Active Inference: Demystified and Compared

Noor Sajid
noor.sajid.18@ucl.ac.uk
Wellcome Centre for Human Neuroimaging, UCL Queen Square Institute of Neurology, London, WC1N 3AR, U.K.

Philip J. Ball
ball@robots.ox.ac.uk
Machine Learning Research Group, Department of Engineering Science, University of Oxford, Oxford OX1 3PJ, U.K.

Thomas Parr
thomas.parr.12@ucl.ac.uk
Karl J. Friston
k.friston@ucl.ac.uk
Wellcome Centre for Human Neuroimaging, UCL Queen Square Institute of Neurology, London, WC1N 3AR, U.K.

Active inference is a first principle account of how autonomous agents operate in dynamic, nonstationary environments. This problem is also considered in reinforcement learning, but limited work exists on comparing the two approaches on the same discrete-state environments. In this letter, we provide (1) an accessible overview of the discrete-state formulation of active inference, highlighting natural behaviors in active inference that are generally engineered in reinforcement learning, and (2) an explicit discrete-state comparison between active inference and reinforcement learning on an OpenAI gym baseline. We begin by providing a condensed overview of the active inference literature, in particular viewing the various natural behaviors of active inference agents through the lens of reinforcement learning. We show that by operating in a pure belief-based setting, active inference agents can carry out epistemic exploration—and account for uncertainty about their environment—in a Bayes-optimal fashion. Furthermore, we show that the reliance on an explicit reward signal in reinforcement learning is removed in active inference, where reward can simply be treated as another observation we have a preference over; even in the total absence of rewards, agent behaviors are learned through preference learning. We make these properties explicit by showing two scenarios in which active inference agents can infer behaviors in reward-free environments compared to both Q-learning and Bayesian model-based reinforcement learning agents and by placing

Neural Computation **33, 674–712** (2021)
https://doi.org/10.1162/neco_a_01357

zero prior preferences over rewards and learning the prior preferences over the observations corresponding to reward. We conclude by noting that this formalism can be applied to more complex settings (e.g., robotic arm movement, Atari games) if appropriate generative models can be formulated. In short, we aim to demystify the behavior of active inference agents by presenting an accessible discrete state-space and time formulation and demonstrate these behaviors in a OpenAI gym environment, alongside reinforcement learning agents.

1 Introduction

Active inference provides a framework (derived from first principles) for solving and understanding the behavior of autonomous agents in situations requiring decision-making under uncertainty (Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2017; Friston, Rosch, Parr, Price, & Bowman, 2017). It uses the free energy principle to describe the properties of random dynamical systems (such as an agent in an environment), and by minimizing expected free energy over time, Bayes-optimal behavior can be obtained for a given environment (Friston et al., 2014; Friston, 2019). More concretely, this optimal behavior is determined by evaluating evidence (i.e., marginal likelihood) under an agent's generative model of outcomes (Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2016). The agent's generative model of the environment is an abstraction, which assumes that certain internal (hidden) states give rise to these outcomes. One goal of the agent is to infer what these hidden states are, given a set of outcomes. The generative model also provides a way, through searching and planning, to form beliefs about the future. Thus, the agent can make informed decisions over which sequence of actions (i.e., policies) it is most likely to choose. In active inference, due to its Bayesian formulation, the most likely policies lead to Bayes-optimal outcomes (i.e., those most coherent with prior beliefs). This formulation has two complementary objectives: infer Bayes-optimal behavior, and optimize the generative model based on the agent's ability to infer which hidden states gave rise to the observed data. Both can be achieved, simultaneously, by minimizing free energy functionals. This free energy formulation gives rise to realistic behaviors, such as natural exploration-exploitation trade-offs, and, by being fully Bayesian, is amenable to online learning settings, where the environment is nonstationary. This follows from the ability to model uncertainty over contexts (Friston et al., 2015; Parr & Friston, 2017).

Active inference can also be seen as providing a formal framework for jointly optimizing action and perception (Millidge, Tschantz, Seth, & Buckley, 2020). In the context of machine learning, this is often referred to as planning as inference (Attias, 2003; Botvinick & Toussaint, 2012; Baker & Tenenbaum, 2014; Millidge, Tschantz, Seth, et al., 2020), and in the case of non-equilibrium physics, it is analogous to self-organization or self-assembly (Crauel & Flandoli, 1994; Seifert, 2012; Friston, 2019).

The main contributions of active inference, in contrast to analogous reinforcement learning (RL) frameworks, follow from its commitments to a pure belief-based scheme. Reinforcement learning is a broad term used in different fields. To make meaningful comparisons between active inference and reinforcement learning, we commit to the definition of reinforcement learning in Sutton and Barto (1998, 2018): *"Reinforcement learning is learning what to do—how to map situations to actions—so as to maximize a numerical reward signal."* RL algorithms, under this definition, can be model based or model free. Model-based methods learn a model of environmental dynamics, which is used to infer a policy that maximizes long-term reward, while model-free RL estimates a policy that maximizes long-term reward directly from trajectory data. Throughout this letter, RL refers to both model based and model free unless stated otherwise. This definition rests on the reward hypothesis: that *"any goal or purpose can be well thought of as maximization of the expected value of the cumulative sum of a received scalar signal (reward)"* (Sutton & Barto, 2018). Here, reward is, by definition, some outcome that reinforces behavior, and hence a circular definition of reward behavior. For example going to the cafe and buying coffee can be explained from two perspectives: (1) a cup of coffee in the morning is intrinsically rewarding and therefore I go out to the cafe to get the coffee, and (2) going to the cafe to get coffee means, tautologically, that coffee is rewarding. In short, rewards reinforce behaviors that secure rewards. Traditionally, these (model-free) RL algorithms operate directly on the state of the environment, but (model-based) RL algorithms that operate on beliefs also represent an active area of research (Igl, Zintgraf, Le, Wood, & Whiteson, 2018).

Conversely, in active inference, an agent's interaction with the environment is determined by action sequences that minimize expected free energy (and not the expected value of a reward signal). Additionally, unlike in reinforcement learning, the reward signal is not differentiated from other types of sensory outcomes. That is, any type of outcome may be more or less preferred. This means that the implicit reward associated with any outcome is a feature of the creature seeing the observation, not the environment they inhabit. This may be different for different agents or even for the same agent at different points in time. This highlights that the two frameworks have fundamentally different objectives: reward maximization in reinforcement learning and free energy minimization in active inference.

In this letter, we reveal circumstances in which behavior might be the same and when it may differ under these two distinct objectives. We show that the main contributions of active inference, in comparison to reinforcement learning, are that (1) reward functions (i.e., prior preferences) are not always necessary because any policy has an epistemic value, even in the absence of prior preferences; (2) agents can learn their own reward function and this becomes a way of describing how the agent expects itself to behave, as opposed to getting something from the environment; (3) a principled account of epistemic exploration and intrinsic motivation as minimizing uncertainty (Parr & Friston, 2017; Schwartenbeck et al., 2019); and (4) incorporating uncertainty as a natural part of belief updating (Parr & Friston, 2017). Why are these contributions of interest? In standard reinforcement learning, the reward function defines the agent's goal and allows it to learn how to best act within the environment (Sutton & Barto, 1998). However, defining the reward function is difficult; if it is a specific signal from the environment based on action, then is it unanimously good in that environment—for example, whether an agent controlling the thermostat should get a positive reward only for turning on the heating during winter. Consequently, crafting appropriate reward functions is not easy, and it is possible for agents to learn suboptimal actions if the reward function is poorly specified (Amodei et al., 2016). However, active inference bypasses this problem by replacing the traditional reward function, used in reinforcement learning, with prior beliefs about preferred outcomes. This causes the agent to act in a way—via the beliefs it holds—such that the observed outcomes match prior preferences. This is useful when we have imprecise or no prior preferences since active inference endows agents with the ability to learn prior preferences from interacting with the environment itself by learning and empirical prior distribution over preferred outcomes. In other words, an agent can learn the kinds of outcomes it can achieve, and these become its prior preferences (in virtue of the fact that these outcomes are achievable, they underwrite the agent's viability in that environment). This way of defining the reward function (prior preferences) highlights that whether a state is rewarding (or not) is a function of the agent, not the environment. This reward function conceptualization is distinct from reward functions under reinforcement learning.

Another challenge within reinforcement learning is balancing the ratio between exploration and exploitation: What actions should the agent take at any given point in time? Should the agent continue to explore and find more valuable actions or exploit its (current) most valuable action sequence? Many different algorithms have been used to address this, including ϵ-greedy (Vermorel & Mohri, 2005; Mnih et al., 2013, 2016), action selection based on action-utility (Sutton, 1990) and counter-based strategies (Wiering & Schmidhuber, 1998; Tijsma, Drugan, & Wiering, 2016). However, even with these exploratory mechanisms in place, most reinforcement learning formulations call on a temperature hyperparameter to weight extrinsic reward (from the environment) against the intrinsic motivation (from the agent). There is no such hyperparameter in active inference, although the precision of various priors plays an analogous role, because the distinction between extrinsic value (i.e., expected reward) and intrinsic value (i.e., intrinsic motivation) is just one way of decomposing expected free energy. In active inference, everything minimizes free energy, including hyperparameters. Usually, these hyperparameters transpire to be precise over various beliefs, which allows for a natural trade-off between epistemic exploration and pragmatic behavior. This means that all the required machinery is in play from the start but should be added to the reinforcement framework (by definition) only if it helps maximize expected return. Consequently, state-of-the-art reinforcement learning approaches can be regarded as a series of refinements to the base algorithm that help resolve problems as they are encountered, including the need to marginalize out hyperparameters instead of defining particular values (Cesa-Bianchi, Gentile, Lugosi, & Neu, 2017).

Here, we unpack these properties of active inference, with appropriate ties to the reinforcement learning literature, under the discrete state-space and time formulation, thereby providing a brief overview of the theory. Furthermore, we demonstrate these properties, and points of contact, with reinforcement learning agents on a series of experiments using a modified FrozenLake OpenAI baseline. This is purely an illustration of the conceptual premises, not a demonstration of their implications. Thus, while our simulations of reinforcement learning could have included more complex (i.e., context-aware) aspects, such as in Cao and Ray (2012), Lloyd and Leslie (2013), and Padakandla and Bhatnagar (2019). These approaches describe various ways to perform inference without explicit reference to their impact on behavior. Indeed, as O'Donoghue, Osband, and Ionescu (2020) showed, state-of-the-art approaches, which can be seen as framing reinforcement learning as probabilistic inference, make simplifications for practical reasons. These subtle modeling choices that trade off tractability for accuracy can result in suboptimal behavior (e.g., failing to account for Q-value uncertainty and ensuing "dithering" behavior). Active inference avoids this ambiguity by clearly defining how latent variable models are constructed to solve partially observable Markov decision process problems and how inference should proceed, using gradient descent on expected free energy. This allows the agent to consider the effect of its own actions on future rewards (i.e., preferred outcomes) when evaluating the expected free energy of all plausible policies (i.e., action trajectories), based on the anticipated consequences of those policies. However, these are Bellman–optimal for one time step but Bayes–optimal (i.e., realize prior beliefs and minimize free energy) for distal time horizons. Additionally, by minimizing expected free energy, the agent balances exploration, and exploitation resulting in a Bayes-optimal arbitration between the two, which may not be reward maximizing from an RL perspective. In contrast to RL, active inference accounts for epistemic uncertainty by operating in an explicitly belief-based framework (Levine, 2018). Additionally, the conceptual approach of active inference means that all the appropriate terms, relating to the intrinsic value of information, are in play from the start but should be added to the reinforcement framework (by definition) only if they help maximize long-term reward. Consequently, state-of-the-art reinforcement learning approaches can be regarded as a series of refinements to the base algorithm that help resolve problems as they are experienced. Therefore, for explicit behavioral comparison, we simulate the performance of the two frameworks after removal of the reward signal from the FrozenLake environment, that is, a flat value function. In this setting, there is no motivation for adding any new information gain terms because they cannot be justified in terms of increasing expected value of reward.

This letter has four sections. Section 2 considers the discrete state-space and time formulation of active inference and provides commentary on its derivation, implementation, and connections to reinforcement learning. Section 3 provides a concrete example of the key components of the generative model and update rules in play, using a modified version of OpenAI's FrozenLake environment. Through these simulations, we compared the performance of three types of agents: active inference, Q-learning (Watkins & Dayan, 1992) using ϵ-greedy exploration, and Bayesian model-based reinforcement learning using Thompson sampling (Poupart, 2018) in stationary and non-stationary environments. We note that while all agents are able to perform appropriately in a stationary setting, active inference's ability to carry out online planning allows for Bayes-optimal behavior in the nonstationary environment. The simulations demonstrate that in the absence of a reward signal, the active inference exhibits information-seeking behavior (to build a better model of its environment), in contrast to Q-learning agents but similar to the Bayesian reinforcement-learning agent. We make explicit the conceptual differences in reward function under active inference and reinforcement learning through learning of prior preferences that enable the agent to settle into its niche. We highlight that from the perspective of reinforcement learning, this niche might be counterintuitive, that is, reward minimizing. We conclude with a brief discussion of how this formalism could be applied in (more complex) engineering applications (e.g., robotic arm movement, Atari games) if the appropriate underlying probability distribution or generative model can be formulated.

2 Active Inference

2.1 Motivation.

Active inference describes how (biological or artificial) agents navigate dynamic, nonstationary environments (Friston, FitzGerald, et al., 2017; Friston, Rosch, et al., 2017). It postulates that agents maintain homeostasis by residing in (attracting) states that minimize surprise (Friston, Mattout, & Kilner, 2011; Bogacz, 2017).

Definition 1. *Surprise is defined as the negative log probability of an outcome. For this, we introduce a random variable, $o \in O$, that corresponds to a particular outcome received by the agent, and O is a finite set of all possible outcomes,*

$$S(o) = -log\, P(o), \tag{2.1}$$

where P denotes a probability distribution.

In active inference, the agent determines how to minimize surprise by maintaining a generative model of the (partially observable) world. This is necessary because the agent does not have access to a true measurement of its current state (i.e., the state of the actual generative process). Instead, it can only perceive itself and the world via outcomes (Friston, FitzGerald, et al., 2017; Friston, Parr, et al., 2017). This allows the problem to be framed as a partially observable Markov decision process (POMDP) (Astrom, 1965), where the generative model allows us to make inferences about true states given outcomes. In active inference, the agent makes choices based on its beliefs about these states of the world and not based on the value of the states (Friston et al., 2016). This distinction is key: in standard model-based reinforcement learning frameworks, the agent is interested in optimizing the *value function of the states* (Sutton & Barto, 1998) that is, making decisions that maximize expected value. In active inference, we are interested in optimizing a *free energy functional of beliefs about states*, that is, making decisions that minimize expected free energy. Put even more simply, in reinforcement learning, we are interested in residing in high-value states under a reward function, while in active inference, we wish to reside in states that give rise to outcomes that match our prior preferences (i.e., a target distribution). In one sense, this is a false distinction, as we could interpret a reward function as a log prior preference or vice versa. However, ensuring consistency with a distribution is different from maximizing reward. The latter implies we try to spend all of our time in the most rewarding state, while the former nuances this with an imperative to spend time in states that generate outcomes proportionate to the prior probability associated with those outcomes. However, particular RL algorithms, such as SMM (Lee et al., 2019) also optimize similar objectives of matching state marginal distributions to some target distribution.

From an implementation perspective, this means replacing the traditional reward function used in reinforcement learning with prior beliefs about preferred outcomes. The agent's prior preferences, $\log P(o)$, are defined only to within an additive constant (i.e., a single negative or positive number). This means that the prior probability of an outcome is a softmax function of utility, $P(o) = \sigma(U(o))$, and therefore depends on relative differences between rewarding (Chong, Familiar, & Shim, 2016) and unrewarding (surprising) outcomes. Additionally, by incorporating learning over priors, active inference agents can be equipped with the ability to learn their own preferences over outcomes, thereby bypassing the need for an explicit reward signal from the environment and the agent to exhibit self-evidencing behavior (see section 3).

2.2 Variational Free Energy. Starting from a simple generative model for outcomes, it is possible to derive a variational free energy formulation, as motivated by Figure 1. This gives the starting point for the full active inference derivation. First, we introduce the random variable, $s \in S$, to

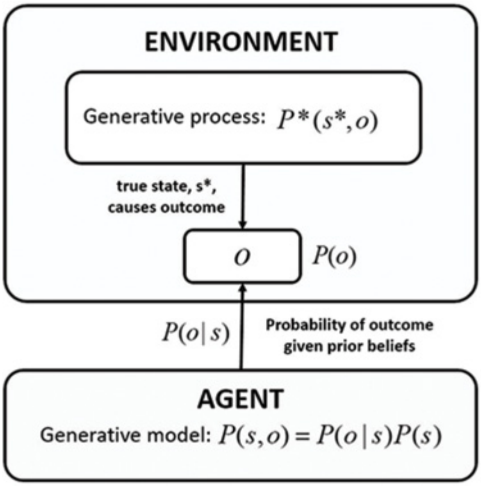
ENVIRONMENT
Generative process: $P^*(s^*,o)$
true state, s*, causes outcome
o
$P(o)$
$P(o|s)$
Probability of outcome given prior beliefs
AGENT
Generative model: $P(s,o) = P(o|s)P(s)$

Figure 1: Graphical representation of the generative process (based on true states, $s*$) in the world and the corresponding (internal) generative model (based on probabilistic beliefs random variables, s, that stand in for true states that are hidden) that best explain the outcomes, o. This graphic highlights that the outcomes are shared between the generative process and model.

represent a particular hidden state of the world, where S is a finite set of all possible (hidden) states.

The generative model abstraction asserts that the world has a true (hidden) state $s*$, which results in the outcomes o (via the generative process): $s* \in S$. The agent correspondingly has an internal representation of (or distribution over) s, which it infers from o (via its generative model). The hidden state is a combination of features relevant to the agent (e.g., location, color), and the outcome is the information from the environment (e.g., feedback, velocity, reward). By the reverse process of mapping from its hidden state to the outcomes (through Bayesian model inversion), the agent can explain the outcomes in terms of how they were caused by hidden states. This is Bayesian model inversion or inference.

Definition 2. *Generative model is defined as a partially observable MDP that rests on the following (simplified) joint probability: $P(o, s)$ where $o \in O$ and $s \in S$ as stated previously. The joint probability can be factorized into a likelihood function $P(o|s)$ and prior over internal states $P(s)$ (see the supplementary materials for a full specification of the generative model):*

$$P(o, s) = P(o|s)P(s). \tag{2.2}$$

We know that for the agent to minimize its surprise, we need to marginalize over all possible states that could lead to a given outcome. This can be achieved by using the above factorization:

$$P(o) = \sum_{s \in S} P(o|s)P(s). \tag{2.3}$$

This is not a trivial task, since the dimensionality of the hidden state (and sequences of actions) space can be extremely large. Instead, we utilize a variational approximation of this quantity, $P(o)$, which is tractable and allows us to estimate quantities of interest.

Definition 3. *Variational free energy, F, is defined as the upper bound on surprise (see definition 1). It is derived using Jensen's inequality and commonly known as the (negative) evidence lower bound (ELBO) in the variational inference literature (Blei, Kucukelbir, & McAuliffe, 2017):*

$$\begin{aligned} -\log P(o) &= -\log \sum_{s \in S} P(o, s) && (2.4)\\ &\leq -\sum_{s \in S} Q(s) \log \frac{P(o, s)}{Q(s)} && (2.5)\\ &= \sum_{s \in S} Q(s) \log \frac{Q(s)}{P(o, s)}. && (2.6)\end{aligned}$$

Here, $Q(.)$ is the variational distribution.

To make the link more concrete, we further manipulate the variational free energy quantity, F:

$$\begin{aligned} F &= \sum_{s \in S} Q(s) \log \frac{Q(s)}{P(o, s)} && (2.7)\\ &= \sum_{s \in S} Q(s) \log \frac{Q(s)}{P(s|o)P(o)} && (2.8)\\ &= \sum_{s \in S} Q(s) \left(\log \frac{Q(s)}{P(s|o)} - \log P(o) \right) && (2.9)\\ &= D_{\mathrm{KL}}[Q(s)||P(s|o)] - \log P(o). && (2.10)\end{aligned}$$

By rearranging the last equation, the connection between surprise and variational free energy is made explicit:

$$-\log P(o) = F - D_{\mathrm{KL}}[Q(s)||P(s|o)]. \tag{2.11}$$

Additionally, we can express variational free energy as a function of these posterior beliefs in many forms:

$$F = \underbrace{D_{\mathrm{KL}}[Q(s|\pi)||P(s|o,\pi)]}_{\textit{evidence bound}} - \underbrace{\log P(o)}_{\textit{log evidence}} \tag{2.12}$$

$$= \underbrace{D_{\mathrm{KL}}[Q(s|\pi)||P(s|\pi)]}_{\text{complexity}} - \underbrace{\mathbb{E}_{s\sim Q(s)}[\log P(o|s)]}_{\text{accuracy}}. \tag{2.13}$$

Since KL divergences cannot be less than zero, from equation 2.12, we see that free energy is minimized when the approximate posterior becomes the true posterior. In that instance, the free energy would simply be the negative log evidence for the generative model (Beal, 2003). This shows that minimizing free energy is equivalent to maximizing (generative) model evidence. In other words, it is minimizing the complexity of accurate explanations for observed outcomes, as seen in equation 2.13. Note that we have conditioned the probabilities in equations 2.12 and 2.13 on policies, π. These policies can be regarded as hypotheses about how to act that, as we will see, pertain to probabilistic transitions among hidden states. For the moment, the introduction of policies simply means that the variational free energy above can be evaluated for any given action sequence.

2.3 Expected Free Energy. Variational free energy gives us a way to perceive the environment (i.e., determine s from o) and addresses one part of the problem: making inferences about the world (i.e., the "inference" in active inference). However, the "active" part of the formulation is still lacking; we have not accounted for the fact that the agent can take actions. To motivate this, we would like to minimize not only our variational free energy F calculated from past and present observations, but also our expected free energy G, which depends on anticipated observations in the future. Minimization of expected free energy allows the agent to influence the future by taking actions in the present, which are selected from policies. We first consider the definition of a policy and later determine how to evaluate their likelihoods from the generative model, which ultimately leads to the action selected by the agent.

Definition 4. *Policy is defined as a sequence of actions at time τ that enable an agent to transition between hidden states: $\tau \in [1, 2, \ldots, T]$, where T is the total number of time steps in a given experiment under the generative model. For this, we introduce two random variables: (1) $u_\tau \in U$ to represent a particular action at time τ where U is a finite set of all possible actions, and (2) $\pi \in \Pi$ to represent a particular policy, where Π is a finite set of allowable policies, that is, sequences of actions in the sense of sequential policy optimization (Alagoz et al., 2010),*

$$\pi = \{u_1, u_2, \ldots, u_\tau\}, \tag{2.14}$$

up to a given time horizon, τ. The explicit link between policy and action is

$$u_\tau = \pi(\tau). \tag{2.15}$$

Remark 1 (Connections to State-Action Policies). From definition 4, in active inference, a policy is simply a sequence of choices for actions through time (i.e., a sequential policy). This contrasts with state-action policies in reinforcement learning, where π_{RL} denotes a mapping of states to actions (Bellman, 1952; Sutton, 1990):

$$\pi_{RL}(u, s) = P(u \mid s).$$

The two policy types (sequential, π and state-action π_{RL}) are equivalent, under a POMDP formulation when $\tau = 1$ that is, $\pi = \{u_1\}$ (Friston et al., 2016). Note that the Bayes-optimal policy is selected from these policies. For the remainder of the letter, a policy (π) refers to sequential policy.

To derive the expected free energy, we first extend the variational free energy definition to be dependent on time (τ) and policy (π) (and present its matrix formulation, equation 2.18):

$$F(\tau, \pi) = \sum_{s_\tau^\pi} Q(s_\tau|\pi)Q(s_{\tau-1}|\pi) \log \frac{Q(s_\tau|\pi)}{P(o_\tau, s_\tau|s_{\tau-1}, \pi)} \tag{2.16}$$

$$= \mathbb{E}_{Q(s_{\tau-1}|\pi)}\big[D_{\mathrm{KL}}[Q(s_\tau|\pi)||P(s_\tau|s_{\tau-1}, \pi)]\big] - \mathbb{E}_{Q(s_\tau|\pi)}\big[\ln P(o_\tau|s_\tau)\big] \tag{2.17}$$

$$= s_\tau^\pi \cdot \big(\log s_\tau^\pi - \log \boldsymbol{B}_{\tau-1}^\pi s_{\tau-1}^\pi - \log \boldsymbol{A} \cdot o_\tau\big). \tag{2.18}$$

Here s_τ^π is the expected state conditioned on each policy; $\boldsymbol{B}_\tau^\pi$ is the transition probability for hidden states, contingent on pursuing a given policy, at a particular time; $\boldsymbol{A}$ is the expected likelihood matrix mapping from hidden states to outcomes, and o_τ represents the outcomes. These are simply vectors (s_τ^π) or matrices ($\boldsymbol{B}_\tau^\pi$ and $\boldsymbol{A}$) specifying a probability for each alternative state or outcome. For the matrices, each column corresponds to a different value to the variable we condition on (here, hidden states), while rows give the probability of each hidden state at the next time or the outcome at the current time step, respectively. Now having developed this functional dependency on time, we simply take an expectation with respect to the posterior distribution of outcomes from our generative model, $P(o_\tau|s_\tau)$.

Definition 5. *Expected free energy is defined as a free energy functional of future trajectories, G. It effectively evaluates evidence for plausible policies based on outcomes that have yet to be observed (Parr & Friston, 2019b). Heuristically, we can obtain G from equation 2.16 by making two moves. The first is to include beliefs about future outcomes in the expectation, that is, supplementing the expectation* *under the approximate posterior with the likelihood, resulting in a predictive distribution given by $P(o_\tau|s_\tau)Q(s_\tau|\pi)$. The second is to (implicitly or explicitly) condition the joint probabilities of states and observations in the generative model on some desired state of affairs (C), as opposed to a specific policy. These two moves ensure (1) we can evaluate this quantity before the observations are obtained and (2) minimization of G encourages policies whose result is consistent with C.*[1]

$$G(\tau, \pi) = \sum_{s_\tau, o_\tau} P(o_\tau|s_\tau)Q(s_\tau|\pi)Q(s_{\tau-1}|\pi)\log\frac{Q(s_\tau|\pi)}{P(o_\tau, s_\tau|s_{\tau-1}, C)} \tag{2.19}$$

$$= \mathbb{E}_{\tilde{Q}}\left[\log(Q(s_\tau|\pi) - \log(P(o_\tau, s_\tau|s_{\tau-1}, C))\right] \tag{2.20}$$

$$= \mathbb{E}_{\tilde{Q}}\left[\log(Q(s_\tau|\pi) - \log(P(s_\tau|o_\tau, s_{\tau-1})) - \log(P(o_\tau|C))\right] \tag{2.21}$$

$$\geq \underbrace{\mathbb{E}_{\tilde{Q}}\left[\log(Q(s_\tau|\pi) - \log(Q(s_\tau|o_\tau, s_{\tau-1}, \pi))\right]}_{\text{-ve mutual information}} - \underbrace{\mathbb{E}_{\tilde{Q}}\left[\log(P(o_\tau|C))\right]}_{\text{expected log evidence}} \tag{2.22}$$

$$= \underbrace{\mathbb{E}_{\tilde{Q}}\left[\log(Q(o_\tau|\pi) - \log(Q(o_\tau|s_\tau, s_{\tau-1}, \pi))\right]}_{\text{-ve epistemic value}} - \underbrace{\mathbb{E}_{\tilde{Q}}\left[\log(P(o_\tau|C))\right]}_{\text{extrinsic value}} \tag{2.23}$$

$$= \underbrace{D_{\mathrm{KL}}[Q(o_\tau|\pi)||P(o_\tau|C)]}_{\text{expected cost}} + \underbrace{E_{Q(s_\tau|s_{\tau-1},\pi)}\left[H[P(o_\tau|s_\tau)]\right]}_{\text{expected ambiguity}} \tag{2.24}$$

$$= o_\tau^\pi \cdot (o_\tau^\pi - \boldsymbol{C}_\tau) + s_\tau^\pi \cdot \boldsymbol{H}, \tag{2.25}$$

where the following notation is used: $\tilde{Q} = P(o_\tau|s_\tau)Q(s_\tau|\pi)$; $Q(o_\tau|s_\tau, \pi) = P(o_\tau|s_\tau)$; $\boldsymbol{C}_\tau = \log P(o_\tau|C)$ is the logarithm of prior preference over outcomes, o_τ is the vector of posterior predictive outcomes (i.e., $\boldsymbol{A}s_\tau^\pi$), and $\boldsymbol{H} = -diag\big(\mathbb{E}_Q[\boldsymbol{A}_{i,j}].\mathbb{E}_Q[\boldsymbol{A}]\big)$ is the vector encoding the ambiguity over outcomes for each hidden state.

When minimizing expected free energy, we can regard equation 2.23 as capturing the imperative to maximize the amount of information gained, from observing the environment, about the hidden state (i.e., maximizing epistemic value), while maximizing expected value, as scored by the

[1] There are other ways in which we could consider constructing free energy functionals to deal with outcomes that have yet to be observed. While some of the alternatives are plausible from a theoretical perspective, they tend to dispense with aspects of observed behavior that we seek to capture and therefore do not apply to the kinds of system we are interested in here. For example, Millidge, Tschantz, and Buckley, (2020) propose an alternative that subtracts the conditional entropy of the likelihood (i.e., ambiguity) from the expected free energy shown here. This leads to agents that are less-ambiguity-averse agents and do not seek information. As active inference deals with curious agents, we retain this ambiguity in the expected free energy functional.

(log) preferences (i.e., extrinsic value). This entails a clear trade-off: the former (epistemic) component promotes curious behavior, with exploration encouraged as the agent seeks out salient states to minimize uncertainty about the environment, and the latter (pragmatic) component encourages exploitative behavior through leveraging knowledge that enables policies to reach preferred outcomes. In other words, the expected free energy formulation enables active inference to treat exploration and exploitation as two different ways of tackling the same problem: minimizing uncertainty.

This natural curiosity can be contrasted with handcrafted exploration in reinforcement learning schemes, where curiosity is replaced by random action selection (Mnih et al., 2013) or through the use of ad hoc novelty bonuses, which are appended to the reward function (Pathak, Agrawal, Efros, & Darrell, 2017). Information-theoretic approaches have also been explored in a reinforcement learning context (Still & Precup, 2012; Mohamed & Rezende, 2015; Blau, Ott, & Ramos, 2019). Some of these approaches leverage beliefs about latent states (Blau et al., 2019; Sekar et al., 2020). For example, Blau et al. (2019) is a model-free algorithm that implicitly accounts for beliefs over the latent states. Additionally, Seker et al. (2020) is very close in its treatment of latent states to active inference but leverages an ensemble of belief states to inform epistemic exploration, rather than a true Bayesian posterior. Additionally, curiosity as formulated under active inference can emerge in reinforcement learning under POMDP formulations. Example algorithms may incorporate inductive biases (Igl, Zintgraf, Le, Wood, & Whiteson, 2018) and uncertainty over state transitions or outcomes (Ross, Chaib-draa, & Pineau, 2008; Kolter & Ng, 2009; Zintgraf et al., 2019). The key aspect of these belief-POMDP schemes is that they deal with belief states (i.e., a space of probability distributions over hidden states). This is crucial for exploration and minimizing uncertainty, because uncertainty is an attribute of a belief about hidden states, not the hidden states per se.

Normatively speaking, active inference dispenses with the Bellman optimality principle and replaces it with a (variational) principle of least action (see Friston, Samothrakis, & Montague, 2012, for further discussion). However, recent Bayesian RL schemes have used variational principles, for example, maintaining latent over the MDP (Zintgraf et al., 2019) or the explicit beliefs (Igl et al., 2019). While in many of these settings, the distance between the two schools of thought may seem to be closing, a fundamental distinction, which has yet to be bridged, is the situation in which there are no rewards or, in active inference, when prior preferences are uninformative. In a scheme motivated by reward maximization, no meaningful behavior can be generated in this setting. This is not a criticism of such schemes but a statement of their scope and the problems they are designed to solve. In contrast, the intrinsic value of seeking information—regardless of its potential to evince reward—in active inference means that in the absence of any rewarding outcomes, agents are still driven by a curiosity that helps them build a better model of their world. Furthermore, active inference agents can learn their priors over observations and will exhibit ambiguity-minimizing behavior in order to fulfill these prior expectations (Friston et al., 2016). In short, they can learn epistemic habits in the absence of extrinsic rewards.

Equation 2.24 offers an alternative perspective on the same objective: an agent wishes to minimize the ambiguity and the degree to which outcomes (under a given policy) deviate from prior preferences $P(o_\tau|C)$. Thus, ambiguity is the expectation of the conditional entropy, or uncertainty about outcomes, under the current policy. Low entropy suggests that outcomes are salient and uniquely informative about hidden states (e.g., visual cues in a well-lit environment, as opposed to the dark). In addition, the agent would like to pursue policy-dependent outcomes ($Q(o_\tau|\pi)$) that resemble its preferred outcomes ($P(o_\tau|C)$). This is achieved when the KL divergence between predicted and preferred outcomes (i.e., expected cost) is minimized by a particular policy. Furthermore, prior beliefs about future outcomes equip the agent with goal-directed behavior (i.e., toward states they expect to occupy and frequent).

It is now also possible to specify priors over policies using the expected free energy. Policies, a priori, minimize the expected free energy term, G (Friston, FitzGerald, et al., 2017). This has sometimes been framed in terms of a heuristic reductio ad absurdum argument that if selected policies realize prior beliefs and minimize free energy, then the only tenable prior beliefs are policies that will minimize free energy (Parr & Friston, 2019b). If this were not true, then an active inference agent would not have prior belief that it selects policies that minimize expected free energy and it would infer (and pursue) policies that were not free energy minimizing. As such, it would not be an active inference (i.e., free energy minimizing) agent, which is a contradiction. This leads to the agent's prior belief that it will select policies that minimize the free energy expected under that policy. There are some subtleties to this argument that leave some room for maneuver. Specifically, this does not tell us how to construct an expected free energy functional, but this is typically chosen to be consistent with definition 5. This choice ensures both exploratory and exploitative behavior and is therefore sufficiently flexible to deal with the kind of problem we are interested in for this letter.

This can be realized by expressing the probability of any policy with a softmax function (i.e., normalized exponential) of expected free energy,

$$P(\pi) = \sigma[-\beta^{-1} \cdot G(\pi)], \tag{2.26}$$

where σ denotes a softmax function and β is a temperature parameter.

This illustrates the "self-evidencing" behavior of active inference. Action sequences (policies) that result in lower expected free energy are more likely. Intuitively this makes sense; since all notions of how to act in the world (i.e., exploration, exploitation) are wrapped up in the expected free energy G, policy selection simply becomes a matter of determining (through search) the set of actions that get us closest to this goal (i.e., the attracting set defined by prior preferences $P(o|C)$).

Note the similarities to Dyna-style/planning model-based reinforcement learning (Sutton, 1990): hypothetical rollouts are used to model the consequences of each policy. However, the actual controller in active inference is derived through an approach similar to model predictive control (Camacho & Alba, 2013), where a search is performed over possible action sequences at each time step.

Now that we have priors over policies, we can incorporate these into the generative model and into the free energy. This gives

$$F = \mathbb{E}_{Q(\pi)}[F(\pi)] + D_{\mathrm{KL}}[Q(\pi)||P(\pi)]. \tag{2.27}$$

Here, the free energy of the model conditioned on the model plays the role of a negative log marginal likelihood, giving this the form of an accuracy and complexity term in the space of beliefs about policies. Often the temperature parameter β is also equipped with priors (normally using a gamma distribution) and posteriors, which add another complexity term to the free energy.

2.4 Optimizing Free Energy. From this free energy formulation, we can optimize expectations about hidden states, policies, and precision through inference and optimize model parameters (likelihood, transition states) through learning (via a learning rate, η). This optimization requires finding sufficient statistics of posterior beliefs that minimize variational free energy (Friston, Parr, et al., 2017). Under variational Bayes, this would mean iterating the appropriate formulations (for inference and learning) until convergence. Under the active inference scheme, we calculate the solution by using a gradient descent (with a default step size, ζ, of 4) on free energy F, which allows us to optimize both action-selection and inference simultaneously, using a mean-field approximation (Beck, Pouget, & Heller, 2012; Parr, Markovic, Kiebel, & Friston, 2019). The gradients of the negative free energy with respect to states and precisions, respectively, are

$$\varepsilon_\tau^\pi = (\log \boldsymbol{A} \cdot o_\tau + \log \boldsymbol{B}_{\tau-1}^\pi s_{\tau-1}^\pi + \log \boldsymbol{B}_\tau^\pi \cdot s_{\tau+1}^\pi) - \log s_\tau^\pi, \tag{2.28}$$

$$\varepsilon^\gamma = (\beta - \beta_\tau) + (\pi - \pi_0) \cdot G, \tag{2.29}$$

where $\beta_\tau = \beta + (\pi - \pi_0).G$; $\beta = \frac{1}{\gamma}$ encodes posterior beliefs about (inverse) precision (i.e., temperature); π represents the policies specifying action sequences; and $\pi_0 = \sigma(-\gamma.G)$. G in this equation is a vector whose elements are the expected free energies for each policy under consideration.

This involves converting the discrete updates, defined in equations 2.28 and 2.29, into dynamics for inference that minimize state and precision prediction errors: $\varepsilon_\tau^\pi = -\partial_s F$ and $\varepsilon^\gamma = -\partial_\gamma F$. These prediction errors are free energy gradients. Gradient flows then produce posterior expectations that minimize free energy to provide Bayesian estimates of hidden variables. This particular optimization scheme means expectations about hidden variables are updated over several timescales. During each outcome or trial, beliefs about each policy are evaluated based on prior beliefs about future outcomes, which get in through the expected free energy. This is determined by updating posterior beliefs about hidden states (i.e., state estimation under each policy, $Q(s|\pi)$) on a fast timescale, while posterior beliefs find new extrema (i.e., as new outcomes are sampled, $P(s|o, \pi)$). These posterior beliefs are then used to compute the posterior predictive probabilities of future outcomes, $Q(o|\pi)$, which themselves contribute to the expected free energy and, through this, the priors over policies. This introduces an unusual feature of the online inference schemes under active inference rarely seen in Bayesian accounts; the priors over policies change with new outcomes. This is a key distinction between active inference and standard Bayesian RL perspectives and can occur only when the best policies to engage in are functionals of beliefs that may be updated.

Using this kind of belief updating, we can calculate the posterior beliefs about each policy—namely, a softmax function based on expected free energy and the likelihood of past observations under that policy (approximated with $F(\pi)$), extending the definition of the prior covered in equation 2.26. The softmax function is a generalized sigmoid for vector input and can, in a neurobiological setting, be regarded as a firing rate function of neuronal depolarization (Friston, Rosch, et al., 2017). Having optimized posterior beliefs about policies, they are used to form a Bayesian model average of the next outcome (i.e., under these beliefs about what I will do next, which observations I would expect on average), which is realized through action selected to conform to this. Practically, action selection is often achieved more simply by sampling from the distribution over actions at a given time implied by posterior beliefs about policies.

In active inference, the scope and depth of the policy search are exhaustive, in the sense that any policy entertained by the agent is encoded explicitly and any hidden states over the sequence of actions entailed by policy are continuously updated. However, in practice, this can be computationally expensive; therefore, a policy is no longer evaluated if its log evidence is ζ (default 20) times less likely than the (current) most plausible policy. This, ζ, can be treated as an adjustable hyperparameter. Additionally, at the end of each sequence of outcomes, the expected parameters are updated to allow for learning across trials. This is like Monte Carlo reinforcement learning, where model parameters are updated at the end of each trial. Finally, temporal discounting emerges naturally from active inference, where the generative model determines the nature of discounting (based on parameter-capturing precision), with predictions in the distant future being less precise and thus discounted in a Bayes-optimal fashion (Friston, FitzGerald, et al., 2017). Practically, this involves the inclusion of a hyperparameter $\beta = \gamma^{-1}$ that can be regarded as the inverse precision or temperature of posterior beliefs over policies. This is marginalized out during belief updating, enabling the exploration-exploitation trade-off to be inferred. This highlights the flexibility of active inference in the sense that it can be applied to any generative model. It is unclear how similar parameterizations could be adopted in conventional reinforcement learning, where optimizing temperature parameters typically involves a grid search over value. Having said this, sophisticated formulations of reinforcement learning, such as soft actor-critic (Haarnoja, Zhou, Abbeel, & Levine, 2018; Cesa-Bianchi et al., 2017), have addressed this issue.

This discussion suggests that from a generic generative model, we can derive Bayesian updates that clarify how perception, policy selection, and actions shape beliefs about hidden states and subsequent outcomes in a dynamic (nonstationary) environment. This formulation can be extended to capture a more representative generative process by defining a hierarchical (deep temporal) generative model (described in Friston, FitzGerald, et al., 2017; Friston, Parr, et al., 2017; Parr & Friston, 2017), continuous state-space models (Buckley, Kim, McGregor, & Seth, 2017; Parr & Friston, 2019a; Ueltzhöffer, 2018) or mixed models with both discrete and continuous states as described in Friston, Parr, et al. (2017) and Parr and Friston (2018). In the case of a continuous formulation, the generative model state-space can be defined in terms of generalized coordinates of motion (i.e., the coefficients of a Taylor series expansion in time, as opposed to a series of discrete time steps), which generally have a nonlinear mapping to the observed outcomes. Additionally, future work looks to evaluate how these formulations (agents) may interact with each other to emulate multi-agent exchanges.

The implicit variational updates presented here have previously been used to simulate a wide range of neuronal processing (using a gradient descent on variational free energy): ranging from single cell responses (including place-cell activity) (Friston, FitzGerald, et al., 2017) and midbrain dopamine activity (Friston et al., 2014) to evoked potentials, including those associated with mismatch negative (MMN) paradigms (Friston, FitzGerald, et al., 2017). Additionally, there has been some evidence implicating these variational inferences with neuromodulatory systems: action selection (dopaminergic), attention and expected uncertainty (cholinergic), and volatility and unexpected uncertainty (noradrenergic) (Parr & Friston, 2017, 2019b). (For a detailed overview, see Friston, FitzGerald, et al., 2017; Parr & Friston, 2019b, and Da Costa et al., 2020.)

In what follows, we provide a simple worked example to show precisely the behaviors that emerge—naturally—under active inference.

3 Simulations

This section considers inference using simulations of a modified version of OpenAI gym's FrozenLake environment. For simplicity, we have chosen this paradigm. Note that more complex simulations have been explored in the literature—for example, behavioral economics trust games (Moutoussis, Trujillo-Barreto, El-Deredy, Dolan, & Friston, 2014; Schwartenbeck et al., 2015), narrative construction and reading (Friston, Rosch, et al., 2017), saccadic searches and scene construction (Mirza, Adams, Mathys, & Friston, 2016), and Atari games (Cullen, Davey, Friston, & Moran, 2018). In closely related work, Cullen, Davey, Friston, and Moran (2018) demonstrated that active inference agents perform better in another OpenAI gym environment, Doom, compared to reward-maximizing agents. Their reward-maximizing agents are active inference agents without the epistemic value term (G) and can therefore be considered distinct from standard reinforcement learning agents.

We first describe the environment setup and then simulate how an agent learns to navigate the lake to successfully reach the goal. The simulations involve searching for the reward (a Frisbee) in a 3 × 3 frozen lake and avoiding falling in a hole. The purpose of these simulations is to provide an accessible overview (and accompanying code; see the Software Note at the end of the main text of this letter) of the conceptual (and practical) differences between active inference and standard reinforcement learning.

3.1 Setup.

The frozen lake has a gridlike structure with four patches: starting point (S), frozen surface (F), hole (H), and goal (G), where the Frisbee is located. All patches, except for (H), are safe. The agent starts each episode at (S)—position 1. From there, to reach the Frisbee location, the agent needs to take a series of actions: left, right, down, or up. The agent is allowed to continue moving around the frozen lake, with multiple revisits to the same positions, but each episode ends when either (H) or (G) is visited. (G) and (H) can be located in one of two locations: positions 8 and 6 or 6 and 8, respectively. The objective is to reach (G), the Frisbee location, ideally in as few steps as possible, while avoiding the hole (H). If the agent is able to reach the Frisbee without falling in the hole, it receives a score of 100 at the end of trial. This scoring metric is framework agnostic and allows us to compare active inference to reinforcement learning methods. However, it is important to note that maximizing reward is not the definition of Bayes-optimal behavior for an active inference agent, where information gain is also of value. This will become important later. Finally, we limit the maximum number of time steps (i.e., the horizon) to 15.

3.2 Active Inference Agents.

For this paradigm, we define the generative model for the active inference agent as follows (see Figure 2): four action states that encode direction of movement (left, right, down, and up), 18 hidden states (9 locations factorized by 2 contexts), and outcome modalities include grid position (9) and score (3). The action states control the transitions between the hidden state location factors; for example, when at location 4, the agent can move to location 5 (right), 7 (down), or 1 (up) or stay at 4 (left). The hidden state factor, *location*, elucidates the agents'

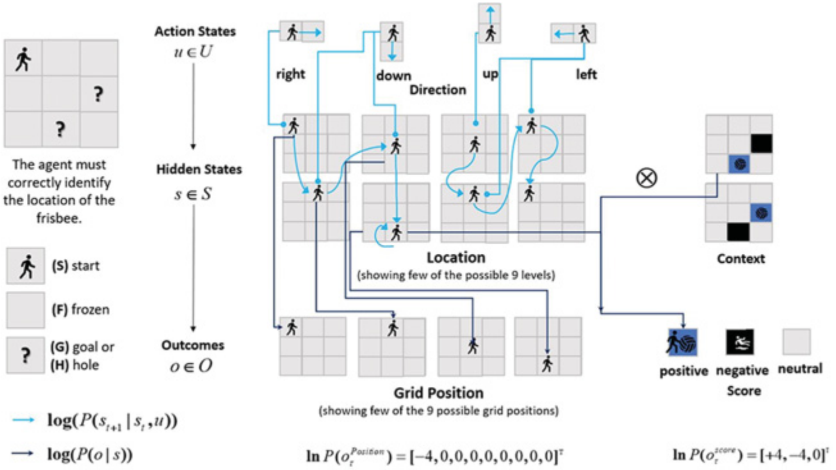
?
?
The agent must correctly identify the location of the frisbee.
(S) start
(F) frozen
?
(G) goal or
(H) hole
Action States
$u \in U$
Hidden States
$s \in S$
Outcomes
$o \in O$
right
down
up
left
Direction
Location
(showing few of the possible 9 levels)
Context
⊗
Grid Position
(showing few of the 9 possible grid positions)
positive
negative
neutral
Score
$\log(P(s_{t+1} | s_t, u))$
$\log(P(o | s))$
$\ln P(o_t^{Position}) = [-4,0,0,0,0,0,0,0,0]^T$
$\ln P(o_t^{score}) = [+4,-4,0]^T$

Figure 2: Graphical representation of the active inference generative model. The model contains four action states that encode direction of movement: right, down, up, and left. They control the ability to transition between hidden state location factors (one of the nine locations going from 1 → 9; only a few states are shown). Each action navigates the agent to a different location (a select few are shown)—for example, if the agent starts in position 1 and chooses to turn right, it will end up in state 2 at the next time step. However, if the agent started in location 5 and goes up, it would end up in location 2 instead. Note that both 8 and 6 are absorbing states; only 8 is denoted as such by the circular arrow. Additionally, if an agent makes an improbable move by trying, for example, to go left from location 1, it will remain in location 1 (as shown). The hidden states have a Kronecker tensor product (⊗) form with two factors: location and context (one of the two goal locations). The context cannot be changed by the agent and corresponds to the associated Frisbee location: 8 if context 1 or 6 if context 2. Note that in context 1, the hole location is 6. From each of the two hidden state factors (location and context) an outcome is generated. The agent observes two types of outcomes at each time point: its grid position and score. Categorical parameters, that define the generative model, $\boldsymbol{A}$ (likelihood $- P(o|s)$): have an identity mapping between hidden state location and outcome grid position with some uncertainty: for example. "If I have beliefs that I am in position 6, then I will observe myself in position 6, irrespective of context." The score likelihood, given the hidden states, is determined by the context: for context 1, a positive score received at location 8, and negative or nothing elsewhere. $P(o)$ corresponds to prior preference: the agent expects to find a positive score and not remain at the starting location.

beliefs about its location in the frozen lake. The context hidden state factor elucidates the agent's beliefs about the location of (G) and (H): if context is 1, then (G) location is 8 and (H) location is 6. The outcomes correspond to the following: being at any of the 9 possible grid positions and receiving 3 types of potential reward (positive, negative, or neutral). Positive reward is received if the agent correctly navigates to the (G) location, negative if to the (H) location, and neutral otherwise (F, S).

We define the likelihood $P(o|s)$ as follows: an identity mapping between hidden state location and outcome grid position-for example, if I have beliefs that I am located in position 6, then I will observe myself in position 6 regardless of context. However, the likelihood for scoring, given the hidden states, is determined by the context: if the context is 1 (2), then the positive score will be received at location 8 (6), and negative or nothing elsewhere. The action-specific transition probabilities $P(s_{t-1}|s_t, u)$ encode allowable moves, except for the sixth and eighth locations, which are absorbing latent states that the agent cannot leave. We define the agent as having precise beliefs about the contingencies (i.e., large prior concentration parameters = 100). The utility of the outcomes, $\boldsymbol{C}$, is defined by $\ln P(o)$: 4 and -4 *nats* for rewarding and unrewarding outcome: this can be regarded as a replacement for writing out an explicit reward function. This means, that the agent expects to be rewarded e^8 times more at (G) than (H). Notice that rewards and losses are specified in terms of *nats* or natural units, because we have stipulated reward in terms of the natural logarithms of some outcome. The prior beliefs about the initial state were initialized: location state ($D = 1$) for the first location and zero otherwise, with uniform beliefs for context state. We equip the agent with deep policies: these are potential permutations of action trajectories—for example, (Left, Left, Right) or (Down, Right, Up). Practically, policies (action sequences) are removed if the relative posterior probability is of 1/128 or less than the most likely policy. After each episode, the posteriors about the current state are carried forward as priors for the next episode. By framing the paradigm in this way, we treat solving the POMDP as a *planning-as-inference* problem; in order to act appropriately, the agent needs to correctly update internal beliefs about the current context.

Having specified the state-space and contingencies, we can solve the belief updating equations 2.28 and 2.29 to simulate appropriate behavior. Pseudocode for the belief updating and action selection for this particular type of discrete state-space and time formulation is presented in the supplementary materials. To provide a baseline for purely exploratory behavior, we also simulated a null active inference agent, which had no prior preferences (i.e., was insensitive to the reward).

3.3 Reinforcement Learning Agents. We compared the active inference agents' performance against two reinforcement learning algorithms: Q-Learning using ϵ-greedy exploration (Watkins, 1989; Sutton & Barto, 1998) and Bayesian model-based reinforcement learning using standard Thompson sampling (Poupart, 2018; Ghavamzadeh, Mannor, Pineau, & Tamar, 2016). Thompson sampling is an appropriate procedure here because it entails the optimization of dual objectives: reward maximization and information gain. This is achieved by having a distribution over a particular function that is parameterized by a prior distribution over it.

We evaluate two permutations of the Q-learning algorithm: an agent with fixed exploration ($\epsilon = 0.1$) and an agent with decaying exploration ($\epsilon = 1$ decaying to 0). The pseudocode is presented in the supplementary materials. For both Q-learning agents, we specify the learning rate as 0.5 and discount factor as 0.99.

The Bayesian RL approach is a standard Dyna-style (Sutton, 1990) approach, where we train Q-learning agents in a belief-based internal model (planning), which accounts for uncertainty over both the transition model and reward function (i.e., separate prior distribution over both functions); the pseudocode is presented in the supplementary materials. The transition model encodes the probability for the next state, given the current state and action. These transition probability distributions are the same as the active inference generative model: high probability for intended move and extremely low probability for an implausible move. The reward function encodes the uncertainty about the reward location (an implicit contextual understanding about the environment). The likelihoods, for the transition model and reward function, are modeled via two separate Bernoulli distributions, with beta distributions as the conjugate prior over their parameters. The beta distribution pseudocounts for the reward and transition model—are initialized as 1. The posterior for the reward and that for the transition model distribution are evaluated by updating the prior ($Beta(\alpha, \beta)$). Thus, by treating them as pseudocounts, the evidence for intended move (likely reward location), x, is added to α, and an implausible move (unlikely reward location), y, is added to β: the posterior is $Beta(\alpha + x, \beta + y)$. The discount factor is specified as 0.9.

The Bayesian RL agent is a planning-based RL algorithm that parameterizes the transition and reward model using two separate Bernoulli distribution with a beta prior. At each episode, we sample k (=50), θ, from each of the beta distributions. Using the sampled priors, we define k MDPs and solve them using value iteration. This simulation gives us the Q-value function, which is averaged out to get the optimal Q-value. The optimal Q-value function is used to determine the next action and move to the next state. This procedure continues until the agent reaches the goal or falls down the hole. The process of solving the k MDP to determine the next action is similar to hypothetical rollouts in other planning-based algorithms.

Note that more sophisticated reinforcement learning schemes may have been more apt for solving this task (Daw, Courville, & Touretzky, 2006; Fuhs & Touretzky, 2007; Gershman & Niv, 2010; Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Gershman & Daw, 2017). However, our aim was to compare standard formulations of active inference and reinforcement learning. The rationale for this will become clearer when we compare the behavioral performance in the absence of reward: no motivation for adding

Table 1: Average Reward (and 95% Confidence Interval) for Each Agent, across Both Deterministic and Stochastic Environments.

		Average Score [95% CI]	
Algorithm	Belief-Based	Deterministic Environment	Stochastic Environment
Q-Learning ($\epsilon = 0.1$)	N	97.79 [97.41, 98.16]	66.08 [63.28, 68.88]
Q-Learning ($\epsilon = 1$ decaying to 0)	N	80.44 [78.96, 81.93]	65.13 [62.57, 67.68]
Bayesian RL	Y	99.76 [99.45, 100.00]	64.39 [60.33, 68.44]
Active Inference	Y	99.88 [99.64, 100.00]	98.90 [98.00, 99.79]
Active Inference (null model)	Y	50.03 [49.70, 50.35]	50.22 [49.89, 50.22]

Note: The results are calculated from the 200 trials across 500 episodes.

more heuristics because they cannot be justified in terms of increasing expected value of reward. Thus, the adopted RL agents are suited for our purpose.

3.4 Learning to Navigate the Frozen Lake. We evaluate how well the different agents are able to navigate the frozen lake in both stationary and nonstationary environments, as described below. Each of the environments was simulated for 200 trials with 500 episodes for the five agents: Q-learning ($\epsilon = 0.1$), Q-learning ($\epsilon = 1$ decaying to 0), Bayesian model-based reinforcement learning, active inference (see Figure 2) and active inference (null model; without any prior outcome preferences, i.e., $\log P(o|C) = 0$ for all outcomes). To aid intuition, the flattening of the prior preferences in the active inference model is equivalent to reclassifying reward as just another state or observation in a reinforcement learning scheme. While an agent would still be "told" whether it had encountered a rewarding stimulus, this would have no impact on the value function. As noted, this is not an exact equivalence: there is a philosophical distinction between making a change to the environment (in the reinforcement learning setting) and to the agent (in the active inference setting).

3.4.1 Stationary Environment. For this setup, the goal (G) exists at position 6 and hole (H) location at 8 for the entire experiment. We then evaluate the agent performance online and make no distinction between offline and online behavior modes. This is to better simulate exploration and exploitation in the real world, where we use the same policy to gather training data and act; indeed, it is this exact paradigm that is one of the major motivators for active inference. The average score (see Table 1) for all agents, except the null model specification of the active inference model, was considerably high at over 80, showing that all frameworks were able to solve the MDP.

The low score for the null (active inference) model reflects the lack of prior preferences for the type of outcomes the agent would like to observe—that is, it does not differentiate among any of the different patches (S, F, G, and H) in the frozen lake. The null model does not seek out the goal state (i.e., reward), which it does not prefer over other states. Instead, it either falls in the hole or reaches the goal with equal probability. From the perspective of information seeking, this is a very sensible policy, as the same amount of information can be obtained by finding either the hole (which tells us where the goal is) or the goal (which tells us where the hole is). As such, there is nothing to disambiguate between the two. Over subsequent exposures to the environment, given that it is stationary, this agent will be left with little uncertainty to resolve, as it will know everything at the start of a trial based on experience. Over time, this will lead to a loss of purposeful behavior, resembling what we might expect from a reinforcement learning agent in the absence of any environmental rewards (even in the presence of uncertainty).

The learning curve, shown in Figure 3, highlights that the active inference and Bayesian model-based reinforcement learning agent learn reward-maximizing behavior (and resolve uncertainty about reward location) in a short amount of time (fewer than 10 episodes). They are able to maintain this for the remaining trials. This is reflected by the tight confidence intervals around the average reward for both agents. In contrast, Q-learning ($\epsilon = 0.1$), while also quickly learning appropriate state-action pairing, has slightly larger confidence intervals for the average reward due to the 10% of selecting a random action.

3.4.2 Nonstationary Environment. We introduce nonstationarity into the environment; the location of the (G) and (H) are flipped after a certain number of episodes. Initially (G) is located at position 6 and (H) at position 8, and then we swap (G) and (H) at the following time steps: 21, 121, 141, 251, 451. This means that after episode 451, (G) remains at position 8 until the end of the simulation. These changes in the reward location test how quickly the agent can relearn the correct (G) location. The average score for all agents is presented in Table 1.

As in the stationary setup, all agents are initially uncertain about the reward location. Figure 4 illustrates the performance of these agents. This is quickly resolved, and by episode 20, active inference, Bayesian RL, and Q-learning ($\epsilon = 0.1$) exhibit appropriate behavior for solving the task. The null (active inference) model and Q-learning ($\epsilon = 1$ decaying to 0.00) exhibit fairly poor performance (consistent with stationary). However, at episode 21, the performance for all agents drops to zero due to the change in reward location—except for the agent with no preferences, who persists with achieving the reward half of the time. For the reinforcement learning (Q-learning and Bayesian RL) agents, this drop in performance persists for the next 40 or so episodes. This is because by treating this as a "learning"

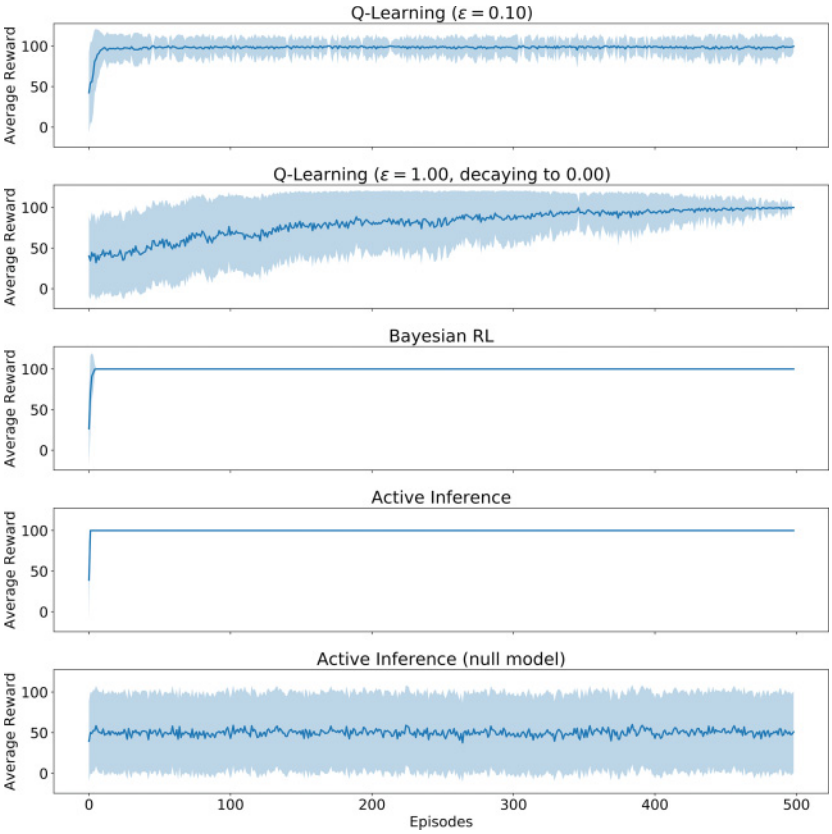

Figure 3: Learning curve for deterministic environment. The x-axis denotes the episode number and y-axis the average (online) reward. The results presented are calculated from 200 trials.

problem, the agent has to do the following: (1) reverse the learning of its previous understanding of the reward location and (2) relearn the current reward location. In contrast, by treating this as a planning-as-inference problem, the active inference agent is able to quickly recover performance after a single episode, as the generative model takes into account the context switch. In other words, the agent simply infers that a switch has happened, and acts accordingly. This quick performance recovery is persistent for all changes in reward location across the 500 episodes (see Figure 4). However, for Bayesian RL, the ability to adapt its behavior to the changing goal locations continues to prove difficult; each time a greater number of episodes are required to reverse the learning of the prior distribution over the reward function due to the accumulation of pseudocounts. This contrasts with

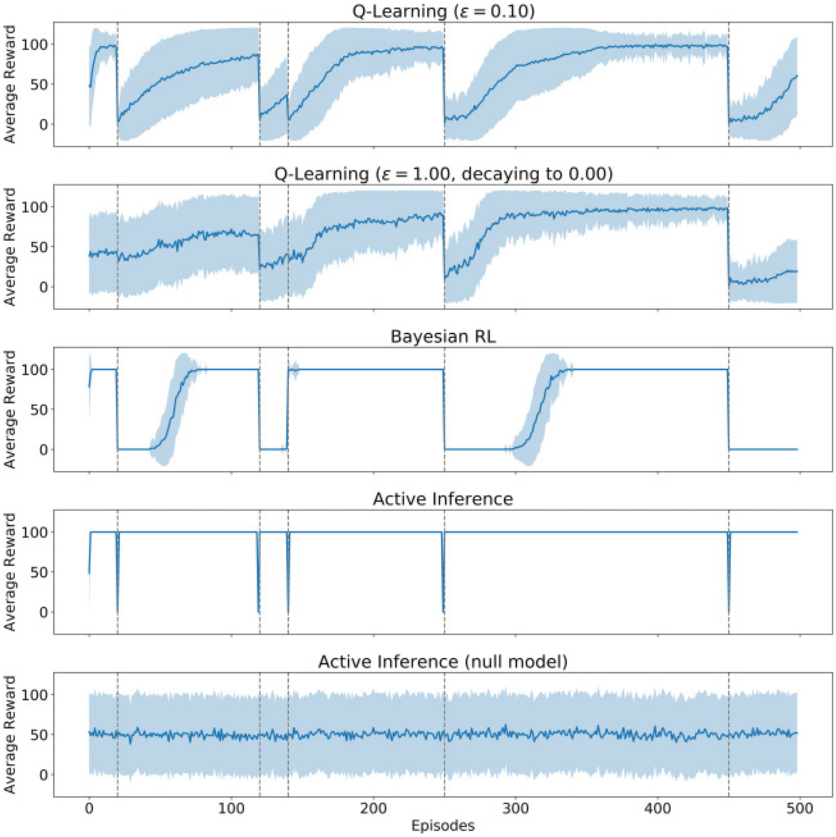

Figure 4: Learning curve for stochastic environment. The x-axis denotes the episode number and y-axis the average (online) reward. The results presented are calculated from 200 trials. The dotted gray lines represent the change in (G) (and (H)) location.

Q-learning ($\epsilon = 0.1$), which adapts fairly quickly to these fluctuating reward locations because it needs to only update the appropriate state and action Q-values.

Therefore, for nonstationary environments, active inference offers an attractive, natural adaptation mechanism—for training artificial agents—due to its Bayesian model updating properties. This is in contrast to standard reinforcement learning, where issues of environmental nonstationarity are not accommodated properly, as shown through the above simulations. They can be dealt with using techniques that involve the inclusion of inductive biases, such as, importance sampling of experiences in multi-agent environments (Foerster et al., 2017) or using metalearning to adapt

Table 2: Reward Shaping: Average Score and Number of Moves across 100 Episodes for 100 Agents.

Rewards			Average Score (Average Number of Moves)		
(G)	(H)	(F)	Q-Learning* ($\epsilon = 0.1$)	Bayesian RL	Active Inference
0.00	0.00	0.00	0.00 (15.00)	39.94 (9.17)	44.00 (8.67)
0.00	−100	0.00	0.00 (15.00)	0.00 (15.00)	0.00 (15.00)
100	−100	0.00	95.56 (3.53)	99.77 (3.02)	99.52 (3.03)
100	0.00	−10.0	96.00 (3.48)	99.89 (3.00)	99.47 (3.00)
100	−100	−10.0	96.47 (3.42)	99.79 (3.01)	99.58 (3.00)
100	0.00	0.00	95.32 (3.58)	99.74 (3.00)	99.50 (3.07)

Note: For this experiment, we evaluate under $\epsilon = 0.0$ (i.e., on-policy).

gradient-update approaches more quickly (Al-Shedivat et al., 2017). Finally, we acknowledge that the simulations presented limit the comparison between active inference and standard (i.e., naive) reinforcement learning schemes. We could have introduced further complexity (i.e., additional distributions to be learned) within the Bayesian reinforcement learning model—for example, explicit beliefs about latent contexts instead of implicit context encoding via the reward location (Gershman & Niv, 2010; Gershman, Norman, & Niv, 2015; Rakelly, Zhou, Finn, Levine, & Quillen, 2019). To evaluate the comparison of more complex reinforcement learning agents, to active inference, in nonstationary environments remains an outstanding research question. We appreciate that with additional design choices the Bayesian reinforcement learning agent may exhibit similar behavioral performance to the active inference agent.

3.5 Comparing Prior Preferences and Rewards. In reinforcement learning, goals are defined through reward functions, that is, explicit scalar signal from the environment. In contrast, in active inference, goals are defined through the agent's prior preferences over outcomes. We now illustrate the link between these definitions of goal-directed behavior by presenting experiments that show the effect of reward shaping (Ng, 2003) in the FrozenLake stationary environment (see Table 2).

We apply the following shaping: modifying the reward for reaching the goal (G), modifying the reward for falling down the hole (F), and modifying the reward for any state that isn't a goal (H) (this can be considered a "living cost"). In order to convert the shaped reward into prior preferences, we manipulate the prior preferences such that their relative weighting matches that introduced through the reward shaping for example, a reward of -100 is equivalent to prior preferences of $-\log(5)$.

As our experiments show, when we define a prior preference through a reward function, the behaviors of the belief-based policies (i.e., Bayesian

RL and active inference) are nearly identical and learn to solve the environment as soon as a positive reward is defined for the goal. On the other hand, the nonprobabilistic Q-learning approach appears more sensitive to reward shaping, with living costs causing greedier behavior (i.e., taking fewer steps per episode). A possible explanation for this is that the construction of the generative models for both Bayesian RL and active inference clearly define that the location of the goal/hole is in either state 6 or 8 hence, Bayes-optimal behavior (i.e., getting to the goal in as few steps as possible) can be learned even in the absence of negative rewards/preferences over certain states. All that is required is some notion of where the goal state might exist—hence the ability to learn Bayes-optimal policies by only specifying the goal location (see the last row of Table 2).

Another interesting behavior is when there is an absence of preferences/rewards (see the first row of Table 2). The Q-learning approach learns a deterministic circular policy with little exploration despite the ϵ term since it does not update its parameters due to the lack of reward signal. The belief-based approaches on the other hand maintain exploration throughout, represented by the average score ranging between 40 and 45, as their probabilistic models remain uniform over the beliefs of which transitions produce preferred outcomes. This suggests that by having an objective function that is optimizing a dual objective, the agent exhibits some sort information gain (i.e., exploring the world is intrinsically motivated because it helps the agent build a better model of the world). We will see more purposeful exploration, under active inference, in the next section.

Finally, we observe that all three approaches learn the same circular behavior when only a negative preference or reward is specified (see the second row of Table 2). This is because all the approaches learn to avoid the hole state (H), but since there is no notion of goal-seeking behavior, the agents do not learn to go to the goal state. Interestingly, in the case of the belief-based approaches (Bayesian RL and active inference), since the generative model defines the presence of hole states in either state 6 or 8 and since it receives no preference for goal states, the generative model assigns nonzero probability with the hole state being in either state 6 or 8. As a result, policies derived from these generative models learn to avoid both states, therefore terminating only when the time limit is reached.

Through this brief study, we have illustrated an implicit equivalence between Bayesian model-based reinforcement learning and active inference. This equivalence rests on treating prior preferences as a reward function. In other words, by expressing an arbitrary reward function as a potential function (i.e., a log probability over future outcomes), reward functions can be absorbed into expected free energy. This means one can elicit identical behaviors from reinforcement learning and active inference. Indeed, if one removes uncertainty, in the form of epistemic value, we are left with pragmatic value: expected future reward. This shows that reinforcement learning can be regarded as a limiting or special case of model-based approaches in general—or active inference in particular. However, the FrozenLake environment is by no means representative of all discrete environments, and this merits further research. It is important to note that behavioral equivalences are a result purely of the environmental setup and the accompanying reward signal (e.g., changing the FrozenLake environment for a maze with noisy attractor state and no reward) might reveal additional behavioral differences between Bayesian model-based reinforcement learning and active inference.

3.6 Learning Prior Outcome Preferences. In some settings, explicitly defining prior outcome preferences might be challenging due to time-dependent preferences, an inability to disambiguate between different types of outcomes, or simply lack of domain knowledge. In those instances, the appropriate distribution of prior outcome preferences can be learned via the agent's interaction with the environment. This difficulty extends to reinforcement learning, where defining a reward function may not be possible, and in its vanilla formulation, reinforcement learning offers no natural way to learn behaviors in the absence of a reward function (see the first row of Table 2).

In order to demonstrate the ability of active inference to select policies in the absence of prespecified prior preferences, we allow both the likelihood distribution ($\log P(o|s)$) and outcome preferences ($\log P(o|C)$) to be learned. This allows us to make explicit that whether a state is rewarding or not is determined by the agent learning its prior preferences and it is not a specific signal from the environment. For this, the generative model is extended to include prior beliefs about the parameters of these two distributions (a prior over priors in the case of ($\log P(o|C)$), which are learned through belief updates (Friston, FitzGerald, et al., 2017). The natural choice for the conjugate prior for both distributions is a Dirichlet distribution, given that the probability distributions are specified as a categorical distribution. This means that the probability can be represented simply in terms of Dirichlet concentration parameters. We define the Dirichlet distribution (for both likelihood and prior preferences) as completely flat (initialized as 5 for likelihood and 1 for prior preferences for all possible options). This is in contrast to row 1 of Table 2, where we specify flat prior preferences but the agent is not equipped with (Dirichlet) hyperpriors that enable the agent to learn about the kind of outcomes it prefers.

Incrementally, we enabled learning of these parameters. First, all outcome preferences (and their Dirichlet priors) are removed. Therefore, the agent can only learn the likelihood. As a result, there is no behavioral imperative other than pure exploration (Schmidhuber, 2006). This setup was simulated 15 times, and likelihood was learned in an experience-dependent fashion. This results in an initial (exploratory) trajectory that covers all unchartered territory in the most efficient way possible, that is, there is no revisiting of locations that have already been encountered (see Figure 5.1).

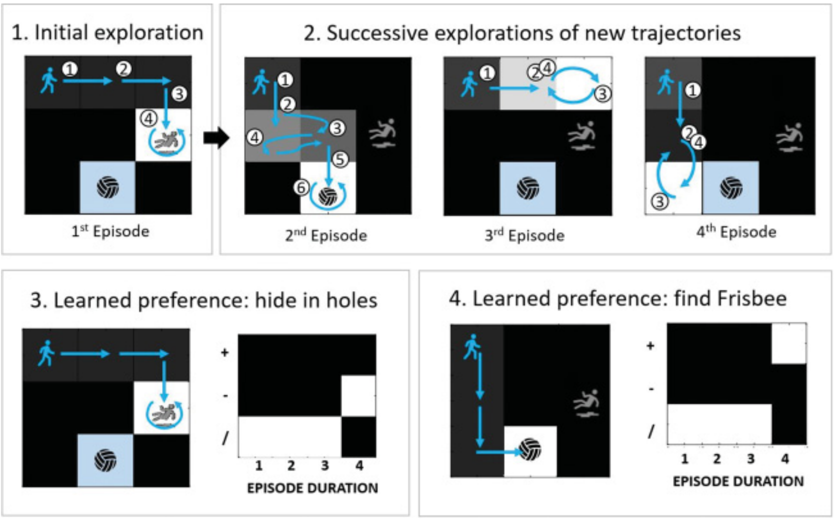
1. Initial exploration
1st Episode
2. Successive explorations of new trajectories
2nd Episode
3rd Episode
4th Episode
3. Learned preference: hide in holes
4. Learned preference: find Frisbee
+
-
/
1 2 3 4
EPISODE DURATION

Figure 5: Parameter learning for a single reward location: results for likelihood learning presented in panels 1 and 2 and prior preference learning presented in panels 3 and 4. Blue arrows denote the trajectory taken, and numbers in the circles denote the trajectory sequence. Circular arrows represent loops: once in that state, the same outcome is observed until the maximum number of moves reached (15). Panel 1 is a pictorial representation of the first episode trajectory, with no prior preference: right(1 → 2), right (2 → 3), down(3 → 6), right(6 ↔ 6). Panel 2 depicts the next four episodes from the trial. Panel 3 has two figures: a pictorial representation of the trajectory to the hole and a heat map of the accumulated Dirichlet parameters for score (+is positive, — is negative, and/is neutral). For this trial, there is a strict preference for holes at time step 4. Panel 4 presents similar information but for a goal-preferring agent: a pictorial representation of the trajectory to the goal and a heat map of the accumulated Dirichlet parameters for score. There is a strict preference for goals at time step 4.

Furthermore, this behavior persists past the initial exploration, with continuous explorations via new (nonoverlapping) trajectories (see Figure 5.2). This represents "true" exploratory behavior, distinct from random action selection, of the sort possible only in a belief-based scheme. Furthermore, as there are no rewards, this behavior would be impossible to motivate from a reinforcement learning perspective, as this learning is for its own sake, not to improve reward seeking. While such an imperative could plausibly be introduced to a belief-based reinforcement learning scheme, it would have to appeal to heuristic arguments like the potential for a reward function to be introduced in the future.

Next, we equip the agent with the ability to learn outcome preferences (rather than learn about the environment). This entails updating the outcome preferences via accumulation of Dirichlet parameters, without learning the likelihood. The setup was simulated 10 times, for two separate kinds of outcome. During the first kind, in the absence of negative preferences, holes become attractive because they are encountered first, and this is what the agent learns about its behavior (and implicit preferences). In other words, because holes (H) are absorbing states and the agent observes itself falling in a hole recurrently, it learns to prefer this outcome (see Figure 5.3). Similarly, in the second kind of trial, the agent finds itself recurrently acquiring the Frisbee. This causes it to exhibit preferences for acquiring Frisbees (see Figure 5.4). These represent the capacity of active inference agents to develop into hole-seeking or Frisbee-seeking agents. As one of these outcomes becomes more familiar, the agent observes its own behavior and concludes, "I am the sort of creature that enjoys spending time in holes [or with Frisbees]," and adjusts future behavior to be consistent with this.

This capacity is another important point of distinction with reinforcement learning approaches, where the problem is defined in terms of a prespecified reward function. If this is the problem one hopes to solve, it is clearly undesirable for agents to develop ulterior motives. This speaks to the fundamental differences in the problems being solved by the two approaches. Under active inference, the ultimate goal is to maintain a coherent phenotype and persist over time. Hole-seeking agents achieve this despite their behavior deviating from what an observer, or the designer of an AI gym game, might regard as appropriate.

Finally, we look at the interaction between the epistemic imperatives to resolve uncertainty about the likelihood mapping and uncertainty about prior preferences. This setup was simulated 10 times and both likelihood distribution and prior outcome preferences learned. By parameterizing both the likelihood and prior outcome preferences with Dirichlet distributions, we induce a contribution to expected free energy that makes visiting every location attractive (i.e., every location acquires epistemic affordance or novelty). However, after a sufficient number of trials, the agent has learned (i.e., reduced its uncertainty) that it prefers to hide in holes (see Figure 6). This causes the agent to exhibit the exploitative behavior of hiding rather than continue exploring. After five trials, the agent goes straight to the hole.

This is an interesting example of how by observing one's own behavior, preference formation contextualizes the fundamental imperative to explore.

It is important to note that the learned outcome preferences are time dependent: the agent prefers to visit safe (F) patches for the first three time points and then visit goal (G) patches with a high preference (see Figure 6). As noted, these are learned by accumulating experience (in the form of Dirichlet concentration parameters) such that uniform priors over outcomes become precise posteriors. These precise posteriors then become the

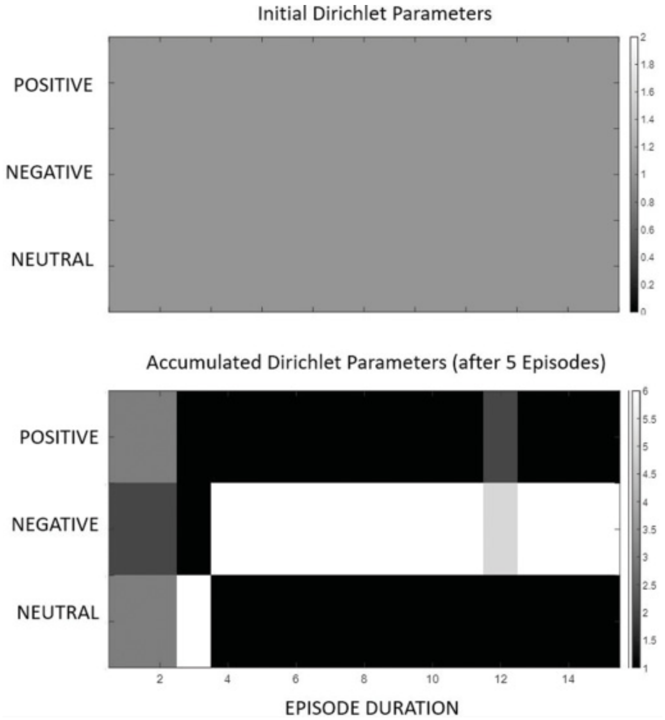

Figure 6: Learning prior outcome preferences for outcome modality, score: initial (top) and after five episodes (bottom) for a single reward location.

agent's preferences. Put simply, it has learned that this is the kind of creature it is.

We have observed that even in the absence of clearly defined prior preferences, active inference agents are able to learn these preferences naturally; since prior preferences are defined in terms of probability distributions, we simply define a distribution over distributions and learn these from the data using the standard inference/gradient updates (see section 2.4). However, it is important to highlight that these learned prior preference might be at odds with the reward from the environment. This conceptualization flips rewarding states on its head—a matter of preference, not a specific scalar signal from the environment. Concretely, we can indeed encourage AI agents to solve RL environments by placing a prior preference that maximizes the observation corresponding to reward, but definitionally, active inference does not require the resultant reward-maximizing behavior to be considered a successful agent. As long as it can learn and then maintain a consistent set of behaviors over time—through free energy minimization—we consider such an agent to be successful under the active inference problem definition.

Furthermore, by allowing various parts of the active inference framework to be learned from the environment (i.e., $\log P(o|s)$), we can infer time-dependent preferences from the environment. This is in contrast to vanilla reinforcement learning, where it is less clear how to naturally account for learning an intrinsic reward function, with many competing approaches (Still & Precup, 2012; Mohamed & Rezende, 2015; Pathak et al., 2017).

4 Discussion

We have described active inference, and the underlying minimization of variational and expected free energy, using a (simplified) discrete state-space and time formulation. Throughout this review, we have suggested that active inference can be used as framework to understand how agents (biological or artificial) operate in dynamic, nonstationary environments (Friston, Rosch, et al., 2017), via a standard gradient descent on a free energy functional. More generally, active inference can be thought of as a formal way of describing the behavior of random dynamical systems with latent states.

As noted in the formulation of active inference (see equation 2.23), epistemic foraging (or exploration) emerges naturally. This is captured by the desire to maximize the mutual information between outcomes and the hidden states on the environment. Exploration means that the agent seeks out states that afford outcomes, which minimize uncertainty about (hidden) states of affairs. In the FrozenLake simulation, this was highlighted by the initial exploratory move made by the agent due to uncertainty about reward location. The move resolved the agent's uncertainty about the reward location, and all subsequent episodes (when the reward location remained consistent) exploited this information. Note that in the formulation presented, we discussed model parameter exploration that might also be carried out by the agent, when learning either the likelihood or prior preferences, by having priors over the appropriate probability distributions and applying the expected free energy derivations to those parameters (Schwartenbeck et al., 2019). The simulations showed that in the absence of a reward signal from the environment, the agent could learn a niche and exhibit self-evidencing behavior. Additionally, it highlighted that due to the fundamental differences in the conceptual approach, active inference agents may exhibit Bayes-optimal behavior that is counterintuitive from the perspective of reinforcement learning (i.e., reward minimization). However, from an active inference perspective, reward is simply the sort of outcome that is preferred, and an agent can learn to prefer other sorts of outcomes.

The canonical properties presented with respect to decision making under uncertainty are usually engineered in conventional reinforcement learning schemes. However, more sophisticated formulations of reinforcement learning define a central role for uncertainty over Q-value functions (Dearden, Friedman, & Russell, 1998; Dearden, Friedman, & Andre, 2013; Osband, Blundell, Pritzel, & Van Roy, 2016; O'Donoghue, Osband, Munos, & Mnih, 2018), MDP (Dearden et al., 2013; Osband, Blundell, Pritzel, & Van Roy, 2016; Zintgraf et al., 2019) or even the reward function (Sorg, Singh, & Lewis, 2012; Fürnkranz, Hüllermeier, Cheng, & Park, 2012; Zintgraf et al., 2019). The formulation presented in Zintgraf et al. (2019) incorporates uncertainty over both the model parametrization and reward function. This suggests that there is potential to build on (and remove components of) Bayesian reinforcement learning algorithms to render them formally equivalent to active inference. However, this may come at increased algorithmic complexity cost and loss of generalization. Additionally, these algorithmic design choices are nontrivial and may demonstrate counterintuitive behavior (O'Donoghue, Osband, & Ionescu, 2020). In contrast, active inference enables decision making under uncertainty with no heuristics in play.

The simulations reveal that once the reward signal is removed, the active inference exhibits information-seeking behavior (to build a better model of its environment), similar to the Bayesian reinforcement-learning agent. This type of reward-free learning has been central to the curiosity literature in reinforcement learning despite by definition not being true to the definition of reinforcement learning. Concretely, these approaches induce an intrinsic reward using some heuristic—dynamics prediction (Pathak et al., 2017), random feature prediction (Burda, Edwards, Storkey, & Klimov, 2018), or information gain (Mohamed & Rezende, 2015)—but this does not necessarily align with the axiomatic goal of maximizing a numerical reward signal; they are simply tools (i.e., inductive biases) that may lead us to achieve this, and in the case of completely absent rewards, it is unclear what the goal of reinforcement learning is (i.e., What behaviors are we reinforcing?).

Our treatment emphasizes that, via a belief-based scheme, active inference enables us to specify prior beliefs over preferred outcomes or not (to produce purely epistemic behavior). Practically, these can produce similar outcomes and have behavioral equivalences to the reward function in reinforcement learning by assigning high and low prior preferences to outcomes with positive and negative rewards, respectively. Moreover, this highlights a conceptual distinction between prior beliefs over preferred outcomes in active inference and reward functions in standard reinforcement learning. While a reward function specifies how an agent should interact with the environment, prior beliefs over preferred outcomes are a description (via some particular instantiation) of how the agent wishes to behave. Crucially, this description can be learned over time, based on relative frequencies of outcomes encountered. This speaks to an eliminative use of Bayes optimality, which replaces the notion of reward as a motivator of behavior with prior beliefs about the outcomes an agent works toward. Conceptually, this dissolves the tautology of reinforcement learning, that rewards reinforce behaviors that secure rewards. Having said this, related formulations can be found in reinforcement learning—for example, belief-based reward functions in Sorg et al. (2012), Fürnkranz et al. (2012), and Zintgraf et al. (2019).

In active inference, the an agent is likely to maximize extrinsic value (cf. expected reward) by having prior preferences about unsurprising outcomes (see equation 2.22) via the minimization of expected free energy. It is important to note that the minimization of expected free energy is achieved by choosing appropriate policies (sequences of actions). We accounted for this in the initial setup of the FrozenLake simulation, where the agent had strong positive preference for finding the Frisbee. Additionally, hole locations were associated with strong negative preferences. In contrast, the active inference null model with no prior preferences and no ability to learn them encouraged exploratory behavior, and the agent ended in the (G) location 44.0% of the time.

However, it is worth noting that these properties follow from the form of the underlying generative model. The challenge is to identify the appropriate generative model that best explains the generative process (or the empirical responses) of interest (Gershman & Beck, 2017). In the FrozenLake simulation, by equipping the agents with beliefs about the current context, we were able (via the generative model and its belief updating process) to convert a learning problem into a planning-as-inference problem. However, this can be treated as a learning problem by specifying a hierarchical MDP with learning capacity over the problem space. This would allow for slow-moving dynamics at a higher level that account for changes in context and fast-moving dynamics at the lower level that equip the agent with the ability to navigate the given instantiation of the FrozenLake (Friston, Rosch, et al., 2017). When comparing prior preferences and rewards, we highlighted that due to no explicit prior preference for goal states, the belief-based (active inference and Bayesian RL) agents exhibit conservative behaviors; choosing to avoid the (G) state. This behavior is a caveat of the underlying generative model form—uncertainty modeled over the location of the (G) & (H) state—and manipulating the prior probability distributions (or the factorization of the states) might lead to policies where agents chooses to not avoid the (G) location. Additionally, the generative models underlying this active inference formulation can be equipped with richer forms (e.g., via amortization) or learned via structural learning (Gershman & Niv, 2010; Tervo, Tenenbaum, & Gershman, 2016). Thus, if one was to find the appropriate generative model, active inference could be used for a variety of problems: robotic arm movement, dyadic agents, and playing Atari games, for example. We note that the task of defining the appropriate generative model (discrete or continuous) might be difficult. Thus, future work should look to incorporate implicit generative models (based on feature representation from empirical data) or shrinking hidden state-spaces by defining transition probabilities based on likelihood (rather than latent states).

Software Note

The simulations presented in this letter are available at https://github.com/ucbtns/dai.

Acknowledgments

N.S. is funded by the Medical Research Council (MR/S502522/1). P.J.B. is funded by the Willowgrove Studentship. K.J.F. is funded by the Wellcome Trust (088130/Z/09/Z). We thank the anonymous reviewers for their suggestions and insightful comments on the manuscript.

Disclosure Statement

We have no disclosures or conflict of interest.

References

Al-Shedivat, M., Bansal, T., Burda, Y., Sutskever, I., Mordatch, I., & Abbeel, P. (2017). *Continuous adaptation via meta-learning in nonstationary and competitive environments.* arXiv:1710.03641.

Alagoz, O., Hsu, H., Schaefer, A. J., & Roberts, M. S. (2010). Markov decision processes: A tool for sequential decision making under uncertainty. *Medical Decision Making*, *30*(4), 474–483.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete problems in AI safety*. arXiv:1606.06565.

Astrom, K. J. (1965). Optimal control of Markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, *10*(1), 174–205.

Attias, H. (2003). Planning by probabilistic inference. In *AISTATS*. Citeseer.

Baker, C. L., & Tenenbaum, J. B. (2014). Modeling human plan recognition using Bayesian theory of mind. In G. Sukthankar, C. Geib, H. Bui, D. Pynadath, & R. Goldman (Eds.), *Plan, activity, and intent recognition: Theory and practice* (pp. 177–204). San Mateo, CA: Morgan Kaufmann.

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

Beck, J., Pouget, A., & Heller, K. A. (2012). Complex inference in neural circuits with probabilistic population codes and topic models. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems*, *25* (pp. 3059–3067). Red Hook, NY: Curran.

Bellman, R. (1952). On the theory of dynamic programming. In *Proceedings of the National Academy of Sciences of the United States of America*, *38*(8), 716.

Blau, T., Ott, L., & Ramos, F. (2019). *Bayesian curiosity for efficient exploration in reinforcement learning*. arXiv:1911.08701.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, *112*(518), 859–877.

Bogacz, R. (2017). A tutorial on the free-energy framework for modelling perception and learning. *Journal of Mathematical Psychology*, *76*, 198–211.

Botvinick, M., & Toussaint, M. (2012). Planning as inference. *Trends in Cognitive Sciences*, *16*(10), 485–488.

Buckley, C. L., Kim, C. S., McGregor, S., & Seth, A. K. (2017). The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology*, *81*, 55–79.

Burda, Y., Edwards, H., Storkey, A., & Klimov, O. (2018). *Exploration by random network distillation.* arXiv:1810.12894.

Camacho, E. F., & Alba, C. B. (2013). *Model predictive control*. New York: Springer Science & Business Media.

Cao, F., & Ray, S. (2012). Bayesian hierarchical reinforcement learning. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems*, *25* (pp. 73–81). Red Hook, NY: Curran.

Cesa-Bianchi, N., Gentile, C., Lugosi, G., & Neu, G. (2017). Boltzmann exploration done right. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems*, *25* (pp. 6284–6293).

Chong, E., Familiar, A. M., & Shim, W. M. (2016). Reconstructing representations of dynamic visual objects in early visual cortex. In *Proceedings of the National Academy of Sciences*, *113*(5), 1453–1458.

Crauel, H., & Flandoli, F. (1994). Attractors for random dynamical systems. *Probability Theory and Related Fields*, *100*(3), 365–393.

Cullen, M., Davey, B., Friston, K. J., & Moran, R. J. (2018). Active inference in OpenAI gym: A paradigm for computational investigations into psychiatric illness. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*(9), 809–818.

Da Costa, L., Parr, T., Sajid, N., Veselic, S., Neacsu, V., & Friston, K. (2020). *Active inference on discrete state-spaces: A synthesis.* arXiv:2001.07203.

Daw, N. D., Courville, A. C., & Touretzky, D. S. (2006). Representation and timing in theories of the dopamine system. *Neural Computation*, *18*(7), 1637–1677.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215.

Dearden, R., Friedman, N., & Andre, D. (2013). *Model-based Bayesian exploration.* arXiv:1301.6690.

Dearden, R., Friedman, N., & Russell, S. (1998). Bayesian q-learning. In *Proceedings of the 15th National Conference on Artificial Intelligence* (pp. 761–768). Palo Alto, CA: AAAI.

Foerster, J. N., Chen, R. Y., Al-Shedivat, M., Whiteson, S., Abbeel, P., & Mordatch, I. (2017). *Learning with opponent-learning awareness*. arXiv:1709.04326.

Friston, K. (2019). *A free energy principle for a particular physics.* arXiv:1906.10184.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2016). Active inference and learning. *Neuroscience and Biobehavioral Reviews*, *68*, 862–879.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: A process theory. *Neural Computation*, *29*(1), 1–49.

Friston, K., Mattout, J., & Kilner, J. (2011). Action understanding and active inference. *Biological Cybernetics*, *104*(1–2), 137–160.

Friston, K. J., Parr, T., & de Vries, B. (2017). The graphical brain: Belief propagation and active inference. *Network Neuroscience*, *1*(4), 381–414.

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive Neuroscience*, *6*(4), 187–214.

Friston, K. J., Rosch, R., Parr, T., Price, C., & Bowman, H. (2017). Deep temporal models and active inference. *Neuroscience and Biobehavioral Reviews*, *77*, 388–402.

Friston, K., Samothrakis, S., & Montague, R. (2012). Active inference and agency: Optimal control without cost functions. *Biological Cybernetics*, *106*(8–9), 523–541.

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., & Dolan, R. J. (2014). The anatomy of choice: Dopamine and decision-making. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1655), 20130481.

Fuhs, M. C., & Touretzky, D. S. (2007). Context learning in the rodent hippocampus. *Neural Computation*, *19*(12), 3173–3215.

Fürnkranz, J., Hüllermeier, E., Cheng, W., & Park, S.-H. (2012). Preference-based reinforcement learning: A formal framework and a policy iteration algorithm. *Machine Learning*, *89*(1–2), 123–156.

Gershman, S. J., & Beck, J. M. (2017). Complex probabilistic inference. In A. Mopustafa (Ed.), *Computational models of brain and behavior*. Hoboken, NJ: Wiley.

Gershman, S. J., & Daw, N. D. (2017). Reinforcement learning and episodic memory in humans and animals: An integrative framework. *Annual Review of Psychology*, *68*, 101–128.

Gershman, S. J., & Niv, Y. (2010). Learning latent structure: Carving nature at its joints. *Current Opinion in Neurobiology*, *20*(2), 251–256.

Gershman, S. J., Norman, K. A., & Niv, Y. (2015). Discovering latent causes in reinforcement learning. *Current Opinion in Behavioral Sciences*, *5*, 43–50.

Ghavamzadeh, M., Mannor, S., Pineau, J., & Tamar, A. (2016). Bayesian reinforcement learning: A survey. *Foundations and Trends in Machine Learning*, *8*(5–6):359–483.

Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). *Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor*. arXiv:1801.01290.

Igl, M., Ciosek, K., Li, Y., Tschiatschek, S., Zhang, C., Devlin, S., & Hofmann, K. (2019). Generalization in reinforcement learning with selective noise injection and information bottleneck. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems*, *32* (pp. 13978–13990). Red Hook, NY: Curran.

Igl, M., Zintgraf, L., Le, T. A., Wood, F., & Whiteson, S. (2018). *Deep variational reinforcement learning for POMDPs*. arXiv:1806.02426.

Kolter, J. Z., & Ng, A. Y. (2009). Near-Bayesian exploration in polynomial time. In *Proceedings of the 26th Annual International Conference on Machine Learning* (pp. 513–520). New York: ACM.

Lee, L., Eysenbach, B., Parisotto, E., Xing, E., Levine, S., & Salakhutdinov, R. (2019). *Efficient exploration via state marginal matching*. arXiv:1906.05274.

Levine, S. (2018). *Reinforcement learning and control as probabilistic inference: Tutorial and review*. arXiv:1805.00909.

Lloyd, K., & Leslie, D. S. (2013). Context-dependent decision-making: A simple Bayesian model. *Journal of the Royal Society Interface*, *10*(82), 20130069.

Millidge, B., Tschantz, A., & Buckley, C. L. (2020). *Whence the expected free energy?* arXiv:2004.08128.

Millidge, B., Tschantz, A., Seth, A. K., & Buckley, C. L. (2020). *On the relationship between active inference and control as inference*. arXiv:2006.12964.

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

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., . . . Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *Proceedings of the International Conference on Machine Learning* (pp. 1928–1937). Madison, WI: Omnipress.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). *Playing Atari with deep reinforcement learning*. arXiv:1312.5602.

Mohamed, S., & Rezende, D. J. (2015). *Variational information maximisation for intrinsically motivated reinforcement learning*. arXiv:1509.08731.

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

Ng, A. Y., (2003). *Shaping and policy search in reinforcement learning*. PhD diss., University of California, Berkeley.

O'Donoghue, B., Osband, I., & Ionescu, C. (2020). *Making sense of reinforcement learning and probabilistic inference*. arXiv:2001.00805.

O'Donoghue, B., Osband, I., Munos, R., & Mnih, V. (2018). The uncertainty Bellman equation and exploration. In *Proceedings of the International Conference on Machine Learning* (pp. 3836–3845). New York: ACM.

Osband, I., Blundell, C., Pritzel, A., & Van Roy, B. (2016). Deep exploration via bootstrapped DQN. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems*, *29* (pp. 4026–4034). Red Hook, NY: Curran.

Padakandla, S., & Bhatnagar, S. 2019. *Reinforcement learning in non-stationary environments*. arXiv:1905.03970.

Parr, T., & Friston, K. J. (2017). Uncertainty, epistemics and active inference. *Journal of the Royal Society Interface*, *14*(136), 20170376.

Parr, T., & Friston, K. J. (2018). The discrete and continuous brain: From decisions to movement—and back again. *Neural Computation*, *30*(9), 2319–2347.

Parr, T., & Friston, K. J. (2019a). The computational pharmacology of oculomotion. *Psychopharmacology*, *236*(8), 2473–2484.

Parr, T., & Friston, K. J. (2019b). Generalised free energy and active inference. *Biological Cybernetics*, *113*(5–6), 495–513.

Parr, T., Markovic, D., Kiebel, S. J., & Friston, K. J. (2019). Neuronal message passing using mean-field, Bethe, and marginal approximations. *Scientific Reports*, *9*(1), 1–18.

Pathak, D., Agrawal, P., Efros, A. A., & Darrell, T. (2017). Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 16–17). Piscataway, NJ: IEEE.

Poupart, P. (2018). Lecture slides on Bayesian reinforcement learning from cs885: https://cs.uwaterloo.ca/ppoupart/teaching/cs885-spring18/slides/cs885-lecture10.pdf

Rakelly, K., Zhou, A., Finn, C., Levine, S., & Quillen, D. (2019). Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *Proceedings of the International Conference on Machine Learning* (pp. 5331–5340). New York: ACM.

Ross, S., Chaib-draa, B., & Pineau, J. (2008). Bayes-adaptive POMDPs. In J. C. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in neural information processing systems*, *20* (pp. 1225–1232). Cambridge, MA: MIT Press.

Schmidhuber, J. (2006). Developmental robotics, optimal artificial curiosity, creativity, music, and the fine arts. *Connection Science*, *18*(2):173–187.

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

Schwartenbeck, P., Passecker, J., Hauser, T. U., FitzGerald, T. H., Kronbichler, M., & Friston, K. J. (2019). Computational mechanisms of curiosity and goal-directed exploration. *Elife*, *8*, e41703.

Seifert, U. (2012). Stochastic thermodynamics, fluctuation theorems and molecular machines. *Reports on Progress in Physics*, *75*(12), 126001.

Sekar, R., Rybkin, O., Daniilidis, K., Abbeel, P., Hafner, D., & Pathak, D. (2020). *Planning to explore via self-supervised world models*. arXiv:2005.05960.

Sorg, J., Singh, S., & Lewis, R. L. (2012). *Variance-based rewards for approximate Bayesian reinforcement learning*. arXiv:1203.3518.

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

Sutton, R. S. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Machine Learning Proceedings 1990* (pp. 216–224). Amsterdam: Elsevier.

Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning*. Cambridge, MA: MIT Press.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Tervo, D. G. R., Tenenbaum, J. B., & Gershman, S. J. (2016). Toward the neural implementation of structure learning. *Current Opinion in Neurobiology*, *37*, 99–105.

Tijsma, A. D., Drugan, M. M., & Wiering, M. A. (2016). Comparing exploration strategies for q-learning in random stochastic mazes. In *Proceedings of the IEEE Symposium Series on Computational Intelligence* (pp. 1–8). Piscataway, NJ: IEEE.

Ueltzhöffer, K. (2018). Deep active inference. *Biological Cybernetics*, *112*(6), 547–573.

Vermorel, J., & Mohri, M. (2005). Multi-armed bandit algorithms and empirical evaluation. In *Proceedings of theEuropean Conference on Machine Learning* (pp. 437–448). Berlin: Springer.

Watkins, C. J. C. H. (1989). *Learning from delayed rewards*. PhD diss., University of Cambridge.

Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*(3–4), 279–292.

Wiering, M., & Schmidhuber, J. (1998). Efficient model-based exploration. In *Proceedings of the Sixth International Conference on Simulation of Adaptive Behavior: From Animals to Animats*, *6* (pp. 223–228). Cambridge, MA: MIT Press.

Zintgraf, L., Shiarlis, K., Igl, M., Schulze, S., Gal, Y., Hofmann, K., & Whiteson, S. (2019). *Varibad: A very good method for Bayes-adaptive deep Rl via meta-learning*. arXiv:1910.08348.

Received July 10, 2020; accepted October 7, 2020.