LEARNING ACTION-ORIENTED MODELS THROUGH ACTIVE INFERENCE

Alexander Tschantz
Sackler Centre for Consciousness Science
School of Engineering & Informatics
University of Sussex
tschantz.alec@gmail.com

Anil K. Seth
Sackler Centre for Consciousness Science
Canadian Institute for Advanced Research
School of Engineering & Informatics
University of Sussex
A.K.Seth@sussex.ac.uk

Christopher L. Buckley
Evolutionary and Adaptive Systems Research Group
School of Engineering & Informatics
University of Sussex
C.L.Buckley@sussex.ac.uk

September 10, 2019

ABSTRACT

Converging theories suggest that organisms learn and exploit probabilistic models of their environment. However, it remains unclear how such models can be learned in practice. The open-ended complexity of natural environments means that it is generally infeasible for organisms to model their environment comprehensively. Alternatively, *action-oriented* models attempt to encode a parsimonious representation of adaptive agent-environment interactions. One approach to learning action-oriented models is to learn online in the presence of goal-directed behaviours. This constrains an agent to behaviourally relevant trajectories, reducing the diversity of the data a model need account for. Unfortunately, this approach can cause models to prematurely converge to sub-optimal solutions, through a process we refer to as a *bad-bootstrap*. Here, we exploit the normative framework of *active inference* to show that efficient action-oriented models can be learned by balancing goal-oriented and epistemic (*information-seeking*) behaviours in a principled manner. We illustrate our approach using a simple agent-based model of bacterial chemotaxis. We first demonstrate that learning via goal-directed behaviour indeed constrains models to behaviorally relevant aspects of the environment, but that this approach is prone to sub-optimal convergence. We then demonstrate that epistemic behaviours facilitate the construction of accurate and comprehensive models, but that these models are not tailored to any specific behavioural niche and are therefore less efficient in their use of data. Finally, we show that active inference agents learn models that are parsimonious, tailored to action, and which avoid bad bootstraps and sub-optimal convergence. Critically, our results indicate that models learned through active inference can support adaptive behaviour in spite of, and indeed *because of*, their departure from veridical representations of the environment. Our approach provides a principled method for learning adaptive models from limited interactions with an environment, highlighting a route to sample efficient learning algorithms.

1 Introduction

In order to survive, biological organisms must be able to efficiently adapt to and navigate in their environment. Converging research in neuroscience, biology, and machine learning suggests that organisms achieve this feat by exploiting probabilistic models of their world (Doll et al., 2012; Dayan and Berridge, 2014; Botvinick and Weinstein, 2014; Dolan and Dayan, 2013; Conant and Ashby, 1970; Friston, 2013; Kuvayev and Sutton, 1996; Deisenroth, 2011). These models encode statistical representations of the states and contingencies in an environment and agent-environment interactions. Such models plausibly endow organisms with several advantages. For instance, probabilistic models can be used to perform perceptual inference, implement anticipatory control, overcome sensory noise and delays, and generalize existing knowledge to new tasks and environments. While encoding a probabilistic model can be advantageous in these and other ways, natural environments are extremely complex and it is infeasible to model them in their entirety. Thus it is unclear how organisms with limited resources could exploit probabilistic models in rich and complex environments.

One approach to this problem is for organisms to selectively model their world in a way that supports action (Seth, 2015; Seth and Tsakiris, 2018; Baltieri and Buckley, 2017; Clark, 2015; Pezzulo et al., 2017; Gibson, 2014). We refer to such models as *action-oriented*, as their functional purpose is to enable adaptive behaviour, rather than to represent the world in a complete or accurate manner. An action-oriented representation of the world can depart from a veridical representation in a number of ways. First, because only a subset of the states and contingencies in an environment will be relevant for behaviour, action-oriented models need not exhaustively model their environment (Baltieri and Buckley, 2017). Moreover, specific *misrepresentations* may prove to be useful for action (Wiese, 2017; McKay and Dennett, 2009; Mendelovici, 2013; M. Zehetleitner and Schönbrodt, 2015), indicating that action-oriented models need not be accurate. By reducing the need for models to be isomorphic with their environment, an action-oriented approach can increase the tractability of the model learning process (Verschure et al., 2003; Montúfar et al., 2015; Thornton, 2010; Ruesch et al., 2011; Lungarella and Sporns, 2005, 2006), especially for organisms with limited resources.

Within an action-oriented approach, an open question is how action-oriented models can be learned from experience. The environment, in and of itself, provides no distinction between states and contingencies that are relevant for behaviour and those which are not. However, organisms do not receive information passively. Rather, organisms *actively* sample information from their environment, a process which plays an important role in both perception and learning (Yang et al., 2018; Gottlieb and Oudeyer, 2018; Lungarella and Sporns, 2005; Friston et al., 2012b). One way that active sampling can facilitate the learning of efficient action-oriented models is to learn online in the presence of *goal-directed* actions. Performing goal-directed actions restricts an organism to behaviourally relevant trajectories through an environment. This, in turn, structures sensory data in a behaviorally relevant way, thereby reducing the diversity and dimensionality of the sampled data (see figure 1). Therefore, this approach offers an effective mechanism for learning parsimonious models that are tailored to an organism's adaptive requirements (Montúfar et al., 2015; Barandiaran, 2017; Verschure et al., 2003; Lungarella and Sporns, 2005, 2006; Egbert and Barandiaran, 2014).

Learning probabilistic models to optimise behaviour has been extensively explored in the model-based reinforcement learning (RL) literature (Polydoros and Nalpantidis, 2017; Atkeson and Santamaria, 1997; Ha and Schmidhuber, 2018; Deisenroth, 2011). A significant drawback to existing methods is that they tend to prematurely converge to sub-optimal solutions (Chua et al., 2018; McAllister and Rasmussen, 2016). One reason this occurs is due to the inherent coupling between action-selection and model learning. At the onset of learning, agents must learn from limited data, and this can lead to models that initially overfit the environment and, as a consequence, make sub-optimal predictions about the consequences of action. Subsequently using these models to determine goal-oriented actions can result in biased and sub-optimal samples from the environment, further compounding the model's inefficiencies, and ultimately entrenching maladaptive cycles of learning and control, a process we refer to as a "bad-bootstrap" (see figure 1).

One obvious approach to resolving this problem is for an organism to perform some actions, during learning, that are not explicitly goal-oriented. For example, heuristic methods, such as ε-greedy (Watkins, 1989), utilise noise to enable exploration at the start of learning. However, random exploration of this sort is likely to be inefficient in rich and complex environments. In such environments, a more powerful method is to utilize the uncertainty quantified by

probabilistic models to determine *epistemic* (or *intrinsic, information-seeking, uncertainty reducing*) actions that attempt to minimize the model uncertainty in a directed manner (Stadie et al., 2015; Houthooft et al., 2016; Sun et al., 2011; Friston et al., 2015; Burda et al., 2018; Friston et al., 2017). While epistemic actions can help avoid bad-bootstraps and sub-optimal convergence, such actions necessarily increase the diversity and dimensionality of sampled data, thus sacrificing the benefits afforded by learning in the presence of goal-directed actions. Thus, a principled and pragmatic method is needed to learn action-oriented models in the presence of both goal-directed *and* epistemic actions.

In this paper, we develop an effective method for learning action-oriented models. This method balances goal-directed and epistemic actions in a principled manner, thereby ensuring that an agent's model is tailored to goal-relevant aspects of the environment, while also ensuring that epistemic actions are contextualized by and directed towards an agent's adaptive requirements. To achieve this, we exploit the theoretical framework of active inference, a normative theory of perception, learning and action (Friston and Stephan, 2007; Friston, 2010; Friston et al., 2016a). Active inference proposes that organisms maintain and update a probabilistic model of their typical (habitable) environment and that the states of an organism change to maximize the evidence for this model. Crucially, both goal-oriented and epistemic actions are complementary components of a single imperative to maximize model evidence - and are therefore evaluated in a common (information-theoretic) currency (Friston et al., 2015, 2016a, 2017).

We illustrate this approach with a simple agent-based model of bacterial chemotaxis. This model is not presented as a biologically-plausible account of chemotaxis, but instead, is used as a relatively simple behaviour to evaluate the hypothesis that adaptive action-oriented models can be learned via active inference. First, we confirm that learning in the presence of goal-directed actions leads to parsimonious models that are tailored to specific behavioural niches. Next, we demonstrate that learning in the presence of goal-directed actions *alone* can cause agents to engage in maladaptive cycles of learning and control - 'bad bootstraps' - leading to premature convergence on sub-optimal solutions. We then show that learning in the presence of epistemic actions allows agents to learn accurate and exhaustive models of their environment, but that the learned models are not tailored to any behavioural niche, and are therefore inefficient and unlikely to scale to complex environments. Finally, we demonstrate that balancing goal-directed and epistemic actions through active inference provides an effective method for learning efficient action-oriented models that avoid maladaptive patterns of learning and control. 'Active inference' agents learn well-adapted models from a relatively limited number of agent-environment interactions and do so in a way that benefits from systematic representational inaccuracies. Our results indicate that probabilistic models can support adaptive behaviour in spite of, and moreover, *because of*, the fact they depart from veridical representations of the external environment.

The structure of the paper is as follows. In section two, we outline the active inference formalism, with a particular focus on how it prescribes both goal-directed and epistemic behaviour. In section three, we present the results of our agent-based simulations, and in section four, we discuss these results and outline some broader implications. In section five, we outline the methods used in our simulations, which are based on the Partially Observed Markov Decision Process (POMDP) framework, a popular method for modelling choice behaviour under uncertainty.

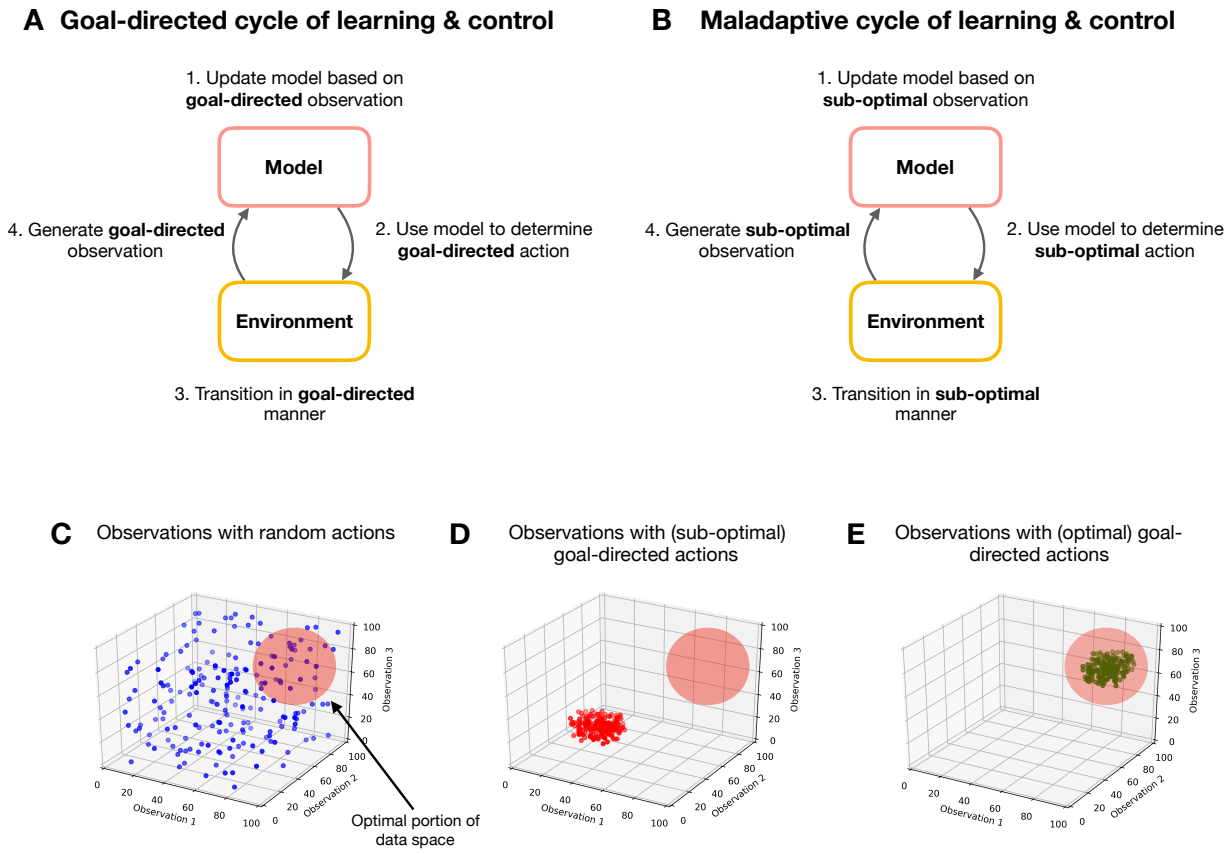

Figure 1: **(A) Goal-directed cycle of learning and control**: a schematic overview of the coupling between a model and its environment when learning takes place in the presence of goal-directed actions. Here, a model is learned based on sampled observations. This model is then used to determine goal-directed actions, causing goal-relevant transitions in the environment, which in turn generate goal-relevant observations. **(B) Maladaptive cycle of learning and control**: a schematic overview of the model-environment coupling when learning in the presence of goal-directed actions, but for the case where a maladaptive model has been initially learned. The feedback inherent in the online learning scheme means that the model samples sub-optimal observations, which are subsequently used to update the model, thus entrenching maladaptive cycles of learning and control (bad bootstraps). **(C) Observations sampled from random actions**: The spread of observations covers the space of possible observations uniformly, meaning that a model of these observations must account for a diverse and distributed set of data, increasing the model's complexity. The red circle in the upper right quadrant indicates the region of observation space associated with optimal behaviour, which is only sparsely sampled. Note these are taken from a fictive simulation and are purely illustrative. **(D) Observations sampled from sub-optimal goal-directed actions**: Only a small portion of observation space is sampled. A model of this data would, therefore, be more parsimonious in its representation of the environment. However, the model prescribes actions that cause the agent to selectively sample a sub-optimal region of observation space (i.e outside the red circle in the upper-right quadrant). As the agent only samples this portion of observation space, the model does not learn about more optimal behaviours. **(E) Observations sampled from optimal goal-directed actions**: Here, as in **D**, the goal-directed nature of action ensures that only a small portion of observation space is sampled. However, unlike **D**, this portion is associated with optimal behaviours.

2 Formalism

Active inference is a normative theory that unifies perception, action and learning under a single imperative - the minimization of variational *free energy* (Friston, 2010; Friston et al., 2016a). Free energy $\mathcal{F}(\phi, o)$ is defined as:

$$\begin{aligned}\mathcal{F}(\phi, o) &= \mathbb{KL}[Q(x|\phi)||P(x, o)] \\ &= \mathbb{KL}[Q(x|\phi)||P(x|o)] - \ln P(o)\end{aligned} \tag{1}$$

where $\mathbb{KL}$ is the Kullback-Libeler divergence (KL-divergence) between two probability distributions, both of which are parameterized by the internal states of an agent. The first is the approximate posterior distribution, $Q(x|\phi)$[1], which is a distribution over unknown or latent variables x with sufficient statistics ϕ. This distribution encodes an agent's 'beliefs' about the unknown variables x [2]. The second distribution is the generative model, $P(x, o)$, which is the joint distribution over unknown variables x and observations o. This distribution encodes an agent's probabilistic model of its (internal and external) environment. We provide two additional re-arrangements of equation 1 in appendix 1.

Minimizing free energy has two functional consequences. First, it minimizes the divergence between the approximate posterior distribution $Q(x|\phi)$ and the true posterior distribution $P(x|o)$, thereby implementing a tractable form of approximate Bayesian inference known as variational Bayes (Hinton and van Camp, 1993; Beal, 2003). On this view, perception can be understood as the process of maintaining and updating beliefs about hidden state variables s, where $s \in \mathcal{S}$. The hidden state variables can either be a compressed representation of the potentially high-dimensional observations (i.e. representing an object), or they can represent quantities that are not directly observable (i.e. velocity). This casts perception as a process of approximate inference, connecting active inference to influential theories such as the Bayesian brain hypothesis (Knill and Pouget, 2004; Gregory, 1980) and predictive coding (Rao and Ballard, 1999). Under active inference, *learning* can also be understood as a process of approximate inference (Friston et al., 2016a). This can be formalized by assuming that agents maintain and update beliefs over the parameters θ of their generative model, where $\theta \in \Theta$. Finally, action can be cast as a process of approximate inference by assuming that agents maintain and update beliefs over control states u, where $u \in \mathcal{U}$, which prescribe actions a, where $a \in \mathcal{A}$ [3]. Together, this implies that $x = (s, \theta, u)$. Approximate inference, encompassing perception, action, and learning, can then be achieved through the following scheme:

$$\phi^* = \arg\min_{\phi} \mathcal{F}(\phi, o) \tag{2}$$

In other words, as new observations are sampled, the sufficient statistics ϕ are updated in order to minimize free energy (see the Methods section for the implementation used in the current simulations, or (Buckley et al., 2017) for an alternative implementation based on the Laplace approximation). Once the optimal sufficient statistics ϕ^* have been identified, the approximate posterior will become an approximation of the true posterior distribution $Q(x|\phi^*) \approx P(x|o)$, meaning that agents will encode approximately optimal beliefs over hidden states s, model parameters θ and control states u.

The second consequence of minimizing free energy is that it maximizes the Bayesian *evidence* for an agents generative model, or equivalently, minimizes 'surprisal' $-\ln P(o)$, which is the information-theoretic *surprise* of sampled observations (see appendix 1). Active inference proposes that an agent's goals, preferences and desires are encoded in the generative model as a prior preference for favourable observations (e.g. blood temperature at 37°) (Friston et al., 2009). In other words, it proposes that an agent's generative model is biased towards favourable states of affairs. The process of actively minimizing free energy will, therefore, ensure that these favourable (i.e. probable) observations are preferentially sampled (Friston et al., 2012a). However, model evidence cannot be directly maximized through the inference scheme described by equation 2, as the marginal probability of observations $P(o)$ is independent of the

[1] This distribution is often referred to as the *recognition* or *proposal* distribution.

[2] Here, the term 'belief' does not necessarily refer to beliefs in the cognitive sense but instead implies a probabilistic representation of unknown variables.

[3] Separating control states from actions is necessary as actions are a physically realized phenomena, whereas control states are unknown variables that must be inferred.

sufficient statistics ϕ. Therefore, to maximize model evidence, agents must *act* in order to change their observations. This process can be achieved in a principled manner by selecting actions in order to minimize *expected* free energy, which is the free energy that is expected to occur from executing some (sequence of) actions (Friston et al., 2015, 2014).

2.1 Expected free energy

To ensure that actions minimize (the path integral of) free energy, an agent's generative model should specify that control states are *a-priori* more likely if they are expected to minimize free energy in the future, thus ensuring that the process of approximate inference assigns a higher posterior probability to the control states that are expected to minimize free energy (Parr and Friston, 2018). The expected free energy for a candidate control state $\mathbf{G}_\tau(\phi_\tau, u_t)$ quantifies the free energy expected at some future time τ given the execution of some control state u_t, where t is the current time point and:

$$\begin{aligned}\mathbf{G}_\tau(\phi_\tau, u_t) &= \mathbb{E}_{Q(o_\tau, x_\tau|u_t,\phi_\tau)}[\ln Q(x_\tau|u_\tau,\phi_\tau) - \ln P(o_\tau, x_\tau|u_t)] \\ &\approx \underbrace{\mathbb{E}_{Q(o_\tau, x_\tau|u_t,\phi_\tau)}\big[\ln Q(x_\tau|u_t,\phi_\tau) - \ln Q(x_\tau|o_\tau,u_t,\phi_\tau)\big]}_{\text{(Negative) epistemic value}} \\ &\quad \underbrace{-\mathbb{E}_{Q(o_\tau, x_\tau|u_t,\phi_\tau)}\big[\ln P(o_\tau)\big]}_{\text{(Negative) instrumental value}}\end{aligned} \tag{3}$$

We describe the formal relationship between free energy and expected free energy in appendix 2. In order to evaluate expected free energy, agents must first evaluate the expected consequences of control, or formally, evaluate the predictive approximate posterior $Q(o_\tau, x_\tau|u_t, \phi_\tau)$. We refer readers to the Methods section for a description of this process.

The second (approximate) equality of equation 3 demonstrates that expected free energy is composed of an *instrumental* (or *extrinsic*, *pragmatic*, *goal-directed*) component and an *epistemic* (or *intrinsic*, *uncertainty-reducing*, *information-seeking*) component[4]. We provide a full derivation of the second equality in appendix 3, but note here that the decomposition of expected free energy into instrumental and epistemic value affords an intuitive explanation. Namely, as free energy quantifies the divergence between an agent's current beliefs and its model of the world, this divergence can be minimized via two methods: by changing beliefs such that they align with observations (associated with maximizing epistemic value), or by changing observations such that they align with beliefs (associated with maximizing instrumental value).

Formally, instrumental value quantifies the degree to which the predicted observations o_τ - given by the predictive approximate posterior $Q(o_\tau, x_\tau|u_t, \phi_\tau)$ - are consistent with the agents prior beliefs $P(o_\tau)$. In other words, this term will be maximized when an agent expects to sample observations that are consistent with its prior beliefs. As an agent's generative model assigns a higher prior probability to favourable observations (i.e. goals and desires), maximizing instrumental value can be associated with promoting 'goal-directed' behaviours. This formalizes the notion that, under active inference, agents seek to maximize the evidence for their (biased) model of the world, rather than seeking to maximize reward as a separate construct (as in, e.g., reinforcement learning) (Friston et al., 2009).

Conversely, epistemic value quantifies the expected reduction in uncertainty in the beliefs over unknown variables x. Formally, it quantifies the expected information gain for the predictive approximate posterior $Q(x_\tau|u_t, \phi_\tau)$. By noting

[4]Note that under active inference, agents are mandated to *minimize* expected free energy, and as both the instrumental and epistemic terms are in a negative form in equation 3, expected free energy will be minimized when instrumental and epistemic value are maximized.

that that x can be factorized into hidden states s and model parameters θ[5], we can rewrite *positive* epistemic value (i.e. the term to be maximized) as:

$$\underbrace{\mathbb{E}_{Q(o_\tau, s_\tau, \theta | u_t, \phi_\tau)}\big[\ln Q(s_\tau | o_\tau, u_t, \phi_\tau) - \ln Q(s_\tau | u_t, \phi_\tau)\big]}_{\text{State epistemic value}} + \underbrace{\mathbb{E}_{Q(o_\tau, s_\tau, \theta | u_t, \phi_\tau)}\big[\ln Q(\theta | s_\tau, o_\tau, u_t, \phi_\tau) - \ln Q(\theta | \phi_\tau)\big]}_{\text{Parameter epistemic value}} \tag{4}$$

We provide a full derivation of equation 4 in appendix 4 and discuss its relationship to several established formalisms. Here, we have decomposed epistemic value into *state* epistemic value, or *salience*, and *parameter* epistemic value, or *novelty* (Schwartenbeck et al., 2018). State epistemic value quantifies the degree to which the expected observations o_τ reduce the uncertainty in an agent's beliefs about the hidden states s_τ. In contrast, parameter epistemic value quantifies the degree to which the expected observations o_τ and expected hidden states s_τ reduce the uncertainty in an agent's beliefs about model parameters θ. Thus, by maintaining a distribution over model parameters, the uncertainty in an agent's generative model can be quantified, allowing for 'known unknowns' to be identified and subsequently acted upon (Friston et al., 2017). Maximizing parameter epistemic value, therefore, causes agents to sample novel agent-environment interactions, promoting the exploration of the environment in a principled manner.

2.2 Summary

In summary, active inference proposes that agents learn and update a probabilistic model of their world, and act to maximize the evidence for this model. However, in contrast to previous 'perception-oriented' approaches to constructing probabilistic models (Baltieri and Buckley, 2017), active inference requires an agent's model to be intrinsically biased towards certain (favourable) observations. The goal is not, therefore, to construct a model that accurately captures the true causal structure underlying observations, but is instead to learn a model that is tailored to a specific set of prior preferences, and thus tailored to a specific set of agent-environment interactions. Moreover, by ensuring that actions maximize evidence for a (biased) model of the world, active inference prescribes a trade-off between instrumental and epistemic actions. Crucially, the fact that actions are selected based on both instrumental *and* epistemic value means that epistemic foraging will be contextualized by an agent's prior preferences. Specifically, epistemic foraging will be biased towards parts of the environment that also provide instrumental value, as these parts will entail a lower expected free energy relative to those that provide no instrumental value. Moreover, the degree to which epistemic value determines the selection of actions will depend on instrumental value. Thus, when the instrumental value afforded by a set of actions is low, epistemic value will dominate action selection, whereas if actions afford a high degree of instrumental value, epistemic value will have less influence on the action selection. Finally, as agents maintain beliefs about (and thus quantify the uncertainty of) the hidden state of the environment *and* the parameters of their generative model, epistemic value promotes agents to actively reduce the uncertainty in both of these beliefs.

3 Results

To test our hypothesis that acting to minimize expected free energy will lead to the learning of well-adapted action-oriented models, we empirically compare the types of model that are learned under four different action strategies. These are (i) minimization of expected free energy, (ii) maximization of instrumental value, (iii) maximization of epistemic value, and (iv) random action selection. For each strategy, we assess model performance after a range of model learning durations. We assess model performance across several criteria, including whether or not the models can prescribe well-adapted behaviour, the complexity and accuracy of the learned models, whether the models are tailored to a behavioural niche, and whether or not the models become entrenched in maladaptive cycles of learning and control ('bad-bootstraps').

[5]Epistemic value for control states u cannot be considered as the distributions in expected free energy are conditioned on control states.

We implement a simple agent-based model of bacterial chemotaxis that infers and learns based on the active inference scheme described above. Specifically, our model implements the 'adaptive gradient climbing' behaviour of *E. coli* [6] (Berg and Brown, 1972). This behaviour depends on the chemical gradient at the bacteria's current orientation. In positive chemical gradients, bacteria 'run' forward in the direction of their current orientation. In negative chemical gradients, bacteria 'tumble', resulting in a new orientation being sampled. This behaviour, therefore, implements a rudimentary biased random-walk towards higher concentrations of chemicals. To simulate the adaptive gradient climbing behaviour of *E. coli*, we utilize the partially observed Markov Decision Process (POMDP) framework (Puterman, 1994). This framework implies that agents do not have direct access to the true state of the environment, that the state of the environment only depends on the previous state and the agent's previous action, and that all variables and time are discrete. Note that while agents operate on discrete representations of the environment, the true states of the environment (i.e the agent's position, the location of the chemical source, and the chemical concentrations) are continuous.

At each time step t, agents receive one of two observations, either a positive chemical gradient $o^{\texttt{pos}}$ or a negative chemical gradient $o^{\texttt{neg}}$ [7]. After receiving an observation, agents update their beliefs in order to minimize free energy. In the current simulations, agents maintain and update beliefs over three variables. The first is the hidden state variable s, which represents the agent's belief about the local chemical gradient, and which has a domain of $\{s^{\texttt{pos}}, s^{\texttt{neg}}\}$, representing positive and negative chemical gradients, respectively. The second belief is over the parameters θ of the agent's generative model, which describe the probability of transitions in the environment, given action. The final belief is over the control variable u, which has the domain of $\{u^{\texttt{run}}, u^{\texttt{tumble}}\}$, representing running and tumbling respectively. Agents are also endowed with the prior belief that observing positive chemical gradients $o^{\texttt{pos}}$ is *a-priori* more likely, such that the evidence for an agent's model is maximized (and free energy minimized) when sampling positive chemical gradients.

Once beliefs have been updated, agents execute one of two actions, either run $a^{\texttt{run}}$ or tumble $a^{\texttt{tumble}}$, depending on which of the corresponding control states was inferred to be more likely. Running causes the agent to move forward one unit in the direction of their current orientation, whereas tumbling causes the agent to sample a new orientation at random. The environment is then updated and a new time step begins. We refer the reader to the Methods section for a full description of the agents generative model, approximate posterior, and the corresponding update equations for inference, learning and action.

3.1 Agents

All of the action strategies we compare infer posterior beliefs over hidden states, model parameters and control states via the minimization of free energy. However, they differ in how they assign prior (and thus posterior) probability to control states. The first strategy we consider is based on the minimization of *expected free energy*, which entails the following prior over control states:

$$P_{\texttt{EFE}}(u_t) = \sigma\Big(\mathbb{E}_{Q(o_\tau, s_\tau, \theta | u_t, \phi_\tau)}\big[\ln Q(\theta | s_\tau, o_\tau, u_t, \phi_\tau) - \ln Q(\theta | \phi_\tau)\big] + \mathbb{E}_{Q(o_\tau, s_\tau, \theta | u_t, \phi_\tau)}[\ln P(o_\tau)]\Big) \tag{5}$$

where $\sigma(\cdot)$ is the softmax function, which ensures that $P_{\texttt{EFE}}(u_t)$ is a valid distribution. The first term corresponds to *parameter* epistemic value, or 'novelty', and quantifies the amount of information the agent expects to gain about their (beliefs about their) model parameters θ. The second term corresponds to instrumental value and quantifies the degree to which the expected observations conform to prior beliefs. Therefore, the expected free energy agent selects actions that

[6] Note that we do not propose our model as a biologically realistic account of bacterial chemotaxis. Instead, we use chemotaxis as a relatively simple behaviour that permits a thorough analysis of the learned models. However, the active inference scheme described in this paper has a degree of biological plausibility (Friston et al., 2018), and there is some evidence to suggest that bacteria engage in model-based behaviours (Mitchell et al., 2009; Mitchell and Lim, 2016; Freddolino and Tavazoie, 2012)

[7] The chemical gradient is computed spatially rather than temporally (Thar and Kuhl, 2003), and thus only depends on the agent's current position and orientation, and the position of the chemical source.

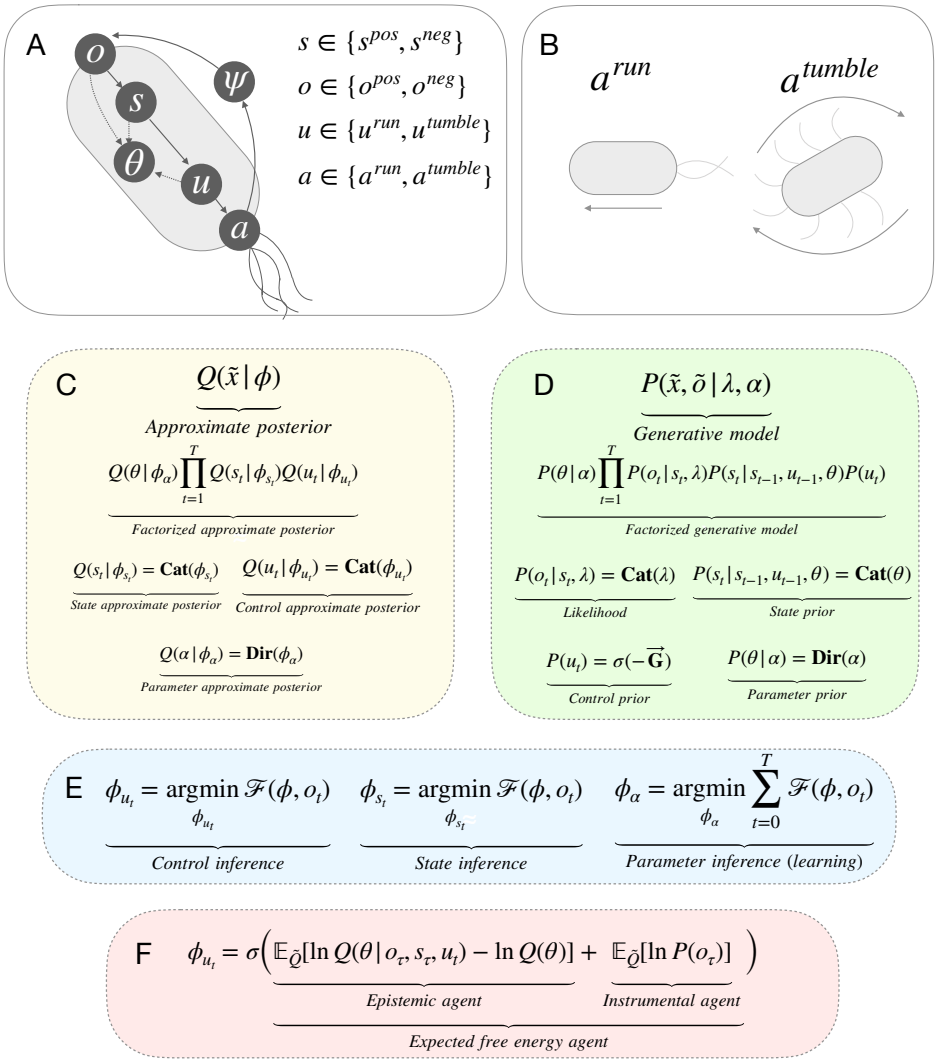

Figure 2: **(A) Agent overview:** Agents act in an environment which is described by states ψ, which are unknown to the agent but generate observations o. The agent maintains beliefs about the state of the environment s, however, s and ψ need not be homologous. Agents also maintain beliefs about control states u, which in turn prescribe actions a. Finally, the agent maintains beliefs over model parameters θ, which describe the probability of transitions in s under different control states u. **(B) Actions:** at each time step, agents can either *run*, which moves them forward one unit in the direction of their current orientation, or *tumble*, which causes a new orientation to be sampled at random. **(C) Approximate posterior:** the factorization of the approximate posterior, and the definition of each factor. We refer readers to Methods section for a full description of these distributions. **(D) Generative model:** the factorization of the generative model and the definition of each factor. **(E) Free energy minimization:** the general scheme for free energy minimization under the mean-field assumption. We refer readers to the Methods section for further details. **(F) Control state inference:** the update equation for control state inference, where $\tilde{Q} = Q(o_\tau, s_\tau, \theta | u_t)$. This equation highlights the difference between the three action-strategies considered in the following simulations.

are expected to result in probable ('favourable') observations, and that are expected to disclose maximal information about the consequences of action. Note that in the following simulations, agents have no uncertainty in their likelihood

distribution, which describes the relationship between the hidden state variables s and the observations o (see Methods). As such, the expected free energy agent does not assign probability to control states based on state epistemic value [8].

The second strategy is the *instrumental*, or 'goal-directed', strategy, which utilizes the following prior over control states:

$$P_{\texttt{Instrumental}}(u_t) = \sigma\Big(\mathbb{E}_{Q(o_\tau, s_\tau, \theta | u_t, \phi_\tau)}[\ln P(o_\tau)]\Big) \tag{6}$$

The instrumental agent, therefore, selects actions that are expected to give rise to favourable observations. The third strategy is the *epistemic*, or 'information-seeking', strategy, which is governed by the following prior over control states:

$$P_{\texttt{Epistemic}}(u_t) = \sigma\Big(\mathbb{E}_{Q(o_\tau, s_\tau, \theta | u_t, \phi_\tau)}\big[\ln Q(\theta | s_\tau, o_\tau, u_t, \phi_\tau) - \ln Q(\theta | \phi_\tau)\big]\Big) \tag{7}$$

The epistemic agent selects actions that are expected to disclose maximal information about model parameters. The final strategy is the *random* strategy, which assigns prior probability to actions at random. Implementation details for all four strategies are provided in the Methods section.

3.2 Model performance

We first assess whether the learned models can successfully generate chemotactic behaviour. We quantify this by measuring an agent's distance from the source after an additional (i.e., post-learning) testing phase. Each testing phase begins by placing an agent at a random location and orientation 400 units from the chemical source. The agent is then left to act in the environment for 1000 time steps, utilizing the model that was learned during the preceding learning phase. No additional learning takes place during the testing phase. As the epistemic and random action strategies do not assign any instrumental (goal-oriented) value to actions, there is no tendency for them to navigate towards the chemical source. Therefore, to ensure a fair comparison between action strategies, all agents select actions based on the minimization of expected free energy during the testing phase. This allows us to assess whether the epistemic and random strategies can learn models that can support chemotactic behaviour, and ensures that any observed differences are determined solely by attributes of the learned models.

Figure 3a shows the final distance from the source at the end of the testing phase, plotted against the duration of the preceding learning phase, and averaged over 300 learned models for each action strategy and learning duration. The final distance of the expected free energy, epistemic and random strategies decreases with the amount of time spent learning, meaning that these action strategies were able to learn models which support chemotactic behaviour. However, the instrumental strategy shows little improvement over baseline performance[9], irrespective of the amount of time spent learning. Models learned by the expected free energy strategy consistently finish close to the chemical source, and learn chemotactic behaviour after fewer learning steps relative to the other strategies.

3.3 Model accuracy

We now move on to consider whether learning in the presence of goal-oriented behaviour leads to models that are tailored to a behavioural niche. First, we assess how each action strategy affects the overall *accuracy* of the learned models. To test this, we measure the KL-divergence between the learned models and a 'true' model of agent-environment dynamics. Here, a 'true' model describes a model that has the same variables, structure and fixed parameters, but which has had infinite training data over all possible action-state contingencies [10]. Figure 4a shows the average accuracy of the learned models for each action strategy, plotted against the amount of time spent learning. These results demonstrate that the epistemic and random strategies consistently learn the most accurate models while the instrumental strategy consistently

[8] Formally, when there is no uncertainty in the likelihood distribution, state epistemic value reduces to the entropy of the predictive approximate posterior over s, see (Friston et al., 2015). For simplicity, we have omitted this term from the current simulations.

[9] Note that the first learning period consists of zero learning steps, meaning that the corresponding distance gives the (averaged) baseline performance for a randomly initialized model. This is less than the initial distance (400 units) as some of the randomly initialized models can support chemotaxis without any learning.

[10] We measure the accuracy of the *expectation* of the approximate posterior distribution over parameters θ, i.e. $\mathbb{E}[Q(\theta|\phi_\alpha)]$

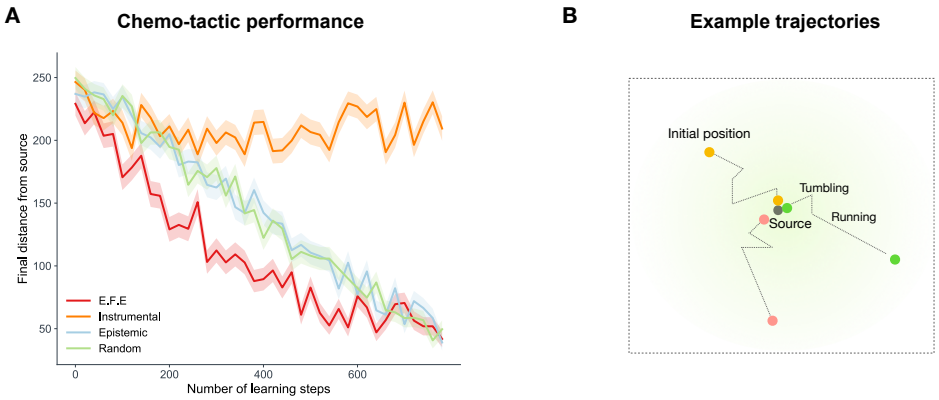

Figure 3: **(A) Chemotactic performance:** The average final distance from the chemical source after an additional testing phase, in which agents utilized the models learned in the corresponding learning phase. The average distance is plotted against the number of steps in the corresponding learning phase and is averaged over 300 models for each strategy and learning duration. Filled regions show +-SEM. **(B) Examples trajectories:** The spatial trajectories of agents who successfully navigated up the chemical gradient towards the chemical source.

learns the least accurate models. However, the expected free energy strategy learns a model that is significantly less accurate than both the epistemic and random strategies, indicating that the most well-adapted models are not necessarily the most accurate.

Figure 4a additionally suggests that the epistemic and random strategies learn equally accurate models. This result may appear surprising, as the epistemic strategy actively seeks out transitions that are expected to improve model accuracy. However, given the limited number of possible state transitions in the current simulation, it is plausible that a random strategy offers a near-optimal solution to exploration. To confirm this, we evaluated the accuracy of models learned by the epistemic and random strategies in high-dimensional state space. The results of this experiment are given in appendix 6, where it can be seen that the epistemic strategy does indeed learn models that are considerably more accurate than the random strategy.

We hypothesized that the expected free energy and instrumental strategies learned less accurate models because they were acting in a goal-oriented manner while learning. This, in turn, may have caused these strategies to selectively sample particular (behaviourally-relevant) transitions, at the cost of sampling other (behaviourally-irrelevant) transitions less frequently. To confirm this, we measured the distribution of state transitions sampled by each of the strategies after 1000 time steps learning, averaged over 300 agents. Because agents learn an *action-conditioned representation* of state transitions, i.e. $P(s_t|s_{t-1}, u_{t-1}, \theta)$, we separate state transitions that follow agents running from those that follow agents tumbling. These results are shown in figure 4b. For the epistemic and random strategies, the distribution is uniformly spread over (realizable) state transitions (running-induced transitions from positive to negative and negative to positive gradients are rare for all strategies, as such transitions can only occur in small portions of the environment). In contrast, the distributions sampled by the expected free energy and instrumental strategies are heavily biased towards a running-induced transition from positive to positive gradients. This is the transition that occurs when an agent is 'running up the chemical gradient', i.e., performing chemotaxis. The bias means that the remaining state transitions are sampled less, relative to the epistemic and random strategies.

How do the learned models differ, among the four action strategies? To address this question, we measured the post-learning change in different distributions of the full model. This change reflects a measure of 'how much' an agent has learned about that particular distribution. As described in the Methods, the full transition model $P(s_t|s_{t-1}, u_{t-1}, \theta)$

is composed of four separate categorical distributions. The first describes the effects of tumbling in negative gradients, the second describes the effects of tumbling in positive gradients, the third describes the effects of running in negative gradients, and fourth describes the effects of running in positive gradients. Figure 4c plots the KL-divergence between each of the original (randomly-initialized) distributions and the subsequent (post-learning) distributions. These results show that the expected free energy and instrumental strategies learn substantially less about three of the distributions, compared to the epistemic and random agents, explaining the overall reduction of accuracy displayed in figure 4a. However, for the distribution describing the effects of running in positive gradients, the instrumental strategy learns as much as the epistemic and random strategies, while the expected free energy strategy learns substantially more. These results, therefore, demonstrate that acting in a goal-oriented manner biases an agent to preferentially sample particular (goal-relevant) transitions in the environment and that this, in turn, causes agents to learn more about these (goal-relevant) transitions.

To further verify this result, we repeated the analysis described in figure 4b and 4c, but for the case where agents learn in the presence of reversed prior preferences (i.e. the agents believe that observing *negative* chemical gradients is *a-priori* more likely, and thus preferable). The results for these simulations are shown in 4d and 4e, where it can be seen that the expected free energy and instrumental strategy now preferentially sample running-induced transitions from negative to negative gradients, and learn more about the distribution describing the effects of running in negative gradients. This is the distribution relevant to navigating *down* the chemical gradient, a result that is expected if the learned models are biased towards prior preferences. By contrast, the models learned by the epistemic and random agents are not dependent on their prior beliefs or preferences.

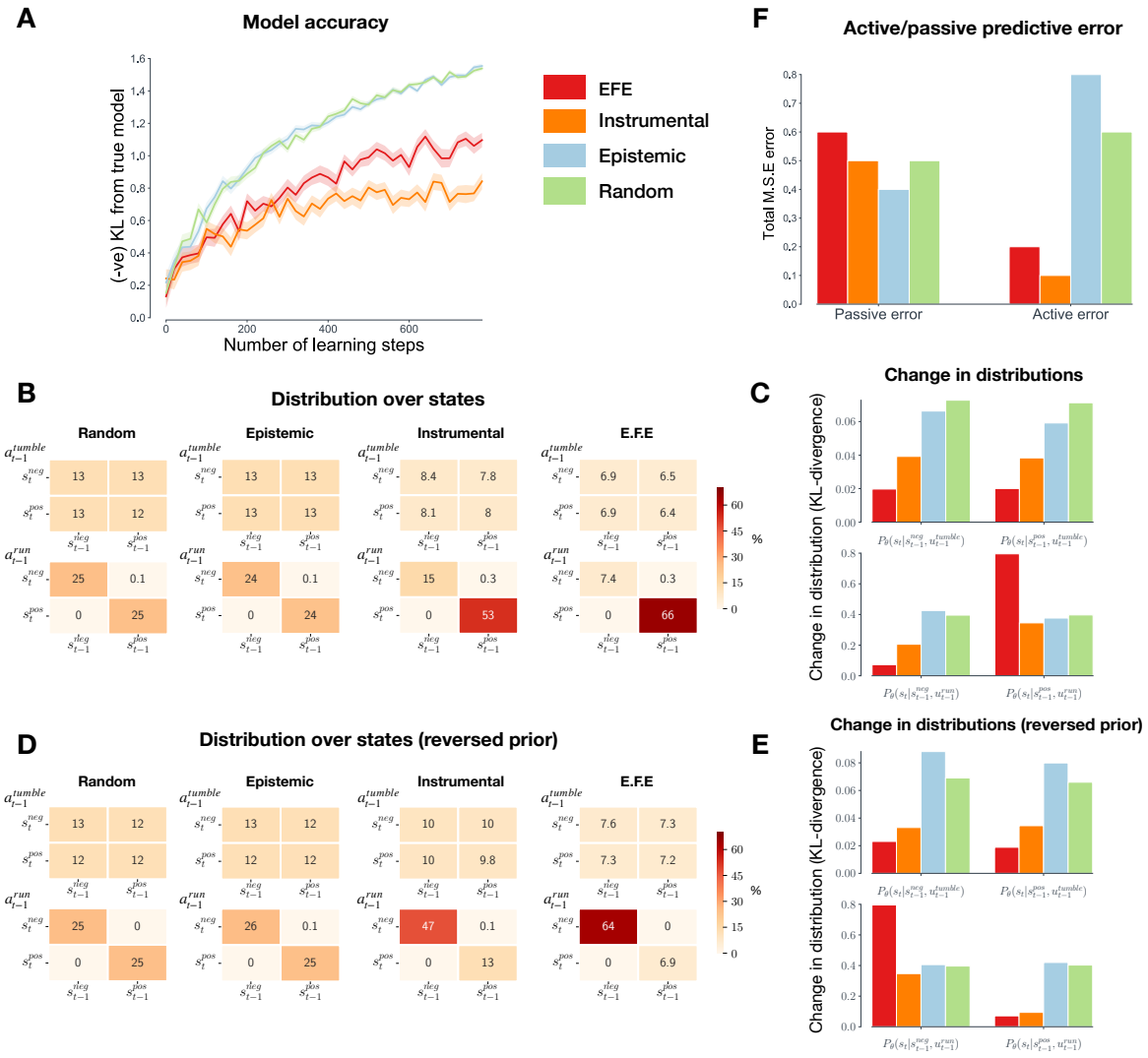

Figure 4: **(A) Model accuracy:** The average *negative* model accuracy, measured as the KL-divergence from a 'true' model of agent-environment dynamics. The accuracy is plotted against the number of steps in the corresponding learning phase and is averaged over 300 models for each strategy. Filled regions show +-SEM. **(B) Distributions of state transitions:** The distribution of action-dependent state transitions for each strategy over 1000 learning steps, averaged over 300 models for each strategy. Here, columns indicate the state at the previous time step, whereas rows indicate the state following the transition. The top matrices display transitions that follow from tumbling, whereas the bottom matrices display transitions that follow from running. **(C) Change in distributions:** The average change in each of the distributions of the full learned model, measured as the KL-divergence between the original (randomly-initialized) distributions and the final (post-learning) distribution. Refer to Methods section for a description of these distributions. **(D & E) Reversed preferences:** These results are the same as for panels B & C, but for the case where agents have reversed preferences (i.e. priors). Here, agents believe running *down* chemical gradients to be more likely. The results demonstrate that the models of expected free energy and instrumental agent are sensitive to prior preferences. **(F) Active/passive prediction error:** The cumulative mean squared error of counterfactual predictions about state transitions, over 1000 steps learning and averaged over 300 agents. The active condition describes predictions of state-transitions following self-determined actions, whereas the passive condition describes predictions following random actions.

3.4 Active and passive accuracy

The previous results suggest that learning in the presence of goal-directed behaviour leads to models that are biased towards certain patterns of agent-environment interaction. To further elucidate this point, we distinguish between *active accuracy* and *passive accuracy*. We define active accuracy as the accuracy of a model in the presence of the agents own self-determined actions (i.e. the actions chosen according to the agent's strategy), and passive accuracy as the accuracy of a model in the presence of random actions. We measured both the passive and active accuracy of the models learned under different action strategies following 300 time-steps of learning. To do this, we let agents act in their environment for an additional 1000 time steps according to their action strategy, and, at each time step, measured the accuracy of their counterfactual predictions about state transitions. In the active condition, agents predicted the consequence of a self-determined action, whereas, in the passive condition, agents predicted the consequence of a randomly selected action. We then measured the mean squared error between the agents' predictions and the 'true' predictions (i.e. the predictions given by the 'true' model, as described for figure 4a). The accumulated prediction errors for the passive and active conditions are shown in figure 4f, averaged over 300 learned models for each strategy. As expected, there is no difference between the passive and active condition for the random strategy, as this strategy selects actions at random. The epistemic strategy shows the highest active error, which is due to the fact that the epistemic strategy seeks out novel (and thus less predictable) transitions. The instrumental strategy has the lowest active prediction error, and therefore the highest active accuracy. This is consistent with the view that learning in the presence of goal-directed behaviour allows agents to learn models that are accurate in the presence of their self-determined behaviour. Finally, the expected free energy strategy has an active error that is lower than the epistemic and random strategies, but higher than the instrumental strategy. This arises from the fact that the expected free energy strategy balances both goal-directed and epistemic actions. Note that, in the current context, active accuracy is improved at the cost of passive accuracy. While the instrumental strategy learns the least accurate model, it is the most accurate at predicting the consequences of its self-determined actions

3.5 Pruning parameters

We now consider whether learning in the presence of goal-directed behaviour leads to *simpler* models of agent-environment dynamics. A principled way to approach this question is to ask whether each of the model's parameters are increasing or decreasing the Bayesian *evidence* for the overall model, which provides a measure of both the *accuracy* and the *complexity* of a model. In brief, if a parameter decreases model evidence, then removing - or 'pruning' - that parameter results in a model with higher evidence. This procedure can, therefore, provide a measure of how many 'redundant' parameters a model has, which, in turn, provides a measure of the complexity of a model (assuming that redundant parameters can, and should, be removed). We utilise the method of *Bayesian model reduction* (Friston et al., 2016b) to evaluate the evidence for models with removed parameters. This procedure allows us to evaluate the evidence for reduced models without having to refit the model's parameters.

We first let each of the strategies learn a model for 500 time-steps. The parameters optimized during this learning period are then treated as priors for an additional (i.e., post-learning) testing phase. During this testing phase, agents act according to their respective strategies for an additional 500 time-steps, resulting in posterior estimates of the parameters.

Given the prior parameters α and posterior parameters ϕ_α, we can evaluate an approximation for the change in model evidence under a reduced model through the equation:

$$\Delta\mathcal{F} = \ln \mathbf{B}(\phi_\alpha) + \ln \mathbf{B}(\alpha') - \ln \mathbf{B}(\alpha) - \ln \mathbf{B}(\phi_\alpha + \alpha' - \alpha) \tag{8}$$

where $\ln \mathbf{B}(\cdot)$ is the beta function, α' are the prior parameters of the reduced model, and $\mathcal{F}$ is the variational free energy, which provides a tractable approximation of the Bayesian model evidence. See (Friston et al., 2017) for a derivation of equation 8. If $\Delta\mathcal{F}$ is positive, then the reduced model - described by the reduced priors α' - has less evidence than the full model, and *vice versa*. We remove each of the prior parameters individually by setting their value

to zero and evaluate equation 8. Figure 5a shows the percentage of trials that each parameter was pruned for each of the action strategies, averaged over 300 trials for each strategy. For the instrumental and epistemic agents, the parameters describing the effects of running in negative gradients and tumbling in positive gradients are most often pruned, as these are the parameters that are irrelevant to chemotaxis (which involves running in positive chemical gradients and tumbling in negative chemical gradients). In figure 5b we plot the total number of parameters pruned, averaged over 300 agents. These results demonstrate that the expected free energy strategy entails models that have the highest number of redundant parameters, followed by the instrumental strategy. Under the assumption that redundant parameters can, and should, be pruned, the expected free energy and instrumental strategies learn simpler models, compared to the epistemic and random strategies.

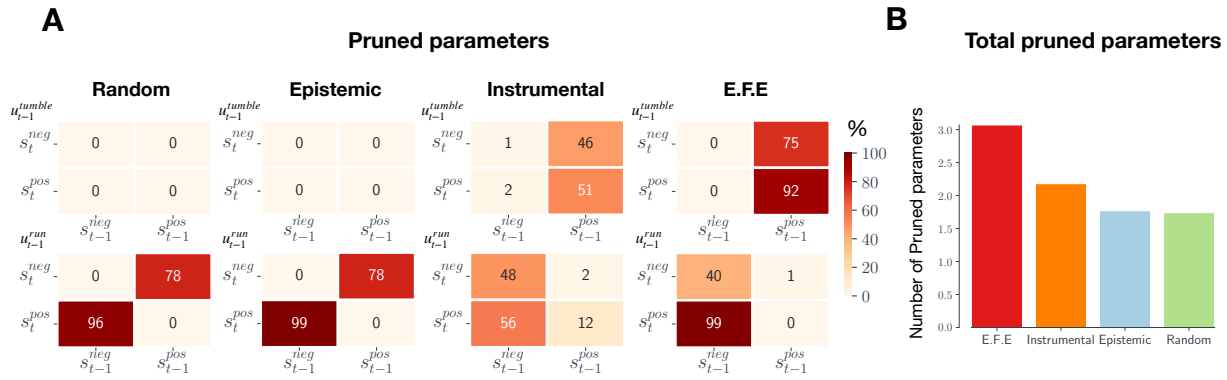

Figure 5: **(A) Number of pruned parameters:** Percentage of times each parameter was pruned, averaged over 300 agents. A parameter was pruned if it *decreased* the evidence for agents model. **(B) Total pruned parameters:** The average number of total number of pruned parameters, averaged over 300 agents.

3.6 Bad bootstraps and sub-optimal convergence

In the Introduction, we hypothesized that 'bad-bootstraps' occur when agents (and their models) become stuck in maladaptive cycles of learning and control, resulting in an eventual failure to learn well-adapted models. To test for the presence of bad-bootstraps, we allowed agents to learn models over an extended period of 4,000-time steps. We allowed this additional time to exclude the possibility that opportunities to learn had not been fully exploited by agents. (We additionally conducted the same experiment with 10,000-time steps; results were unchanged). We then tested the learned models on their ability to support chemotaxis, by allowing them to interact with their environment for an additional 1,000 time-steps using the expected free energy action strategy. To quantify whether the learned models were able to perform chemotaxis in any form, we measured whether the agent had moved more than 50 units towards the source by the end of the testing period.

After 4,000 learning steps, all the agents that had learned models using the expected free energy, epistemic or random strategies were able to perform at least some chemotaxis. In contrast 36% of the agents that had learned models under maximization of instrumental value did not engage in any chemotaxis at all. To better understand why instrumental agents frequently failed to learn well-adapted models, even after significant learning, we provide an analysis of a randomly selected failed model. This model prescribes a behavioural profile whereby agents continually tumble, even in positive chemical gradients. This arises from the belief that tumbling is more likely to give rise to positive gradients, even when the agent is in positive gradients. In other words, the model encodes the erroneous belief that, in positive gradients, running will be less likely to give rise to positive chemical gradients, relative to tumbling. Given this belief,

the agent continually tumbles, and therefore never samples information that disconfirms this maladaptive belief. This exemplifies a 'bad bootstrap' arising from the goal-directed nature of the agent's action strategy.

Finally, we explore how assigning epistemic value to actions can help overcome bad bootstraps. We analyse an agent which acts to minimize expected free energy, quantifying the relative contributions of epistemic and instrumental value to running and tumbling. We initialize an agent with a randomly selected maladapted model and allow the agent to interact with (and learn from) the environment according to the expected free energy action strategy. In Figure 6a, we plot the (negative) expected free energy of the running and tumbling control states over time, along with the relative contributions of instrumental and epistemic value. These results show that the (negative) expected free energy for the tumble control state is initially higher than that of the running control state because the agent believes there is less instrumental value in running. This causes the agent to tumble, which in turn causes the agent to gather information about the effects of tumbling. Consequently, the model becomes less uncertain about the expected effects of tumbling, thereby decreasing the epistemic value of tumbling (and thus the (negative) expected free energy of tumbling). This continues until the negative expected free energy of tumbling becomes less than that of running, which has remained constant (since the agent has not yet gained any new information about running). At this point, the agent infers running to be the more likely action, which causes the agent to run. The epistemic value of running now starts to decrease, but as it does so the new sampled observations disclose information that running is very likely to cause transitions from positive to positive gradients (i.e., to maintain positive gradients). The instrumental value of running (and thus the negative expected free energy of running) therefore sharply increases in positive gradients, causing the agent to continue to run in positive gradients. Note that this agent did not fully resolve its uncertainty about tumbling. This highlights the fact that, under active inference, the epistemic value of an action is contextualized by current instrumental imperatives.

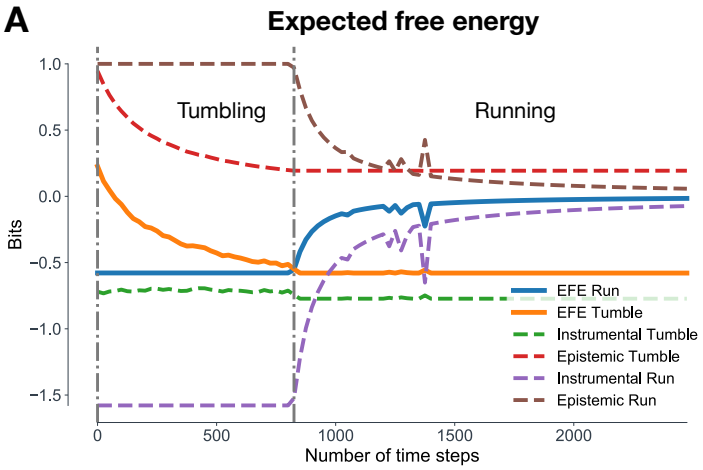

Figure 6: **(A) Expected free energy:** a plot of expected free energy for run and tumble control states overtime for an agent with an initially maladapted model. This model encodes the erroneous belief that running is less likely to give rise to positive chemical gradients, relative to tumbling. Therefore, at the start of the trial, the instrumental value of tumbling (green dotted line) is higher than the instrumental value of running (purple dotted line). The epistemic value of both running and tumbling (brown and red dotted lines, respectively) is initially the same. As the (negative) expected free energy for tumbling (orange line) is higher than the (negative) expected free energy for running (blue line), the agent tumbles for the first 900 time steps. During this time, agents gain information about the effects of tumbling, and the epistemic value of tumbling decreases, causing the negative expected free energy for tumbling to also decrease. This continues until the negative expected free energy is for tumbling is lower than the negative expected free energy for running, which has remained constant. Agents then run and gather information about the effects of running. This causes the epistemic value of running to decrease, but also causes the instrumental value of running to sharply increase, as the new information disconfirms their erroneous belief that running will not give rise to positive gradients.

4 Discussion

Equipping agents with generative models provides a powerful solution to prescribing well-adapted behaviour in structured environments. However, these models must, at least in part, be learned. For behaving agents - i.e., biological agents - the learning of generative models necessarily takes place in the presence of actions; i.e., in an 'online' fashion, during ongoing behaviour. Such models must also be geared towards prescribing actions that are useful for the agent. How to learn such 'action-oriented' models poses significant challenges for both computational biology and model-based reinforcement learning (RL).

In this paper, we have demonstrated that the active inference framework provides a principled and pragmatic approach to learning adaptive action-oriented models. Under this approach, the minimization of expected free energy prescribes an intrinsic and context-sensitive balance between goal-directed (instrumental) and information-seeking (epistemic) behaviours, thereby shaping the learning of the underlying generative models. After developing the formal framework, we illustrated its utility using a simple agent-based model of bacterial chemotaxis. We compared three situations. When agents learned solely in the presence of goal-directed actions, the learned models were specialized to the agent's behavioural niche but were prone to converging to sub-optimal solutions, due to the instantiation of 'bad-bootstraps'. Conversely, when agents learned solely in the presence of epistemic (information-seeking) actions, they learned accurate models which avoided sub-optimal convergence, but at the cost of reduced sample efficiency due to the lack of behavioural specialisation.

Finally, we showed that the minimisation of expected free-energy effectively-balanced goal-directed and information-seeking actions, and that the models learned in the presence of these actions were tailored to the agent's behaviours and goal, and were also robust to bad-bootstraps. Learning took place efficiently, requiring fewer interactions with the environment. The learned models were also less complex, relative to other strategies. Importantly, models learned via active inference departed in systematic ways from a veridical representation of the environment's true structure. For these agents, the learned models supported adaptive behaviour not only in spite of, but *because of*, their departure from veridicality.

4.1 Learning action-oriented models: good and bad bootstraps

When learning generative models online in the presence of actions, there is a circular dynamic in which learning is coupled to behaviour. The (partially) learned models are used to specify actions, and these actions provide new data which is then used to update the model. This circular dynamic (or 'information self-structuring' (Montúfar et al., 2015)) raises the potential for both 'good' and 'bad' bootstraps.

If actions are selected based purely on (expected) instrumental value, then the resulting learned models will be biased towards an agent's behavioural profile and goals (or prior preferences under the active inference framework - see Figure 4c & 4e), but will also be strongly constrained by the model's initial conditions. In our simulations, we showed that learning from instrumental actions was prone to the instantiation of 'bad-bootstraps'. Specifically, we demonstrated that these agents typically learned an initially maladapted model due to insufficient data or sub-optimal initialisation, and then subsequently used this model to determine goal-directed actions. This resulted in agents engaging with the environment in a sub-optimal and biased manner, thereby reintroducing sub-optimal data and causing models to become entrenched within local minima. Recent work in model-based RL has identified this coupling to be one of the major obstacles facing current model-based RL algorithms (Wang et al., 2019). More generally, it is likely that bad-bootstraps are a prevalent phenomenon whenever parameters are used to determine the data from which the parameters are learned. Indeed, this problem played a significant role in motivating the (now common) use of 'experience replay' in model-free RL (Mnih et al., 2013). Experience replay describes the method of storing past experiences to be later sampled from for learning, thus breaking the tight coupling between learning and behaviour.

In the context of online learning, one way to avoid bad-bootstraps is to select actions based on (expected) epistemic value (Schwartenbeck et al., 2018; Friston et al., 2017; Sun et al., 2011), where agents seek out novel interactions based on counterfactually informed beliefs about which actions will lead to informative transitions. By utilising the

uncertainty encoded by (beliefs about) model parameters, this approach can proactively identify optimally informative transitions. In our simulations, we showed that agents using this strategy learned models that asymptoted towards veridicality and, as such, were not tuned to any specific behavioural niche. This occurred because pure epistemic exploration treats all uncertainties as equally important, meaning that agents were driven to resolve uncertainty about all possible agent-environment contingencies. While models learned using this strategy were able to support chemotactic behaviour (Figure 3a), learning was highly sample-inefficient.

We have argued that a more suitable approach is to balance instrumental and epistemic actions in a principled way during learning. This is what is achieved by the active inference framework, via minimization of expected free energy. Minimizing expected free energy means that the model uncertainties associated with an agent's goals and desires are prioritised over those which are not. Furthermore, it means that model uncertainties are only resolved until an agent (believes that it) is sufficiently able to achieve its goals, such that agents need not resolve all of their model uncertainty. In our simulations, we showed that active inference agents learned models in a sample-efficient way, avoided being caught up in bad bootstraps, and generated well-adapted behaviour in our chemotaxis setting. Our data, therefore, support the hypothesis that learning via active inference provides a principled and pragmatic approach to the learning of well-adapted action-oriented generative models.

4.2 Exploration vs. exploitation

Balancing epistemic and instrumental actions recalls the well-known trade-off between exploration and exploitation in reinforcement learning. In this context, the simplest formulation of this trade-off can be construed as a model-free notion in which exploration involves random actions. One example of this simple formulation is the *ε-greedy* algorithm which utilises noises in the action selection process to overcome premature sub-optimal convergence (Watkins, 1989).

The balance between epistemic and instrumental actions in our active inference agents is more closely connected to the exploration-exploitation trade-off in model-based RL. As in our agents, model-based RL agents often employ exploratory actions that are selected to resolve model uncertainty. As we have noted, such actions can help avoid sub-optimal convergence (bad bootstraps), especially at the early stages of learning where data is sparse. However, in model-based RL it is normally assumed that, in the limit, a maximally comprehensive and maximally accurate (i.e., veridical) model would be optimal. This is exemplified by approaches that conduct an initial 'exploration' phase - in which the task is to construct a veridical model from as few samples as possible - followed by a subsequent 'exploitation' phase. By contrast, our approach highlights the importance of 'goal-directed exploration', in which the aim is not to resolve all uncertainty to construct a maximally accurate representation of the environment, but is instead to selectively resolve uncertainty until adaptive behaviour is (predicted to be) possible.

This kind of goal-directed exploration highlights an alternative perspective on the exploration-exploitation trade-off. We demonstrated that "exploitation" - traditionally associated with exploiting the agent's current knowledge to accumulate reward - can also lead to a type of constrained learning that leads to 'action-oriented' representations of the environment. In other words, our results suggest that, in the context of model-learning, the "explore-exploit" dilemma additionally entails an "explore-constrain" dilemma. This is granted a formal interpretation under the active inference framework - as instrumental actions are associated with soliciting observations that are consistent with the model's prior expectations. However, given the formal relationship between instrumental value in active inference and the Bellman equations (Friston et al., 2016a), a similar trade-off can be expected to arise in any model-based RL paradigm.

4.3 Model non-veridicality

In our simulations, models learned through active inference were able to support adaptive behaviour even when their *structure* and *variables* departed significantly from an accurate representation of the environment. By design, the models utilized a severely impoverished representation of the environment. An exhaustive representation would have required models to encode information about the agent's position, orientation, the position of the chemical source, as well as a spatial map of the chemical concentrations so that determining an adaptive action would require a complex transformation of these variables. In contrast, our model was able to support adaptive behaviour by simply encoding a

representation of the instantaneous effects of action on the local chemical gradient. Therefore, rather than encoding a rich and exhaustive internal mirror of nature, the model encoded a parsimonious representation of sensorimotor couplings that were relevant for enabling action (Baltieri and Buckley, 2019b). While this particular 'action-oriented' representation was built-in through the design of the generative model, it nonetheless underlines that models need not be homologous with their environment if they are to support adaptive behaviour.

By evaluating the number of 'redundant' model parameters (as evaluated through Bayesian model reduction), we further demonstrated that learning in the presence of goal-directed behaviour leads to models that were more parsimonious in their representation of the environment, relative to other strategies (Figure 5b). Moreover, we showed that this strategy leads to models that did not asymptote to veridicality, in terms of the accuracy of the model's parameters (Figure 4a). Interestingly, these agents nevertheless displayed high 'active accuracy' (i.e., the predictive accuracy in the presence of self-determined actions), highlighting the importance of contextualising model accuracy in terms of an agent's actions and goals.

While these results demonstrate that models can support adaptive behaviour in spite of their misrepresentation of the environment and that these misrepresentations afforded benefits in terms of sample efficiency and model complexity, the active inference framework additionally provides a mechanism whereby misrepresentation *enables* adaptive behaviour. Active inference necessarily requires an organism's model to include systematic misrepresentations of the environment, by virtue of the organism's existence. Specifically, an organism's generative model must encode a set of prior beliefs that distinguish it from its external environment. For instance, the chemotaxis agents in the current simulation encoded the belief that observing positive chemical gradients was *a-priori* more likely. From an objective and passive point of view, these prior beliefs are, by definition, false. However, these systematic misrepresentations can be realized through action, thereby giving rise to apparently purposeful and autopoietic behaviour. Thus, under active inference, adaptive behaviour is achieved *because of*, and not just in spite of, a models departure from veridicality (Wiese, 2017).

Encoding frugal and parsimonious models plausibly afford organism's several evolutionary advantages. First, the number of model parameters will likely correlate with the metabolic cost of that model. Moreover, simpler models will be quicker to deploy in the service of action and perception and will be less likely to overfit the environment. This perspective, therefore, suggests that the degree to which exhaustive and accurate models are constructed should be mandated by the degree to which they are necessary for on-going survival. If the mapping between the external environment and allostatic responses is complex and manifold, then faithfully modelling features of the environment may pay dividends. However, in the case that frugal approximations and rough heuristics can be employed in the service of adaptive behaviour, such faithful modelling should be avoided. We showed that such "action-oriented" models arise naturally under ecologically valid learning conditions, namely, learning online in the presence of goal-directed behaviour. However, action-oriented behaviour that was adapted to the agent's goals only arose under the minimisation of expected free energy.

4.4 Active inference

While any approach to balancing exploration and exploitation is amenable to the benefits described in the previous sections, we have focused on the normative principle of active inference. From a purely theoretical perspective, active inference re-frames the exploration-exploitation dilemma by suggesting that both exploration and exploitation are complementary perspectives on a single objective function - the minimization of expected free energy. However, an open question remains as to whether this approach provides a practical solution to balancing exploration and exploitation. On the one hand, it provides a practically useful recipe by casting both epistemic and instrumental value in the same (information-theoretic) currency. However, the balance will necessarily depend on the shape of the agent's beliefs about hidden states, beliefs about model parameters, and prior beliefs about preferable observations. In the current work, we introduced an artificial weighting term to keep the epistemic and instrumental value within the same range. The same effect could have been achieved by constructing the shape (i.e. variance) of the prior preferences $P(o)$.

Active inference also provides a suitable framework for investigating the emergence of action-oriented models. Previous work has highlighted the fact that active inference is consistent with, and necessarily prescribes, frugal and parsimonious

generative models, thus providing a potential bridge between 'representation-hungry' approaches to cognition espoused by classical cognitivism and the 'representation-free' approaches advocated by embodied and enactive approaches (Kiverstein, 2018; Kirchhoff and Robertson, 2018; Ramstead et al., 2019; Negru, 2018; Linson et al., 2018; Pezzulo et al., 2017; Clark, 2015; Williams, 2018; Kirchhoff Michael et al., 2018; Kirchhoff and Froese, 2017; Friston, 2013; Baltieri and Buckley, 2019a,c,b).

This perspective has been motivated by at least three reasons. First, active inference is proposed as a description of self-organization in complex systems (Friston, 2013). Deploying generative models and minimizing free energy are construed as emergent features of a more fundamental drive towards survival. On this account, the purpose of representation is not to construct a rich internal world model, but instead to capture the environmental regularities that allow the organism to act adaptively.

The second reason is that minimizing free energy implicitly penalizes the complexity of the generative model (see appendix 1). This implies that minimizing free energy will reduce the complexity (or parameters) required to go from prior beliefs to (approximately) posterior beliefs, i.e. in explaining some observations. This occurs under the constraint of accuracy, which makes sure that the inferred variables can sufficiently account for the observations. In other words, minimizing free energy ensures that organism's maximize the accuracy of their predictions while minimizing the complexity of the models that are used to generate those predictions.

As discussed in the previous section, active inference also *requires* agents to encode systematic misrepresentations of their environment. Our work has additionally introduced a fourth motivation for linking active inference to adaptive action-oriented models, namely, that the minimization of expected free energy induces a balance between self-sustaining (and thus constrained) patterns of agent-environment interaction and goal-directed exploration.

4.5 Conclusion

In this paper, we have demonstrated that the minimization of expected free energy (through active inference) provides a principled and pragmatic solution to learning action-oriented probabilistic models. These models can make the process of learning models of natural environments tractable, and provide a potential bridge between 'representation-hungry' approaches to cognition and those espoused by enactive and embodied disciplines. Moreover, we showed how learning online in the presence of behaviour can give rise to 'bad-bootstraps' - a phenomenon that has the potential to be problematic whenever learning is coupled with behaviour. Epistemic or information-seeking actions provide a plausible mechanism for overcoming bad-bootstraps. However, to exploration to be efficient, the epistemic value of actions must be contextualized by agents goals and desires. The ability to learn adapted models that are tailored to action provides a potential route to tractable and sample efficient learning algorithms in a variety of contexts, including computational biology and model-based RL.

Acknowledgements

AT is funded by a PhD studentship from the Sackler Foundation and the School of Engineering and Informatics at the University of Sussex. CLB is supported by BBRSC grant number BB/P022197/1. We are grateful to the Dr. Mortimer and Theresa Sackler Foundation, which supports the Sackler Centre for Consciousness Science. AKS is additionally grateful to the Canadian Institute for Advanced Research (Azrieli Programme on Brain, Mind, and Consciousness). For helpful comments on earlier versions of this manuscript we are grateful to Conor Heins and Brennan Klein.

5 Methods

5.1 The Generative Model

The agent's generative model specifies the joint probability over observations o, hidden state variables s, control variables u and parameter variables θ. To account for temporal dependencies among variables, we consider a generative model that is over a sequence of variables through time, i.e. $\tilde{x} = \{x_1, ..., x_t\}$, where tilde notation indicates a sequence from time $t = 0$ to the current time t, and x_t denotes the value of x at time t. The generative model is given by the joint probability distribution $P(\tilde{o}, \tilde{s}, \tilde{u}, \theta|\lambda, \alpha)$, where:

$$P(\tilde{o}, \tilde{s}, \tilde{u}, \theta|\lambda, \alpha) = P(\theta|\alpha) \prod_{t=1}^{T} P(o_t|s_t, \lambda) P(s_t|s_{t-1}, u_{t-1}, \theta) P(u_t)$$

$$\begin{aligned} P(o_t|s_t, \lambda) &= \mathbf{Cat}(\lambda) \\ P(s_t|s_{t-1}, u_{t-1}, \theta) &= \mathbf{Cat}(\theta) \\ P(\theta|\alpha) &= \mathbf{Dir}(\alpha) \\ P(u_t) &= \sigma(-\tilde{\mathbf{G}}) \end{aligned} \tag{9}$$

where $\sigma(\cdot)$ is the softmax function. For simplicity, we initialize $P(s_{t=0})$ as a uniform distribution, and therefore exclude it from equation 9.

The likelihood distribution specifies the probability of observing some chemical gradient o_t given a belief about the chemical gradient s_t. This distribution is described by a set of categorical distributions, denoted $\mathbf{Cat}(\cdot)$, where each categorical distribution is a distribution over k discrete and exclusive possibilities. The parameters of a categorical distribution can be represented as a vector with each entry describing the probability of some event p_i, with $\sum_{i=1}^{k} p_i = 1$. As the likelihood distribution is a conditional distribution, a separate categorical distribution is maintained for each hidden state in $\mathcal{S}$, (i.e. $s^{\texttt{pos}}$ and $s^{\texttt{neg}}$), where each of these distributions specifies the conditional probability of observing some chemical gradient (either $o^{\texttt{pos}}$ and $o^{\texttt{neg}}$). The parameters of the likelihood distribution can therefore be represented as a 2 x 2 matrix where each column j is a categorical distribution that describes $P(o_t|s_t = j, \lambda)$. For the current simulations, we provide agents with the parameters λ and do not require these parameters to be learned. The provided parameters encode the belief that there is an unambiguous mapping between $s^{\texttt{pos}}$ and $o^{\texttt{pos}}$, and between $s^{\texttt{neg}}$ and $o^{\texttt{neg}}$, meaning that λ can be encoded as an identity matrix.

The prior probability over hidden states s_t is given by the transition distribution $P(s_t|s_{t-1}, u_{t-1}, \theta)$, which specifies the probability of the current hidden state, given beliefs about the previous hidden state and the previous control state. In other words, this distribution describes an agent's beliefs about how running and tumbling will cause changes in the chemical gradient. Following previous work (Friston et al., 2015), we assume that agents know which control state was executed at the previous time step. As with the likelihood distribution, the prior distribution is described by a set of categorical distributions. Each categorical distribution j specifies the probability distribution $P(s_t|s_{t-1} = j, \theta)$, such that $P(s_t|s_{t-1}, \theta)$ can again be represented as a 2 x 2 matrix. However, the transition distribution is also conditioned on control states u, meaning a separate transition matrix is maintained for both $u^{\texttt{run}}$ and $u^{\texttt{tumble}}$, such that the transition distribution can be represented as a 2 x 2 x 2 tensor. Agents, therefore, maintain separate beliefs about how the environment is likely to change for each control state.

We require agents to learn the parameters θ of the transition distribution. At the start of each learning period, we randomly initialize θ, such that agents start out with random beliefs about how actions cause transitions in the chemical gradient. To enable these parameters to be learned, the generative model encodes (time-invariant) prior beliefs over θ in the distribution $P(\theta|\alpha)$. This distribution is modelled as Dirichlet distribution, denoted $\mathbf{Dir}(\cdot)$, where α are the parameters of this distribution. A Dirichlet distribution represents a distribution *over* the parameters of a distribution.

In other words, sampling from this distribution returns a vector of parameters, rather than a scalar. By maintaining a distribution over θ, the task of learning about the environment is transformed into a task of inferring unknown variables.

Finally, the prior probability of control states is proportional to a softmax transformation of $-\tilde{\mathbf{G}}$, which is a vector of (negative) expected free energies, with one entry for each control state. This formalizes the notion that control states are *a-priori* more likely if they are expected to minimize free energy. We provide a full specification of expected free energy in the following sections.

5.2 The Approximate Posterior

The approximate posterior encodes an agent's current approximately posterior beliefs about the chemical gradient s, the control state u and model parameters θ. As with the generative model, the approximate posterior is over a sequence of variables $Q(\tilde{s}, \tilde{u}, \theta|\phi)$, where ϕ are the sufficient statistics of the distribution.

In order to make inference tractable, we utilize the mean-field approximation to factorize the approximate posterior. This approximation treats a potentially high-dimensional distribution as a product of a number of simpler marginal distributions. Heuristically, this treats certain variables as statistically independent. Practically, it allows us to infer individual variables while keeping the remaining variables fixed. This approximation makes inference tractable, at the (potential) price of making inference sub-optimal. For inference to be optimal, the factorization of the approximate posterior must match the factorization of the true posterior.

Here, we factorize over time, the beliefs about the chemical gradient, the beliefs about model parameters and the beliefs about control states:

$$
\begin{aligned}
Q(\tilde{s}, \tilde{u}, \theta|\phi) &= Q(\theta|\phi_\alpha) \prod_{t=0}^{T} Q(s_t|\phi_{s_t}) Q(u_t|\phi_{u_t}) \\
Q(\theta|\phi_\alpha) &= \mathbf{Dir}(\phi_\alpha) \\
Q(s_t|\phi_{s_t}) &= \mathbf{Cat}(\phi_{s_t}) \\
Q(u_t|\phi_{u_t}) &= \mathbf{Cat}(\phi_{u_t})
\end{aligned}
\tag{10}
$$

5.3 Inference, Learning and Action

Having defined the generative model and the approximate posterior, we can now specify how free energy can be minimized. In brief, this involves updating the sufficient statistics of the approximate posterior ϕ as new observations are sampled. To minimize free energy, we identify the derivative of free energy with respect to the sufficient statistics $\frac{\partial \mathcal{F}(\phi,o)}{\partial \phi}$, solve for zero, i.e. $\frac{\partial \mathcal{F}(\phi,o)}{\partial \phi} = 0$, and rearrange to give the variational updates that minimize free energy. Given the mean-field assumption, we can perform this scheme separately for each of the partitions of ϕ, i.e ϕ_{s_t}, ϕ_{u_t} and ϕ_α

For the current scheme, the update equations for the hidden state parameters ϕ_s are (see Appendix 5 for a full derivation):

$$
\phi_{s_t} = \sigma(\ln P(o_t|s_t, \lambda) + \ln P(s_t|s_{t-1}, u_{t-1}, \theta)) \tag{11}
$$

This equation corresponds to state estimation or 'perception' and can be construed as a Bayesian filter that combines the likelihood of the current observation with a prior belief that is based on the previous hidden state and the previous

control state. To implement this update in practice, we rewrite equation 11 in terms of the relevant parameters and sufficient statistics (see Appendix 5):

$$\phi_{s_t} = \sigma(\ln \lambda \cdot \vec{o_t} + \bar{\theta}^{u_{t-1}} \cdot \phi_{s_{t-1}})$$

$$\begin{aligned} \bar{\theta}^{u_{t-1}} &= \mathbb{E}_{Q(\theta|\phi_\alpha)}[\ln \theta^{u_{t-1}}] \\ &= \psi(\phi_{\alpha_{ij}}^{u_{t-1}}) - \psi(\sum_{i=1}^{n} \phi_{\alpha_j}^{u_{t-1}}) \end{aligned} \tag{12}$$

Here, $\vec{o_t}$ is a one-hot encoded vector specifying the current observation, θ^u specifies the transition distribution corresponding to control state u, and $\psi(\cdot)$ is the digamma function. Note that the parameters of the likelihood distribution λ are point-estimates of a categorical distribution, meaning it is possible to straightforwardly take the logarithm of this distribution. However, the beliefs about θ are described by the Dirichlet distribution $Q(\theta|\alpha)$, meaning that the mean of the logarithm of this distribution (denoted $\bar{\theta}$) must be evaluated (leading to lines two and three of equation 12).

Learning can be conducted in a similar manner by updating the parameters ϕ_α (see Appendix 5 for a full derivation):

$$\phi_\alpha^u = \alpha^u + \sum_{t=1}^{T} [a_{t-1} = u_{t-1}] \cdot \xi \phi_{s_t} \phi_{s_{t-1}} \tag{13}$$

where $[\cdot]$ is an inversion bracket that returns one if the statement inside the bracket is true and zero otherwise, and ξ is an artificial learning rate, set to 0.001 for all simulations. Note that we update the parameters ϕ_α after each iteration, but use a small learning rate to simulate the difference in time scales implied by the factorization of the generative model and approximate posterior. This update bears a resemblance to Hebbian plasticity, in the sense that the probability of each parameter increases if the corresponding transition is observed (i.e. 'fire together wire together').

Finally, actions can be inferred by updating the parameters ϕ_{u_t}, where the update is given by (see Appendix 5 for a full derivation):

$$\phi_{u_t} = \sigma(-\tilde{\mathbf{G}}) \tag{14}$$

This equation demonstrates that the (approximately) posterior beliefs over control states are proportional to the vector of negative expected free energies. In other words, the posterior and prior beliefs about control states are identical.

5.4 Expected free energy

In this section, we describe how to evaluate the vector $-\tilde{\mathbf{G}}$. This is a vector of negative expected free energies, with one for each control state $u \in \mathcal{U}$. As specified in the formalism, the negative expected free energy for a single control state is defined as $-\mathbf{G}_\tau(u_t)$, where τ is some future time point, and, for the current simulations:

$$\begin{aligned} -\mathbf{G}_\tau(u_t) = &\underbrace{\mathbb{E}_{Q(o_\tau, s_\tau, \theta|u_t, \phi_\tau)}\big[\ln Q(\theta|s_\tau, o_\tau, u_t, \phi_\tau) - \ln Q(\theta|\phi_\tau)\big]}_{\text{Parameter epistemic value}} \\ &+ \underbrace{\mathbb{E}_{Q(o_\tau, s_\tau, \theta|u_t, \phi_\tau)}[\ln P(o_\tau)]}_{\text{Instrumental value}} \end{aligned} \tag{15}$$

As described in the results section, we ignore the epistemic value for hidden states, as there is no uncertainty in the likelihood distribution. Moreover, for all simulations, $\tau = t + 1$, such that we only consider the immediate the immediate effects of action. This scheme is, however, entirely consistent with a sequence of actions, i.e. a policy.

In order to evaluate expected free energy, we rewrite equation 15 in terms of parameters. By noting that $\mathbb{E}_{Q(o_\tau,s_\tau,\theta|u_t,\phi_\tau)}[\ln P(o_\tau)] = \mathbb{E}_{Q(o_\tau|u_t,\phi_\tau)}[\ln P(o_\tau)]$, we can write instrumental value as:

$$\mathbb{E}_{Q(o_\tau|u_t,\phi_\tau)}[\ln P(o_\tau)] = \phi_{o_\tau} \cdot \rho \tag{16}$$

where ϕ_{o_τ} are the sufficient statistics of $Q(o_\tau|u_t,\phi_\tau)$, and ρ are the parameters of $P(o_\tau)$, which is a categorical distribution, such that ρ is a vector with one entry for each $o \in \mathcal{O}$. In order to evaluate parameter epistemic value, we utilise the following approximation:

$$\begin{aligned}\mathbb{E}_{Q(o_\tau,s_\tau,\theta|u_t,\phi_\tau)}\big[\ln Q(\theta|s_\tau,o_\tau,u_t,\phi_\tau) - \ln Q(\theta|\phi_\tau)\big] &\approx \phi_{s_\tau} \cdot \mathbf{W}^{u_t} \cdot \phi_{s_t} \\ \mathbf{W}^{u_t} &= \sum_{i=1}^{n} \phi_{\alpha_j}^{-1} - \phi_\alpha^{-1}\end{aligned} \tag{17}$$

For details of this approximation, we refer the reader to (Friston et al., 2017). For a given control state u_t, negative expected free energy can, therefore, be calculated as:

$$-\mathbf{G}_\tau(u_t) = \phi_{s_\tau} \cdot \mathbf{W}^{u_t} \cdot \phi_{s_t} + \delta(\phi_{o_\tau} \cdot \rho) \tag{18}$$

where ϕ_{s_τ} are the sufficient statistics of $Q(s_\tau|u_t,\phi_\tau)$ and δ is an optional weighting term. For all simulations, this is set to $1/10$. To calculate equation 18, it is first necessary to evaluate the expected beliefs $Q(s_\tau|u_t,\phi_\tau)$ and $Q(o_\tau|u_t,\phi_\tau)$. The expected distribution over hidden states $Q(s_\tau|u_t,\phi_\tau)$ is given by $\mathbb{E}_{Q(s_t|u_t,\phi_\tau)}\big[P(s_\tau|s_t,u_t,\theta)\big]$. Given these beliefs over future hidden states, we can evaluate $Q(o_\tau|u_t,\phi_\tau)$ as $\mathbb{E}_{Q(s_\tau|u_t,\phi_\tau)}\big[P(o_\tau|s_\tau,\lambda)\big]$.

The full update scheme for the agents is provided in algorithm 1:

Algorithm 1 Active inference MDP algorithm

Require: parameters of likelihood distribution λ, parameters of prior distribution over transition distribution parameters α, prior probability of observations ρ

1: **for** t in T **do**
2: $o_t \leftarrow$ `env.observe()` ▷ Sample observation from environment
3: $\phi_{s_t} = \sigma(\ln \lambda \cdot \vec{o_t} + \bar{\theta}^{u_{t-1}} \cdot \phi_{s_{t-1}})$ ▷ Hidden state inference
4: $\phi_{u_t} = \sigma(-\vec{\mathbf{G}})$ ▷ Control state inference
5: **where** $-\mathbf{G}_\tau(u_t) = \underbrace{\underbrace{\phi_{s_\tau} \cdot \mathbf{W}^{u_t} \cdot \phi_{s_t}}_{\text{Epistemic agent}} + \underbrace{\phi_{o_\tau} \cdot \rho}_{\text{Instrumental agent}}}_{\text{Expected free energy agent}}$
6: $\phi_\alpha^u = \alpha^u + \sum_{t=1}^{T}[a_{t-1} = u_{t-1}] \cdot \xi \phi_{s_t} \phi_{s_{t-1}}$ ▷ Learning inference
7: $a_t \sim Q(u_t|\phi_{u_t})$ ▷ Sample action
8: `env.update(`a_t`)` ▷ Perform action
9: **end for**

6 Appendix 1

In this appendix, we provide three rearrangements of the free energy functional $\mathcal{F}(\phi, o)$, thus providing an intuition as to what the minimization of free energy entails. By defining free energy $\mathcal{F}(\phi, o)$ as the KL-divergence between an approximate posterior $Q(x|\phi)$ and a generative model $P(x, o)$, we can write:

$$\begin{aligned}\mathcal{F}(\phi, o) &= \mathbb{KL}[Q(x|\phi)||P(x, o)] \\ &= \mathbb{E}_{Q(x|\phi)}[\ln Q(x|\phi) - \ln P(x, o)]\end{aligned} \tag{19}$$

where the second equality is simply the definition of the KL-divergence, $\mathbb{KL}[Q||P] = \mathbb{E}_Q[\ln Q - \ln P]$. We can factorize the generative model $P(x, o)$ as $P(x|o)P(o)$, allowing us to rewrite equation 19 as:

$$\begin{aligned}\mathcal{F}(\phi, o) &= \mathbb{E}_{Q(x|\phi)}[\ln Q(x|\phi) - \ln P(x|o)] - \ln P(o) \\ &= \mathbb{KL}[Q(x|\phi)||P(x|o)] - \ln P(o)\end{aligned} \tag{20}$$

Note that the negative log-likelihood of observations $-\ln P(o)$ remains outside of the expectation in the first equality as $P(o)$ does not depend on $Q(x|\phi)$. The second equality demonstrates that free energy can be expressed as the KL-divergence between the approximate posterior $Q(x|\phi)$ and the true posterior distribution $P(x|o)$, minus the log-likelihood of observations $\ln P(o)$. As the KL-divergence is a strictly non-negative quantity, free energy will always be greater than or equal to the negative log-likelihood of observations, $\mathcal{F}(\phi, o) \geq -\ln P(o)$. This means that free energy will be equal to to the negative log-likelihood of observations when the posterior divergence term is equal to zero. Therefore, free energy is an *upper bound* on the negative log-likelihood of observations, a quantity sometimes referred to as *surprisal*. Minimizing free energy will, therefore, minimize surprisal, or equivalently, maximize Bayesian model evidence $P(o)$.

An alternative expression of free energy can be derived through an alternative factorization of the generative model, $P(x, o) = P(o|x)P(x)$, allowing us to rewrite equation 19 as:

$$\begin{aligned}\mathcal{F}(\phi, o) &= \mathbb{E}_{Q(x|\phi)}[\ln Q(x|\phi) - \ln P(o|x) - \ln P(x)] \\ &= \mathbb{E}_{Q(x|\phi)}[\ln Q(x|\phi) - \ln P(x)] - \mathbb{E}_{Q(x|\phi)}[\ln P(o|x)] \\ &= \mathbb{KL}[Q(x|\phi)||P(x)] - \mathbb{E}_{Q(x|\phi)}[\ln P(o|x)]\end{aligned} \tag{21}$$

The final equality of equation 21 demonstrates that free energy can be expressed as the KL-divergence between the approximate posterior $Q(x|\phi)$ and the prior probability of unknown variables $P(x)$, minus the conditional log-probability of observations $P(o|x)$ expected under the approximate posterior. The first of these terms quantifies the *complexity* of the approximate posterior, as it measures how much the approximate posterior changed in order to account for some new observations (i.e. in going from prior to approximately posterior beliefs). The second term measures the *accuracy* of the approximate posterior, as it quantifies how likely the observations are, given the beliefs encoded by the approximate posterior. Therefore, minimizing free energy entails a trade-off between minimizing the complexity of the beliefs encoded by the approximate posterior and maximizing the accuracy of those beliefs.

Finally, we can rearrange free energy as:

$$\begin{aligned}\mathcal{F}(\phi, o) &= \mathbb{E}_{Q(x|\phi)}[\ln Q(x|\phi)] - \mathbb{E}_{Q(x|\phi)}[\ln P(x, o)] \\ &= -\mathbf{H}[Q(x|\phi)] - \mathbb{E}_{Q(x|\phi)}[\ln P(x, o)]\end{aligned} \tag{22}$$

where $\mathbf{H}[Q(x|\phi)]$ is the Shannon entropy of the approximate posterior, defined as $-\mathbb{E}_{Q(x|\phi)}[\ln Q(x|\phi)]$. The final equality demonstrates that minimizing free energy entails maximizing the entropy of the approximate posterior, while also maximizing the expected energy $\mathbb{E}_{Q(x|\phi)}[\ln P(x, o)]$. Maximizing the entropy of the approximate posterior ensures that the approximate posterior provides a generic and parsimonious explanation of the observed data, thereby ensuring that those explanations are not based on highly-specific (i.e. low-entropy) beliefs.

7 Appendix 2

In what follows, we formally describe the relationship between free energy $\mathcal{F}(\phi, o)$ and expected free energy $\mathbf{G}_\tau(\phi_\tau, u_t)$. To help clarify this relationship, we rewrite free energy as $\mathcal{F}_t(\phi_t, o_t)$, where subscript x_t implies the value at time t, thus making explicit the fact that free energy corresponds to the current time t, and where (following Appendix 1):

$$\mathcal{F}_t(\phi_t, o_t) = \mathbb{E}_{Q(x_t|\phi_t)}[\ln Q(x_t|\phi_t) - \ln P(x_t, o_t)] \tag{23}$$

Expected free energy $\mathbf{G}_\tau(\phi_\tau, u_t)$ differs from the free energy functional in equation 23 in three respects. First, it looks to quantify the free energy that is expected to occur at some future time τ. We can attempt to define the free energy for time τ as:

$$\mathcal{F}_\tau(\phi_\tau, o_\tau) = \mathbb{E}_{Q(x_\tau|\phi_\tau)}[\ln Q(x_\tau|\phi_\tau) - \ln P(o_\tau, x_\tau)] \tag{24}$$

However, equation 24 poses a problem. The free energy at the current time point $\mathcal{F}_t(\phi_t, o_t)$ is a *function* of observations o_t because the observations at time t are known. In contrast, the observations at time τ are unknown, meaning that the free energy at time τ cannot be a function of (unobserved) observations. Instead, *beliefs* over observations at time τ are required to evaluate free energy at the future time point τ. By assuming that observations o_τ depend on the unknown variables x_τ (which will be formalized once the generative model has been defined), we can introduce a distribution (or beliefs) over future observations $Q(o_\tau|x_\tau, \phi_\tau)$. Free energy at future time τ can then be evaluated under the expectation of this distribution:

$$\begin{aligned}\mathcal{F}_\tau(\phi_\tau) &= \mathbb{E}_{Q(x_\tau|\phi_\tau)}\Big[\mathbb{E}_{Q(o_\tau|x_\tau,\phi_\tau)}[\ln Q(x_\tau|\phi_\tau) - \ln P(x_\tau, o_\tau)]\Big] \\ &= \mathbb{E}_{Q(o_\tau,x_\tau|\phi_\tau)}\big[\ln Q(x_\tau|\phi_\tau) - \ln P(x_\tau, o_\tau)\big]\end{aligned} \tag{25}$$

Finally, we note that the functional purpose of expected free energy is to quantify the free energy that is expected to occur at time τ *given* the execution of some particular action (or sequence of actions). We can therefore specify the free energy that is expected to occur at time τ *given* some control state u_t by conditioning all of the distributions in equation 25 on u_t:

$$\mathcal{F}_\tau(\phi_\tau, u_t) = \mathbb{E}_{Q(o_\tau,x_\tau|u_t,\phi_\tau)}[\ln Q(x_\tau|u_t, \phi_\tau) - \ln P(x_\tau, o_\tau|u_t)] \tag{26}$$

This equation defines the expected free energy for time τ given control state u_t, which we denote $\mathbf{G}_\tau(\phi_\tau, u_t)$, i.e. $\mathbf{G}_\tau(\phi_\tau, u_t) = \mathcal{F}_\tau(\phi_\tau, u_t)$. Equation 26 is the form of expected free energy used in the main text.

8 Appendix 3

In this appendix, we demonstrate that expected free energy is composed of both instrumental (goal-directed) and epistemic (uncertainty-reducing) components. In the main text, we defined expected free energy $\mathbf{G}_\tau(\phi_\tau, u_t)$ as:

$$\mathbf{G}_\tau(\phi_\tau, u_t) = \mathbb{E}_{Q(o_\tau,x_\tau|u_t,\phi_\tau)}[\ln Q(x_\tau|u_\tau, \phi_\tau) - \ln P(x_\tau, o_\tau|u_t)] \tag{27}$$

We can factorize the generative model over future variables as $P(x_\tau, o_\tau|u_t) = P(x_\tau|o_\tau, u_t)P(o_\tau)$, allowing us to rewrite equation 27 as:

$$\mathbf{G}_\tau(\phi_\tau, u_t) = \mathbb{E}_{Q(o_\tau,x_\tau|u_t,\phi_\tau)}[\ln Q(x_\tau|u_t, \phi_\tau) - \ln P(x_\tau|o_\tau, u_t) - \ln P(o_\tau)] \tag{28}$$

This factorization is desirable as it exposes the prior probability of future observations $P(o_\tau)$, a distribution that describes an agents 'preferences' in a manner that is independent of any particular control states u or unknown

variables x. However, it additionally exposes the (intractable) posterior distribution $P(x_\tau|o_\tau, u_t)$. The need to evaluate this distribution can be circumvented by noting that it is approximated by the predictive approximate posterior $Q(x_\tau|o_\tau, u_t, \phi_\tau)$ i.e. $Q(x_\tau|o_\tau, u_t, \phi_\tau) \approx P(x_\tau|o_\tau, u_t)$. This is due to the fact that the approximate posterior is implicitly conditioned on the previous history of observations, due to the belief update scheme inherent in the active inference framework.

Applying this approximation to equation 28, we can derive:

$$\begin{aligned}\mathbf{G}_\tau(\phi_\tau, u_t) &\approx \mathbb{E}_{Q(o_\tau, x_\tau|u_t, \phi_\tau)}[\ln Q(x_\tau|u_t, \phi_\tau) - \ln Q(x_\tau|o_\tau, u_t, \phi_\tau) - \ln P(o_\tau)] \\ &= \underbrace{\mathbb{E}_{Q(o_\tau, x_\tau|u_t, \phi_\tau)}\big[\ln Q(x_\tau|u_t, \phi_\tau) - \ln Q(x_\tau|o_\tau, u_t, \phi_\tau)\big]}_{\text{(Negative) epistemic value}} \\ &\quad \underbrace{-\mathbb{E}_{Q(o_\tau, x_\tau|u_t, \phi_\tau)}\big[\ln P(o_\tau)\big]}_{\text{(Negative) instrumental value}}\end{aligned} \tag{29}$$

In the second equality, we have taken the prior probability of observations $P(o)$ into its own expectation in order to clarify the separation between epistemic value (first term) and instrumental value (second term). The second equality of equation 29 is the form of expected free energy used in the main text.

9 Appendix 4

In this section we demonstrate that epistemic value - a component of the expected free energy functional - is equivalent to a number of established formalisms. In the main text, (positive) epistemic value was defined as $\mathbb{E}_{Q(o_\tau, x_\tau|u_t)}[\ln Q(x_\tau|o_\tau, u_t) - \ln Q(x_\tau|u_t)]$, an equation which can be rewritten as:

$$\begin{aligned}&\mathbb{E}_{Q(o_\tau, x_\tau|u_t)}[\ln Q(x_\tau|o_\tau, u_t) - \ln Q(x_\tau|u_t)] \\ &= \mathbb{E}_{Q(o_\tau|u_t)Q(x_\tau|o_\tau, u_t)}[\ln Q(x_\tau|o_\tau, u_t) - \ln Q(x_\tau|u_t)] \\ &= \mathbb{E}_{Q(o_\tau|u_t)}\Big[\mathbb{E}_{Q(x_\tau|o_\tau, u_t)}[\ln Q(x_\tau|o_\tau, u_t) - \ln Q(x_\tau|u_t)]\Big] \\ &= \mathbb{E}_{Q(o_\tau|u_t)}\Big[\mathbb{KL}[Q(x_\tau|o_\tau, u_t)||\ln Q(x_\tau|u_t)]\Big]\end{aligned} \tag{30}$$

The first equality is derived by factorizing the predictive approximate posterior as $Q(o_\tau, x_\tau|u_t) = Q(o_\tau|u_t)Q(x_\tau|o_\tau, u_t)$, the second equality is derived through the standard property of expectations over random variables, and the final equality is reached through the standard definition of the KL-divergence.

The final equality of equation 30 quantifies the KL-divergence between the agents predictive approximate posterior after taking into observations into account $Q(x_\tau|o_\tau, u_t)$ and before taking observations into account $Q(x_\tau|u_t)$. It is therefore an (expected) KL-divergence between posterior and prior beliefs, a quantity known as (expected) Bayesian surprise (Baldi and Itti, 2010). Maximizing Bayesian surprise entails sampling observations that will lead to the greatest change in beliefs, or in other words, sampling observations that will disclose the maximum amount of information.

We can rearrange the final equality of equation 30 as:

$$\begin{aligned}&\mathbb{E}_{Q(o_\tau|u_t)}\Big[\mathbb{KL}[Q(x_\tau|o_\tau, u_t)||\ln Q(x_\tau|u_t)]\Big] \\ &= \mathbb{E}_{Q(o_\tau|u_t)}\Big[\mathbb{E}_{Q(x_\tau|o_\tau, u_t)}[\ln Q(x_\tau|o_\tau, u_t) - \ln Q(x_\tau|u_t)]\Big] \\ &= \mathbb{KL}[Q(x_\tau, o_\tau|u_t)||Q(x_\tau|u_t)Q(o_\tau|u_t)] \\ &= \mathbf{I}(\mathcal{X}_\tau; \mathcal{O}_\tau|u_t)\end{aligned} \tag{31}$$

where $\mathbf{I}(\cdot)$ is the mutual information. To reach the final equality we have utilized the fact that the third equality is the definition of mutual information, i.e. $\mathbf{I}(X;Y) = \mathbb{KL}[P(X,Y)||P(X)P(Y)]$. Epistemic value therefore quantifies the

mutual information between beliefs and observations, conditioned on control. Mutual information scores the amount of information one gains about a random variable X from observing some other random variable Y. Acting in order to maximize the mutual information between beliefs and observations is equivalent to acting in order to reduce the uncertainty in those beliefs, which can be seen by rewriting equation 31 as:

$$\mathbf{I}(\mathcal{X}_\tau, \mathcal{O}_\tau | u_t) = \mathbf{H}[Q(x_\tau|u_t)] - \mathbb{E}_{Q(o_\tau|x_\tau,u_t)}\Big[\mathbf{H}[Q(x_\tau|o_\tau,u_t)]\Big] \tag{32}$$

where $\mathbf{H}[\cdot]$ is the Shannon entropy, a standard measure of uncertainty. Expected free energy will therefore be maximized when observations minimize the entropy (i.e. uncertainty) in the beliefs encoded by $Q(x_\tau)$.

10 Appendix 5

In this section we derive the update equations for (beliefs over) hidden states, control states and model parameters. We first write out the free energy functional in terms of the sufficient statistics and parameters, before finding the derivative of this functional with respect to the sufficient statistics ϕ of the recognition distribution. We then solve these resulting equations for 0 (i.e. where the gradient of free energy is 0 with respect to the sufficient statistics), and rearrange to find the updates for ϕ that minimize free energy.

The free energy for the current time point $\mathcal{F}_t(\phi, o_t)$ is given by:

$$\begin{aligned}\mathcal{F}_t(\phi, o_t) &= -\mathbb{E}_{Q(x_t|\phi_t)}[\ln P(o_t|x_t)] + \mathbb{KL}[Q(x_t|\phi_t)||P(x_t)] \\ &= -\mathbb{E}_{Q(s_t|\phi_{s_t})}[\ln P(o_t|s_t,\lambda)] + \mathbb{KL}[Q(s_t|\phi_{s_t})||P(s_t|s_{t-1},u_{t-1},\theta)] \\ &\quad + \mathbb{KL}[Q(u_t|\phi_{u_t})||P(u_t)] + \mathbb{KL}[Q(\theta|\phi_\alpha)||P(\theta|\alpha)]\end{aligned} \tag{33}$$

where the first equality was derived in Appendix 1, and the second equality utilizes the factorization of the generative model and recognition distribution. We can rewrite the second equality of equation 33 in terms of parameters of the generative model and sufficient statistics of the recognition distribution:

$$\begin{aligned}\mathcal{F}_t(\phi, o_t) = &\underbrace{\phi_{s_t} \cdot (-\lambda \cdot \vec{o_t})}_{-\mathbb{E}_{Q(s_t|\phi_{s_t})}[\ln P(o_t|s_t,\lambda)]} + \underbrace{\phi_{s_t} \cdot (\ln \phi_{s_t} - \bar{\theta}^{u_{t-1}} \cdot \phi_{s_{t-1}})}_{\mathbb{KL}[Q(s_t|\phi_{s_t})||P(s_t|s_{t-1},u_{t-1},\theta)]} + \\ &\underbrace{\phi_{u_t} \cdot (\ln \phi_{u_t} - (-\vec{\mathbf{G}} \cdot \phi_{u_t}))}_{\mathbb{KL}[Q(u_t|\phi_{u_t})||P(u_t)]} + \\ &\underbrace{\sum_{ij}\big((\phi_{\alpha_{ij}} - \alpha_{ij})\bar{\theta}_{ij} - \ln\Gamma(\phi_{\alpha_{ij}})\big) + \sum_j \ln\Gamma(\phi_{\alpha_j^0})}_{\mathbb{KL}[Q(\theta|\phi_\alpha)||P(\theta|\alpha)]}\end{aligned} \tag{34}$$

where $\phi_{\alpha_j^0} = \sum_k \phi_{\alpha_{kj}}$, $\bar{\theta}_{ij} = \mathbb{E}[\ln\theta_{ij}] = \psi(\phi_{\alpha_{ij}}) - \psi(\phi_{\alpha_j^0})$ and $\vec{o_t}$ is a one-hot encoded vector specifying the current observation. This derivation follows from $\mathbb{E}(x) = \sum_x P(x)x = P \cdot x$ when P is a vector of probabilities for the various values of x held in the vector x. We now collect terms from equation 34 to give:

$$\begin{aligned}\mathcal{F}_t(\phi, o_t) = &\phi_{s_t} \cdot (\ln \phi_{s_t} - \lambda \cdot \vec{o_t} - \bar{\theta}^{u_{t-1}} \cdot \phi_{s_{t-1}} + \vec{\mathbf{G}} \cdot \phi_{u_t}) + \phi_{u_t} \cdot \ln \phi_{u_t} + \\ &\sum_{ij}\big((\phi_{\alpha_{ij}} - \alpha_{ij})\bar{\theta}_{ij} - \ln\Gamma(\phi_{\alpha_{ij}})\big) + \sum_j \ln\Gamma(\phi_{\alpha_j^0})\end{aligned} \tag{35}$$

We can now differentiate free energy with to the individual sufficient statistics of ϕ, giving:

$$\begin{aligned}
\frac{\partial \mathcal{F}(\phi, o)}{\partial \phi_{s_t}} &= \mathbf{1} + \ln \phi_{s_t} - \ln \lambda \cdot \vec{o_t} - \bar{\theta}^{u_{t-1}} \cdot \phi_{s_{t-1}} \\
\frac{\partial \mathcal{F}(\phi, o)}{\partial \phi_{u_t}} &= \mathbf{1} + \ln \phi_{u_t} - \tilde{\mathbf{G}} \\
\frac{\partial \mathcal{F}(\phi, o)}{\partial \phi_{\alpha}} &= \frac{\partial \bar{\theta}}{\partial \phi_{\alpha}} (\phi_{\alpha} - \alpha - \phi_{s_t} \phi_{s_{t-1}})
\end{aligned} \tag{36}$$

Where we have ignored the derivative of $\tilde{\mathbf{G}}$ with respect to ϕ_{s_t} for simplicity. Including this would lead to an additional 'optimism bias' term (Friston et al., 2015) in the first equality. Finally, the variational updates that minimize free energy can be obtained by solving for zero and rearranging, giving:

$$\begin{aligned}
\ln \phi_{s_t} &= \ln \lambda \cdot \vec{o_t} + \bar{\theta}^{u_{t-1}} \cdot \phi_{s_t} \\
\ln \phi_{u_t} &= -\tilde{\mathbf{G}} \\
\phi_{\alpha} &= \alpha + \phi_{s_t} \phi_{s_{t-1}}
\end{aligned} \tag{37}$$

To obtain the updates provided in the main text, we apply the softmax operator on $\ln \phi_{s_t}$ and $\ln \phi_{u_t}$, giving the updates for ϕ_{s_t} and ϕ_{u_t}, respectively.

11 Appendix 6

In simulations presented in the main text, the random and the epistemic agent learn models at a similar rate. We hypothesized that this may have been due to the simplicity of the transition distribution used in the current simulations, meaning that randomly choosing actions gave a near-optimal set of samples to learn from. To test this, we implemented a simple simulation in a larger state space. This simulation was of a 15 x 15 grid world, where agents sense the world directly and simply had to determine the transition dynamics. There were 225 different control states (one for each grid position), and each control state had a corresponding transition distribution, leading to 255 x (15 x 15) = 57375 parameters. We investigated the learning of an epistemic and random agent within a single trial (with an altered learning rate of 1). Figure 7 shows the KL-divergence of the transition distributions from the true transition distribution over the course of this trial. It is evident that the epistemic agent learns in a more efficient manner than the random agent. The size of the parameter space means that towards the start of the trial, the random agent searches in a near optimal manner. However, the consistently linear nature of the epistemic agents learning curve demonstrates that the agent searched the state space in a principled manner. This suggests that sampling to reduce epistemic uncertainty provides a principled method for learning a model of environment dynamics.

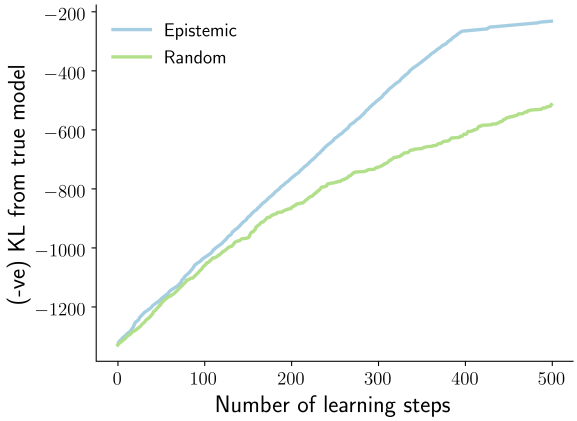

Figure 7: **High dimensional model accuracy:** A comparison of the model accuracy of the epistemic and random strategies in a high dimensional state space.

References

Christopher G. Atkeson and Juan Carlos Santamaria. A Comparison of Direct and Model-Based Reinforcement Learning. In *In International Conference on Robotics and Automation*, pages 3557–3564. IEEE Press, 1997.

Pierre Baldi and Laurent Itti. Of bits and wows: A Bayesian theory of surprise with applications to attention. *Neural Networks: The Official Journal of the International Neural Network Society*, 23(5):649–666, June 2010. ISSN 1879-2782. doi: 10.1016/j.neunet.2009.12.007.

Manuel Baltieri and Christopher Buckley. The dark room problem in predictive processing and active inference, a legacy of cognitivism? preprint, PsyArXiv, May 2019a. URL https://osf.io/p4z8f.

Manuel Baltieri and Christopher L. Buckley. An active inference implementation of phototaxis. *The 2018 Conference on Artificial Life: A Hybrid of the European Conference on Artificial Life (ECAL) and the International Conference on the Synthesis and Simulation of Living Systems (ALIFE)*, 29:36–43, September 2017. doi: 10.1162/isal_a_011. URL https://www.mitpressjournals.org/doi/abs/10.1162/isal_a_011.

Manuel Baltieri and Christopher L. Buckley. Generative models as parsimonious descriptions of sensorimotor loops. *arXiv:1904.12937 [cs, q-bio]*, April 2019b. URL http://arxiv.org/abs/1904.12937. arXiv: 1904.12937.

Manuel Baltieri and Christopher L. Buckley. Nonmodular architectures of cognitive systems based on active inference. *arXiv:1903.09542 [cs, q-bio]*, March 2019c. URL http://arxiv.org/abs/1903.09542. arXiv: 1903.09542.

Xabier E. Barandiaran. Autonomy and Enactivism: Towards a Theory of Sensorimotor Autonomous Agency. *Topoi*, 36 (3):409–430, September 2017. ISSN 1572-8749. doi: 10.1007/s11245-016-9365-4. URL https://doi.org/10.1007/s11245-016-9365-4.

Matthew J. Beal. Variational algorithms for approximate Bayesian inference. Technical report, 2003.

H. C. Berg and D. A. Brown. Chemotaxis in Escherichia coli analysed by three-dimensional tracking. *Nature*, 239 (5374):500–504, October 1972. ISSN 0028-0836.

Matthew Botvinick and Ari Weinstein. Model-based hierarchical reinforcement learning and human action control. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 369(1655), November 2014. ISSN 1471-2970. doi: 10.1098/rstb.2013.0480.

Christopher L. Buckley, Chang Sub Kim, Simon McGregor, and Anil K. Seth. The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology*, 81:55–79, December 2017. ISSN 0022-2496. doi: 10.1016/j.jmp.2017.09.004. URL http://www.sciencedirect.com/science/article/pii/S0022249617300962.

Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A. Efros. Large-Scale Study of Curiosity-Driven Learning. *arXiv:1808.04355 [cs, stat]*, August 2018. URL http://arxiv.org/abs/1808.04355. arXiv: 1808.04355.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep Reinforcement Learning in a Handful of Trials using Probabilistic Dynamics Models. *arXiv:1805.12114 [cs, stat]*, May 2018. URL http://arxiv.org/abs/1805.12114. arXiv: 1805.12114.

Andy Clark. Radical Predictive Processing. *Southern Journal of Philosophy*, 53(S1):3–27, 2015.

Roger C. Conant and W. Ross Ashby. Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, 1(2):89–97, October 1970. ISSN 0020-7721. doi: 10.1080/00207727008920220. URL https://doi.org/10.1080/00207727008920220.

Peter Dayan and Kent C. Berridge. Model-based and model-free Pavlovian reward learning: revaluation, revision, and revelation. *Cognitive, Affective & Behavioral Neuroscience*, 14(2):473–492, June 2014. ISSN 1531-135X. doi: 10.3758/s13415-014-0277-8.

Marc Peter Deisenroth. A Survey on Policy Search for Robotics. *Foundations and Trends in Robotics*, 2(1-2):1–142, 2011. ISSN 1935-8253, 1935-8261. doi: 10.1561/2300000021. URL http://www.nowpublishers.com/articles/foundations-and-trends-in-robotics/ROB-021.

Ray J. Dolan and Peter Dayan. Goals and Habits in the Brain. *Neuron*, 80(2):312–325, October 2013. ISSN 0896-6273. doi: 10.1016/j.neuron.2013.09.007. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3807793/.

Bradley B Doll, Dylan A Simon, and Nathaniel D Daw. The ubiquity of model-based reinforcement learning. *Current opinion in neurobiology*, 22(6):1075–1081, December 2012. ISSN 0959-4388. doi: 10.1016/j.conb.2012.08.003. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3513648/.

Matthew D. Egbert and Xabier E. Barandiaran. Modeling habits as self-sustaining patterns of sensorimotor behavior. *Frontiers in Human Neuroscience*, 8, 2014. ISSN 1662-5161. doi: 10.3389/fnhum.2014.00590. URL https://www.frontiersin.org/articles/10.3389/fnhum.2014.00590/full.

Peter L. Freddolino and Saeed Tavazoie. Beyond homeostasis: a predictive-dynamic framework for understanding cellular behavior. *Annual Review of Cell and Developmental Biology*, 28:363–384, 2012. ISSN 1530-8995. doi: 10.1146/annurev-cellbio-092910-154129.

Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews. Neuroscience*, 11(2):127–138, February 2010. ISSN 1471-0048. doi: 10.1038/nrn2787.

Karl Friston. Life as we know it. *Journal of the Royal Society, Interface*, 10(86):20130475, September 2013. ISSN 1742-5662. doi: 10.1098/rsif.2013.0475.

Karl Friston, Rick Adams, and Read Montague. What is value-accumulated reward or evidence? *Frontiers in Neurorobotics*, 6:11, 2012a. ISSN 1662-5218. doi: 10.3389/fnbot.2012.00011.

Karl Friston, Rick A. Adams, Laurent Perrinet, and Michael Breakspear. Perceptions as Hypotheses: Saccades as Experiments. *Frontiers in Psychology*, 3, May 2012b. ISSN 1664-1078. doi: 10.3389/fpsyg.2012.00151. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3361132/.

Karl Friston, FitzGerald Thomas, Moutoussis Michael, Behrens Timothy, and Dolan Raymond J. The anatomy of choice: dopamine and decision-making. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1655):20130481, November 2014. doi: 10.1098/rstb.2013.0481. URL https://royalsocietypublishing.org/doi/full/10.1098/rstb.2013.0481.

Karl Friston, Francesco Rigoli, Dimitri Ognibene, Christoph Mathys, Thomas Fitzgerald, and Giovanni Pezzulo. Active inference and epistemic value. *Cognitive Neuroscience*, 6(4):187–214, 2015. ISSN 1758-8936. doi: 10.1080/17588928.2015.1020053.

Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, John ODoherty, and Giovanni Pezzulo. Active inference and learning. *Neuroscience & Biobehavioral Reviews*, 68:862–879, September 2016a. ISSN

0149-7634. doi: 10.1016/j.neubiorev.2016.06.022. URL http://www.sciencedirect.com/science/article/pii/S0149763416301336.

Karl J. Friston and Klaas E. Stephan. Free-energy and the brain. *Synthese*, 159(3):417–458, December 2007. ISSN 1573-0964. doi: 10.1007/s11229-007-9237-y. URL https://doi.org/10.1007/s11229-007-9237-y.

Karl J. Friston, Jean Daunizeau, and Stefan J. Kiebel. Reinforcement Learning or Active Inference? *PLOS ONE*, 4(7): e6421, July 2009. ISSN 1932-6203. doi: 10.1371/journal.pone.0006421. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0006421.

Karl J. Friston, Vladimir Litvak, Ashwini Oswal, Adeel Razi, Klaas E. Stephan, Bernadette C. M. van Wijk, Gabriel Ziegler, and Peter Zeidman. Bayesian model reduction and empirical Bayes for group (DCM) studies. *NeuroImage*, 128:413–431, March 2016b. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2015.11.015. URL http://www.sciencedirect.com/science/article/pii/S105381191501037X.

Karl J. Friston, Marco Lin, Christopher D. Frith, Giovanni Pezzulo, J. Allan Hobson, and Sasha Ondobaka. Active Inference, Curiosity and Insight. *Neural Computation*, 29(10):2633–2683, 2017. ISSN 1530-888X. doi: 10.1162/neco_a_00999.

Karl J. Friston, Richard Rosch, Thomas Parr, Cathy Price, and Howard Bowman. Deep temporal models and active inference. *Neuroscience & Biobehavioral Reviews*, 90:486–501, July 2018. ISSN 0149-7634. doi: 10.1016/j.neubiorev.2018.04.004. URL http://www.sciencedirect.com/science/article/pii/S0149763418302525.

James J. Gibson. *The Ecological Approach to Visual Perception : Classic Edition*. Psychology Press, November 2014. ISBN 978-1-317-57938-0. doi: 10.4324/9781315740218. URL https://www.taylorfrancis.com/books/9781317579380.

Jacqueline Gottlieb and Pierre-Yves Oudeyer. Towards a neuroscience of active sampling and curiosity. *Nature Reviews Neuroscience*, 19(12):758, December 2018. ISSN 1471-0048. doi: 10.1038/s41583-018-0078-0. URL https://www.nature.com/articles/s41583-018-0078-0.

R. L. Gregory. Perceptions as hypotheses. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 290(1038):181–197, July 1980. ISSN 0962-8436. doi: 10.1098/rstb.1980.0090.

David Ha and Jürgen Schmidhuber. World Models. *arXiv:1803.10122 [cs, stat]*, March 2018. doi: 10.5281/zenodo.1207631. URL http://arxiv.org/abs/1803.10122. arXiv: 1803.10122.

Geoffrey E. Hinton and Drew van Camp. Keeping the Neural Networks Simple by Minimizing the Description Length of the Weights. In *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, COLT '93, pages 5–13, New York, NY, USA, 1993. ACM. ISBN 978-0-89791-611-0. doi: 10.1145/168304.168306. URL http://doi.acm.org/10.1145/168304.168306. event-place: Santa Cruz, California, USA.

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. VIME: Variational Information Maximizing Exploration. *arXiv:1605.09674 [cs, stat]*, May 2016. URL http://arxiv.org/abs/1605.09674. arXiv: 1605.09674.

Michael D. Kirchhoff and Tom Froese. Where There is Life There is Mind: In Support of a Strong Life-Mind Continuity Thesis. *Entropy*, 19(4):169, April 2017. doi: 10.3390/e19040169. URL https://www.mdpi.com/1099-4300/19/4/169.

Michael D. Kirchhoff and Ian Robertson. Enactivism and predictive processing: a non-representational view. *Philosophical Explorations*, 21(2):264–281, May 2018. ISSN 1386-9795. doi: 10.1080/13869795.2018.1477983. URL https://aap.tandfonline.com/doi/full/10.1080/13869795.2018.1477983.

Kirchhoff Michael, Parr Thomas, Palacios Ensor, Friston Karl, and Kiverstein Julian. The Markov blankets of life: autonomy, active inference and the free energy principle. *Journal of The Royal Society Interface*, 15(138):20170792, January 2018. doi: 10.1098/rsif.2017.0792. URL https://royalsocietypublishing.org/doi/full/10.1098/rsif.2017.0792.

Julian Kiverstein. Free Energy and the Self: An Ecological–Enactive Interpretation. *Topoi*, April 2018. ISSN 1572-8749. doi: 10.1007/s11245-018-9561-5. URL https://doi.org/10.1007/s11245-018-9561-5.

David C. Knill and Alexandre Pouget. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, 27(12):712–719, December 2004. ISSN 0166-2236. doi: 10.1016/j.tins.2004.10.007.

Leonid Kuvayev and Richard S. Sutton. Model-Based Reinforcement Learning with an Approximate, Learned Model. In *in Proceedings of the Ninth Yale Workshop on Adaptive and Learning Systems*, pages 101–105, 1996.

Adam Linson, Andy Clark, Subramanian Ramamoorthy, and Karl Friston. The Active Inference Approach to Ecological Perception: General Information Dynamics for Natural and Artificial Embodied Cognition. *Frontiers in Robotics and AI*, 5, 2018. ISSN 2296-9144. doi: 10.3389/frobt.2018.00021. URL https://www.frontiersin.org/articles/10.3389/frobt.2018.00021/full.

M. Lungarella and O. Sporns. Information Self-Structuring: Key Principle for Learning and Development. In *Proceedings. The 4th International Conference on Development and Learning, 2005*, pages 25–30, July 2005. doi: 10.1109/DEVLRN.2005.1490938.

Max Lungarella and Olaf Sporns. Mapping Information Flow in Sensorimotor Networks. *PLOS Computational Biology*, 2(10):e144, October 2006. ISSN 1553-7358. doi: 10.1371/journal.pcbi.0020144. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.0020144.

M. Zehetleitner and F. B. Schönbrodt. When misrepresentations are successful. In *Epistemological Dimensions of Evolutionary Psychology*. New York: Springer, 2015.

Rowan McAllister and Carl E. Rasmussen. Improving PILCO with Bayesian Neural Network Dynamics Models. 2016.

Ryan T. McKay and Daniel C. Dennett. The evolution of misbelief. *The Behavioral and Brain Sciences*, 32(6):493–510; discussion 510–561, December 2009. ISSN 1469-1825. doi: 10.1017/S0140525X09990975.

Angela Mendelovici. Reliable Misrepresentation and Tracking Theories of Mental Representation. *Philosophical Studies*, 165(2):421–443, 2013.

Amir Mitchell and Wendell Lim. Cellular perception and misperception: Internal models for decision-making shaped by evolutionary experience. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, 38(9): 845–849, 2016. ISSN 1521-1878. doi: 10.1002/bies.201600090.

Amir Mitchell, Gal H. Romano, Bella Groisman, Avihu Yona, Erez Dekel, Martin Kupiec, Orna Dahan, and Yitzhak Pilpel. Adaptive prediction of environmental changes by microorganisms. *Nature*, 460(7252):220–224, July 2009. ISSN 1476-4687. doi: 10.1038/nature08112.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing Atari with Deep Reinforcement Learning. *arXiv:1312.5602 [cs]*, December 2013. URL http://arxiv.org/abs/1312.5602. arXiv: 1312.5602.

Guido Montúfar, Keyan Ghazi-Zahedi, and Nihat Ay. A Theory of Cheap Control in Embodied Systems. *PLOS Computational Biology*, 11(9):e1004427, September 2015. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004427. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004427.

Teodor Negru. Self-organization, Autopoiesis, Free-energy Principle and Autonomy. *Organon F*, 25(2):215–243, 2018. ISSN 1335-0668, 2585-7150. URL https://www.ceeol.com/search/article-detail?id=692450.

Thomas Parr and Karl J. Friston. Generalised free energy and active inference: can the future cause the past? *bioRxiv*, page 304782, April 2018. doi: 10.1101/304782. URL https://www.biorxiv.org/content/10.1101/304782v1.

Giovanni Pezzulo, Francesco Donnarumma, Pierpaolo Iodice, Domenico Maisto, and Ivilin Stoianov. Model-Based Approaches to Active Perception and Control. *Entropy*, 19:266, 2017. doi: 10.3390/e19060266.

Athanasios S. Polydoros and Lazaros Nalpantidis. Survey of Model-Based Reinforcement Learning: Applications on Robotics. *Journal of Intelligent & Robotic Systems*, 86(2):153–173, May 2017. ISSN 1573-0409. doi: 10.1007/s10846-017-0468-y. URL https://doi.org/10.1007/s10846-017-0468-y.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1994. ISBN 978-0-471-61977-2.

Maxwell J. D. Ramstead, Michael D. Kirchhoff, and Karl J. Friston. A tale of two densities: Active inference is enactive inference, 2019. URL http://philsci-archive.pitt.edu/16167/.

R. P. Rao and D. H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, January 1999. ISSN 1097-6256. doi: 10.1038/4580.

J. Ruesch, R. Ferreira, and A. Bernardino. A measure of good motor actions for active visual perception. In *2011 IEEE International Conference on Development and Learning (ICDL)*, volume 2, pages 1–6, August 2011. doi: 10.1109/DEVLRN.2011.6037355.

Philipp Schwartenbeck, Johannes Passecker, Tobias Hauser, Thomas H. B. FitzGerald, Martin Kronbichler, and Karl J. Friston. Computational mechanisms of curiosity and goal-directed exploration. *bioRxiv*, page 411272, September 2018. doi: 10.1101/411272. URL https://www.biorxiv.org/content/10.1101/411272v1.

Anil K Seth. The cybernetic Bayesian brain. In *Open MIND*. 2015.

Anil K. Seth and Manos Tsakiris. Being a Beast Machine: The Somatic Basis of Selfhood. *Trends in Cognitive Sciences*, 22(11):969–981, 2018.

Bradly C. Stadie, Sergey Levine, and Pieter Abbeel. Incentivizing Exploration In Reinforcement Learning With Deep Predictive Models. *arXiv:1507.00814 [cs, stat]*, July 2015. URL http://arxiv.org/abs/1507.00814. arXiv: 1507.00814.

Yi Sun, Faustino Gomez, and Juergen Schmidhuber. Planning to Be Surprised: Optimal Bayesian Exploration in Dynamic Environments. *arXiv:1103.5708 [cs, stat]*, March 2011. URL http://arxiv.org/abs/1103.5708. arXiv: 1103.5708.

Roland Thar and Michael Kuhl. Bacteria are not too small for spatial sensing of chemical gradients: an experimental evidence. *Proceedings of the National Academy of Sciences of the United States of America*, 100(10):5748–5753, May 2003. ISSN 0027-8424. doi: 10.1073/pnas.1030795100.

Chris Thornton. Gauging the value of good data: Informational embodiment quantification. *Adaptive Behavior*, 18(5): 389–399, October 2010. ISSN 1059-7123. doi: 10.1177/1059712310383914. URL https://doi.org/10.1177/1059712310383914.

Paul F. M. J. Verschure, Thomas Voegtlin, and Rodney J. Douglas. Environmentally mediated synergy between perception and behaviour in mobile robots. *Nature*, 425(6958):620–624, October 2003. ISSN 1476-4687. doi: 10.1038/nature02024. URL https://www.nature.com/articles/nature02024.

Tingwu Wang, Xuchan Bao, Ignasi Clavera, Jerrick Hoang, Yeming Wen, Eric Langlois, Shunshi Zhang, Guodong Zhang, Pieter Abbeel, and Jimmy Ba. Benchmarking Model-Based Reinforcement Learning. *arXiv:1907.02057 [cs, stat]*, July 2019. URL http://arxiv.org/abs/1907.02057. arXiv: 1907.02057.

C. J. C. H. Watkins. Learning form delayed rewards. *Ph. D. thesis, King's College, University of Cambridge*, 1989. URL https://ci.nii.ac.jp/naid/10007782517/.

Wanja Wiese. Action Is Enabled by Systematic Misrepresentations. *Erkenntnis*, 82(6):1233–1252, December 2017. ISSN 1572-8420. doi: 10.1007/s10670-016-9867-x. URL https://doi.org/10.1007/s10670-016-9867-x.

Daniel Williams. Predictive Processing and the Representation Wars. *Minds and Machines*, 28(1):141–172, March 2018. ISSN 1572-8641. doi: 10.1007/s11023-017-9441-6. URL https://doi.org/10.1007/s11023-017-9441-6.

Scott Cheng-Hsin Yang, Daniel M. Wolpert, and Máté Lengyel. Theoretical perspectives on active sensing. *Current opinion in behavioral sciences*, 11:100–108, October 2018. ISSN 2352-1546. doi: 10.1016/j.cobeha.2016.06.009. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6116896/.