality. *Lecture Notes on Artificial Intelligence*, *6830*, 269–274.

Ortega, P. A., & Braun, D. A. (2013). Thermodynamics as a theory of decision-making with information-processing costs. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *469*, 2153. doi:10.1098/rspa.2012.0683

Oudeyer, P.-Y., & Kaplan, F. (2007). What is intrinsic motivation? A typology of computational approaches. *Frontiers in Neurorobotics*, *1*, 6. doi:10.3389/neuro.12.006.2007

Penny, W., Zeidman, P., & Burgess, N. (2013). Forward and backward inference in spatial cognition. *Plos Computational Biology*, *9*(12), e1003383. doi:10.1371/journal.pcbi.1003383

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., & Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, *442*(7106), 1042–1045. doi:10.1038/nature05051

Pezzementi, Z., Plaku, E., Reyda, C., & Hager, G. (2011). Tactile-object recognition from appearance information. *IEEE Transactions on Robotics*, *27*, 473–487. doi:10.1109/TRO.2011.2125350

Pezzulo, G., & Castelfranchi, C. (2009). Thinking as the control of imagination: A conceptual framework for goal-directed systems. *Psychological Research Psychologische Forschung*, *73*, 559–577. doi:10.1007/s00426-009-0237-z

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

Frontiers in Psychology, *4*, 92. doi:10.3389/fpsyg.2013.00092
Pezzulo, G., Van Der Meer, M., Lansink, C., & Pennartz, C. (2014). Internally generated sequences in learning and executing goal-directed behavior. *Trends in Cognitive Sciences*, *18*(2), pp. 647–657.
Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2* (1), 79–87. doi:10.1038/4580
Redgrave, P., & Gurney, K. (2006). The short-latency dopamine signal: A role in discovering novel actions? *Nature Reviews. Neuroscience*, *7*(12), 967–975. doi:10.1038/nrn2022
Roy, N., Burgard, W., Fox, D., & Thrun, S. (1999). Coastal navigation: Robot navigation under uncertainty in dynamic environments. In *Proceedings of the IEEE International Conference on Robotics andAutomation (ICRA)*. Washington, DC: IEEE Computer Society.
Ryan, R., & Deci, E. (1985). *Intrinsic motivation and self-determination in human behavior*. New York, NY: Plenum.
Schembri, M., Mirolli, M., & Baldassare, G. (2007). Evolving internal reinforcers for an intrinsically motivated reinforcement learning robot. In Y. Demiris, B. Scassellati, & D. Mareschal (Eds.) *Proceedings of the 6th IEEE International Conference on Development and Learning (ICDL2007)*. IEEE.
Schmidhuber, J. (1991). Curious model-building control systems. In *Proceeding of International Joint Conference on Neural Networks*. Singapore: IEEE. **2**: 1458–1463.
Schneider, A., Sturm, J., Stachniss, C., Reisert, M., Burkhardt, H., & Burgard, W. (2009). Object identification with tactile sensors using bag-of-features. *IROS 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2009*. p. 243. IEEE.
Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, *80*(1), 1–27.
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.
Schwartenbeck, P., FitzGerald, T., Mathys, C., Dolan, R., & Friston, K. (2014). The dopaminergic midbrain encodes the expected certainty about desired outcomes. *Cerebral Cortex*. pii: bhu159. doi:10.1093/cercor/bhu159
Shackle, G. (1972). *Epistemics and economics*. Cambridge, MA: Cambridge University Press.
Singh, A., Krause, A., Guestrin, C., & Kaiser, W. (2009). Efficient informative sensing using multiple robots. *Journal of Artificial Intelligence Research*, *34*(2), 707.
Solway, A., & Botvinick, M. (2012). Goal-directed decision making as probabilistic inference: A com- putational framework and potential neural correlates. *Psychological Review*, *119*, 120–154. doi:10.1037/a0026435
Sornkarn, N., Nanayakkara, T., & Howard, M. (2014). Internal impedance control helps information gain in embodied perception. In *IEEE International Conference on Robotics and Automation (ICRA)*, IEEE.
Still, S. (2009). Information-theoretic approach to interactive learning. *EPL (Europhysics Letters)*, *85*(2), 28005. doi:10.1209/0295-5075/85/28005
Still, S., & Precup, D. (2012). An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, *131*(3), 139–148. doi:10.1007/s12064-011-0142-z
Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.
Tani, J., & Nolfi, S. (1999). Learning to perceive the world as articulated: An approach for hierarchical learning in sensory–motor systems. *Neural Networks*, *12*, 1131–1141. doi:10.1016/S0893-6080(99)00060-X
Tishby, N., & Polani, D. (2010). Information theory of decisions and actions. In V. Cutsuridis, A. Hussain, & J. Taylor (Eds.), *Perception-reason-action cycle: Models, algorithms and systems*. Springer: Berlin.
Toussaint, M., & Storkey, A. (2006). Probabilistic inference for solving discrete and continuous state Markov Decision Processes. In *Proceeding of the 23nd International Conference on Machine Learning*. ACM. pp. 945–952.
Van den Broek, J. L., Wiegerinck, W. A. J. J., & Kappen, H. J. (2010). Risk-sensitive path integral control. *UAI*, *6*, 1–8.
Verschure, P. F., Voegtlin, T., & Douglas, R. J. (2003). Environmentally mediated synergy between perception and behavior in mobile robots. *Nature*, *425*, 620–624. doi:10.1038/nature02024
Vijayakumar, S., Toussaint, M., Petkos, G., & Howard, M. (2009). Planning and moving in dynamic environments. In B. Sendhoff (Ed.), *Creating brain-like intelligence* (pp. 151–191). Berlin: Springer-Verlag.
Viswanathan, G., Buldyrev, S., Havlin, S., Da Luz, M., Raposo, E., & Stanley, H. (1999). Optimizing the success of random searches. *Nature*, *401*(6756), 911–914. doi:10.1038/44831
Walther, D., Rutishauser, U., Koch, C., & Perona, P. (2005). Selective visual attention enables learning and recognition of multiple objects in cluttered scenes. *Computer Vision and Image Understanding*, *100*, 41–63. doi:10.1016/j.cviu.2004.09.004
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. doi:10.1016/j.neuron.2008.04.027

APPENDIX A

Lemma (predictive free energy): *Under a generative model $P(s_\tau, o_\tau|\pi) = Q(s_\tau|o_\tau, \pi)P(o_\tau|m)$ and policy π, the negative free energy of the approximate posterior predictive density is $\forall : \tau > t$*

$$\mathbf{Q}_\tau(\pi) = -\underbrace{E_{Q(s_\tau|\pi)}[H[P(o_\tau|s_\tau)]]}_{\text{Predicted uncertainty}} - \underbrace{D[Q(o_\tau|\pi)||P(o_\tau|m)]}_{\text{Predicted divergence}} \tag{A1.1}$$

Proof: The expected free energy of the approximate posterior predictive distribution over hidden states (under policy π at $\tau > t$ in the future) is the expected energy minus its entropy (where the energy of a hidden state $G(s_\tau, \pi)$ is itself an expectation over outcomes):

$$\begin{aligned} G(s_\tau, \pi) &= -E_{P(o_\tau|s_\tau)}[\ln P(o_\tau, s_\tau|\pi)] \\ F_\tau(\pi) &= E_{Q(s_\tau|\pi)}[G(s_\tau, \pi)] - H[Q(s_\tau|\pi)] \end{aligned} \tag{A1.2}$$

This means the quality or value of the policy is:

$$\begin{aligned} \mathbf{Q}_\tau(\pi) &= -F_\tau(\pi) \\ &= E_{Q(o_\tau, s_\tau|\pi)}[\ln P(o_\tau, s_\tau|\pi) - \ln Q(s_\tau|\pi)] \\ &= E_{Q(o_\tau, s_\tau|\pi)}[\ln Q(s_\tau|o_\tau, \pi) + \ln P(o_\tau|m) - \ln Q(s_\tau|\pi)] \\ &= E_{Q(o_\tau, s_\tau|\pi)}[\ln Q(o_\tau|s_\tau, \pi) + \ln P(o_\tau|m) - \ln Q(o_\tau|\pi)] \\ &= -\underbrace{E_{Q(s_\tau|\pi)}[H[P(o_\tau|s_\tau)]]}_{\text{Predicted uncertainty}} - \underbrace{D[Q(o_\tau|\pi)||P(o_\tau|m)]}_{\text{Predicted divergence}} \end{aligned}$$

$$\begin{aligned} Q(s_\tau|\pi) &= E_{Q(s_t)}[P(s_\tau|s_t, \pi)] \\ Q(o_\tau|\pi) &= E_{Q(s_\tau|\pi)}[P(o_\tau|s_\tau)] \\ Q(o_\tau, s_\tau|\pi) &= P(o_\tau|s_\tau)Q(s_\tau|\pi) \end{aligned} \tag{A1.3}$$

Where $Q(o_\tau|s_\tau, \pi) = P(o_\tau|s_\tau)$ is the (predictive) likelihood of the (predictive) generative model

Remarks: Intuitively, the generative model of future states encodes beliefs that certain outcomes in the future are surprising (irrespective of the current state or policy), while future hidden states (given those outcomes) are surprising when they are not predicted. This generative model is defined in terms of a posterior (predictive) distribution over hidden states and a prior over outcomes. This contrasts with the usual construction of a generative model of past outcomes, in terms of a likelihood and prior over hidden states. Heuristically, this reflects the fact that current outcomes are caused by past transitions among hidden states but future outcomes can cause current state transitions (through policy selection).

Note that when $\tau = t$, the outcome is observed and the expected free energy reduces to the free energy of approximate posterior beliefs about hidden states:

$$\begin{aligned} G(s_t) &= -\ln P(o_t, s_t) \\ F_t &= E_{Q(s_t)}[G(s_t)] - H[Q(s_\tau)] \end{aligned} \tag{A1.4}$$

Optimizing this free energy corresponds to Bayes optimal state estimation; however, because this free energy functional has no concept of the future it cannot support purposeful behavior or active inference.

APPENDIX B

The variational updates are a self-consistent set of equalities that minimize variational free energy. Let $\tilde{x} = \tilde{s}_t, \tilde{u}, \gamma$ denote the hidden variables and $\hat{x} = \hat{s}_t, \hat{\pi}, \hat{\gamma}$ denote their sufficient statistics. Using the dot notation $A \cdot B = A^T B$, the variational free energy can be expressed in terms of its energy and entropy (with $\mathbf{B}(a_0)\hat{s}_0 = \mathbf{D}$):

$$\begin{aligned} F(\tilde{o}, \hat{x}) &= -E_Q[\ln P(\tilde{o}, \tilde{x}|m)] - H[Q(\tilde{x}|\hat{x})] \\ &= \hat{s}_t \cdot (\ln \hat{s}_t - \ln \mathbf{A} \cdot o_t - \ln(\mathbf{B}(a_{t-1})\hat{s}_{t-1})) \\ &\quad + \hat{\pi} \cdot (\ln \hat{\pi} - \hat{\gamma}\mathbf{Q}) + \beta\hat{\gamma} \\ &\quad + \alpha(\ln \alpha - \ln \hat{\gamma} - \ln \beta - 1) \end{aligned}$$

$$\begin{aligned} E_Q[\ln P(\tilde{o}, \tilde{x}|m)] &= E_Q[\ln P(o_t|s_t) + \ln P(s_t|s_{t-1}, a_{t-1})\hat{s}_{t-1} \\ &\quad + \ln P(\tilde{u}|\gamma) + \ln P(\gamma|\beta)] \\ &= \hat{s}_t \cdot (\ln \mathbf{A} \cdot o_t + \ln(\mathbf{B}(a_{t-1})\hat{s}_{t-1})) \\ &\quad + \hat{\gamma}\mathbf{Q} \cdot \hat{\pi} + (\alpha - 1)(\psi(\alpha) - \ln \hat{\beta}) - \beta\hat{\gamma} \\ &\quad + \alpha \ln \beta - \ln \Gamma(\alpha) \end{aligned}$$

$$\begin{aligned} H[Q(\tilde{x}|\hat{x})] &= \alpha - \ln \hat{\beta} + \ln \Gamma(\alpha) \\ &\quad + (1 - \alpha)\psi(\alpha) - \hat{\pi} \cdot \ln \hat{\pi} - \hat{s}_t \cdot \ln \hat{s}_t \end{aligned} \tag{A2.1}$$

Differentiating the variational free energy with respect to the sufficient statistics gives

$$\begin{aligned} \frac{\partial F}{\partial \hat{s}_t} &= \mathbf{1} + \ln \hat{s}_t - \ln \mathbf{A} \cdot o_t - \ln(\mathbf{B}(a_{t-1})\hat{s}_{t-1}) - \hat{\gamma} \cdot \nabla_{\hat{s}} \mathbf{Q} \cdot \hat{\pi} \\ \frac{\partial F}{\partial \hat{\pi}} &= \mathbf{1} + \ln \hat{\pi} - \hat{\gamma} \cdot \mathbf{Q} \\ \frac{\partial F}{\partial \hat{\gamma}} &= \beta - \mathbf{Q} \cdot \hat{\pi} - \hat{\beta} \end{aligned} \tag{A2.2}$$

Finally, we obtain the variational updates by solving for zero and rearranging to give:

$$\begin{aligned} \ln \widehat{s}_t &= \ln \mathbf{A} \cdot \widehat{o}_t + \ln(\mathbf{B}(a_{t-1})\widehat{s}_{t-1}) - \mathbf{1} \\ \ln \widehat{\pi} &= \widehat{\gamma} \cdot \mathbf{Q} - 1 \\ \widehat{\beta} &= \beta - \mathbf{Q} \cdot \widehat{\pi} \end{aligned} \qquad \text{A2.3}$$

For simplicity, we have ignored the derivative of value with respect to the hidden states (numerically, this simplification appears to make little difference in the Markov decision processes considered in this and previous papers). Including this term leads to an additional term in the (Bayesian filter) updates of expected states corresponds to an optimism bias (Friston et al., 2014).

The variational updates for precision can be multiplied by $(1 - \lambda)$ and rearranged to give:

$$\widehat{\beta} = \lambda\widehat{\beta} + (1 - \lambda)(\beta - \mathbf{Q} \cdot \widehat{\pi}) \qquad \text{A2.4}$$

This effectively slows the updates to provide a more time-resolved model of the implicit (e.g., dopamine) dynamics. In this paper, we used $\lambda = \frac{1}{4}$.

Commentaries

What is epistemic value in free energy models of learning and acting? A bounded rationality perspective

Pedro A. Ortega[1]
and Daniel A. Braun[2,3]
[1]School of Engineering and Applied Sciences, University of Pennsylvania, Philadelphia, USA
[2]Max Planck Institute for Biological Cybernetics, Tübingen, Germany
[3]Max Planck Institute for Intelligent Systems, Tübingen, Germany
E-mail: daniel.braun@tuebingen.mpg.de

http://dx.doi.org/10.1080/17588928.2015.1051525

Abstract: Free energy models of learning and acting do not only care about utility or extrinsic value, but also about intrinsic value, that is, the information value stemming from probability distributions that represent beliefs or strategies. While these intrinsic values can be interpreted as epistemic values or exploration bonuses under certain conditions, the framework of bounded rationality offers a complementary interpretation in terms of information-processing costs that we discuss here.

In the information-theoretic model of bounded rationality (Braun, Ortega, Theodorou, & Schaal, 2011; Ortega & Braun, 2011, 2013), a bounded rational decision-maker has a prior strategy $P_0(a)$ and a probabilistic model $P_0(s|a)$ about the states $s \in \mathrm{S}$ that might result from taking action $a \in \mathrm{A}$. The decision-maker plans to optimize a real-valued utility function $U : \mathrm{S} \times \mathrm{A} \rightarrow \mathrm{R}$ that can be evaluated for any (s, a) pair. As the decision-maker is bounded, the posterior strategy $P(a)$ after deliberation can only deviate from the prior strategy $P_0(a)$ by a certain number of bits of information, that is $D_{KL}(P(a)||P_0(a)) \leq K$. Similarly, the decision-maker might have model uncertainty (Hansen & Sargent, 2008) and thus consider any model $P(s|a)$ that is within a given deviation from the prior model $P_0(s|a)$, that is, $D_{KL}(P(s|a)||P_0(s|a)) \leq C$. Mathematically, the bounded rational decision-maker behaves as if solving the following variational problem

$$\max_{P(a)} \operatorname*{ext}_{P(s|a)} \sum_a P(a) \sum_s P(s|a) \left(U(s,a) - \frac{1}{\alpha} \log \frac{P(a)}{P_0(a)} - \frac{1}{\beta} \log \frac{P(s|a)}{P_0(s|a)} \right) \quad (1)$$

with the solution $P^*(s|a) = P_0(s|a)\exp\{\beta U(s,a)\}/Z_\beta(a)$ and $P^*(a) = P_0(a)\exp\{\frac{\alpha}{\beta}\log Z_\beta(a)\}/Z_\alpha$, where $\alpha \in R_+$ and $\beta \in R$ are the boundedness parameters of the constrained decision problem and $Z_\alpha, Z_\beta(a)$ are normalizing constants. For $\beta < 0$ we have the extremum operator ext = min and for $\beta > 0$ we have ext = max. The perfectly rational expected utility maximizer is obtained in the limit $\alpha \rightarrow \infty$ (perfect choice of action) and $\beta \rightarrow 0$ (perfect trust in prior beliefs).

We now attempt to derive Equation (5) in the Discussion Paper by Friston et al. (this issue) from (1) to gain further insight into their assumptions. The expression (1) can be rewritten as

$$\max_{P(a)} \operatorname*{ext}_{P(s|a)} \sum_a P(a) \left(-\frac{1}{\beta} D_{KL}[P(s|a)||P_{des}(s|a)] - \frac{1}{\alpha} \log \frac{P(a)}{P_0(a)} - \frac{1}{\beta} \log Z_\beta(a) \right), \quad (2)$$

where we have defined the desired target probability $P_{des}(s|a) = P_0(s|a)\exp\{\beta U(s,a)\}/Z_\beta(a)$ in which $Z_\beta(a)$ is a normalizing constant that depends on the action. Here, the utility is expressed as an informational difference between the actual and the desired distribution over observations. Accordingly, the utility-maximizing decision-maker of Equation (1) can be equivalently thought to minimize surprise as described by Equation (2). To obtain Equation (5) of the Discussion Paper, we must make two additional assumptions: (1) we neglect the cost of action selection ($\alpha \rightarrow \infty$); and (2) we assume that $P_{des}(s|a) = P_{des}(s)$ does not depend on the action a, which is the case, for example, in active inference models where

$P_{des}(s)$ is thought to represent the decision-maker's preferences over outcomes. Then, the variational problem in Equation (2) becomes equivalent to Equation (5) of the Discussion paper, that is,

$$-D_{KL}[P(s|a)||P_{des}(s)] = \underbrace{\mathrm{E}_{P(s|a)}[\log P_{des}(s)]}_{extrinsic-value} + \underbrace{H[P(s|a)]}_{intrinsic-value}, \quad (3)$$

which is to be maximized with respect to $P(s|a)$. From the point of view of bounded rationality, maximizing the intrinsic and extrinsic values corresponds to choosing an action that maximizes the expected utility, but subject to minimizing the information costs of deviating from a prior strategy and a prior belief model.

Since both actions and observations follow essentially the same variational principle, the distinction between extrinsic and epistemic value does not hinge on the presence of hidden or observable states, but appears already in the simplest scenario with a single action variable a with

$$\max_{P(a)} \sum_a P(a)\left(U(a) - \frac{1}{\alpha}\log\frac{P(a)}{P_0(a)}\right) = \max_{P(a)}\left(\sum_a P(a)\tilde{U}(a) + \frac{1}{\alpha}H[P(a)]\right) \quad (4)$$

where we have defined a modified utility $\tilde{U}(a) := U(a) + \frac{1}{\alpha}\log P_0(a)$. Written in this form, the free energy value of the policy $P(a)$ consists of an expected utility and an entropy. While this could be interpreted as searching for a policy that "leaves options open" or encourages exploration, the bounded rational interpretation is that the decision-maker has limited resources and cannot deviate too much from the prior. If the prior $P_0(a)$ is not explicitly considered in $\tilde{U}$, from the point of view of bounded rationality it is still implied as a uniform prior. The information processing costs given by the informational deviation from the (implicit) prior can also be interpreted in computational terms as a sampling complexity (Daniel, 2014; Ortega, Braun, & Tishby, 2014). In summary, the bounded rationality approach offers a different perspective on epistemic value in terms of intrinsic information processing costs.

REFERENCES

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* (pp. 202–209). doi:10.1109/ADPRL.2011.5967366

Daniel, A. B., & Pedro A. O. (2014). Information-theoretic bounded rationality and ε-optimality. *Entropy*, *16*(8), 4662–4676. doi:10.3390/e16084662

Hansen, L. P., & Sargent, T. J. (2008). *Robustness*. Princeton, NJ: Princeton University Press.

Ortega, P. A., & Braun, D. A. (2011). Information, utility and bounded rationality. *Lecture Notes on Artificial Intelligence*, *6830*, 269–274.

Ortega, P. A., Braun, D. A., & Tishby, N. (2014). Monte Carlo methods for exact & efficient solution of the generalized optimality equations. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on* (pp. 4322–4327). doi:10.1109/ICRA.2014.6907488

Ortega, P. A., & Braun, D. A. (2013). Thermodynamics as a theory of decision-making with information-processing costs. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science*, *469*(2153). doi:10.1098/rspa.2012.0683

* * *

Free energy minimization and information gain: The devil is in the details

Johan Kwisthout and Iris van Rooij
Radboud University Nijmegen, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands
E-mail: j.kwisthout@donders.ru.nl

http://dx.doi.org/10.1080/17588928.2015.1051014

Abstract: Contrary to Friston's previous work, this paper describes free energy minimization using categorical probability distributions over discrete states. This alternative mathematical framework exposes a fundamental, yet unnoticed challenge for the free energy principle. When considering discrete state spaces one must specify their granularity, as the amount of information gain is defined over this state space. The more detailed this state space, the lower the precision of the predictions will be, and consequently, the higher the prediction errors. Hence, an optimal trade-off between precision and detail is needed, and we call for incorporating this aspect in the free energy principle.

Keywords: free energy; predictive coding; prediction error minimization; information gain.

"If you take care of the small things,
the big things take care of themselves.

You can gain more control over your life by paying closer attention to the little things."
Emily Dickinson, 1830–1886

There is much value in Dickinson's advice. In this commentary, we are particularly interested in the epistemic value of *detailed predictions* ("paying closer attention to the little things") for *free energy minimization* ("gaining more control over your life"). We will show that specifying the granularity of state spaces is crucial for minimizing free energy: When the granularity is too low, little information is gained from correct predictions; if it is too high, prediction errors will be needlessly high.

In the target article, Friston and colleagues bring the exploration-exploitation trade-off under the free energy minimization regime, by assuming that the agent's prior beliefs are such that they expect to minimize future free energy and plan their actions accordingly. Formally, they describe their theory using partially observable Markov decision processes (POMDPs) with discrete states and actions; consequently, the generative models are described using categorical probability distributions. This approach overlooks the fact that "state" and "action" in these models depend on the granularity (or *level of detail*) of the state space and the actions operating on them. Given that the required granularity cannot be assumed to be fixed, as it may be context dependent, any discrete free energy account will also need to address the question of how the right level of detail is determined.

For example, one may plan to shop for groceries. The action "shop for groceries" is fairly abstract and may be described more in detail as "first pick up some croissants at the bakery, then head for the produce market to get vegetables, and don't forget to buy cat food". Note that the more detailed we make these predictions, the more information they carry; however, they are also more prone to prediction errors. When one expects to buy *this-and-that* flavour of cat food from brand *such-and-so*, then any other flavour or brand would result in a prediction error. If, on the other hand, we expect merely "to buy cat food", then as long as we end up buying some brand or flavour of cat food, regardless which one, there would be no prediction error. Hence, increasing the level of detail of predicted and actual outcomes will —everything else being equal—increase average uncertainty, simply because it will increase the entropy of the probability distribution over possible outcomes.

A now classic objection of the free energy principle is that it seems to predict that organisms would seek shelter in a dark cave to defer from any sensory experiences and hence minimize prediction errors (Thornton, 2010). Even though a satisfactory answer may have been given to this objection (Friston, Thornton & Clark, 2012), we raise a novel problem that seems to generalize that idea and follows naturally from the consideration of the level of detail: Consider that one stands in the middle of Times Square during rush hour, one can minimize prediction errors by simply predicting that "stuff happens around me" and interpret the sensory inputs accordingly. This very low level of detail of expected and actual outcomes will, by definition, lead to low free energy (prediction errors), simply because there are fewer categories in the probability distribution.

The example illustrates that predicting and interpreting all our sensory experiences as "stuff happens" is equally ineffective as staying in a dark cave forever. Arguably, an individual that makes more fine-grained predictions, e.g., by discriminating between cars that are parked and cars that are driving, will be more successful in the long run. Making more fine-grained predictions than "stuff happens" induces potential uncertainty and excessive prediction errors, but it allows us to benefit by making more informative predictions. In contrast, to assess whether we should wait or whether we can walk the street, it is seldom beneficial to make predictions that are too detailed. It is of little use to predict the car type and brand in order to prevent getting run over. The added information from such a prediction is outweighed by the increased prediction error when the prediction turns out to be wrong. These considerations show that somehow a trade-off needs to be made between precision and detail.

As Friston et al. acknowledge, hyperpriors on the (expected) precision are crucial for weighting prediction errors (Clark, 2013; Friston, 2010). As we highlighted with our examples, it is necessary to extend the notion of hyperpriors to govern also level of detail, as precision is a property of predictions at every level of detail. Such an enhanced theory may shed light on why and how we are able to make predictions that trade off information gain and prediction error, and how this fits in with free energy minimization.

REFERENCES

Clark, A. (2013). The many faces of precision (Replies to commentaries on "Whatever next? Neural prediction, situated agents, and the future of cognitive science"). *Frontiers in Theoretical and Philosophical Psychology*, *4*, 270.

Friston, K. L. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*, 127–138. doi:10.1038/nrn2787

Friston, K. L., Thornton, C., & Clark, A. (2012). Free-energy minimization and the dark-room problem. *Frontiers in Psychology*, *3*, 130. doi:10.3389/fpsyg.2012.00130

Thornton, C. (2010). Some puzzles relating to the free-energy principle: Comment on Friston. *Trends in Cognitive Sciences*, *14*, 53–54. doi:10.1016/j.tics.2009.11.010

* * *

About the goal of a goals' goal theory

Ivan Herreros[1]
and Paul Fmj Verschure[1,2]
[1]Synthetic Perceptive Emotive and Cognitive Systems Lab, Universitat Pompeu Fabra, Barcelona, Spain
[2]Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain
E-mail: paul.verschure@upf.edu

http://dx.doi.org/10.1080/17588928.2015.1051952

The Free-Energy Principle (FEP) provides a powerful normative framework to explain perception and behavior. The all-encompassing ambition of the endeavor becomes clear by the diversity of topics covered from the "predictive coding" interpretation of perception to the more recent focus on behavior and "active inference." However, for the broader audience of the uninitiated, the goal of FEP, its theoretical framework, and leverage may still be elusive. We will try to identify some of the challenges FEP is facing from the perspective of the partially initiated, focusing on the somewhat equivocal role that "goals" (as priors) have in the theory.

The main contribution in Friston et al. (this issue) stems from reconciling two apparently opposed views on the dopaminergic response as reflecting either reward prediction (e.g., Schultz, 1998) or surprise (Redgrave & Gurney, 2006). This putative reconciliation follows from "casting reward or utility in terms of log-probabilities" which allows for measuring them with a common "information currency." This step allows an elegant blending of goal-directed and uncertainty-reducting behavior without the use of specific parameters to control the exploitation/exploration trade-off. However, as the paper acknowledges, this has already been done in risk-sensitive control (Van den Broek, Wiegerinck, & Kappen, 2010; Ortega & Braun, 2011), thus, this paper's main contribution can also be seen as extending risk-sensitive control to a model that includes non-observable states. However, even if that is a significant step forward from a computational perspective, one may ask what does the prior step of casting goals as probabilistic priors entail? Is it a falsifiable statement? And what leverage would this reformulation give us in terms of explanations and predictions? To provide a counter example, the bottom-up embodied Distributed Adaptive Control (DAC) theory of the brain bootstraps goals from the foundation of the self (Verschure, Pennartz, & Pezzulo, 2014); that is, goals emerge from needs that emerge to reduce drives. DAC expands across a number of layers (reactive, adaptive, and contextual) providing the system with the means to achieve goals that ultimately serve drive reduction, sustaining the physically instantiated self from feeding to fighting and from reproducing to self-realization. Here a large and variable set of goals emerges, in turn comprising a multitude of states: Perception, value, and action. In contrast, FEP seems to replace all drives with a single one: The minimization of surprise. If one now looks at where goals, formulated as probabilistic priors, originate, FEP's single drive claim might be a *trompe-l'oeil*: Instead of *explaining* goals, it adds an additional meta-goal. Hence, the elegance of FEP comes for a price in terms of its assumptions and this cost is not sufficiently considered. In addition, FEP seems to strive toward the super power of explaining everything. This, however, will make it transcend the obligation of each theory to be testable. Hence, with FEP the devil is in the priors.

Ever since Helmholtz's work, we can look at the brain as a prediction machine. We can note, however, that Plato has already struggled with the issue of beliefs as predictions in his *Theaetetus* (369 BC). Be that as it may, the question now becomes: If computation can be unified to such an extent in terms of FEP, why is the brain so diverse in its implementation of prediction-based mechanisms? The specific predictive mechanisms that have been identified seem to differ markedly across brain areas. For instance, prediction in neocortical area A1 is mediated through recurrent inhibition (Sánchez-Montañés, Konig, & Verschure, 2002), the hippocampus seems to use attractor dynamics in coupled excitatory neurons for the same purpose (Rennó-Costa, Lisman, & Verschure, 2014), and the cerebellum relies on a well-defined tri-synaptic nucleo-olivo-cortical loop to adjust its predictions based on negative feedback (Herreros & Verschure, 2013), to mention three distinct neuronal systems where prediction modulates plasticity that we have directly

This work was supported by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) [grant number 341196].

studied in the context of the DAC theory. This variability in the implementation in distinct systems raises the fundamental question of how this apparent conflict between the synthesis of a unified computational framework of FEP and the diversity of computational realizations in biology can be resolved. Even if active inference succeeds in providing a unified explanation of behavior, then FEP will still have to backtrack and explain why evolution has de-unified its implementation or it will be shown to be incomplete.

Concepts such as predictive coding, sensory prediction-based motor control, risk-sensitive control, and planning as inference have huge explanatory power, and may end up reshaping our understanding of the brain and directing future neuroscience research. Free energy seems to provide a canvas integrating all of them, but at present it is still difficult to see which is its specific contribution. For instance, if such a contribution stems from the self-consistent prior at the heart of the active inference formulation, namely that "any system that does not (believe it will) minimize the long-term average of surprise does not (believe it will) exist," how can the neuroscience community benefit from this insight or actually anyone who has the wish to explain natural phenomena and derive predictions from that explanation? In addition, FEP also faces the risk of panpsychism by expanding the explanation to "any system."

In summary, the free energy and active inference theories anticipate a Copernican shift in theoretical neuroscience where commonly accepted concepts in the twentieth century, like the perceive-think-act cycle and the classical interpretation of a neuron's receptive field, are to be replaced by a more powerful framework. It is, however, not about claiming to have found a descendent of Copernicus, but rather whether one has better explanatory and predictive power combined with an enhanced ability to control nature. We expect that in the future, as FEP and active inference is applied to more scenarios, it will eventually jump to the domain of controlling real-world artifacts, as that is, we believe, the definitive test for any theory of behavior and, as such, it will facilitate the identification of the value of FEP in understanding mind and brain.

REFERENCES

Herreros, I., & Verschure, P. F. (2013). Nucleo-olivary inhibition balances the interaction between the reactive and adaptive layers in motor control. *Neural Networks*, *47*, 64–71. doi:10.1016/j.neunet.2013.01.026

Ortega, P. A., & Braun, D. A. (2011). Information, utility and bounded rationality. *Lecture Notes on Artificial Intelligence*, *6830*, 269–274.

Redgrave, P., & Gurney, K. (2006). The short-latency dopamine signal: A role in discovering novel actions? *Nature Reviews. Neuroscience*, *7*, 967–975. doi:10.1038/nrn2022

Rennó-Costa, C., Lisman, J. E., & Verschure, P. F. (2014). A signature of attractor dynamics in the CA3 region of the hippocampus. *Plos Computational Biology*, *10*(5), e1003641.

Sánchez-Montañés, M. A., Konig, P., & Verschure, P. F. (2002). Learning sensory maps with real-world stimuli in real time using a biophysically realistic learning rule. *IEEE Transactions on Neural Networks*, *13*(3), 619–632. doi:10.1109/TNN.2002.1000128

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, *80*, 1–27.

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

Verschure, P. F., Pennartz, C. M., & Pezzulo, G. (2014). The why, what, where, when and how of goal-directed choice: Neuronal and computational principles. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1655). doi:10.1098/rstb.2013.0483

* * *

Exploration-exploitation: A cognitive dilemma still unresolved

Russell N. James III

Texas Tech Neuroimaging Institute, Texas Tech University, Lubbock, USA

E-mail: russell.james@ttu.edu

http://dx.doi.org/10.1080/17588928.2015.1051012

Abstract: The solution to the exploration-exploitation dilemma presented essentially subsumes exploitation into an information-maximizing model. Such a single-maximization model is shown to be (1) more tractable than the initial dual-maximization dilemma, (2) useful in modeling information-maximizing subsystems, and (3) profitably applied in artificial simulations where exploration is costless. However, the model fails to resolve the dilemma in ethological or practical circumstances with objective outcomes, such as inclusive fitness, rather than information outcomes, such as lack of surprise.

Keywords: Exploration; Exploitation; Simulation.

The exploration-exploitation dilemma presents the conflict between the need to obtain new knowledge and the need to use that knowledge to improve outcomes or performance (Laureiro-Martínez, Brusoni, & Zollo, 2010). Friston et al. (this issue) "offer a solution to the exploration-exploitation dilemma that rests solely on the minimization of expected free energy." The minimization of expected free energy is initially described as maximizing extrinsic value while maximizing information gain. Although this may initially appear to simply restate the well-known dilemma, a further operationalization of this "free energy" definition leads to a model where the problem of maximizing two outcomes is re-cast as maximizing one of the two outcomes, namely information gain. Expected value can be expressed in terms of information—and expected information gain has value. In other words, information and value have the same currency and can be combined into a single "free energy" imperative. Thus, the minimization of expected free energy is also referred to as minimizing surprise, or equivalently, as maximizing Bayesian model evidence.

How then is the exploitation half of the exploration-exploitation dilemma subsumed within a purely information-gain model? This is done by expressing "(extrinsic) reward in terms of (epistemic) information gain. . . ." Specifically, "preferred outcomes are simply outcomes one expects, *a priori*, to be realized through behavior. . . ." Thus, by definition, a surprising outcome —one where information was lacking—cannot be preferred. This results in a resolution of the exploration-exploitation dilemma by giving primacy to information gain. Where contingencies are unknown, "epistemic value is maximized until there is no further information gain, after which exploitation is assured through maximization of extrinsic value." Such an information-centric model must still account for the need of action. This is done by noting that action is important, because information gain requires sampling the world to resolve uncertainty.

Resolving the dual-priority conflict by subsuming one priority within the other provides enormous advantages for mathematical tractability and theoretical simplicity. It will also be readily confirmed by observations from those parts of a system designed to accomplish one of the two priorities. However, such a resolution works only if the simplifying assumptions are valid. Specifically, *if* preferred outcomes are synonymous with expected outcomes, *then* information gain becomes the primary goal. The model breaks down, however, if the preferred outcomes are objectively distinct from information (i.e., where an outcome may be defined separately from its informational characteristics). If the outcome is, e.g., calories, survival, reproduction, or inclusive fitness, then surprise is no longer independently relevant except to the extent that it impacts those external outcomes. In such cases, the optimal tradeoff between exploitation and exploration depends entirely upon environmental circumstances, and no one strategy will be, a priori, preferred.

Is it possible to construct a simulation in which exploration (information maximization) is the dominant goal? Yes. This is done by creating a situation in which there are no returns to further exploitation. The two-move foraging simulation in Friston et al. (this issue) does this by ending the game (via a trap door) after the first exploitation move. Unlike the typical ethological circumstances where each exploration move comes at the cost of a foregone exploitation move, this simulation prohibits two exploitation moves, and thus makes the initial exploration move costless. Clearly, when exploration is costless, an information-maximizing model will dominate, as in the simulated trials. Just as clearly, such a simulation provides no relevant information for the actual underlying dilemma.

Is it possible to identify information-maximizing neural processes in nature? Yes. This is done by analyzing processes that are themselves information-maximizing functions. Thus, we would fully expect that information-maximizing processes, such as active vision based upon salience or message-passing schemes, would fully conform to an information-maximizing model. However, in the more holistic exploration-exploitation dilemma, relative preferences for information-seeking as compared with immediate experiential outcomes (calories, temperature, sex) may change depending upon the current condition of relevant homeostatic processes (hunger, cold, lust). Thus, relative dopaminergic gains for novel information may change relative to gains for experienced sensation depending upon state conditions (Gros, 2010) that themselves may reflect changing environmental circumstances, leading back to the core environmental-dependent exploration-exploitation dilemma where no single a priori model dominates and prior preferences are likely to be highly context-dependent. Active inference provides an excellent approach to modeling information maximizing systems—and can be applied in simulations—but does not explicitly address the real world exploitation-exploration dilemma inherent in an agent's context-sensitive preferences.

REFERENCES

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (this issue). Active inference and epistemic value. *Cognitive Neuroscience*. doi:10.1080/17588928.2015.1020053

Gros, C. (2010). Cognition and emotion: Perspectives of a closing gap. *Cognitive Computation*, *2*(2), 78–85. doi:10.1007/s12559-010-9034-7

Laureiro-Martínez, D., Brusoni, S., & Zollo, M. (2010). The neuroscientific foundations of the exploration−exploitation dilemma. *Journal of Neuroscience, Psychology, and Economics*, *3*(2), 95–115. doi:10.1037/a0018495

* * *

An active inference and epistemic value view of metacognition

Chris Moulin and Celine Souchay
LEAD CNRS UMR 5022, Université de Bourgogne, Dijon France
E-mail: celine.souchay@u-bourgogne.fr

http://dx.doi.org/10.1080/17588928.2015.1051015

Abstract: Metacognition concerns our monitoring and control of mental operations (*knowing what you know*). Much thinking about metacognition is liable to fall foul of the classic homunculus problem: Nobody can specify who or what does the “metacognition.” We describe how the Active Inference and Epistemic Value (AIEV) model offers an operationalization of epistemic behaviors which can explain two example metacognitive phenomena: Control and monitoring of word learning, and the search for unretrieved information in the feeling of knowing. Curiosity drives a search forward, but it is held in check by considering the utility of what is retrieved from memory.

Nelson and Narens (1990) proposed the most widely cited account of metacognition (*knowing what you know*). In their framework, flows of information between an object level and a meta-level are characterized as representing monitoring and control of cognitive processes. The processes at play in such flows are not well specified (but see Fleming, Dolan, & Frith, 2012) and fall foul of the classic homunculus problem: Nobody can specify who or what does the “metacognition” in their framework. In Friston et al.'s paper we see a concrete operationalization of epistemic behaviors that can explain metacognitive phenomena; an inference-making machine based on the principle of minimizing uncertainty and the efficient expenditure of resources. We briefly outline these ideas here, but the topic in general warrants a much more developed examination.

A common task in assessing metacognition is to learn a set of words over repeated trials. We infer that attention is orientated toward items studied for the first time because they are novel: Study times are longer for items seen for the first time than at subsequent repetitions. If we ask participants to make an explicit declaration of their metacognitive evaluation, they will give higher predictions of subsequent performance for the items they have seen more frequently. Importantly, more study time will be allocated to items thought to be difficult to remember or which are poorly learned. But a bi-directional relationship also exists: It is because we study something for longer, or that learning is non-fluent that we rate things as difficult to remember (e.g., Koriat, Ma'ayan, & Nussinson, 2006).

According to Friston et al., there is an epistemic value to divergences between expected and observed behaviors in an ongoing task such as this. Friston et al.'s system is intrinsically metacognitive: “valuable policies will search out observations, cues, or signs that resolve uncertainty about the state of the world.” The formation and retention of “valuable” policies explains the acquisition, adaption, and implementation of mnemonic strategies (as opposed to, for instance, trial-and-error learning), whereas the search for cues and signs is the process of monitoring the operations of the cognitive system. The metacognitive system acts on *epistemic feelings* (see Moulin & Souchay, 2013) to reduce inefficiencies in the system—which are either unknowns or uncertainties about current or future performance based on current goals. In sum, active inference is not random—because it is “tailored” toward goals and receives feedback through Bayesian processes—we argue that this process encapsulates the flows of information captured in metacognitive notions of control and monitoring. Technically speaking, the active inference scheme described in the target article is quintessentially metacognitive, because epistemic value (the opportunity to minimize uncertainty) of prior beliefs is a function of posterior beliefs about uncertainty. In other words, at the heart of planning through inference, there is a quantity (expected free energy) that rests upon beliefs about beliefs.

In our word-learning example, the principle of minimizing uncertainty will signal when the processing fluency for a word in a list differs from what is expected, and an efficient maximization of information gain will allocate study time appropriately (it will also "know" when to give up when learning is impossible). Thus, Friston et al.'s AIEV framework describes how feedback systems operate to regulate human learning.

One of the strengths of the AIEV approach is that it draws upon Markov decision processes where states are not directly observable and where there is, in short, missing data. The state of "unknowns" in metacognition is a pivotal point, as pointed out by Fleming et al. (2012, p. 1285):

> Object-level representations are often concerned with presence of stimuli in the world; they rarely deal in absence . . . . In contrast, "knowing I do not know" is a meta-level representation of the absence of object-level memory. Investigating this putative function may benefit from greater integration with work quantifying epistemic behaviour—by sampling information over time, an agent can adaptively reduce its uncertainty, achieving a balance between the additional cost of exploration and the benefit of gaining further information

In the Feeling of Knowing (FOK) phenomenon (e.g., Souchay, Isingrini, & Espagnet, 2000) people can accurately gauge the state of their memory system, even when the searched for information cannot be retrieved. When asked "Who was the director of the film Black Swan?" we may find ourselves unable to answer, but a set of information may trigger decision-making as the search for the answer unfolds in time. This search for the answer is well represented by the epistemic feeling of curiosity, and the drive to reduce uncertainty. The FOK is a dynamic state, and according to the AIEV view, based on a series of exploratory searches for information inherent in Markovian processes. The production (or not) of information whilst searching promulgates or terminates the search—and ultimately this can be output as an explicit declaration: *I don't know/It's on the tip-of-my-tongue/Darren Aronofsky*. A certain excitation of the memory system exists at such times, but when the epistemic value of search is small, such as when retrieving irrelevant or repetitive information, the search will be terminated. Curiosity drives the search forward, but it is held in check by considering the utility of what is retrieved from memory (in comparison with what is expected). Thus, the Friston et al. article goes a long way to answering a critical question in metacognition—*what is it we know when we think we do not know something*? We imagine that the answer to this question is that we are aware of an epistemic process akin to that described in the AEIV model.

REFERENCES

Fleming, S., Dolan, R., & Frith, C. (2012). Metacognition: Computation, biology and function introduction. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*, 1280–1286. doi:10.1098/rstb.2012.0021

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology. General*, *135*(1), 36–69. doi:10.1037/0096-3445.135.1.36

Moulin, C. J. A., & Souchay, C. (2013). Epistemic feelings and memory. In T. Perfect, & S. Lindsay (Eds.), *The SAGE handbook of applied memory* (pp. 520–539). London: SAGE Publications Ltd. doi:10.4135/9781446294703.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation*, *26*, 125–173. doi:10.1016/S0079-7421(08)60053-5

Souchay, C., Isingrini, M., & Espagnet, L. (2000). Aging, episodic memory feeling-of-knowing, and frontal functioning. *Neuropsychology*, *14*(2), 299–309. doi:10.1037//0894-4105.14.2.299

* * *

Dopamine and epistemic curiosity in music listening

Diana Omigie

Music Department, Max Planck Institute for Empirical Aesthetics, Frankfurt am Main, Germany

E-mail: diana.omigie@aesthetics.mpg.de

http://dx.doi.org/10.1080/17588928.2015.1051013

This work was supported by a stipend to Diana Omigie from the Max Planck Institute for Empirical Aesthetics.

Abstract: Elucidating the cognitive, affective, and reward processes that take place during music listening is the aim of a growing number of researchers. Several authors have used the Bayesian brain framework and existing models of reward to interpret neural activity observed during musical listening. The claims from Friston and colleagues regarding the role of dopamine, as well as the demonstration that salience-seeking behavior naturally emerges from minimizing free energy, will be of potential interest to those seeking to understand the general principles underlying our motivation to hear music.

In a previous perspective piece (Schwartenbeck, FitzGerald, Dolan, & Friston, 2013), it was suggested that the *Free Energy Principle* formalisms in their current state might not be sufficient "to explain all aspects of higher level activities, such as the appreciation of fine arts." Nevertheless, one cannot help but wonder what the appropriate formalisms would look like. The current article summons similar interest. Not least because the specific claims made are of considerable relevance to the psychological and neural underpinnings of music listening.

A first claim of interest is that of the role of dopamine. Dopamine has been associated with rewarding aspects of music listening via evidence of recruitment of dopaminergic areas and through direct observation of its release during the anticipation and experience of peak emotional responses (e.g., Salimpoor, Benovoy, Larcher, Dagher, & Zatorre, 2011). Interpretations of such data have been made variously in the context of the *Incentive Salience Theory* and/or in the context of the *Reward Prediction Error* hypothesis (Gebauer, Kringelbach, & Vuust, 2012; Salimpoor et al., 2011). The claims in Friston et al. (this issue) that dopamine may be thought of as confidence or belief in an action and as strongly related to so-called *value* would appear to be more in line with the Incentive Salience hypothesis. How this precise account of dopamine may be used to interpret its presence (or not) during enjoyable (or not) acts of musical listening is an interesting question.

Perhaps the most pertinent claim raised by Friston et al., however, is that salience-seeking behavior naturally emerges from the general objective of minimizing free energy. Here it is interesting to note that appreciation of artworks necessitates knowledge accumulation—and this on several time scales. In music, while some knowledge acquisition is implicit (for example, learning the norms of one's native tonal system), others (for instance, learning the structure and form of less popular musical styles) may require more effort. While it is clear that some caution is in order, the attempt of Friston et al. to demonstrate that free energy maximizes *intrinsic* or *epistemic value* or, in other words, exploratory actions, may have implications for music listening specifically and art consumption more generally.

Indeed, it is worth noting that the term *Epistemic emotions* (encompassing emotions like interest, curiosity, and fascination) has been used to describe the affective states engaged while contemplating visual art and music. Further, epistemic emotions have been argued to be distinct from so-called *Utilitarian emotions* in not being triggered by concern for wellbeing or survival (Scherer, Coutinho, Cochrane, & Fantini, 2013). Interestingly, this claim of a relationship between feelings of interest and art consumption resonates with Berlyne's seminal work on curiosity, arousal, and experimental aesthetics (1960). It also resonates with the notion that resolving uncertainty through epistemic or explorative acts "makes the world interesting and exploitable."

At this point, it is useful to emphasize why music listening entails beliefs about policies or action. Here, it is important to remember that attending to music is an active process, and one that we carry out in the hope of resolving any uncertainties elicited by the unfolding musical narrative. An ambitious but relevant question is how future accounts could conceptualize extrinsic and intrinsic value or exploitation and exploration in the context of music listening. One working assumption could be that there is no extrinsic value to be sought in music, therefore leading the agent to always maximize intrinsic value. Another assumption, however, could be that an exploration act involves shifting attention to one of the many possible streams in a complex piece of music, in the knowledge that, once resolved, the given stream along with others already resolved can be "exploited" to bring about explicit value in terms of correct predictions.

In closing, various attempts have already been made to use predictive coding as a framework with which to interpret the role of dopamine in the context of music listening (Gebauer et al., 2012). Accordingly, the new claims by Friston et al. regarding the role of dopamine will be of interest to several workers in the field of music research. Further, while one can certainly agree that the formalisms are at present relatively underspecified for some tasks, it is interesting to consider the patterns that seem to be emerging. In its justification of behaviors with intrinsic value and only long-term benefits, Friston et al.'s free energy formalisms provide support for psychological accounts of the importance of interest, curiosity, and exploratory behavior.

REFERENCES

Berlyne, D. E. (1960). Conflict, arousal, and curiosity. New York, NY: McGraw-Hill.

Gebauer, L., Kringelbach, M. L., & Vuust, P. (2012). The ever changing cycles of musical pleasure: The role of dopamine and anticipation. Psychomusicology: Music, Mind, and Brain, 22, 152–167. doi:10.1037/a0031126

Salimpoor, V., Benovoy, M., Larcher, K., Dagher, A., & Zatorre, R. (2011). Anatomically distinct dopamine release during anticipation and experience of peak emotion to music. Nature Neuroscience, 14, 257–262. doi:10.1038/nn.2726

Scherer, K. & Coutinho, E. (2013). How music creates emotion: a multifactorial process approach. In Cochrane T., Fantini, B., & Scherer, K. R. (Eds.). *The emotional power of music*. Oxford: Oxford University Press.

Schwartenbeck, P., FitzGerald, T., Dolan, R. J., & Friston, K. (2013). Exploration, novelty, surprise, and free energy minimization. Frontiers in Psychology, 4, 710. doi:10.3389/fpsyg.2013.00710