Biol Cybern (2012) 106:523–541
DOI 10.1007/s00422-012-0512-8

PROSPECTS

Biological Cybernetics

Active inference and agency: optimal control without cost functions

Karl Friston · Spyridon Samothrakis · Read Montague

Received: 1 February 2012 / Accepted: 16 July 2012 / Published online: 3 August 2012

Abstract This paper describes a variational free-energy formulation of (partially observable) Markov decision problems in decision making under uncertainty. We show that optimal control can be cast as *active inference*. In active inference, both *action and posterior beliefs* about hidden states minimise a free energy bound on the negative log-likelihood of observed states, under a generative model. In this setting, reward or cost functions are absorbed into prior beliefs about state transitions and terminal states. Effectively, this converts optimal control into a pure inference problem, enabling the application of standard Bayesian filtering techniques. We then consider optimal trajectories that rest on posterior beliefs about hidden states in the future. Crucially, this entails modelling control as a hidden state that endows the generative model with a representation of agency. This leads to a distinction between models with and without inference on hidden control states; namely, agency-free and agency-based models, respectively.

Keywords Partially observable Markov decision processes · Optimal control · Bayesian · Agency · Inference · Action · Free energy

K. Friston (✉)
The Wellcome Trust Centre for Neuroimaging, UCL, Institute of Neurology, 12 Queen Square, London WC1N 3BG, UK
e-mail: k.friston@ucl.ac.uk

S. Samothrakis
School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, UK

R. Montague
Department of Physics, Virginia Tech Carilion Research Institute, Virginia Tech, 2 Riverside Circle, Roanoke, VA 24016, USA

1 Introduction

In this work, we apply variational free-energy minimisation to a well-studied problem in optimal decision theory, psychology and machine learning; namely, Markov decision processes. In brief, we show that the free-energy principle (active inference) and optimal decision theory provide the same solutions when the policies from optimal decision theory are replaced by (prior) beliefs about transitions from one state to another. This is important because specifying behaviour in terms of prior beliefs or policies finesses the difficult problem of optimising behaviour to access distal rewards. Furthermore, it enables one to consider more general notions of optimality in terms of accessing particular states in the future. Bayes-optimal behaviour then depends upon a representation of future behaviours that necessarily entails a model of agency. We illustrate how agency-based decision making can solve quite difficult problems and touch on the possible implications for understanding psychopathology.

This paper considers discrete time (Markov) decision processes of the sort found in optimal control theory, models of behaviour and decision making (Bellman 1952; Watkins and Dayan 1992; Camerer 2003; Daw and Doya 2006; Todorov 2006; Dayan and Daw 2008). Our aim is to establish a link between classical approaches to optimising decisions, in terms of policy optimisation, and the variational free-energy minimisation that underlies active inference (Friston et al. 2009; Beal 2003). Here, classical optimal control schemes are taken to imply that actions (and beliefs about hidden states of the world) are chosen to maximise the expected reward of *future states*. Conversely, in active inference, actions and beliefs about hidden states minimise a variational free energy bound on the (negative log) marginal likelihood of *observed states*, that is, they maximise the marginal likelihood. Linking the two formulations necessarily requires us to formulate

free-energy minimisation in discrete time and think about how reward or cost functions are accommodated.

The key distinction between optimal control and active inference is that in optimal control, action optimises the expected cost associated with the hidden states a system or agent visits. In contrast, active inference requires action to optimise the marginal likelihood (Bayesian model evidence) of observed states, under a generative model. This introduces a distinction between cost-based optimal control and Bayes-optimal control that eschews cost. The two approaches are easily reconciled by ensuring the generative model embodies prior beliefs about state transitions that minimise expected cost. Our purpose is, therefore, not to propose an alternative implementation of optimal control but accommodate optimal control within the larger framework of active inference. In other words, we consider Bayes-optimal control in systems (like the brain) that have to optimise their own actions and beliefs in real-time. This is illustrated by casting terminal cost (the cost of the final state) as prior beliefs about future states that inform posterior beliefs about future control. Crucially, this requires (future or fictive) control to be treated as a hidden state, which means agents have to make inferences about their future behaviour. We associate this inference with planning and a (probabilistic) representation or sense of agency.

Replacing cost functions with prior beliefs allows one to consider optimality in terms of fulfilling prior beliefs about exchanges with the world. This is the basis of active inference, in which action minimises surprise, where surprise is based upon Bayesian predictions about outcomes that are shaped by prior beliefs. In this view, cost functions are replaced by (or absorbed into) prior beliefs about state transitions. At first glance, this may sound untenable; in the sense that we entertain beliefs about particular states irrespective of their value—in the words of one of our reviewers, "how can I express my values in terms of my beliefs without catastrophically eliding the two?" For example, I can believe I am being drenched by rain and yet place a high cost on this state of affairs. However, if I believe that I will seek shelter when it rains, then I will behave optimality, provided I act to fulfil these beliefs. Note that these prior beliefs are not about states of the world but transitions among states (i.e., a policy). So how can one specify optimal behaviour in terms of prior beliefs?

Imagine a (Bayesian) thermostat that infers the ambient temperature through noisy thermoreceptors. This thermostat can position itself in relation to a heat source and is equipped with reflexes that move it towards the source when the predicted temperature is higher than the sensed temperature, and away from the source when the predicted temperature is lower. In the absence of prior beliefs about temperature, the predictions will be an unbiased estimate of average thermoreceptor activity and action (reflexive movement) will not be engaged. However, if the Bayesian thermostat has strong prior beliefs about a particular temperature (its set point), it will move towards or away from the heat source, until the prior predictions and ambient temperature concur. In other words, its prediction errors will be resolved through action. This illustrates, heuristically, how optimal behaviour can be cast as inference. Crucially, both action and perception (estimating the hidden causes of sensory input) are trying to minimise the same thing—roughly speaking, prediction error or the surprise associated with sensations.

This is active inference in its simplest form and has been considered in the context of reinforcement learning (Friston et al. 2009), action selection (Friston et al. 2012) optimal motor control (Friston 2011) and dynamical systems theory (Friston and Ao 2012). However, all of these treatments consider behaviour in continuous time, in contrast to the discrete state space and time formulations that predominate in the literature on optimal decision problems. The motivation for the current work was to bring active inference into a discrete framework, so that it could be compared and contrasted with conventional optimal control and decision theoretic treatments. This allows us to make two important points:

- Optimal control problem formulations can be absorbed into (replaced by) active inference schemes for action and perception.
- Active inference introduces a distinction between action and control that leads to a sense of agency.

The second point is particularly important: in active inference, there is a necessary distinction between action—that couples the agent to its environment—and control, which is a random variable that represents action. This distinction is not usually part of conventional schemes but plays a crucial role in active inference, where agents have to infer their actions using probabilistic representations over hidden control states. In what follows, we define a sense of agency as a probabilistic representation of control that is distinct from the action actually emitted. We hope to show that control (fictive action) plays a key role in realising prior beliefs about the future, and finesses the problem of planning and searching over future options.

It should be noted that this paper is about control not learning. In other words, it is about the prosecution (and planning) of a policy through inference, given a prior belief about desired outcomes—it is not about learning a policy. In active inference, learning a policy corresponds to acquiring (empirical) priors through optimising the parameters of a generative model (with respect to variational free energy). Neurobiologically, this generally reduces to some form of associative plasticity (Friston 2008). One straightforward way to acquire priors—over state transitions—is to marinate an agent in the statistics of an optimal world, as illustrated in Friston et al.

(2009). One might ask where these worlds come from. The answer is that they are created by teachers, parents and conspecifics. In robotics and engineering, the equivalent learning requires the agent to be shown how to perform a task. This form of learning has been used to produce some compelling and animate behaviours (Tani 2003; Namikawa et al. 2011).

This paper comprises three sections: the first reviews Markov decision processes (MDPs), their extensions to partially observable Markov decision processes (POMDPs) and the problem of finding an optimal policy using belief MDPs. We then revisit the problem from the point of view of active inference and demonstrate their formal relationships, active inference separates *inference* about hidden states causing observations from *action*. The motivation for this is pragmatic; in that real agents cannot know how their action affects hidden states (because hidden states have to be inferred). This means that action must be based on a function of observed, as opposed to hidden states. Active inference assumes that this function is the same variational free energy used in approximate Bayesian inference (Hinton and van Camp 1993; MacKay 1995; Neal and Hinton 1998; Dayan et al. 1995; Beal 2003). In other words, active inference extends the minimisation of variational free energy that underlies (approximate) Bayesian inference to *include action* (Friston et al. 2010). However, requiring action to minimise variational free energy appears to contradict optimal control theory, which requires action to minimise expected cost.

The purpose of the second section is to resolve this conflict: in brief, we will see that the cost functions that are used to guide action in optimal control theory can be absorbed into prior beliefs in active inference. Effectively, this means that agents expect their state transitions to minimise cost, while action realises these prior beliefs by maximising the marginal likelihood of observations. Clearly, from the point of view of classical POMDP treatments this does not represent a great advance, because it just establishes a formal equivalence between cost and priors, in terms of ensuing action; however, this equivalence means we can dispense with cost functions and formulate optimal control in terms of approximate Bayesian inference on hidden states, namely Bayes-optimal control. This means one can use standard Bayesian filtering schemes to solve optimal control problems.

The third section illustrates this by showing how optimal policies can be inferred under prior beliefs about future (terminal) states using standard variational Bayesian procedures (Beal 2003). This example leads to a model-based optimisation of behaviour that may provide a useful metaphor for planning, anticipation and a sense of agency in real-world agents. We conclude with an example (the mountain car problem) that illustrates how active inference furnishes online non-linear optimal control, with partially observed (hidden) states that are subject to random fluctuations.

2 Markov decision processes

This section provides a brief summary of Markov decision problems and their solutions based upon cost or reward functions that are an integral part of optimal control theory and reinforcement learning.

Definition A Markov decision process is the tuple $(S, A, \mathbf{T}, \mathbf{r})$, where

- S is a finite set of states.
- A is a finite set of actions.
- $\mathbf{T}(s'|s,a) = \Pr(\{s_{t+1} = s'|s_t = s, a_t = a\})$ is the (transition) probability that the state $s' \in S$ at time $t+1$ follows action $a \in A$ in state $s \in S$ at time t.
- $\mathbf{r}(s)$ is some reward received at state $s \in S$.

Problem The goal is to find a *policy* $\pi : S \to A$ that maximizes cumulative rewards. This can be expressed in terms of the sequence of actions $a_{0:T} := \{a_0, \ldots, a_T\}$ that maximises *value* or negative *cost-to-go*:

$$V(s) = \max_{a_{0:T}} \left\{ \mathbf{r}(s) + \sum_{i=1}^{T} \sum_{s'} \Pr(\{s_i = s'|s_0 = s, a_0, \ldots, a_i\}) \mathbf{r}(s') \right\}. \tag{1}$$

The solution to this equation is a policy or sequence of optimal actions $a_t := \pi(s_t)$ that maximises expected reward in the future, given a probabilistic model of state transitions. In this setting, $(\mathbf{T}, \mathbf{r})$ constitutes a model that comprises a transition matrix and a vector of rewards defined on states. Equation (1) can be expressed as the *Bellman optimality equation* by exploiting the Markovian nature of the problem using recursive substitution (Bellman 1952):

$$V(s) = \max_{a} \left\{ \mathbf{r}(s) + \sum_{s'} \mathbf{T}(s'|s,a) V(s') \right\}. \tag{2}$$

For simplicity, we have assumed a *finite horizon* problem, in which the reward is maximized from $t = 0$ to $t = T$. This allows us to eschew notions of discounting required in infinite horizon problems. Solutions to MDPs can be divided into *reinforcement learning* schemes that compute the value function explicitly and *direct policy searches* that find the optimal policy directly.

In direct policy searches e.g., (Williams 1992; Baxter et al. 2001), a policy is optimised by mapping each state directly to an action, without reference to the value of the state. Policies can also be optimised directly using genetic algorithms

e.g., (Gomez and Miikkulainen 2001; Gomez et al. 2009), where agents are selected to maximise their expected reward through successive exposures to environmental contingencies. Agents that perform well are mutated accordingly, and a new generation of agents are evaluated on the MDP, until the policy is optimised (i.e., the Bellman error is small). Direct policy searches are useful when the value function is hard to learn but the policy is easy to find.

In value-based reinforcement learning there are two general approaches: The first (model based) computes the value function using a model of state transitions and is usually considered when the state space is sufficiently small. This is known as *dynamic programming* and involves iterating the following two steps (Bellman 1952):

$$\pi(s) = \arg\max_a \left\{ \mathbf{r}(s) + \sum_{s'} \mathbf{T}(s'|s,a) V(s') \right\}$$
$$V(s) = \mathbf{r}(s) + \sum_{s'} \mathbf{T}(s'|s,\pi(s)) V(s') \quad (3)$$

This scheme is guaranteed to find the optimal solution, provided all states are visited. In *value iteration* or *backwards induction*, the policy is only calculated when needed. This gives the combined step in (1). In *policy iteration* (Howard 1960), the first step is repeated until convergence, thereby providing a definite stopping condition.

If the transition probabilities or rewards are unknown or the state space is large (precluding a visit to every state), the problem is usually solved with model free reinforcement learning. In these schemes (Rescorla and Wagner 1972; Sutton and Barto 1981; Watkins and Dayan 1992; Friston et al. 1994; Montague et al. 1995) the value function is itself learnt; this enables one to solve Markov decision problems without learning the model (transition probabilities), because the value function acts as a guidance function for action.

2.1 Partially observable Markov decision processes

The formulation above assumes that the agent knows what state it is in. In many scenarios this is unrealistic, in that an agent cannot know the exact state of the world, given noisy or partial observations (Rao 2010). This leads to an extension of the MDP framework to accommodate partially observed states (Kaelbling et al. 1998).

Definition A Partially Observable Markov Decision process is the tuple $(S, A, \mathbf{T}, \mathbf{r}, \Omega, \mathbf{O})$ where

- $(S, A, \mathbf{T}, \mathbf{r})$ is the same tuple as in the MDP formulation.
- Ω is a finite set of observations or outcomes.
- $\mathbf{O}(o|s) = \Pr(\{o_t = o | s_t = s\})$ is the (observation) probability of $o \in \Omega$ given the agent is in state $s \in S$ at time t.

Although it is possible to solve POMDPs using direct policy searches (Gomez et al. 2009), one cannot perform value iteration or reinforcement learning directly, as they require the hidden states. However, a POMDP can be converted to a MDP using beliefs $b(s)$ about the current state: Beliefs are sufficient statistics that can be computed recursively from the observations and actions, where (using Bayes rule):

$$b'(s') = P(s'|o,a,b) = \frac{P(o|s',a,b)P(s'|a,b)}{P(o|a,b)}$$
$$\propto \mathbf{O}(o|s',a) \sum_{s \in S} \mathbf{T}(s'|s,a) b(s). \quad (4)$$

One can then treat the beliefs as states to create a "Belief MDP":

Definition A Belief Markov Decision Process is the tuple $(B, A, \mathbf{T}, \mathbf{r})$ where

- B is the set of belief states over the POMDP states.
- A is a finite set of actions.
- $\mathbf{T}(b'|b,a) = \Pr(\{b_{t+1} = b' | b_t = b, a_t = a\})$ is the probability that the belief state $b' \in B$ at time $t+1$ follows action $a \in A$ in belief state $b \in B$ at time t.
- $\mathbf{r}(b) = \sum_{s \in S} b(s)\mathbf{r}(s)$ is the reward expected in belief state $b \in B$.

Remark Note that a belief MDP is defined over a continuous (belief) state space, which can make them hard to solve using reinforcement learning or dynamic programming (see Oliehoek et al. 2005). However, there are heuristic solutions, which range from ignoring the observation model completely to using function approximators to encode beliefs. The difficult problem of solving large POMDPs is central to Artificial Intelligence and is an important focus of current research (Silver and Veness 2010). See Duff (2002) for a full treatment of POMDPs based on a Bayesian formulation that exploits techniques from reinforcement learning, such as Monte Carlo simulations and parameterised function approximators.

In summary, classical approaches to Markov decision processes rest on the optimization of future rewards and specify an optimal policy in terms of an action from any given state. Note that MDPs appeal to a solipsistic view of the world, in which an agent tries to maximise its future reward against a nature that is governed by laws the agent can infer (one of their key features is the assumption that only the current state matters, hence the 'Markov' label). Partially observed Markov decision processes make inference explicit by introducing a probabilistic mapping between hidden states of the world and observations. Thus, the beliefs that the agent forms (by observing histories of actions and states) can be exploited to optimise behaviour.

2.2 Optimal control as inference

The focus of this paper is an optimal decision making or control as an inference process. Early work in this area addressed dual control problems; for example, Feldbaum (1961) discussed control in the absence of complete information using a Bayesian framework: see Filatov and Unbehauen (2004) for review. The integration of control and inference was pursued by replacing the notion of utility in decision diagrams with an auxiliary random variable conditioned on desired observations. This makes maximizing utility equivalent to maximizing the likelihood of desired observations (Cooper 1988; Pearl 1988; Shachter 1988). Subsequent work focussed on efficient methods to solve the ensuing inference problem (Jensen et al. 1994; Zhang 1998). Later, Dayan and Hinton (1997) proposed an expectation maximization algorithm for reinforcement learning in the case of immediate rewards, while (Toussaint and Storkey 2006) cast the problem of computing optimal policies as a likelihood maximization problem. This generalized the work of Cooper and Shachter to the case of infinite horizons, and cost functions over future states. More recently, this approach has been pursued by applying Bayesian procedures (or minimising Kullback–Leibler divergences) to problems of optimal decision making in MDPs Botvinick and An 2008; Hoffman et al. 2009; Toussaint et al. 2008).

Related work on stochastic optimal control (Kappen 2005a,b; van den Broek et al. 2008; Rawlik et al. 2010), exploits the reduction of control problems to inference problems by appealing to variational techniques to provide efficient and computationally tractable solutions. In particular, formulating the problem in terms of Kullback–Leibler minimization (Kappen 2005a,b) and path integrals of cost functions using the Feynman–Kac formula (Theodorou et al. 2010; Braun et al. 2011).

In summary, current approaches to partially observed MDPs and stochastic optimal control minimise cumulative cost using the same procedures employed by maximum likelihood and approximate Bayesian inference schemes. Indeed, the formal equivalence between optimal control and estimation was acknowledged by Kalman at the inception of Bayesian filtering schemes (Todorov 2008). In the next section, we revisit this equivalence and show how any optimal control problem can be formulated as a Bayesian inference problem, within the active inference framework. The key aspect of this formulation is that action does not minimise cumulative cost but maximises the marginal likelihood of observations under a generative model that entails an optimal policy.

3 Active inference

This section sets up the formalism of active inference, in which the optimisation of action and beliefs about hidden states are treated as two separate processes that both maximise model evidence or the marginal likelihood of observations. In brief, in active inference, action elicits *observations* that are the most plausible under beliefs about (future) states. This is in contrast to conventional formulations, in which actions are chosen to elicit (valuable) states. We will see that active inference can implement any optimal policy; however, it does not solve the optimal control problem, because active inference does not minimise cost-to-go but minimises the self information of observations (aka surprise). This follows from the fact that active inference is a corollary of the free-energy principle:

3.1 The free-energy principle and active inference

The free-energy principle (Friston et al. 2006) tries to explain how agents occupy a small number of attracting states in terms of minimising the Shannon entropy of the probability distribution over their observed states. Under ergodic assumptions, this entropy is (almost surely) the long-term time average of self information or surprise (Birkhoff 1931). Surprise, or more precisely *surprisal*, is a (probability) measure $-\ln P(o_t|m)$ on the states observed by an agent. Minimising the long-term average $\mathbf{E}_t[-\ln P(o_t|m)]$ is assured when agents minimise surprise at each time point. Surprise is just the negative log likelihood of observations, marginalised over hidden states. This marginal likelihood is also known as model evidence. This means that surprise is minimised (approximately or exactly) if agents minimise a variational free energy bound on surprise (Feynman 1972; Hinton and van Camp 1993), given a generative model m of state transitions (Dayan et al. 1995; Friston 2010).

This formulation of behaviour is based on ergodic arguments about the nature of self organising systems (Ashby 1947)—for a fuller discussion please see Friston and Ao (2012) and Friston (2010) for their neurobiological implications. These arguments suggest that the long term average of variational free energy upper bounds the (Shannon) entropy of observations over time; which implies that action must minimise variational free energy to resist the dispersion of its states by random fluctuations (Evans 2003). This is active inference (Friston et al. 2010), which extends the minimisation of variational free energy implicit in approximate Bayesian inference on hidden states to include action per se. There is a fairly developed literature on variational free-energy minimisation and active inference in the neurosciences; covering things from perceptual categorisation of bird songs, through to action observation. Table 1 lists some processes and paradigms we have considered under this framework. The current paper introduces hidden control states that allow one to model agency and planning, using exactly the same principles used previously to explain various aspects of self organisation and perception.

Table 1 Processes and paradigms that have been modelled using variational free-energy minimisation and active inference

Process or paradigm	References
Perceptual categorisation (bird songs)	Friston and Kiebel (2009b)
Novelty and omission-related responses	
Perceptual inference (speech)	Kiebel et al. (2009b)
Pereptual learning (mismatch negativity)	Friston and Kiebel (2009b)
Attention and the Posner paradigm	Feldman and Friston (2010)
Attention and biased competition	
Retinal stabilization and oculomotor reflexes	Friston et al. (2010)
Saccadic eye movements and cued reaching	
Bayes-optimal sensorimotor integration	
Heuristics and dynamical systems theory	Friston (2010)
Action observation and mirror neurons	Friston et al. (2011)
Motor trajectories and place cells	
Goal-directed behaviour	Friston et al. (2009)
The Mountain car problem	

In the present context, active inference unpacks some of the implicit assumptions in Markov decision problems. In particular, it specifies explicitly what the agent knows about the effects of its actions. It does this through a sampling probability that replaces the observation probability of partially observable MDPs. As we will see, this means that probabilistic transitions among observations are conditioned upon action but, in contrast to the MDP formulation, probabilistic transitions among hidden states are not: action simply serves to realise posterior beliefs about state transitions. This conditioning of observations on action (as opposed to conditioning states on action) is not unrelated to treatments based on observer operator models and predictive representations of state: see Jaeger (2000) and Littman et al. (2002).

Definition The free-energy formulation refers to the tuple $(\Omega, S, A, \vartheta, P, Q, R)$ comprising:

- A finite set of observations Ω.
- A finite set of hidden states S.
- A finite set of actions A.
- Real valued parameters $\vartheta \in \mathbb{R}^d$.
- A *sampling probability* $R(o'|o,a) = \Pr(\{o_{t+1} = o'|o_t = o, a_t = a\})$ that observation $o' \in \Omega$ at time $t+1$ follows action $a \in A$, given observation $o' \in \Omega$ at time t.
- A *generative probability* $P(o,s,\theta|m) = \Pr(\{o_0,\ldots,o_t\} = o, \{s_0,\ldots,s_T\} = s, \vartheta = \theta)$ over observations to time t, states at all times and parameters
- A *recognition probability* $Q(s,\theta|\mu) = \Pr(\{s_0,\ldots,s_T\} = s, \vartheta = \theta)$ over states at all times and parameters with sufficient statistics $\mu \in \mathbb{R}^d$.

Here, m denotes the form of a generative model or probability $P_m(o,s,\theta) := P(o,s,\theta|m)$. For clarity, we will omit the conditioning on m for all but prior terms in the generative probability. The sufficient statistics of the recognition probability $Q_\mu(s,\theta) := Q(s,\theta|\mu)$ encode a probability distribution over a sequence of hidden states $s = \{s_0,\ldots,s_T\}$ and the parameters of the model $\theta \in \vartheta$. Crucially, the recognition probability and its sufficient statistics encode hidden states in the future and past, which themselves can change with time: for example, $\mu_k = \{\mu_0^k,\ldots,\mu_T^k\}$, where μ_t^k is the probability over hidden states at time t (in the future or past) under the recognition probability at time k.

Remark It should be noted that the definitions above do not describe a process—in the sense that the sampling and generative probabilities above would not be used to generate a sequence of observations. These probabilities underwrite the action and perception of the agent—they correspond to its formal beliefs about the sensory consequences of action (sampling probability) and the hidden states causing observations (generative probability). In other words, the true states generating observations are unknown and unknowable from the point of view of the agent. This is important when simulating active inference, where one has to make a careful distinction between the (true) states generating observations and those (hidden) states assumed by the agent. We will see an example of this later.

There are three important distinctions between this setup and that used by Markov decision processes. As in partially observed MDPs, there is a distinction between states and observations. However, the transition probability over states has been replaced by a sampling probability over observations. This means, we can formulate everything in terms of observed states (observations) and inference on hidden states. In other words, the agent does not need to know the effect of its actions on the (true) state of the world. It is instead equipped with a probabilistic mapping between its actions and direct sensory consequences—this is the sampling probability. This may sound a bit unusual but is a central tenet of active inference, which separates knowledge about the sensory consequences of action from beliefs about the causes of those consequences. In other words, the agent knows that if it moves it will sense movement (cf. proprioception); however, beliefs about hidden states in the world causing movement have to be inferred. These hidden states may or may not include its own action: the key distinction between the *agency-free* and *agency-based* schemes considered below depends on whether the agent represents its own action or not. Our previous illustrations of active inference

have been agency free; where (in a biological setting) action corresponds to classical motor reflexes, whose set point is determined by proprioceptive predictions. In this context, the sampling probability enables action (reflexes) to fulfil these predictions.

The second distinction is that we have introduced generative and recognition probabilities that are used to infer hidden states. Crucially, these hidden states include future and past states. In other words, the agent represents a sequence or trajectory over states, as opposed to just the current state. The generative probability is over the sequence of sensory states up until the current time, while the recognition probability changes with its time-dependent sufficient statistics. This means that the recognition distribution at any one time is over the sequence or trajectory of states at all times. This enables inference about a particular state in the future to change with time. This will become important later, when we consider planning and agency.

Finally, there are no reward or cost functions. This is an important point and illustrates the fact that active inference does not call upon the notion of reward to optimise behaviour—optimal behaviour minimises variational free energy, which is a functional of observations and the recognition probability distribution or its sufficient statistics. As we will see later, cost functions are replaced by priors over hidden states and transitions, such that costly states are surprising and are avoided by action.

3.2 Perception and action

The free-energy principle states that the sufficient statistics of the recognition probability and action minimise free energy

$$\begin{aligned}\mu_t &= \arg\min_{\mu} \mathcal{F}(\{o_0,\ldots,o_t\},\mu)\\ a_t &= \arg\min_{a} \sum_{\Omega} R(o_{t+1}|o_t,a)\mathcal{F}(\{o_0,\ldots,o_{t+1}\},\mu_t).\end{aligned} \tag{5}$$

This dual optimisation is usually portrayed in terms of perception and action, by associating the sufficient statistics with internal states of the agent (such as neuronal activity or connection strengths) and associating action with the state of effectors or the motor plant. Equation 5 just says that internal states minimise the free energy of currently observed states, while action selects the next observation that, on average, has the smallest free energy.

By factorising the generative probability $P(o,s,\theta|m) = P(o|s,\theta)P(s,\theta|m)$ into likelihood and prior probabilities, one can express the free energy as follows:

$$\begin{aligned}\mathcal{F}(o,\mu) &= \mathbf{E}_Q[-\ln P(o,s,\theta|m)] - \mathbf{E}_Q[-\ln Q(s,\theta|\mu)]\\ &= \mathbf{D}_{KL}[Q(s,\theta|\mu)||P(s,\theta|o)] - \ln P(o|m).\end{aligned} \tag{6}$$

The first equality in (6) expresses free energy as a Gibbs energy (expected under the recognition distribution) minus the entropy of the recognition distribution. The second shows that free energy is an upper bound on surprise, because the first (Kullback–Leibler divergence) term is non-negative by Gibbs inequality (Beal 2003). This also means that when free energy is minimised, the recognition distribution approximates the posterior distribution $Q(s,\theta|\mu) \approx P(s,\theta|o)$ over hidden states and parameters. This formalises the notion of unconscious inference in perception (Helmholtz 1866/1962; Dayan and Hinton 1997; Dayan et al. 1995) and, under some simplifying assumptions, corresponds to predictive coding (Rao and Ballard 1999).

The minimisation of free energy, with respect to action in (5) is active inference. This formulation highlights the fact that action selects observable states (not hidden states) that are the least surprising by virtue of having the smallest free energy. The free energy is determined by the sufficient statistics of the recognition distribution. The optimisation of these sufficient statistics—the first equality in (5)—rests upon the generative model and therefore depends on prior beliefs. It is these that specify what is surprising and reproduces the optimal policies considered above. There are clearly many ways to specify the generative probability. We will consider two forms, both of which respect the Markov property of decision processes. The first reproduces the behaviour under the optimal policy for Markov decision problems and can be regarded as the corresponding free-energy formulation:

3.3 An agency-free formulation of optimal policies

The natural generative model for a partially observable Markov decision process can be expressed in terms of a likelihood plus priors over states and parameters, with the following forms:

$$\begin{aligned}P(o,s,\theta|m) &= P(o|s,\theta)P(s|\theta)P(\theta|m)\\ P(\{o_0,\ldots,o_t\}|s,\theta) &= P(o_0|s_0)P(o_1|s_1)\ldots P(o_t|s_t)\\ P(s|\theta) &= P(s_0|m)\prod_{t=0}^{T-1} P(s_{t+1}|s_t,\theta)\end{aligned} \tag{7}$$

This implies that the current observation depends only on the current hidden state (like a belief MDP), where the hidden states are a Markov process, whose transition probabilities depend upon the parameters (unlike a belief MDP). We will assume that the priors over the parameters $P(\theta|m) = \delta(\theta - \theta_\pi)$ make the priors over state transitions equivalent to the optimal policy of the previous section. In other words, we assume the priors have a point mass over values that render the transition probabilities $P(s_{t+1}|s_t,\theta_\pi) = \mathbf{T}(s_{t+1}|s_t,\pi(s_t))$ optimal in the conventional sense (were the transition probability is defined for Markov decision

processes above and $\pi(s_t)$ corresponds to action under an optimal policy).

The second equality in (6) shows that minimising the free energy, with respect to the sufficient statistics of the recognition distribution, renders it the posterior over hidden states and parameters. This means that the recognition distribution becomes the posterior distribution, where (noting that the posterior and prior over parameters are the same Dirac delta function)

$$Q(s,\theta|\mu_t) \approx P(\{s_0,\ldots,s_T\}|\{o_0,\ldots,o_t\},\theta)\delta(\theta-\theta_\pi). \tag{8}$$

We have used an approximate equality here because we are assuming approximate Bayesian inference. In this context, free-energy minimisation with respect to action becomes, from (5) and (6):

$$\begin{aligned} a_t &= \arg\min_a \sum_\Omega R(o_{t+1}|o_t,a)\mathcal{F}(\{o_0,\ldots,o_{t+1}\},\mu_t) \\ &= \arg\max_a \sum_\Omega R(o_{t+1}|o_t,a)\mathbf{E}_{Q(s_{t+1})}[\ln P(o_{t+1}|s_{t+1})] \\ Q(s_{t+1}) &\approx \sum_S P(s_{t+1}|s_t,\pi(s_t))P(s_t|\{o_0,\ldots,o_t\}) \end{aligned} \tag{9}$$

Note that the free energy of the new observation is just its improbability, expected under posterior beliefs about the hidden states that cause it—these posterior beliefs correspond to the marginal recognition distribution $Q(s_{t+1})$, over the next hidden state.

It can be seen from (9) that action under active inference is exactly the same as action under the optimal policy. This is because action selects the observation that is most likely under the (approximate) posterior distribution. In turn, this is the hidden state that follows the currently inferred state, under the optimal policy. This means that active inference can be considered as a generalisation of optimal control. This is because there are prior beliefs that can reproduce an optimal policy to minimise expected cost. However, there are many other prior beliefs that specify Bayes-optimal control that do not minimise expected cost—see the handwriting simulations in Friston et al. (2011) or the animate behaviours in Tani (2003).

3.4 Optimality and complete class theorems

The fact that one can replace cost functions with priors to produce the same behaviour is related to the complete class theorem (Brown 1981). The complete class theorem states that any admissible decision rule (behaviour) is Bayes-optimal for at least one pair of prior beliefs and cost function (Robert 1992). However, this pair is not necessarily unique: in other words, the same decisions can be reproduced under different combinations of prior and cost functions. In one sense, this duality is resolved by replacing the cost functions of optimal control theory with prior beliefs about state transitions. Casting Bayes-optimal decisions in this way simply means that the agent believes it will move through state space in a way that minimises future costs, while action fulfils these prior beliefs.

Clearly, this does not address the problem of how policies are learned; however, it shows how active inference can be used to implement an optimal policy. From the point of view of active inference, it would be perfectly possible to use solutions of the appropriate Bellman optimality equation (i.e., reinforcement learning) to create controlled environments that enable agents to learn optimal policies: however, value functions *per se* are not learned under active inference; it is the parameters of the prior distributions that are learned. Conversely in reinforcement learning, it is sufficient to learn value functions without having to learn transition probabilities: see (3). It is in this sense that reinforcement learning is referred to as model free: see Dayan and Daw (2008) and Gläscher et al. (2010).

An example of learning policies through priors was presented in Friston et al. (2009), where an agent was immersed in a controlled environment that enforced optimal trajectories through state space. In this example, the trajectories were optimised by minimising the Kullback–Leibler divergence between the ergodic (invariant) probability density function associated with state transitions and a density that minimised the expected terminal cost. The ergodic density was the solution to the appropriate Fokker–Planck or Kolmogorov forward equation (a differential equation describing the evolution of a system's ensemble density). The agent then learned the optimal policy by minimising variational free energy, with respect to posterior beliefs about parameters encoding transition probabilities. In this example, all the heavy lifting was done prior to learning, in the creation of the controlled environment—all the agent had to do was learn optimal state transitions using standard Bayesian learning. In the next section, we consider how agents *infer* the optimal policy online, as opposed to *learning* optimal prior beliefs about state transitions.

4 Bayes-optimal control without cost functions

In this section, we consider agency based optimisation, in which the hidden states are extended to include hidden control states. This is necessary, when inferring optimal state transitions, because transitions depend upon action in the future which is, by definition, hidden from observation. In what follows, we focus on policies that are specified by prior beliefs about specific states that will be occupied at specific times in the future. This corresponds to a finite horizon control problem with terminal costs over states and intermediate

control costs that are specified through prior beliefs about control. Our special focus here is the implication for the timing of optimisation processes—given that real agents have to rehearse their future options before selecting an action.

4.1 Agency-based optimisation

In what follows, we describe a scheme for partially observable Markov decision processes that optimises action in relation to prior beliefs about future states. This scheme uses representations of hidden states in the future to optimise a sequence of fictive actions (policy) before they are enacted. Clearly, this requires the agent to infer (future) actions, which calls for a more sophisticated generative model—a model of agency or control. In other words, the agent must represent its future actions. This leads to Bayesian updates of posterior beliefs about future states that include control. Note that this is not equivalent to solving the optimal control problem at each point in time; because the Bayesian updates are themselves a Markovian process, in which posterior beliefs about future states depend on the corresponding beliefs at the preceding time point. This dependency is exploited to update posterior beliefs *about the future* that are held at the current time.

The heuristic benefit of introducing hidden control states is that putative actions in the future can be optimised, when choosing the best current action. The ensuing solutions are optimal in relation to prior beliefs about states that will be occupied. These are prior beliefs about the final (desired) hidden state and can be expressed in terms of the following generative model:

An agency-based model The generative probability used in this section introduces (a finite set of) control states $u \in U$ and can be expressed in terms of the following likelihood and prior distributions:

$$\begin{aligned} P(o, s, u, \theta|m) &= P(o|s, \theta)P(s, u|\theta)P(\theta|m) \\ P(\{o_0, \ldots, o_t\}|s, \theta) &= P(o_0|s_0, \theta)P(o_1|s_1, \theta)\ldots P(o_t|s_t, \theta) \\ P(s, u|\theta) &= P(s_T|\theta)\prod_{t=1}^{T} P(s_{t-1}|s_t, u_t, \theta)P(u_t|\theta) \end{aligned} \tag{10}$$

Remark There are two important aspects of this generative model: first, control states are not action; they are an internal representation of action that may or may not be related to actions emitted by the agent. In the generative model, control states affect the transitions among hidden states; in other words, they only affect outcomes vicariously through changes in hidden states. It is these control states that represent agency, which may or may not be a veridical representation of what the agent can actually do (or is doing)—in this sense, they can be regarded as fictive action that gives the generative model extra degrees of freedom to model state transitions under prior beliefs. Recall that action only changes observations and is selected on the basis of posterior beliefs about the next observable state. Conversely, control states are modelled as hidden states over time and are inferred. This means they only exist in the mind (posterior beliefs) of the agent.

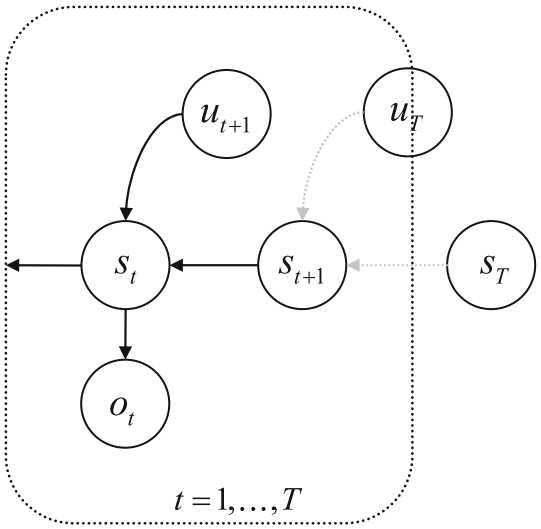


Fig. 1 Probabilistic graphical model illustrating the Markovian dependencies among hidden states generating sensory data. These hidden states (s_t, u_t) are represented explicitly, over all time points: $t = 1, \ldots, T$. This means there is a representation of the past and future that includes hidden states mediating control. Note that the dependency of this hidden Markov model runs backwards in time so that all preceding hidden states are conditioned recursively on the final or terminal goal state

Second, the priors on the hidden states $P(s, u|\theta)$ are formulated in a pullback sense, that is, they run backwards in time. This preserves the Markov dependencies but allows us to specify the prior over a sequence of states in terms of transition probabilities and a prior distribution over the final (terminal) state. Put simply, the parameters of the (transition) model encode where I came from, not where I am going. See Fig. 1. This particular form of prior belief is chosen for convenience, because it accommodates beliefs about the desired final state—of the sort that would be specified with a terminal cost function, $\mathbf{r}(s_T)$.

The generative model in (10) is fairly general and makes no specific assumptions about the implicit cost of inferred control (e.g., it does not assume quadratic control costs) or allowable state transitions. In what follows, we will illustrate inference or model inversion using a particular parameterisation and variational inversion scheme. This example is used to illustrate agency based inference, accepting that there are many different model parameterisations and inversion schemes that could have been used.

Generative probability This model comprises the following likelihood and prior distributions:

$$\begin{aligned} P(o_t|s_t,\theta) &= \mathbf{A}\cdot s_t \\ P(s_{t-1}|s_t,u_t,\theta) &= (\textstyle\prod_i \mathbf{B}_i^{u_{ti}})\cdot s_t \\ P(s_T|\theta) &= \mathbf{c} \\ P(u_t|\theta) &= \textstyle\prod_i \mathbf{d}_i^{u_{ti}} \end{aligned} \tag{11}$$

The parameters $\theta = \{\mathbf{A}, \mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{c}, \mathbf{d}\}$ of this model are

$$\begin{aligned} \mathbf{A} &= \{a_{ij}\} : \sum_j a_{ij} = 1, \quad \forall i \\ \mathbf{B}_k &= \{b_{ijk}\} : \sum_j b_{ijk} = 1, \quad \forall i,k \\ \mathbf{c} &= \{c_i\} : \sum_i c_i = 1 \\ \mathbf{d} &= \{d_i\} : \sum_i d_i = 1 \end{aligned} \tag{12}$$

In this particular model, the parameters in the matrices $\mathbf{B}_k$ encode transition probabilities among hidden states that are engaged when the control state $u_k = 1$, where the control states have a multinomial distribution—only one can be 'on' at any time. The hidden states cause observed states through the mapping encoded by $\mathbf{A}$. The vectors $\mathbf{c}$ and $\mathbf{d}$ encode the prior distribution over the final hidden state and control states respectively; these specify the goal and prior costs on control.

Recognition probability To exploit the Markovian form of the generative model we will use an efficient approximate inference scheme afforded by variational Bayesian learning (Beal 2003); for a tutorial see Fox and Roberts (2011). The efficiency rests on replacing posterior dependencies among hidden states (over time) with mean field effects on the marginal probabilities at each time point. This is achieved using the following *mean-field assumption* for the recognition distribution:

$$\begin{aligned} Q(s,u) &= \prod_{t=1}^{T} Q(s_t)Q(u_t) \\ Q(s_t|\alpha_t) &= \prod_i \alpha_{ti}^{s_i} : \sum_i \alpha_{ti} = 1 \\ Q(u_t|\beta_t) &= \prod_i \beta_{ti}^{u_i} : \sum_i \beta_{ti} = 1 \end{aligned} \tag{13}$$

Standard variational Bayesian learning now provides a recipe for optimising the sufficient statistics (α_t, β_t) of the recognition probability over hidden and control states in a series of variational updates. It is fairly straightforward to show that the marginal recognition distributions that minimise free energy can be expressed in terms of *variational energies* $(I(s_t), I(u_t))$, where

$$\begin{aligned} Q(s_t|\alpha_t) &\propto \exp(I(s_t)) \\ Q(u_t|\beta_t) &\propto \exp(I(u_t)). \end{aligned} \tag{14}$$

The variational energies for the hidden and control states at each time are just the (negative) Gibbs energies in (6) expected under the Markov blanket of each state.

$$\begin{aligned} I(s_t) &= \mathbf{E}_{Q(s_{t-1})Q(s_{t+1})Q(u_t)Q(u_{t+1})}[\ln P(o_t|s_t) \\ &\quad + \ln P(s_t|s_{t+1},u_{t+1}) + \ln P(s_{t-1}|s_t,u_t)] \\ &= [\ln \mathbf{A}^T\cdot o_t] + \sum_j \beta_{(t+1)j}\ln \mathbf{B}_j\cdot\alpha_{t+1} \\ &\quad + \sum_j \beta_{tj}\ln \mathbf{B}_j^T\cdot\alpha_{t-1} \end{aligned} \tag{15}$$

$$\begin{aligned} I(u_{ti}) &= \mathbf{E}_{Q(s_{t-1})Q(s_t)}[\ln P(s_{t-1}|s_t,u_t) + \ln P(u_t|m)] \\ &= \alpha_{t-1}^T\cdot\ln \mathbf{B}_i\cdot\alpha_t + \ln d_i \end{aligned}$$

Here, the square brackets in $[\ln \mathbf{A}^T\cdot o_t]$ indicate that this term is only used when observations are available. This highlights an important aspect of the update scheme; namely, the hidden states at all points during the sequence are updated iteratively at each time point. Although, hidden states in the future are not informed by concurrent sensory information, they are still constrained by prior beliefs about future states. The efficiency of this scheme rests on the fact that the Markov blanket of any state is limited to immediately preceding and past states. This simplicity is due to the way Markov decision problems are set up (see Fig. 1). The ensuing variational updates for the sufficient statistics $\mu_k = \{\alpha_0^k, \ldots, \alpha_T^k, \beta_0^k, \ldots, \beta_T^k\}$ at successive times k are:

$$\begin{aligned} &\text{for}\quad k = 1 \text{ to } T \\ &\text{until}\cdot\text{convergence}: \\ &\quad\text{for}\quad t = (T-1) \text{ to } (k+1) \\ &\qquad \alpha'_t = \exp([\ln \mathbf{A}^T\cdot o_t] \\ &\qquad\quad + \sum_j \beta^k_{(t+1)j}\ln \mathbf{B}_j\cdot\alpha^k_{(t+1)} \\ &\qquad\quad + \sum_j \beta^k_{tj}\ln \mathbf{B}_j^T\cdot\alpha^k_{(t-1)}) \\ &\qquad \alpha_t^{k+1} = \frac{\alpha'_t}{\sum_i \alpha'_{ti}} \\ &\qquad \beta'_{ti} = \exp(\alpha_{t-1}^{kT}\cdot\ln \mathbf{B}_i\cdot\alpha_t^k + \ln d_i) \\ &\qquad \beta_t^{k+1} = \frac{\beta'_t}{\sum_i \beta'_{ti}} \end{aligned} \tag{16}$$

Remark First, note the normalisation of the sufficient statistics in (16). This normalisation is necessary because of the implicit normalisation term (partition function) in (14). This normalisation is a just a re-scaling (all the hard work is done in computing the variational energies). The interesting aspect of these updates is their nested structure over time. At each time point, the variational updates cycle over representations of future states to update the sufficient statistics encoding posterior beliefs. In (16), this cycling continues until convergence, although a fixed (small) number of cycles usually suffice (see next section). Furthermore, the order of the updates can be from the future to the past (cf., backwards induction) as shown above, from the past to the future or both (cf., forward-backward schemes). These update cycles are themselves repeated as time progresses, so that there is convergence both within and between cycles. This means the sufficient statistics change over two timescales; a fast timescale

that updates posterior beliefs about the future and a slow timescale that updates posterior beliefs in the future. Posterior beliefs about the trajectory, at both timescales, ensure that the trajectory convergences on the final (desired) location, where the anticipated trajectory (and control) is realised through action. Anticipated control corresponds to posterior beliefs about future control states that we associate with a sense of agency. In this nested updating, fluctuations or perturbations to the anticipated trajectory are accommodated easily by the implicit online updating.

It is interesting to speculate about neurophysiologic implementations of this sort of scheme, particularly in relation to nested electrophysiological oscillations (Canolty et al. 2006). The notion here is that the electrophysiological correlates of updating may show nested oscillations, with fast (gamma) oscillations reflecting updates in a fictive future and slower (theta) dynamics that reflect updates in real time, with timescales of 25 and 250 ms, respectively.

The treatment above assumes that the parameters of the transition probabilities under different controls are known. If they are not, then it is relatively straightforward to extend the variational Bayesian scheme above to include variational updates for unknown parameters, as described in Chapter 3 of (Beal 2003). The only special consideration here is the use of conjugate (Dirichlet) priors over the parameters, $\theta \supset \{\mathbf{A}, \mathbf{B}_1, \mathbf{B}_2, \ldots\}$ to ensure the recognition distributions retain their multinomial form.

In summary, this section has introduced policy optimisation in terms of active inference that realises prior beliefs about a desired future state. In contrast to agency free models, the desired state can be changed without relearning an optimal policy. This endows the scheme with a context sensitivity that may be important in hierarchical generative models, in which final states are themselves specified by trajectories over slower timescales (Kiebel et al. 2009a). In short, even if prior beliefs about states change, posterior beliefs about the parameters do not and they can be used to access a new goal: this is the essence of agency based control illustrated in the next section:

5 Simulations: the mountain car problem

In the section, we apply the scheme of the previous section to a well known problem in optimal control theory that presents some special challenges: The mountain car problem can be envisaged as follows; one has to park a mountain car halfway up the side of a valley. However, the mountain car is not strong enough to climb directly to the parking place, which means the only way to access the goal is to ascend the other side of the valley to acquire sufficient momentum during the return trip. This represents an interesting problem, when considered in the state space of position and velocity: the agent has to move away from its target location to attain the goal later. In other words, it has to execute a circuitous trajectory through state space (as in avoiding obstacles). We have used this problem previously to illustrate how Bayes-optimal control can be learned in terms of the parameters controlling prior beliefs about trajectories (Friston et al. 2009) and using heuristic policies (Gigerenzer and Gaissmaier 2011) based on the destruction of costly fixed point attractors (Friston 2010).

It should be noted that the mountain car problem is normally cast as a learning problem—in which an optimal policy has to be learned. However, here, we are using it to illustrate optimal behaviour in terms of inference. In other words, we assume the agent has already learned the constraints afforded by the world it operates in—and now has to infer an optimal policy within a single trial. In this setting, the mountain car problem provides a challenging (non-linear) inference problem, particularly when we include random fluctuations in both the states generating observations and the observations themselves. The mountain car problem can be specified with the equations of motion in Fig. 2. Here, we consider a discrete state space and time formulation of this problem and use it to illustrate agency based control.

5.1 Simulation setup

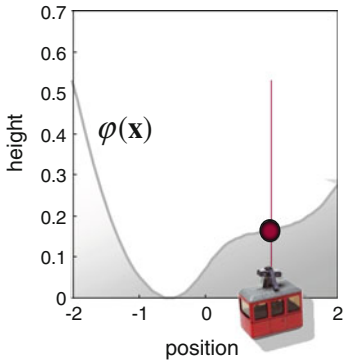
To create a discrete version of this problem, we ensured that expected changes in position and velocity match the equations of motion, when integrated over discrete time intervals (here $\Delta t = 2$ s). The ensuing pullback probabilities for each level of control satisfy (subject to the constraint that only the states adjacent to the expected position and velocity are nonzero)

$$\sum_i \mathbf{x}(s_i) B_{ijk} = \mathbf{x}(s_j) - f(\mathbf{x}(s_j), a(u_k))\Delta t \tag{17}$$

In practice, we actually compute the equivalent forward transition probabilities required for the sampling probability and then normalise their transpose to compute the pullback probability matrices. Here, $\mathbf{x}(s_i) \in \mathbb{R}^2$ returns the continuous position and velocity associated with the i-th hidden state s_i. Similarly, $a(u_k) \in \mathbb{R}$ returns the real valued action associated with the k-th control state u_k. In these simulations, we used five levels of control corresponding to $a(u_k) \in \{-2, -1, 0, 1, 2\}$. This means the agent assumes that strong or intermediate acceleration can be applied in a right or leftward direction.

To simulate random fluctuations in the motion of the mountain car, we convolved the parameter matrix encoding (pullback) probabilities over position and velocity with the kernel: $[\frac{1}{4}, \frac{1}{2}, \frac{1}{4}]$. This just smears the probabilities to augment the uncertainty about the previous states incurred by

$$\varphi(\mathbf{x}) = \begin{cases} -2\mathbf{x} - 1 & : \mathbf{x} \leq 0 \\ -\mathbf{x}^2(1 + 5\mathbf{x}^2)^{-3/2} - \frac{1}{8}\mathbf{x}^4 & : \mathbf{x} > 0 \end{cases}$$



Equations of motion

$$f(\mathbf{x}) = \begin{bmatrix} \dot{\mathbf{x}} \\ \dot{\mathbf{x}}' \end{bmatrix} = \begin{bmatrix} \mathbf{x}' \\ \sigma(a) - \partial_x \varphi - \frac{1}{8}\mathbf{x}' \end{bmatrix}$$

Fig. 2 Schematic of the mountain car problem: the *upper panel* (and associated equations) illustrate the landscape or potential energy function that defines the motion of the car. This has a minima at $\mathbf{x} = -0.5$. The mountain-car is shown at the desired parking position at the top of the hill on the right $\mathbf{x} = 1$ (indicated with a *red ball*). The equations of motion in the *lower panel* describe the forces exerted on the car, which include $\sigma(a)$, a sigmoid (hyperbolic tangent) function of action, gravitational forces and friction.

discretising state space. The state space comprised 32 position (from -2 to 2) and velocity bins (from -3 to 3), giving $32 \times 23 = 1,024$ discrete states.

The resulting parameters are illustrated in Fig. 3 in terms of the probability distributions four and eight time steps before the final position, $\mathbf{x} = (1, 0)$ in this example. The top row is for a control state that pushes the car to the left. This means that preceding states are more likely to be further up the opposite hill, so that the car is accelerating with greater velocity to the final position (the horizontal axis corresponds to position and the vertical axis to velocity). One can see the expected differences after a few time steps, reflecting the different directions in which control forces are applied. After eight time steps, the distributions become increasingly dispersed due to the uncertainty introduced by smoothing the transition probabilities and the fact that motion and velocity are encoded with finite sized bins.

For simplicity, we assumed a one-to-one mapping between hidden and observed states; that is $\mathbf{A} = I$ and placed uniform prior costs over control, such that $\mathbf{d} = [\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}]$. Prior beliefs about the final state specify the goal $\mathbf{x}(s_i : i = \chi(\mathbf{c})) = (1, 0)$; namely, to maintain a position at the parking location with zero velocity (see Fig. 2). Finally, the action-dependent sampling probabilities $R(o_{t+1}|o_t, a_t)$ were the transposed versions of the pullback probabilities in (17). These sampling probabilities were used to select action and to generate the next sensory input. A subtle point here is that we do not need to refer to, or enumerate, real states in the world; everything is cast in terms of quantities that the agent can access. Action used the same five levels as the control states. However, as noted above, there is no requirement that action and control be related formally in this way.

Figure 4 shows the results of a simulation using $T = 16$ time steps and a starting position of $\mathbf{x} = (0, 0)$. In these, and all subsequent simulations, the variational updates were repeated eight times and then an action was selected. This is a fairly arbitrary number of variational cycles but sufficient for the current problem (in which the updates generally converged after a couple of iterations). We chose eight cycles in anticipation of future simulations using continuous time formulations of the scheme in this paper. These simulations produce dynamics that are not dissimilar to the nested oscillations seen in theta-gamma phase coupling between the prefrontal cortex and hippocampal system (Canolty et al. 2006; Axmacher et al. 2010), where there are about eight gamma cycles for each theta cycle.

The upper panel of Fig. 4 shows the trajectories (real and anticipated) through state space, while the lower panels show the inferred control states and action selected as a function of time. The darker line in the upper panel connects the states visited over the 16 time steps, while the grey lines report the anticipated trajectories from the beginning of the trial to the end. The inferred trajectories are shown as the expected position and velocity, based on posterior beliefs over discrete states. One can see that the actual trajectory fulfils, fairly faithfully, the anticipated sequences and that there has been relatively little updating during execution. As anticipated, the mountain car moves away from its target to acquire sufficient momentum to access the goal on the right. Note the similarity between the selected actions (right) and the inferred control states (left). The interesting thing here is that the agent was not always sure about which control state was currently engaged. However, the control state with the highest posterior probability, which corresponds to the action the agent believes it will emit next, is always selected by active inference. In other words, even under uncertainty about hidden and control states, there is sufficient confidence in the next sensory state to inform action.

One can obtain similar results with different combinations of starting and final states (provided the trajectories can be realised): Fig. 5 shows the same results as in Fig. 4 but with a prior that compelled the agent to pass through the parking location with a leftward velocity of one: $\mathbf{x}(s_i : i = \chi(\mathbf{c})) = (1, -1)$. The agent solves this prob-

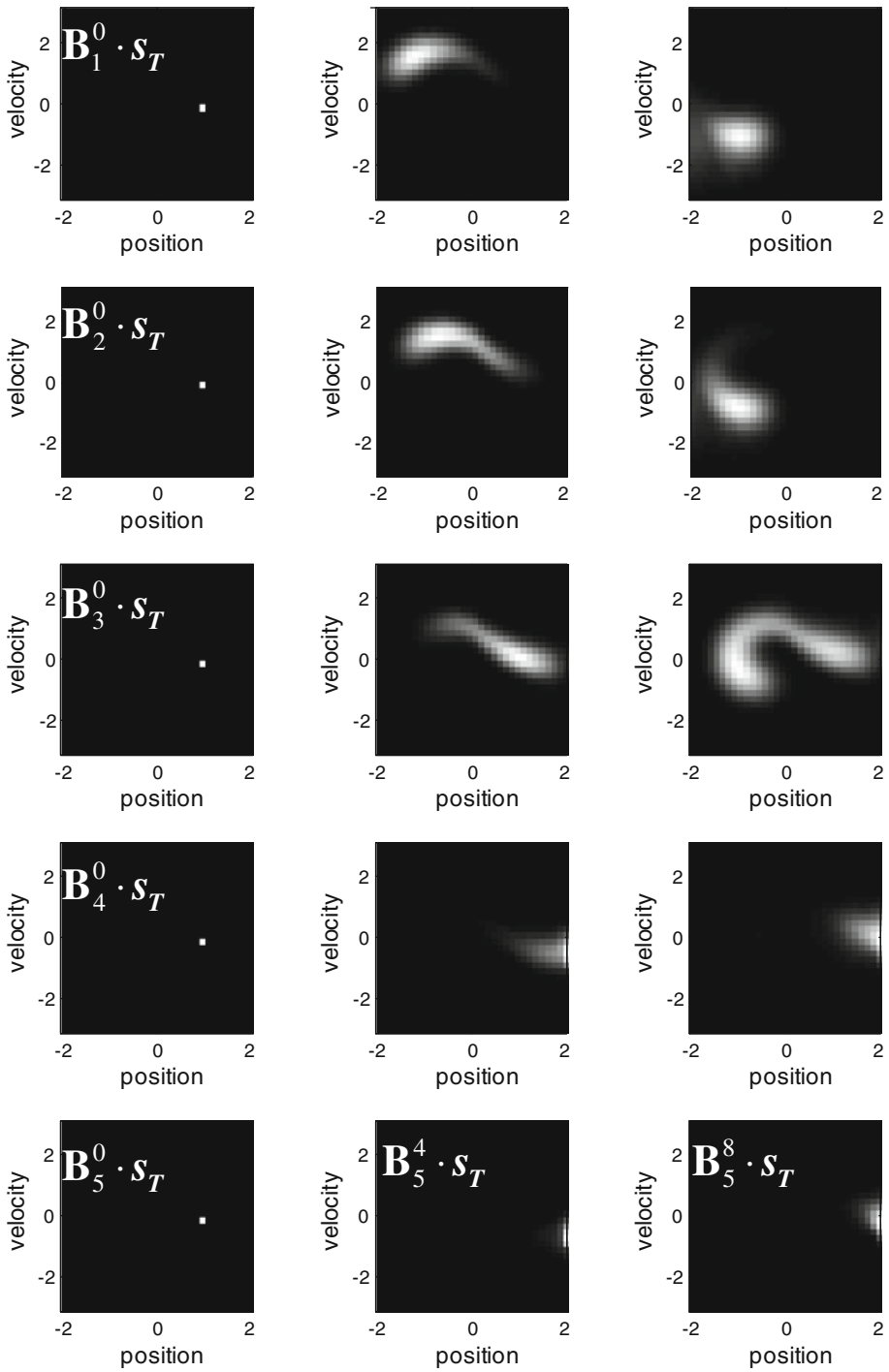


Fig. 3 Illustration of the transitional probabilities describing the discrete formulation of the mountain car problem in the previous figure. These correspond to the probability of occupying a hidden state given a particular final state (*first column*), after four time steps (*second column*) and after eight time steps into the past. The probability distribution functions are shown in image format, over position and velocity. The *top row* shows the transition probabilities for a control that applies leftward forces to the mountain car, the *middle row* shows the corresponding pullback probabilities for a control that exerts no forces and the *last row* shows transition probabilities for control states that accelerate the car towards the right. Note the dispersion of these probability distributions over time, due to the discretisation of the dynamics and random fluctuations on the speed and position

lem by rushing past the goal and then allowing itself to fall downhill, so that it approaches the final destination with the desired velocity. Note that it has to accelerate slightly at the final approach (with a slight leftward acceleration as indicated in the lower panel). As noted above, changing prior beliefs in this way does not require any further learning and could even be accommodated during the execution of a trajectory. This may become important in versions of this scheme that go beyond finite horizon problems. In this context, agents maintain posterior beliefs about a fixed number of hidden states into the future. In other words, they represent a continually evolving trajectory. We will pursue this elsewhere.

Finally, Fig. 6 reproduces the simulation in Fig. 4 but after increasing the number of times steps from $T = 16$ to $T = 32$. This illustrates the difference between trajectories optimised under optimal control theory and Bayes-optimal inference under priors that specify which states will be

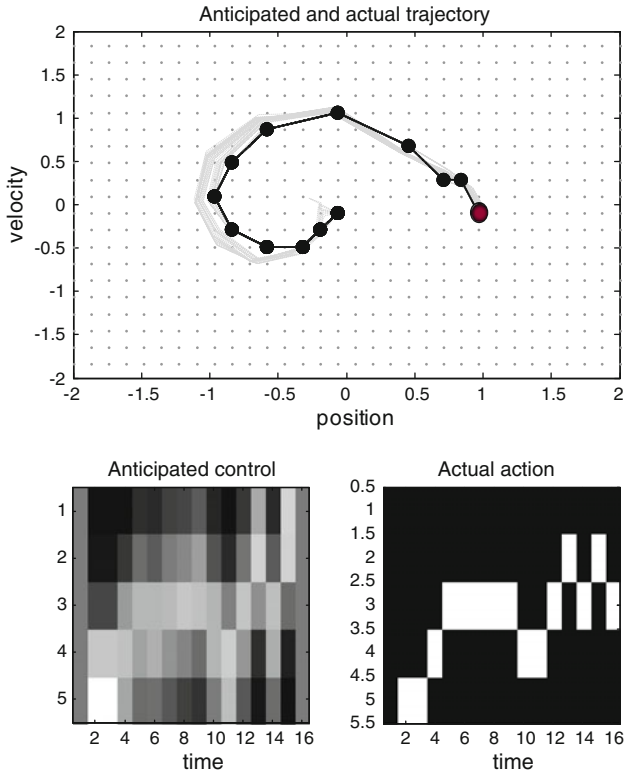


Fig. 4 This figure shows the results of a simulated (agency based) trajectory over $T = 16$ time steps starting at $\mathbf{x} = (0, 0)$ and ending at the goal location $\mathbf{x} = (1, 0)$ (*red ball*) using active inference and explicit representations of the future. The *upper panel* shows the trajectories in the state space of position and velocity. The *grey lines* represent anticipated trajectories accumulated during control, while the *dark* (*dotted*) *lines* show the actual trajectory through state space. The anticipated trajectories are the expected values based upon posterior expectations about past and future states. They are therefore continuous functions of position and velocity. In contrast, the actual trajectory is restricted to the 1,024 discrete states that can be occupied; these are shown as *light grey dots*. The lower panels show the anticipated control and the actual actions selected under active inference (in image format where lighter colours mean a higher probability). Note that there is a high degree of correspondence; however, the posterior beliefs about control are not always absolutely certain: these are the beliefs at the times each action is selected

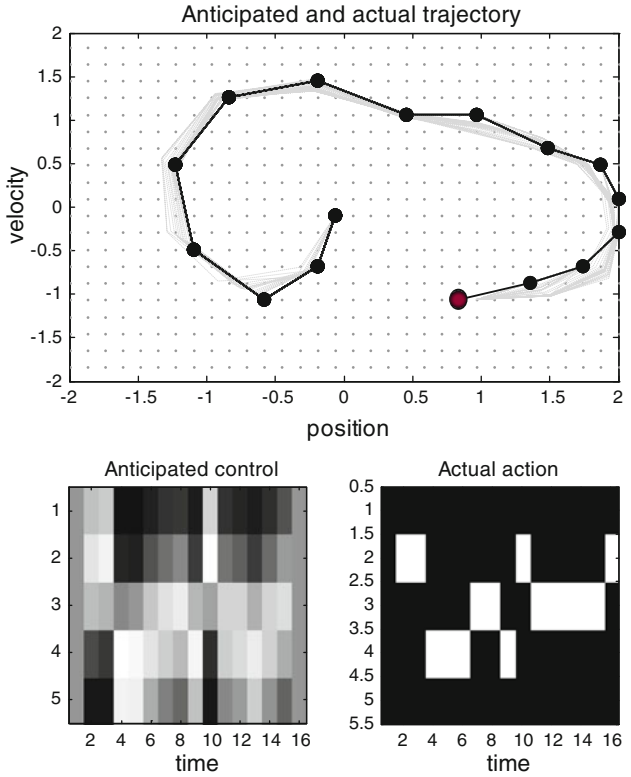


Fig. 5 This figure uses the same format as the previous figure. The only difference here is that we have changed the location of the goal to $\mathbf{x} = (1, -1)$. In other words, we require the mountain car to pass through the goal location from the right, with unit speed. It can be seen that this goal has been obtained, with reasonable accuracy

occupied *and when*. It can be seen from the upper panel of Fig. 6 that the agent spends the first half of the trajectory at the bottom of the valley before ascending to the target location. This trajectory minimises free energy; in other words, it is the least surprising under the agent's posterior beliefs. This example illustrates a key distinction between policies that are optimal in relation to prior expectations about future states and those that maximise expected reward. The latter, guided by value functions, preclude behaviours that access particular states at particular times (unless the value function changes with time). In other words, once an optimal policy has been defined in terms of rewards associated with particular states; there is no opportunity to specify any constraints on when those rewards are accessed. Although this is not particularly important from an engineering perspective, it becomes acute in biological systems with hidden states that evolve over multiple timescales. A simple example here would be the context and time sensitive nature of motivational and physiological drives (Berridge 2004) that render the same states differentially rewarding, depending upon when they were last visited (for example, eating or drinking). In continuous time formulations of active inference, this dependency is usually dealt with in the context of dynamical systems theory, where desired states become attracting sets, which can either be fixed point attractors or more complicated trajectories on attractor manifolds (Friston and Ao 2012). In this setting, the representation of a trajectory over successive states $(s_0, s_1, s_2, \ldots)$ is replaced by representations in generalised coordinates of motion $(s, s', s'', \ldots)$; see Friston (2008) for details.

The trajectory in Fig. 6 is the least surprising given the agent's prior beliefs. These include prior beliefs about control which, in this example, were uninformative. A different trajectory would emerge if, for example, we made acceleration and deceleration more unlikely than doing nothing; e.g., $\mathbf{d} = [\frac{1}{8}, \frac{1}{8}, \frac{1}{2}, \frac{1}{8}, \frac{1}{8}]$ (results not shown). From the perspective of optimal control theory, this is equivalent to prescribing

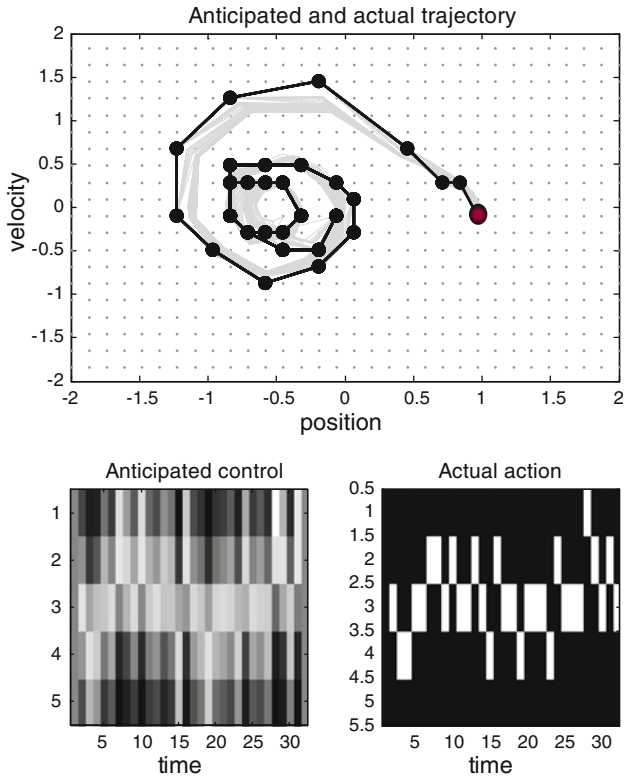


Fig. 6 This figure reports the same results as shown in Fig. 4; however, here we have increased the number of time steps from $T = 16$ to $T = 32$. The optimal (in a Bayesian sense) trajectory now spends its first few steps in the valley before ascending to reach the goal location; $\mathbf{x} = (1, 0)$

differential costs for action that, in the variational free-energy formulation, are absorbed into priors.

In closing, it is worth noting that the scheme above is not robust. This is meant in the sense that it is based upon approximate inference (as opposed to exact inference) that necessarily follows from the mean field approximation that renders the numerics tractable. Furthermore, because the probability distributions associated with Markov decision problems are not convex, the coordinate descent on free energy in (14) is not guaranteed to converge to its global minimum. This means that the example in this section should be taken as a proof of principle that the scheme can work; noting that it took several hours to configure the number of time steps and discretisation of hidden states before the variational updates gave sensible results. Having said this, the failures were often more interesting than the successes reported above (see discussion). Furthermore, the combinatorics the scheme can contend with is enormous. For example, in the final simulation the scheme was able to select a sequence of actions or decisions from among $5^{32} = 2.32 \times 10^{22}$ alternatives. This capacity to deal with high degrees of computational complexity, with approximating assumptions, may help when thinking about how real world agents deal with equivalent decision making problems. The routine Matlab (**spm_MDP.m**) used to generate Figs. 3 through to 3 is available in the DEM Toolbox of the SPM academic freeware (http://www.fil.ion.ucl.ac.uk/spm).

6 Discussion

In summary, we have reviewed classical approaches to (partially observable) Markov decision problems and have recast reward or cost functions in terms of prior beliefs about state transitions. This implicitly resolves the redundancy between cost functions and priors that underlies the complete class theorem. We then went on to exploit this redundancy by specifying optimal policies in terms of prior beliefs about future (terminal) states. The ensuing scheme may provide a metaphor for model based decision making in real agents that has an explicit planning or anticipatory aspect. This solution was based upon approximate (variational) Bayesian inference that respects the Markov nature of decision processes.

The aim of this work was to unpack some of the implications of optimal control for its implementation in real-world (biological) agents. The most important is the representation of hidden control states that are required for accessing distal rewards in the future. This contrasts with the usual problem formulation of MDPs, which is to define a process model and the corresponding notion of optimality, without reference to the internal (representational) states of the agent. Our aim was not to finesse computational problems from a machine learning perspective. Indeed, it is well known that the computational complexity of a problem is not changed when reducing it to an inference problem: see (Littman et al. 2001) for a treatment of this in the setting of stochastic satisfiability problems and probabilistic inference. The equivalence in computational complexity is reflected in the fact that many procedures are found in both approximate solutions to optimal control and Bayesian inference. Examples here include minimisation of Kullback–Leibler divergences (Todorov 2008; Kappen et al. 2009), and expectation maximisation (Toussaint and Storkey 2006), both of which can be formulated as minimising variational free energy (Neal and Hinton 1998). The main contribution of this paper concerns the interpretation of optimality, as opposed to an algorithmic contribution. Having said that, there is a subtle but fundamental difference between classical optimal control and active inference:

6.1 Optimal control and active inference

One could consider optimal control as a special case of active inference. This is because specifying optimal policies directly—in terms of prior beliefs about state transitions—affords a complete specification of a policy. In optimal

control theory, state transitions are specified in terms of value functions that are solutions to the appropriate Bellman optimality equations, given a cost function. The notion that the Bellman optimality principle "can be derived as a limit case" from the variational principles that underlie active inference also emerges in recent information theoretic formulations of bounded rationality (Braun et al. 2011): Braun et al consider control costs in terms of the (cross) entropy of choice probabilities and augment expected utility to produce a free-energy optimality criterion. This free-energy functional (*free utility*) captures bounded rationality by ensuring the divergence between optimal and prior choice probabilities is minimised. As in the current treatment, the generality of this approach "relies on the fact that ultimately any real agent has to be incarnated in a physical system, and the process of information processing must always be accompanied by a pertinent physical process". They show that minimising free utility includes both discrete and continuous stochastic optimal control as special cases and, crucially, can be derived "without invoking the Hamilton–Jacobi–Bellman equation or the Bellman optimality equations". Their treatment of stochastic optimal control uses the Feynman–Kac formula to express the control problem in terms of a Chapman–Kolmogorov equation or, when just considering terminal cost, a Kolmogorov backward equation. They then show that this is the solution to the Hamilton–Jacobi–Bellman equation, under quadratic control costs. See also (Theodorou et al. 2010), who exploit the same formalism but with a more classical motivation. The generalisation of classical optimal control using free utility is compelling and unifies approximate optimal control methods in both the continuous and discrete domain. However, this use of free utility is fundamentally different from the variational free-energy minimisation implied by the free-energy principle and active inference:

Free utility is a functional of choice probabilities over hidden states. In contrast, variational free energy is a functional of the recognition distribution and observed states. Furthermore, free utility depends on a cost function, while free energy does not. This is because the free-energy principle is based on the invariant or ergodic solution $P(o|m)$ to the *Kolmogorov forward equation*, which specifies the value of an observed state $V(o|m) = \ln P(o|m)$ directly, without reference to path integrals or cost (Friston and Ao 2012). Conversely, free utility is based on the *Kolmogorov backward equation*, which can only be solved given terminal costs. In the free-energy formulation, the value of an observed state is prescribed by a generative model in terms of the probability a state will be occupied at (non-equilibrium) steady-state. It can be seen easily that minimising the entropy of the invariant probability distribution over observations maximises expected value: $\mathbf{E}_P[-\ln P(o|m)] = \mathbf{E}_P[V(o|m)]$.

Minimising the entropy of observed states is the *raison d'être* for the free-energy principle, which invokes variational free energy to finesse the (generally) intractable problem of marginalising over hidden states to evaluate value or negative surprise (Beal 2003). This contrasts with the use of free utility to finesse the (generally) intractable problem of solving Bellman optimality equations (Braun et al. 2011). It can be seen from (6) that free energy $\mathcal{F}(o, \mu) \geq -\ln P(o|m) = -V(o|m)$ bounds surprise and can therefore be minimised to maximise value.

In summary, active inference goes beyond noting that there is a formal similarity between cost-based optimal control and Bayesian inference schemes—it suggests that optimal control is a special case of Bayes-optimal inference and that inference is the hard problem. In this setting, optimality reduces to sampling states prescribed by the priors of a generative model that specifies state transitions. The advantages of active inference include:

- A tractable approximate solution to any stochastic, nonlinear optimal control problem to the extent that standard (variational) Bayesian procedures for inference on the system being controlled exist.
- A distinction between (future) control states that are represented probabilistically (because they are necessarily hidden) and action that is a deterministic quantity produced by the system.
- The opportunity to learn and infer environmental constraints; particularly the amplitudes of observation and hidden state noise, in a Bayes-optimal fashion.
- The formalism to handle system or state noise: currently, cost based optimal control schemes are restricted to stochastic control (i.e., random fluctuations on control as opposed to hidden states). One of the practical advantages of active inference is that fluctuations in hidden states are modelled explicitly, rendering control robust to exogenous perturbations. This is seen most easily in continuous time formulations of active inference, as illustrated in (Friston et al. 2009).
- The specification of control costs in terms of priors on control, with an arbitrary form: currently, most approximate stochastic optimal control schemes are restricted to quadratic control costs. In classical schemes that appeal to path integral solutions there are additional constraints that require control costs to be a function of the precision of control noise; e.g.,(Theodorou et al. 2010; Braun et al. 2011). These constraints are not necessary in active inference.

The disadvantage of active inference is that one cannot prescribe optimality in terms of cost functions, because (Bayes) optimal behaviour rests on the generative model that is specified by its likelihood and prior functions. Having said this, for every Bayes-optimal policy there is an associated cost function. This cost function is defined in terms of the expected

change in value, where value is determined uniquely by the invariant solution to the appropriate Fokker Planck or Kolmogorov forward equation (Friston and Ao 2012).

6.2 Learning versus inference

The distinction between model free and model based decision making is based upon the difference between schemes that learn value functions directly (model free) and those that use a generative model of transition probabilities (model based). The active inference formulation makes a similar distinction between schemes based on generative models that do (agency based) and do not (agency free) include hidden control states. Both schemes can exhibit optimal behaviour; however, in agency free schemes the policy has to be *learned* or provided; cf., (Botvinick and An 2008). Agency free policies constitute prior beliefs about the next state transition, which are fulfilled by action and are the same whenever that state is visited. Conversely, in agency based schemes the policy is *inferred* in terms of posterior beliefs about the future. We have associated posterior beliefs about control with a sense of agency to make the treatment a bit more intuitive. However, there are some interesting issues that attend the perspective of agency: crucially, the agency implied by inference on control is not necessarily owned by the agent. In other words, posterior beliefs do not assign agency to any particular agent. This raises the interesting question about where a representation of *self* agency could arise. One might imagine that a sense of self would require (hierarchical) generative models that associate agency with movements of one's own motor plant. This speaks to important questions that relate to theory of mind and the role of things like the mirror neuron system in active inference (Friston et al. 2011).

6.3 Suboptimal control and psychopathology

In this paper, we have limited ourselves to a brief description of the formalism implied by free energy treatments of optimal control. In creating the simulations, many of the more interesting behaviours were failures of optimal behaviour. For example, the behaviour that emerges when prior beliefs about future states cannot be fulfilled, because the state is unattainable over a short sequence of movements. This typically leads to pathological behaviours that bring to mind the phrase 'more haste, less speed'. We hope to present these failures in a subsequent paper and relate them to suboptimal behaviour and its neurochemical mediation (Kishida et al. 2010; Moutoussis et al. 2011). The point here is that approximate inference can fail and the nature of these failures may provide a (principled) model for cognitive and motor pathologies. In short, we are not suggesting that the free-energy formulation will be useful in an engineering context; however, it may be a useful way to think about planning and agency in a behavioural or neuroscience setting. We hope to pursue this in subsequent work using the formalism introduced in this paper.

Acknowledgments We would like to thank Peter Dayan for invaluable comments on this work and also acknowledge the very helpful comments and guidance from anonymous reviewers of this work.

Glossary

(Variational) free energy—a functional of sensory states and a probability distribution over hidden states that cause sensory states. The variational free energy is an upper bound on the surprise (self information) of sensory states, under a (generative) model. Surprise is the negative logarithm of the Bayesian model evidence or marginal likelihood.

Approximate Bayesian inference—minimisation of variational free energy with respect to a probability distribution over (fictive) hidden states causing sensory states (observations). Variational Bayesian inference is approximate because it minimises a (free energy) bound approximation to surprise. When free energy equals surprise, inference is exact.

Free-energy principle—the free energy principle states that a self organising system—that entails a generative model— minimises the free energy of its sensory and internal states; where internal states encode a recognition probability distribution over (fictive) hidden states causing sensory states.

Active inference—the minimisation of free energy through changing internal states (perception) and sensory states by acting on the world (action).

Action—(real valued) variables – associated with an agent – that change hidden states in the world. Action is a set of real states – it is not inferred or represented in the generative model.

Control (states)—(fictive) hidden states that are used to explain the consequences of action. Control states are inferred or represented in the generative model.

Agency based model—a generative model (probability) over hidden states that include control states.

Agency free model—a generative model (probability) over hidden states that preclude control states.

(Sense of) agency—a probabilistic representation of hidden control states, encoded by the internal states (sufficient statistics) of an agency based model.

Optimal control—acting to minimise expected cost.

Bayes-optimal control—acting to minimise the free energy bound on the (negative logarithm) of Bayesian model evidence – with or without agency.

References

Ashby WR (1947) Principles of the self-organizing dynamic system. J Gen Psychol 37:125–128

Axmacher N, Henseler MM, Jensen O, Weinreich I, Elger CE, Fell J (2010) Cross-frequency coupling supports multi-item working memory in the human hippocampus. Proc Natl Acad Sci 107(7):3228–3233

Baxter J, Bartlett PL, Weaver L (2001) Experiments with Infinite-Horizon, Policy-Gradient Estimation. J Artif Intell Res 15:351–381

Beal MJ (2003) Variational algorithms for approximate bayesian inference'. PhD. Thesis, University College London, London

Bellman R (1952) On the theory of dynamic programming. Proc Natl Acad Sci USA 38:716–719

Berridge KC (2004) Motivation concepts in behavioral neuroscience. Physiol Behav 81(2):179–209

Birkhoff GD (1931) Proof of the ergodic theorem. Proc Natl Acad Sci USA 17:656–660

Botvinick MM, An J (2008) Goal-directed decision making in prefrontal cortex: a computational framework. Adv Neural Inf Process Syst (NIPS) 21

Braun DA, Ortega P, Theodorou E, Schaal S (2011) Path integral control and bounded rationality. In: ADPRL 2011, Paris

Brown LD (1981) A complete class theorem for statistical problems with finite sample spaces. Ann Stat 9(6):1289–1300

Camerer CF (2003) Behavioural studies of strategic thinking in games. Trends Cogn Sci 7(5):225–231

Canolty RT, Edwards E, Dalal SS, Soltani M, Nagarajan SS, Kirsch HE, Berger MS, Barbaro NM, Knight R (2006) High gamma power is phase-locked to theta oscillations in human neocortex. Science 313(5793):1626–1628

Cooper G (1988) A method for using belief networks as influence diagrams. In: Proceedings of the Conference on uncertainty in artificial intelligence

Daw ND, Doya K (2006) The computational neurobiology of learning and reward. Curr Opin Neurobiol 16(2):199–204

Dayan P, Daw ND (2008) Decision theory, reinforcement learning, and the brain. Cogn Affect Behav Neurosci 8(4):429–453

Dayan P, Hinton GE (1997) Using expectation maximization for reinforcement learning. Neural Comput 9:271–278

Dayan P, Hinton GE, Neal R (1995) The Helmholtz machine. Neural Comput 7:889–904

Duff M, (2002) Optimal learning: computational procedure for bayes-adaptive markov decision processes. PhD thesis. University of Massachusetts, Amherst

Evans DJ (2003) A non-equilibrium free energy theorem for deterministic systems. Mol Phys 101:15551–15554

Feldbaum AA (1961) Dual control theory, Part I. Autom Remote Control 21(9):874–880

Feldman H, Friston KJ (2010) Attention, uncertainty, and free-energy. Front Hum Neurosci 4:215

Feynman RP (1972) Statistical mechanics. Benjamin, Reading MA

Filatov N, Unbehauen H (2004) Adaptive dual control: theory and applications (lecture notes in control and information sciences. Springer, Berlin

Fox C, Roberts S (2011) A tutorial on variational Bayes. In: Artificial intelligence review. Spinger, Berlin

Friston K (2008) Hierarchical models in the brain. PLoS Comput Biol 4(11):e1000211

Friston K (2010) The free-energy principle: a unified brain theory?. Nat Rev Neurosci 11(2):127–138

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

Friston K, Ao P (2012) Free-energy, value and attractors. In: Computational and mathematical methods in medicine, vol 2012

Friston K, Kiebel S (2009) Cortical circuits for perceptual inference. Neural Netw 22(8):1093–1104

Friston K, Kiebel S (2009) Predictive coding under the free-energy principle. Philos Trans R Soc Lond B Biol Sci 364(1521):1211–1221

Friston KJ, Daunizeau J, Kiebel SJ (2009) Active inference or reinforcement learning?. PLoS One 4(7):e6421

Friston KJ, Daunizeau J, Kilner J, Kiebel SJ (2010) Action and behavior: a free-energy formulation. Biol Cybern 102(3): 227–260

Friston KJST, Fitzgerald T, Galea JM, Adams R, Brown H, Dolan RJ, Moran R, Stephan KE, Bestmann S (2012) Dopamine, affordance and active inference. PLoS Comput Biol 8(1):e1002327

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

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

Friston KJ, Tononi G, Reeke GNJ, Sporns O, Edelman GM (1994) Value-dependent selection in the brain: simulation in a synthetic neural model. Neuroscience 59(2):229–243

Gigerenzer G, Gaissmaier W (2011) Heuristic decision making. Annu Rev Psychol 62:451–482

Gläscher J, Daw N, Dayan P, O'Doherty JP (2010) States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron 66(4):585–595

Gomez F, Miikkulainen R (2001) Learning robust nonlinear control with neuroevolution. Technical Report AI01-292, Department of Computer Sciences, The University of Texas at Austin

Gomez F, Schmidhuber J, Miikkulainen R (2009) Accelerated neural evolution through cooperatively coevolved synapses. J Mach Learn Res 9:937–965

Helmholtz H (1866/1962), Concerning the perceptions in general. In: Treatise on physiological optics, 3rd edn. Dover, New York

Hinton GE, van Camp D (1993) Keeping neural networks simple by minimizing the description length of weights. In: Proceedings of COLT-93,pp 5–13

Hoffman, M, de Freitas, N, Doucet, A, Peters J (2009) An expectation maximization algorithm for continuous markov decision processes with arbitrary rewards. In: Twelfth Int. Conf. on artificial intelligence and statistics (AISTATS 2009)

Howard RA (1960) Dynamic programming and Markov processes. MIT Press Cambridge, MA

Jaeger H (2000) Observable operator models for discrete stochastic time series. Neural Comput 12:1371–1398
Jensen F, Jensen V, Dittmer SL (1994) From influence diagrams to junction trees. In: Proc. of the Tenth Conference on uncertainty in artificial intelligence. Morgan Kaufmann, San Fransisco
Kaelbling LP, Littman ML, Cassandra AR (1998) Planning and acting in partially observable stochastic domains. Artif Intell 101 (1–2):99–134
Kappen HJ (2005) Linear theory for control of nonlinear stochastic systems. Phys Rev Lett 95(20):200201
Kappen HJ (2005) Path integrals and symmetry breaking for optimal control theory. J Stat Mech: Theory Exp 11:P11011
Kappen HJ, Gomez Y, Opper M (2009) Optimal control as a graphical model inference problem. arXiv:0901.0633v2
Kiebel SJ, Daunizeau J, Friston KJ (2009a) Perception and hierarchical dynamics. Front Neuroinf 3:20
Kiebel SJ, von Kriegstein K, Daunizeau J, Friston KJ (2009b) Recognizing sequences of sequences. PLoS Comput Biol 5(8):e1000464
Kishida KT, King-Casas B, Montague PR (2010) Neuroeconomic approaches to mental disorders. Neuron 67(4):543–554
Littman ML, Majercik SM, Pitassi T (2001) Stochastic boolean satisfiability. J Autom Reason 27(3):251–296
Littman ML, Sutton RS, Singh S (2002) Predictive Representations of State. Adv Neural Inf Process Syst 14
MacKay DJ (1995) Free-energy minimisation algorithm for decoding and cryptoanalysis. Electron Lett 31:445–447
Montague PR, Dayan P, Person C, Sejnowski TJ (1995) Bee foraging in uncertain environments using predictive Hebbian learning. Nature 377(6551):725–728
Moutoussis M, Bentall RP, El-Deredy W, Dayan P (2011) Bayesian modelling of Jumping-to-conclusions bias in delusional patients. Cogn Neuropsychiatry 7:1–26
Namikawa J, Nishimoto R, Tani J (2011) A neurodynamic account of spontaneous behaviour. PLoS Comput Biol. 7(10):e1002221
Neal RM, Hinton GE (1998) A view of the EM algorithm that justifies incremental sparse and other variants. In: Jordan M (ed.) Learning in graphical models. Kluwer Academic, Dordrecht
Oliehoek F, Spaan MTJ, Vlassis N (2005) Best-response play in partially observable card games. In: Proceedings of the 14th Annual Machine Learning Conference of Belgium and the Netherlands
Pearl J (1988) Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann, San Fransisco
Rao RP (2010) Decision making under uncertainty: a neural model based on partially observable markov decision processes. Front Comput Neurosci 4:146
Rao RP, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci 2(1):79–87
Rawlik K, Toussaint M, Vijayakumar S (2010) Approximate inference and stochastic optimal control. arXiv:1009.3958
Rescorla RA, Wagner AR (1972) A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In: A Black, W Prokasy (eds.) Classical conditioning II: current research and theory. Appleton Century Crofts, New York
Robert C (1992) L'analyse statistique Bayesienne. In: Economica. Paris, France
Shachter RD (1988) Probabilistic inference and influence diagrams. Operat Res 36:589–605
Silver D, Veness J (2010) Monte-Carlo planning in large POMDPs. In: Proceedings of the Conference on neural information processing systems
Sutton RS, Barto AG (1981) Toward a modern theory of adaptive networks: expectation and prediction. Psychol Rev 88(2):135–170
Tani J (2003) Learning to generate articulated behavior through the bottom-up and the top-down interaction processes. Neural Netw 16(1):11–23
Theodorou E, Buchli J, Schaal S (2010) A generalized path integral control approach to reinforcement learning. J Mach Learn Res 11:3137–3181
Todorov E (2006) Linearly-solvable Markov decision problems. In: Advances in neural information processing systems. MIT Press, Boston
Todorov E (2008) General duality between optimal control and estimation. In: IEEE Conference on decision and control
Toussaint M, Charlin L, Poupart P (2008) Hierarchical POMDP controller optimization by likelihood maximization. In: Uncertainty in artificial intelligence (UAI 2008), AUAI Press, Menlo Park
Toussaint M, Storkey A (2006) Probabilistic inference for solving discrete and continuous state Markov decision processes. In: Proceedings of the 23nd International Conference on machine learning
van den Broek B, Wiegerinck W, Kappen B (2008) Graphical model inference in optimal control of stochastic multi-agent systems. J Artif Int Res 32(1):95–122
Watkins CJ, Dayan P (1992) Q-learning. Mach Learn 8:279–292
Williams RJ (1992) Simple statistical gradient-following algorithms for connectionist reinforcement learning. Mach Learn 8:229–256
Zhang NL (1998) Probabilistic inference in influence diagrams. Comput Intell 14(4):475–497