

Active inference: demystified and compared

Noor Sajid¹, Philip J. Ball² and Karl J. Friston¹

¹The Wellcome Centre for Human Neuroimaging, UCL Queen Square Institute of Neurology, London, UK WC1N 3AR.

²Machine Learning Research Group, Department of Engineering Science, University of Oxford.

Correspondence: Noor Sajid

The Wellcome Centre for Human Neuroimaging,

UCL Queen Square Institute of Neurology,

London, UK WC1N 3AR.

+44 (0)20 3448 4362

noor.sajid.18@ucl.ac.uk

Abstract

Active inference is a first principle account of how autonomous agents operate in dynamic, non-stationary 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 paper, 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; 2) to our knowledge, the first 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. We make these properties explicit by showing that the active inference agent’s ability to carry out online planning, in a pure-belief setting, enables it to act optimally, given the non-stationary dynamics of an environment when compared to both Q-learning and Bayesian model-based reinforcement learning agents. We conclude by noting that this formalism can be applied to more complex settings; e.g., robotic arm movement, Atari games, etc., 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.

Keywords: active inference, variational Bayesian inference, free energy principle, generative models, reinforcement learning

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 et al., 2017; Friston, Rosch et al., 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 the expectation of this quantity over time, optimal behavior can be obtained for a given environment (Friston, Schwartenbeck et al., 2014; Friston, 2019). More concretely, optimal behavior is determined by evaluating evidence (i.e., sensory inputs) under an agent’s generative model of observations (i.e., outcomes) (Friston, FitzGerald et al., 2016). This generative model of the environment is an abstraction, which assumes certain internal (hidden) states give rise to these observations. One goal of the agent is to infer what these hidden states are, given a set of observations. 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 optimal outcomes. This formulation has two complementary objectives: 1) infer optimal behavior, and 2) optimize the generative model based on the agents 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 on-line learning settings, where the environment is non-stationary. This follows from the ability to model uncertainty over contexts (Friston, Rigoli et al., 2015; Parr & Friston, 2017).

Active inference can also be seen as providing a formal framework for jointly optimizing action and perception. In the context of machine learning, this is often referred to as planning as inference (Attias, 2003; Botvinick & Toussaint, 2012; Baker & Tenenbaum, 2014), 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. These contributions include: *a*) not having to explicitly specify a reward function, *b*) a principled account of epistemic exploration and intrinsic motivation (Parr & Friston, 2017; Schwartenbeck, Passecker et al., 2019) and *c*) 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). Crafting appropriate reward functions is not easy, and it is possible for agents to learn sub-optimal actions, if the reward function is poorly specified (Amodei, Dario 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 little or no prior preferences; since the active inference framework naturally gives agents the ability to learn these prior preferences from the environment itself — by placing a distribution over prior preferences, which we demonstrate in our experiments.

Another challenge, within reinforcement learning, is balancing the ratio between exploration and exploitation; i.e., 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, Silver et al., 2013; Mnih, Badia et al., 2016), action selection based on action-utility (Sutton , 1990) and counter-based strategies (Wiering & Schmidhuber , 1998; Tijsma, Drugan, et al., 2016), etc. However, even with these exploratory mechanisms in place, we must then select a temperature hyper-parameter to weight extrinsic reward (from the environment) against the intrinsic curiosity reward (from the agent). In contrast, active inference treats exploration and exploitation as two sides of the same coin: minimizing uncertainty via an expected free energy functional. This allows for a natural trade-off between epistemic exploration and pragmatic behavior. This review paper aims to 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 in comparison with reinforcement learning agents on a modified FrozenLake OpenAI baseline.

The review comprises three sections. The first section considers the discrete state-space and time formulation of active inference, and provides commentary on its derivation, implementation, and connections to reinforcement learning. The second section 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 whilst all agents are able to perform appropriately in a stationary setting, active inference’s ability to carry out online planning allows for optimal behavior in the non-stationary environment. Additionally, through these simulations, we make explicit links between the reward function and prior preferences about outcomes. 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, etc., if the appropriate underlying probability distribution/generative model can be formulated.

2 Active Inference

Motivation

Active inference describes how (biological or artificial) agents navigate dynamic, non-stationary environments (Friston, FitzGerald et al., 2017; Friston, Rosch et al., 2017). It postulates that in any given state, an agent maintains homeostasis by residing in (attracting) states that minimize entropy (or surprising observations) (Friston, Mattout et al., 2011).

Definition 1 (Surprise). *We define entropy — as being related to surprise — from information theory:*

$$S = -\log P(o) \quad (1)$$

o is the set of possible outcomes.

In active inference, the agent determines how to minimize entropy by maintaining a generative model of the 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 around via its sensory observations (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 observations. 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, FitzGerald 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); i.e., making decisions that maximize expected value. In active inference, we are interested in optimizing a *free energy functional of beliefs about states*; i.e., 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, whilst in active inference we wish to reside in states that give rise to observations that match our prior preferences.

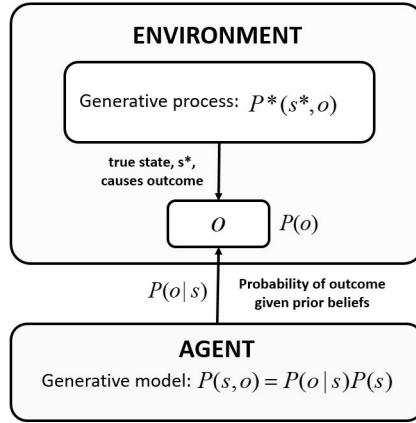


Figure 1: Graphical representation of the generative process (based on true states, s^*) in the world and the corresponding (internal) generative model (based on probabilistic beliefs random variables, s , that stand in for true states that are hidden behind observations) that best explain the outcomes,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

From an implementation perspective, this means replacing the traditional reward function used in reinforcement learning with prior beliefs about preferred outcomes. The agents prior preferences, $\log P(o)$, are defined only to within an additive constant and depend on relative differences between rewarding (familiar) and unrewarding (surprising) outcomes. This needs to be defined a priori, since the agent may otherwise illustrate ‘dark room’ seeking behavior (Friston, Daunizeau, et al , 2009; Baltieri & Buckley , 2019).

Variational Free Energy

Starting from a simple generative model for observations, 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. The generative model abstraction asserts that the world has a true (hidden or latent) state s^* , which results in the observations o (via the generative process). 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, etc.) and the observation is the information from the environment (e.g., feedback, velocity, etc.). By the reverse process of mapping from its hidden state to the observations (through Bayesian model inversion), the agent can explain the observations in terms of how they were caused by hidden states. This is Bayesian model inversion or inference.

Definition 2 (Generative Model). *The joint model of this simple system is defined as $P(o, s)$. This can be factorized, assuming conditional independence, into a likelihood function $P(o|s)$ and prior over internal states $P(s)$ (see Appendix 5 for a full specification of the model):*

$$P(o, s) = P(o|s)P(s). \quad (2)$$

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

$$P(o) = \sum_s P(o, s) \quad (3)$$

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). *Using Jensen's inequality, we can define the variational free energy, F , or the upper bound on surprise. This is, commonly, known as the (negative) evidence lower bound (ELBO) in the variational inference literature (Blei, Kucukelbir et al., 2017):*

$$-\log P(o) = -\log \sum_s P(o, s) \quad (4)$$

$$\leq -\sum_s Q(s) \log \frac{P(o, s)}{Q(s)} \quad (5)$$

$$= \sum_s Q(s) \log \frac{Q(s)}{P(o, s)} \quad (6)$$

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

$$F = \sum_s Q(s) \log \frac{Q(s)}{P(o, s)} \quad (7)$$

$$= \sum_s Q(s) \log \frac{Q(s)}{P(s|o)P(o)} \quad (8)$$

$$= \sum_s Q(s) \left(\log \frac{Q(s)}{P(s|o)} - \log P(o) \right) \quad (9)$$

$$= D_{KL}[Q(s)||P(s|o)] - \log P(o) \quad (10)$$

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

$$-\log P(o) = F - D_{KL}[Q(s)||P(s|o)] \quad (11)$$

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

$$F = \underbrace{D_{KL}[Q(s|\pi)||P(s|o, \pi)]}_{\text{evidence bound}} - \underbrace{\log P(o)}_{\text{log evidence}} \quad (12)$$

$$= \underbrace{D_{KL}[Q(s|\pi)||P(s|\pi)]}_{\text{complexity}} - \underbrace{\mathbb{E}_{s \sim Q(s)}[\log P(o|s)]}_{\text{accuracy}} \quad (13)$$

Since KL divergences cannot be less than zero, from Equation 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 13. Note that we have conditioned the probabilities in Equation 12 and 13 on policies, π . These policies can be regarded as particular priors that, as we will see below, 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 sequence of actions.

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; namely, 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 note that we would like to minimize not only our *instantaneous* variational free energy, F , but also our variational free energy in the *future*; this is called the expected free energy, G . Minimization of expected free energy allows the agent to influence the future by taking actions in the present, which are selected from policies. We will 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, u_τ at time τ , that enable an agent to transition between hidden states. The total number of policies that can be pursued is defined by some arbitrary number, K . Formally this can be written:*

$$u_\tau = \pi(\tau) \text{ where } \pi \in \{0, \dots, K\} \quad (14)$$

Note that policy, as defined in active inference, is inherently different to its reinforcement learning counterpart. In active inference, a policy is simply a concatenation of possible actions through time, but in reinforcement learning it is the mapping of states to actions (i.e., $\pi(s)$).

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 17):

$$F(\tau, \pi) = \sum_{s_\tau^\pi} Q(s_\tau|\pi) \log \frac{Q(s_\tau|\pi)}{P(o_\tau, s_\tau|s_{\tau-1}, \pi)} \quad (15)$$

$$= \mathbb{E}_{Q(s_\tau|\pi)} [D_{\text{KL}}[Q(s_\tau|\pi)||P(s_\tau|s_{\tau-1}, \pi)]] - \mathbb{E}_{Q(s_\tau|\pi)} [\ln P(o_\tau|s_\tau)] \quad (16)$$

$$= s_\tau^\pi (\log s_\tau^\pi - \log \mathbf{B}_{\tau-1}^\pi s_{\tau-1}^\pi - \log \mathbf{A} o_\tau) \quad (17)$$

Here s_τ^π is the expected state conditioned on each policy; \mathbf{B}_τ^π is the transition probability for hidden states, contingent upon pursuing a given policy, at a particular time; \mathbf{A} is the expected likelihood matrix mapping from hidden states to outcomes and o_τ represents the outcomes. Now having developed this functional dependency on time, we simply take an expectation with respect to the posterior distribution of observations from our generative model, $P(o_\tau|s_\tau)$.

Definition 5 (Expected free energy). *is the variational free energy of future trajectories. It effectively evaluates evidence for plausible policies based on outcomes that have yet to be observed (Parr & Friston,*

2018). It can be derived from Equation 15 by taking an expectation under the posterior predictive distribution given by $P(o_\tau|s_\tau)$, then summing over time. This captures the idea of predicting future outcomes, given future hidden states, conditioned on policies.

$$G(\pi) = \sum_{\tau} G(\tau, \pi) \quad (18)$$

The expected free energy summands can be decomposed in complementary ways (and the matrix formulation: Equation 25):

$$G(\tau, \pi) = \sum_{s_\tau, o_\tau} P(o_\tau|s_\tau)Q(s_\tau|\pi) \log \frac{Q(s_\tau|\pi)}{P(o_\tau, s_\tau|s_{\tau-1}, \pi)} \quad (19)$$

$$= \mathbb{E}_{\tilde{Q}} [\log(Q(s_\tau|\pi) - \log(P(o_\tau, s_\tau|s_{\tau-1}, \pi))] \quad (20)$$

$$= \mathbb{E}_{\tilde{Q}} [\log(Q(s_\tau|\pi) - \log(P(s_\tau|o_\tau, s_{\tau-1}, \pi)) - \log(P(o_\tau))] \quad (21)$$

$$\approx \underbrace{\mathbb{E}_{\tilde{Q}} [\log(Q(s_\tau|\pi) - \log(Q(s_\tau|o_\tau, s_{\tau-1}, \pi)))]}_{\text{-ve mutual information}} - \underbrace{\mathbb{E}_{\tilde{Q}} [\log(P(o_\tau))]}_{\text{expected log evidence}} \quad (22)$$

$$= \underbrace{\mathbb{E}_{\tilde{Q}} [\log(Q(o_\tau|\pi) - \log(Q(o_\tau|s_\tau, s_{\tau-1}, \pi)))]}_{\text{-ve epistemic value}} - \underbrace{\mathbb{E}_{\tilde{Q}} [\log(P(o_\tau))]}_{\text{extrinsic value}} \quad (23)$$

$$= \underbrace{D_{KL}[Q(o_\tau|\pi) || P(o_\tau)]}_{\text{expected cost}} + \underbrace{E_{Q(s_\tau|s_{\tau-1}, \pi)} [H[P(o_\tau|s_\tau)]]}_{\text{expected ambiguity}} \quad (24)$$

$$= o_\tau^\pi (o_\tau^\pi - \mathbf{C}_\tau) + s_\tau^\pi \mathbf{H} \quad (25)$$

where the following assumptions are made: $\tilde{Q} = P(o_\tau|s_\tau)Q(s_\tau|\pi)$; $Q(o_\tau|s_\tau, \pi) = P(o_\tau|s_\tau)$; $C_\tau = \log P(o_\tau)$ is the logarithm of prior preference over outcomes and $\mathbf{H} = -\text{diag}(\mathbb{E}_Q[\mathbf{A}_{i,j}] \cdot \mathbb{E}_Q[\mathbf{A}])$ is the vector encoding the ambiguity over outcomes for each hidden state.

When minimizing expected free energy, we can regard Equation 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), whilst maximizing expected value as scored by the (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. The natural curiosity emerging through this formulation, is in contrast to reinforcement learning, where curiosity must be manufactured, either through random action selection (Mnih, Silver et al., 2013) or through additional curiosity terms, which are appended to the reward signal (Pathak, Efros et al., 2017). Information theoretic approaches have also been explored in a reinforcement learning context but do not leverage the (beliefs about) latent states implied by the generative model; see (Still, 2012; Mohamed & Rezende, 2015). Consequently, they do not encourage exploration that would minimize ambiguity over latent states.

Equation 24 offers an alternative perspective on the same objective; i.e., an agent wishes to minimize the ambiguity, while minimizing the degree to which outcomes (under a given policy) deviate from prior

preferences $P(o_\tau)$. 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 ($P(o_\tau)$). 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. towards states they expect to occupy and frequent).

It is now also possible to derive policies given the expected free energy. Policies, a priori, minimize the expected free energy term, G (Friston, FitzGerald et al., 2017). This can be realized by deriving the probability of any policy with a softmax function (i.e., normalized exponential) of expected free energy:

$$P(\pi) = \sigma[-G(\pi)] \quad (26)$$

where σ denotes a softmax function.

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 which get us closest to this goal (i.e., the attracting set defined by prior preferences $P(o)$).

Note that the similarities to Dyna-style/planning model-based reinforcement learning (Sutton , 1990): hypothetical roll-outs 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 & Bordons, 2007), where a search is performed over possible action sequences at each time-step.

Optimizing Expected 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. However, under the active inference scheme, we calculate the solution by using a gradient descent (with a default step size, ζ , of 4) on expected free energy $G(\pi)$, which allows us to optimize both action-selection and inference simultaneously, using a mean-field approximation (Beck, Pouget, et al., 2012; Parr, Markovic, et al., 2019):

$$\varepsilon_\tau^\pi = (\log A.o_\tau + \log B_{\tau-1}^\pi s_{\tau-1}^\pi + \log B_\tau^\pi s_{\tau+1}^\pi) - \log s_\tau^\pi \quad (27)$$

$$\varepsilon^\gamma = (\beta - \beta_\tau) + (\pi - \pi_0).G \quad (28)$$

where $\beta_\tau = \beta + (\pi - \pi_0).G$; $\beta = \frac{1}{\gamma}$ encodes posterior beliefs about precision; π represents the policies specifying action sequences and $\pi_0 = \sigma(-\gamma.G)$.

This involves converting the discrete updates, defined in Equation 27 and 28, 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 time scales: during each observation or trial,

evidence for each policy is evaluated based upon prior beliefs about future outcomes. This is determined by updating posterior beliefs about hidden states (i.e., state estimation under each policy, $P(s|\pi)$) on a fast time scale, while posterior beliefs find new extrema (i.e., as new observations are sampled, $P(s|o)$) to produce a slower evidence accumulation over observations.

Using this kind of belief updating, we can calculate the posterior beliefs about each policy; namely, a softmax function based on expected free energy, as covered in Equation 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, which is realized through action.

In active inference, the scope and depth of the policy search is exhaustive, in the sense that any policy entertained by the agent is encoded explicitly, and any hidden state 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 hyper-parameter. Additionally, at the end of each sequence of observations, 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. Lastly, temporal discounting emerges naturally from the active inference scheme, where the generative model determines the nature of discounting (based on γ parameter capturing precision), with predictions in the distant future being less precise, thus discounted (Friston, FitzGerald et al., 2017).

The discussion above 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 (non-stationary) environment. This formulation can be extended to capture a more representative generative process by defining a hierarchical (deep temporal) generative model as described in (Friston, FitzGerald et al., 2017; Friston, Parr et al., 2017; Parr & Friston, 2017), continuous state spaces models (Buckley, Kim, et al., 2017; Parr & Friston, 2019) or mixed models with both discrete and continuous states as described in (Friston, Parr et al., 2017; Parr & Friston, 2018). In the case of a continuous formulation, the generative model state-space can be defined in terms of generalized coordinates of motion, which generally have a non-linear 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), midbrain dopamine activity (Friston, Schwartenbeck et al., 2014), to evoked potential, 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) with neuromodulatory systems (Parr & Friston, 2017, 2018). Please see (Friston, FitzGerald et al., 2017; Parr & Friston, 2018; Da Costa, Parr et al., 2020), for a detailed overview.

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; e.g., behavioral economics trust games (Moutoussis, Trujillo-Barreto, et al., 2014; Schwartenbeck, FitzGerald, et al., 2015), narrative construction and reading (Friston, Rosch et al., 2017), saccadic searches and scene construction (Mirza, Adams, et al., 2016), Atari games (Cullen, Davey, et al., 2018), etc).

We first describe the environment set-up and then simulate how an agent learns to navigate the lake to successfully reach the goal. The simulations involve searching for the reward (i.e., Frisbee) in a 3×3 frozen lake and avoid falling in a hole.

Set-up

The frozen lake has a grid-like structure with four different 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; e.g. 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: position 8 and 6 or 6 and 8 respectively. The objective is to reach (G), the Frisbee location, ideally in as few steps as possible, whilst 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. Finally, we limit the maximum number of time steps (i.e., the horizon) to 15.

Active inference agents

For this paradigm, we define the generative model for the active inference agent as follows (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 e.g. when at location 4, the agent can move to location 5 (right), 7 (down), 1 (up) or stay at 4 (left). The hidden state factor, *location*, elucidates the agents beliefs about its location in the frozen lake. The context hidden state factor elucidates the agents 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; e.g., if I have beliefs that I am located in position 6, then I will observe myself in position 6, irrespective of context. However, the likelihood for score, given the hidden states, is determined by the context; i.e. if the context is 1 (2) then 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, C , is defined by $\ln P(o) : 4$ and

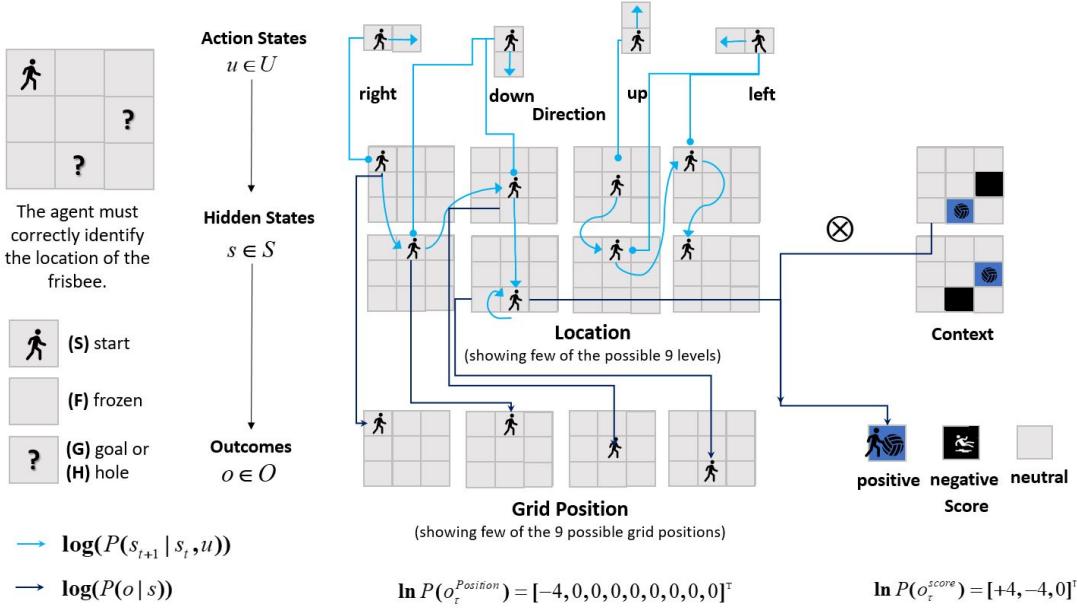


Figure 2: Graphical representation of the active inference generative model: The model contains four action states that encode direction of movement: right, down, up and left. They control the ability to transition between hidden state *location* factors (one of the nine locations going from $1 \rightarrow 9$ - only a few states are shown). Each action navigates the agent to a different location (a select few are shown): e.g., if the agent starts in position 1 and chooses to turn right, then it will end up in state 2 at the next timestep. 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; i.e., tries to go left from location 1, it will remain in location 1 (as shown). The hidden states have a Kronecker tensor product (\otimes) 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 observable outcome is generated. Categorical parameters, that define the generative model, A (likelihood - $P(o|s)$): have an identity mapping between hidden state location and outcome grid position with some uncertainty; e.g. 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; i.e., for context 1, positive score received at location 8, and negative or nothing elsewhere. $\ln P(o)$ corresponds to prior preference: the agent expects to find positive score and not remain at the starting location.

-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 e.g., ('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, with a high precision over action selection (2048). 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 27 and 28 to simulate appropriate behavior. Pseudo-code for the belief updating and action selection for this particular type of discrete state-space and time formulation is presented in Appendix 5. To provide a baseline for purely exploratory behavior, we also simulated a null active inference agent, who had no prior preferences (i.e., was insensitive to the reward).

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 Thompson sampling (Poupart , 2018; Ghavamzadeh, Mannor, et al , 2015).

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 pseudo-code is presented in Appendix 5. 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; the pseudo-code is presented in Appendix 5. 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 above: 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). 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 pseudo-counts — for the reward and transition model — are initialized as 1. The posterior for the reward and transition model distribution are evaluated by updating the prior ($Beta(\alpha, \beta)$). Thus, by treating them as pseudo-counts, 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 : posterior is $Beta(\alpha + x, \beta + y)$. The discount factor is specified as 0.9.

Learning to navigate the frozen lake

We evaluate how well the different agents are able to navigate the frozen lake in both stationary and non-stationary environments, as described below. Each of the environments were 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 (Figure 2) and Active Inference (null model; without any prior outcome preferences i.e. $\ln P(o) = 0$ for all outcomes).

Algorithm	Belief-Based	Average Score [95% CI]	
		Deterministic Env.	Stochastic Env.
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]

Table 1: Average reward (and 95% confidence interval) for each agent, across both deterministic and stochastic environments. The results are calculated from the 200 trials across 500 episodes.

Stationary environment

For this set-up, 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 which is one of the major motivators for active inference. The average score (Table 1) for all agents, except the null model specification of Active inference model, was considerably high at > 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 i.e., it does not differentiate between any of the different patches (S, F, G & H) in the frozen lake. As expected, the null model exhibits ‘dark room’ behavior (Baltieri & Buckley , 2019), preferring to stay in the first few states, and eventually exploring far enough to either fall in the hole or reach the goal, with equal probability.

The learning curve, as shown in (Figure 3), highlights that the active inference and Bayesian model-based reinforcement learning agent learn optimal behavior (and resolve uncertainty about reward location) in a short amount of time (< 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 ($\epsilon = 0.1$), whilst 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.

Non-stationary environment

We introduce non-stationarity 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 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 re-learn the correct (G) location. The average score for all agents is presented in Table 1.

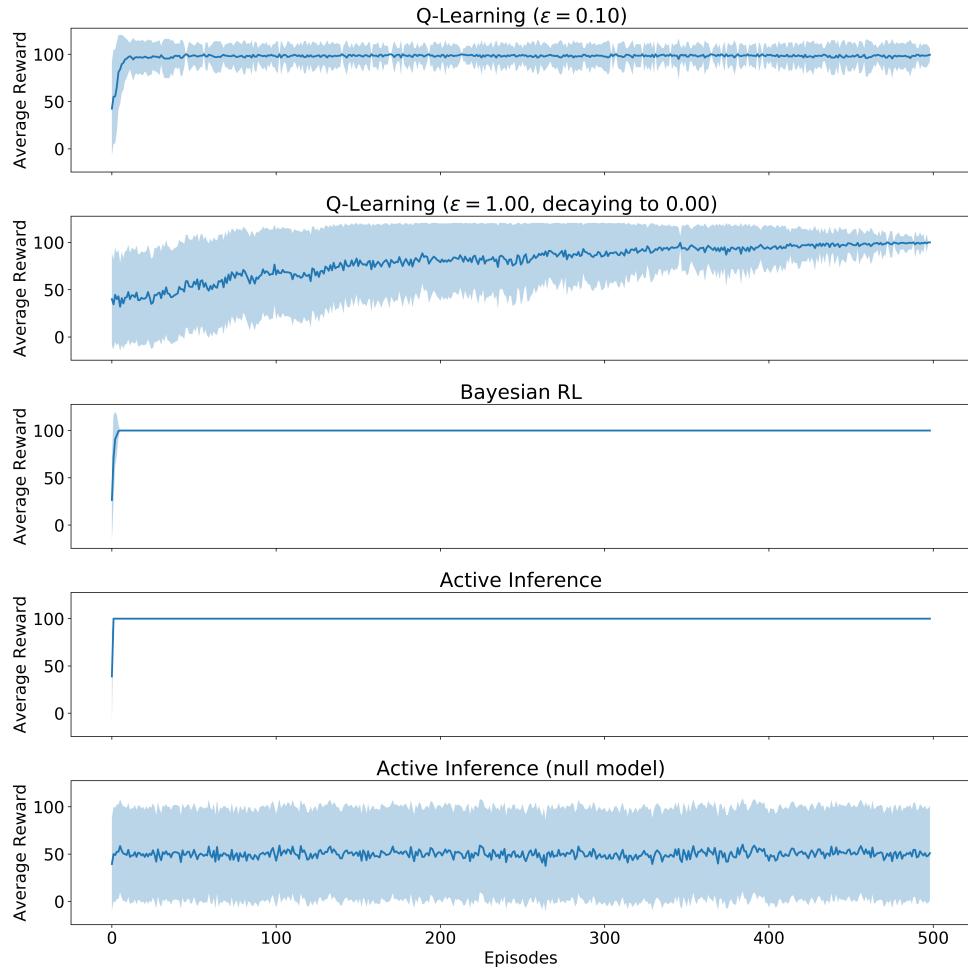


Figure 3: Learning curve for deterministic environment. The x-axis denotes the episode number and y-axis the average (online) reward. The results presented are calculated from 200 trials.

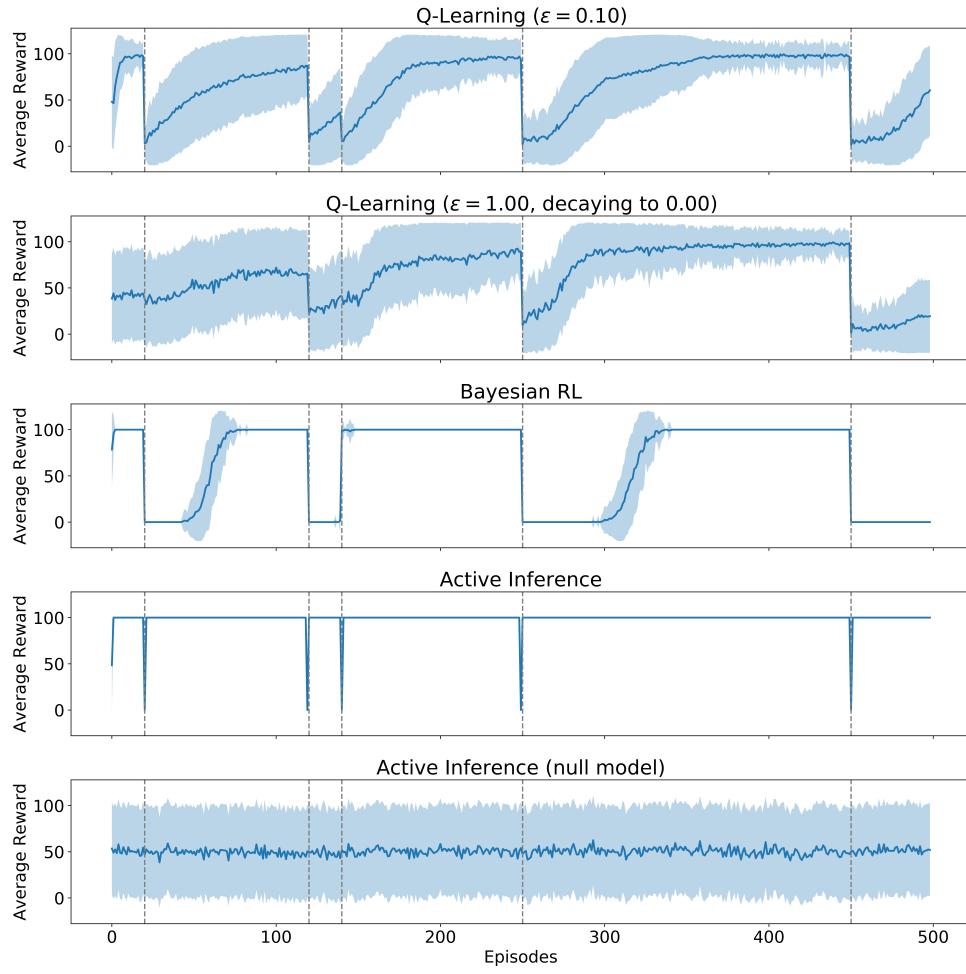


Figure 4: Learning curve for stochastic environment. The x-axis denotes the episode number and y-axis the average (online) reward. The results presented are calculated from 200 trials. The dotted gray lines represent the change in (G) (and (H)) location.

As in the stationary set-up, all agents are initially uncertain about the reward location. This is quickly resolved, and by episode 20, active inference, Bayesian RL and Q-learning ($\epsilon = 0.1$) exhibit optimal behavior. 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 0 due to the change in reward location. For the reinforcement learning (Q-learning and Bayesian RL) agents, this drop in performance persists for the next ~ 40 episodes. This is because by treating this as a ‘learning’ problem, the agent has to do the following: 1) reversal learning of its previous understanding of the reward location and 2) re-learn 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 (Figure 3). 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 reward function due to the accumulation of pseudo-counts. This contrasts with 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 non-stationary 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 non-stationarity are not accommodated properly, as shown through the above simulations. Although, they can be dealt with using techniques that involve the inclusion of inductive biases; e.g., importance sampling of experiences in multi-agent environments (Foerster, Chen, et al., 2017) or using meta-learning to adapt gradient-update approaches more quickly (Al-Shedivat, Bansal, et al., 2018).

Comparing prior preferences and rewards

In reinforcement learning, goals are defined through reward functions, whereas in active inference, goals are defined through prior preferences over observations. 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 (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 e.g., reward of -100 is equivalent to prior preferences of $-\log(5)$, etc.

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 non-probabilistic 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 states 6 or 8, hence 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 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 (i.e., 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, as their probabilistic models remain uniform over the beliefs of which transitions produce preferred behavior.

Finally, we observe that all 3 approaches learn the same circular behavior when only a negative preference/reward is specified (i.e., 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, do not learn to go 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 non-zero 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 only terminating when the time limit is reached.

Through this brief study, we have shown equivalences between belief-based reinforcement learning and active inference, demonstrating that by writing our prior preferences through the reward function, we can illicit identical behaviors from both, as long as we adopt a probabilistic, model-based approach for reinforcement learning. However, the FrozenLake environment is by no means representative of all discrete environments, and this merits further research.

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)

Table 2: Reward shaping: average score and number of moves across 100 episodes for 100 agents. *Note that for this experiment we evaluate under $\epsilon = 0.0$, i.e., on-policy.

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 agents 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 (Bayes optimal) policies in the absence of prior preferences, we allow both the likelihood distribution ($\log P(o|s)$) and outcome preferences ($\log P(o)$) to be learned. 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 $\log P(o)$), 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 one of Table 2, where we specify flat prior preferences, but the agent is not equipped with (Dirichlet) hyper-priors 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 set-up 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 i.e., there is no revisiting of locations that have already been encountered (Figure 5.1). Furthermore, this behavior persists past the initial exploration, with continuous explorations via new (non-overlapping) trajectories (Figure 5.2).

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 set-up 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 behaviour (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 (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 Frisbee's (Figure 5.4).

Finally, we look at the interaction between the epistemic imperatives to resolve uncertainty about the likelihood mapping and uncertainty about prior preferences. This set-up 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 (Figure 6). This causes the agent to exhibit exploitative behavior of hiding, rather than continue exploring. After 5 trials, the agent goes straight to the hole.

This is an interesting example of how — by observing ones own behaviour — habit formation contextualizes the fundamental imperative to explore.

It is important to note that the learned outcome preferences are time-dependent; i.e., the agent prefers to visit safe (F) patches for the first 3 time points and then visit goal (G) patches with high preference (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 agents 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 (Section 2). 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, 2012; Mohamed & Rezende, 2015; Pathak, Efros et al., 2017).

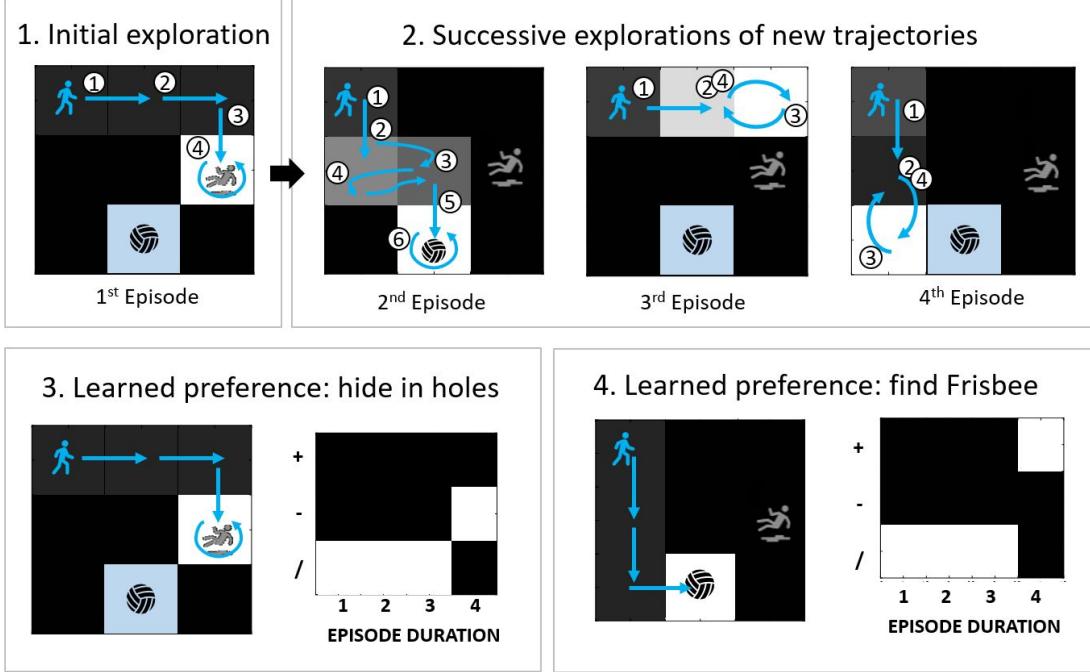


Figure 5: Parameter learning for a single reward location: results for likelihood learning presented in 1&2 and prior preference learning presented in 3&4. Blue arrows denote the trajectory taken and numbers in the circles denote the trajectory sequence. Circular arrows represent loops i.e., once in that state, the same outcome is observed till maximum number of moves reached (15). 5.1 is a pictorial representation of the first episode trajectory, with no prior preference: right($1 \rightarrow 2$), right ($2 \rightarrow 3$), down($3 \rightarrow 6$), right($6 \leftrightarrow 6$). 5.2 depicts the next 4 episodes from the trial. 5.3 has two figures: a pictorial representation of trajectory to hole and 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. 5.4 presents similar information, but for a goal preferring agent; pictorial representation of trajectory to goal and heat-map of the accumulated Dirichlet parameters for score. There is a strict preference for goals at time step 4.

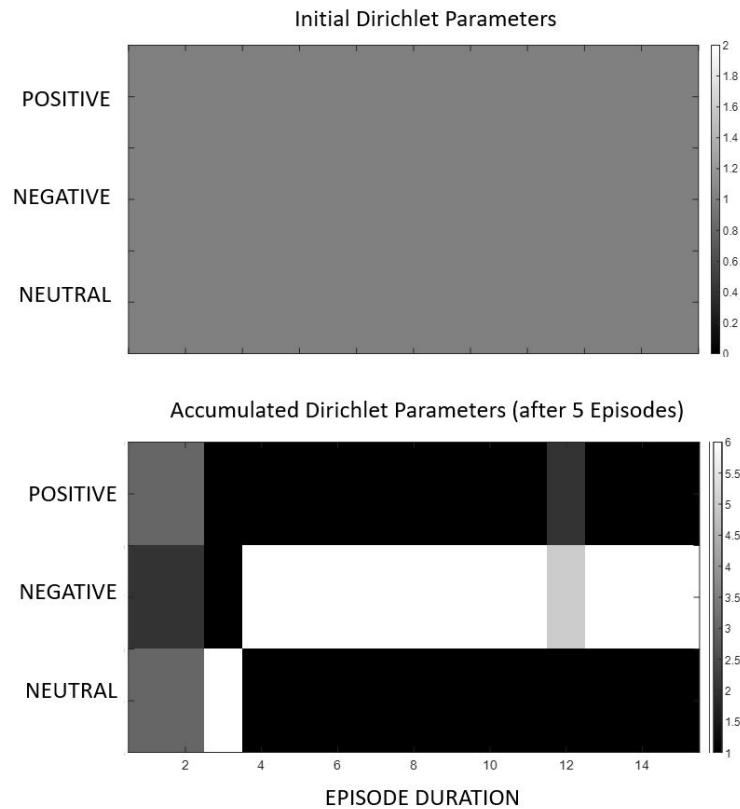


Figure 6: Learning prior outcome preferences for outcome modality, score: initial (top) and after 5 episodes (bottom) for a single reward location

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, non-stationary 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 23), epistemic foraging (or exploration) emerges naturally. This is captured by the desire to maximize the mutual information between observations and the hidden states on the environment. Exploration means that the agent seeks out states that afford observations, 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 briefly discussed parameter exploration that might also be carried out by the agent — when learning either the likelihood or prior outcome preferences — by having priors over the appropriate probability distributions and applying the expected free energy derivations to those parameters (Schwartenbeck, Passeecker et al., 2019). These canonical properties with respect to decision making under uncertainty must be additionally engineered in classical reinforcement learning.

Our treatment has emphasized that, via a belief-based scheme, active inference enables us to specify reward functions in terms of prior beliefs — or not specify rewards at all (to produce purely epistemic behavior). However, if rewards are available as observations or actions, they can be assigned high prior preferences. An agent is likely to maximize reward (or extrinsic value) by having prior preferences about unsurprising outcomes (see Equation 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 set-up 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 navigate 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, et al., 2016). Thus, if one was to find the appropriate generative model, active inference could be used for a variety of different problems; e.g. robotic arm movement, dyadic agents, playing Atari games, etc. 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 paper are available at: <https://github.com/ucbtns/dai>

Acknowledgments

NS is funded by the Medical Research Council (Ref: MR/S502522/1). PJB is funded by the Willowgrove Studentship. KJF is funded by the Wellcome Trust (Ref: 088130/Z/09/Z). We would like to thank the anonymous reviewers for their suggestions and insightful comments on the manuscript.

Disclosure statement

The authors have no disclosures or conflict of interest.

5 Appendix

Explicit parameterization of the generative model

Active inference rests on the tuple (O, S, T, R, P, Q) :

- A finite set of outcomes, O
- A finite set of control states or actions, U
- A finite set of hidden states, S
- A finite set of time-sensitive policies, T
- A generative process $R(\tilde{o}, \tilde{s}, \tilde{u})$ that generates probabilistic outcomes $o \in O$ from (hidden) states $s \in S$ and action $u \in U$
- A generative model $P(\tilde{o}, \tilde{s}, \pi, z)$ with parameters z , over outcomes, states, and policies $\pi \in T$, where $\pi \in 0, \dots, K$ returns a sequence of actions $u_\tau = \pi(\tau)$
- An approximate posterior $Q(\tilde{s}, \pi, z) = Q(s_o|\pi) \cdot Q(s_\tau|\pi)Q(\pi)Q(z)$ over states, policies and parameters with expectations $(s_0^\pi, \dots, s_\tau^\pi, \pi, z)$

The generative process describes transitions between hidden (unobserved) states in the world that generate (observed) outcomes. Their transitions depend on action, which depends on posterior beliefs about the next state. Subsequently, these beliefs are formed using a generative model of how observations are generated. The generative model (based on partially observable MDP) describes what the agent believes about the world, where beliefs about hidden states and policies are encoded by expectations. Here actions are part of the generative process in the world and policies are part of the generative model of the agent.

Pseudo-code for active inference: belief updating and action selection

Initialize the following:

Probability of seeing observations, given states, likelihood: A

Probability of transitioning between states, given an action: B

Log probability of agent's preferences about outcomes: C

Probability of state the agent believes it is at the beginning of each trial: D

for $\tau = 1 : T$ **do**

 Sample state, s based on generative process

 Sample outcome o based on likelihood matrix A

 Variational updates of expected states, s under sequential policies
(gradient descent on F)

 Evaluate expected free energy G of policies π

 Bayesian model averaging of expected states s over policies π

 Select action with the lowest expected free energy

end

Accumulation of (concentration) parameters for learning update based on learning rate

Pseudo-code for Q-Learning

Initialize the following:

Q-value function; $Q(s, a)$

Initialize parameter for exploration; ϵ

Specify learning rate, α and discount factor, γ

for $\tau = 1 : T$ **do**

 Sample exploration rate threshold from a random uniform distribution, $U(0, 1)$

 Choose action based on $\max_a(Q(s, :))$ if exploration rate threshold is greater than ϵ , else choose random action

 Execute a^* and receive r, s'

 Update $Q(s, a) :$ using $(1 - \alpha) * Q(s, a) + \alpha * (r + \gamma * \max_a(Q(s', :)))$

$s = s'$

 Update exploration parameter $\epsilon :$ $\epsilon - decay\ rate$

end

Pseudo-code for Bayesian Model-Based Reinforcement Learning using Thompson Sampling

Initialize the following:

Θ_t, Θ_r as uniform

Probability of transitioning between states, given an action, transition model; Θ_t

Probability of receiving reward, given a state, reward function; Θ_r

Repeat:

Sample $\Theta_{t,1}, \dots, \Theta_{t,k} \sim Pr(\Theta_t) \forall a$

Sample $\Theta_{r,1}, \dots, \Theta_{r,k} \sim Pr(\Theta_r) \forall a$

$Q_{\theta_{t,i}, \theta_{r,i}}^* \Leftarrow \text{solve } MDP_{\theta_{t,i}, \theta_{r,i}}$

$\hat{Q}(s, a) \Leftarrow \frac{1}{k} \sum_{i=1}^k Q_{\theta_{t,i}, \theta_{r,i}}^*(s, a)$

$a^b \Leftarrow \max_a \hat{Q}(s, a)$

Execute a^* and receive r, s'

$b(\Theta_t) \Leftarrow b(\Theta_t) Pr(s'|s, a, \theta_t)$

$b(\Theta_r) \Leftarrow b(\Theta_r) Pr(r|s, a, s', \theta_r)$

$s \Leftarrow s'$

end

References

- Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G.(2017). Active Inference: A Process Theory. *Neural Computation*, 29(1), 1-49.
- Friston, K., Rosch, R., Parr, T., Price, C., & Bowman, H.(2018). Deep temporal models and active inference. *Neurosci Biobehav Rev*, 77, 388-402.
- Pouget, A., Beck, J., Ma, W., & Latham, P.(2013). Probabilistic brains: knowns and unknowns. *Nature neuroscience*, 16(9), 1170.
- Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., & Dolan, R. (2014). The anatomy of choice: dopamine and decision-making. *Philos Trans R Soc Lond B Biol Sci*, 369(1655).
- Friston, K. (2019). A free energy principle for a particular physics. *arXiv preprint, arXiv:1906.10184*.
- Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O'Doherty, J., & Pezzulo, G.(2016). Active inference and learning. *Neurosci Biobehav Rev*, 68, 862-879.
- Friston, K., Rigoli, F., Ognibene, D., Mathys, C., FitzGerald, T., & Pezzulo, G.(2015). Active inference and epistemic value. *Cogn Neurosci*, 1-28.
- Parr, T., & Friston, K. (2017). Uncertainty, epistemics and active inference. *Journal of the Royal Society Interface*, 14(136).
- Tribus, M. (1961). Thermodynamics and Thermostatics: An Introduction to Energy, Information and States of Matter, with Engineering Applications. *New York, USA, D. Van Nostrand Company Inc.*
- Crauel, H., & Flandoli, F. (1994). Attractors for Random Dynamical-Systems. *Probability Theory and Related Fields*, 100(3), 365-393.
- Seifert, U. (2012). Stochastic thermodynamics, fluctuation theorems and molecular machines. *Rep Prog Phys*, 75(12), 126001.
- Attias, H. (2003). Planning by Probabilistic Inference. *Proc. of the 9th Int. Workshop on Artificial Intelligence and Statistics*
- Botvinick, M., & Toussaint M.(2012). Planning as inference. *Trends Cogn Sci*, 16(10), 485-488.
- Baker, C., & Tenenbaum J.(2014). Plan, Activity, and Intent Recognition: Modeling Human Plan Recognition Using Bayesian Theory of Mind. Sukthankar, G., Geib, C., Bui, H., Pynadath, D., & Goldman, R. *Morgan Kaufmann, Boston*, 177-204.
- Schwartenbeck, P., Passecker, J., Hauser, T., FitzGerald, T., Kronbichler, M. & Friston K.(2019). Computational mechanisms of curiosity and goal-directed exploration. *Elife*, 8.
- Friston, K., Mattout, J., & Kilner, J. (2011). Action understanding and active inference. *Biol Cybern*, 104, 137-160.
- Friston, K., Parr, T., & de Vries, B. (2017). The graphical brain: Belief propagation and active inference. *Netw Neurosci*, 1(4), 381-414.

- Astrom, K. J. (1965). Optimal control of Markov processes with incomplete state information. *Journal of mathematical analysis and applications*, 10(1), 174-205.
- Sutton, S. & Barto A. (1998). Introduction to Reinforcement Learning. *MIT Press*.
- Blei, D., Kucukelbir, A., & McAuliffe, J. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518), 859-877.
- Beal, M. (2003). Variational Algorithms for Approximate Bayesian Inference. *PhD. Thesis, University College London*.
- Racanière, S., Reichert, D., Buesing, L., Guez, A., Rezende, D., Badia, A., Vinyals, O., Heess, N., Li, Y., Pascanu, R., Battaglia, P., Hassabis, D., Silver, D., & Wierstra, D. (2017). Imagination-augmented agents for deep reinforcement learning. *International Conference on Neural Information Processing Systems, Long Beach, CA, Curran Associates Inc*.
- Parr, T. & Friston, K. (2018) Generalised free energy and active inference: can the future cause the past?. *bioRxiv*: 304782.
- Mnih, V., Silver, D., Graves, A., Antonoglou, I., Wierstra, D. & Riedmiller, M (2013) Playing Atari with Deep RL. *NIPS Deep Learning Workshop*.
- Mnih, V., Badia, A.P., Mirza, M., Graves, A., Lillicrap, T., Harley T., Silver, D., & Kavukcuoglu, K (2016) Asynchronous methods for deep reinforcement learning. *In International conference on machine learning, 1928-1937*
- Pathak, D., Efros, A., & Darrell, T. (2017) Curiosity-driven Exploration by Self-supervised Prediction. *International Conference on Machine Learning, Sydney*.
- Still, S.(2012) An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, 139148
- Mohamed, S. & Rezende, D. (2015) Variational information maximisation for intrinsically motivated reinforcement learning. *Advances in neural information processing systems*.
- Parr, T., Markovic, D., Kiebel, S. & Friston, K. (2019) Neuronal message passing using Mean-field, Bethe, and Marginal approximations. *Scientific Reports*, 9(1): 1889
- Buckley, C., Kim, C., McGregor, S., & Seth, A., (2017) The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology*, 81, 55-79.
- Parr, T., & Friston, K. (2018) The Discrete and Continuous Brain: From Decisions to Movement-And Back Again. *Neural Comput*, 30(9), 2319-2347.
- Parr, T., & Friston, K. (2019) The computational pharmacology of oculomotion. *Psychopharmacology*.
- Mirza, B., Adams, R., Mathys, C., & Friston, K. (2016) Scene Construction, Visual Foraging, and Active Inference. *Frontiers in Computational Neuroscience*, 10(56).
- Cullen, M., Davey, B., Friston, K., & Moran, R. (2018) Active Inference in OpenAI Gym: A Paradigm for Computational Investigations Into Psychiatric Illness. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 3(9), 809-818.

- Schwartzenbeck, P., FitzGerald, T., 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.
- Moutoussis, M., Trujillo-Barreto, N., El-Deredy, W., Dolan, R., & Friston, K. (2014) A formal model of interpersonal inference. *Front Hum Neurosci*, 8:160.
- Foerster, J., Chen, R., Al-Shedivat, M., Whiteson, S., Abbeel, P., & Mordatch, I. (2017) Learning with Opponent-Learning Awareness. *CoRR, arXiv:1709.04326*.
- Baltieri, M. and Buckley, C. (2019) The dark room problem in predictive processing and active inference, a legacy of cognitivism?. *PsyArXiv, doi:10.31234/osf.io/p4z8f*.
- Friston, K.J., Daunizeau, J. and Kiebel, SJ. (2009) Reinforcement learning or active inference?. *PLoS one*, 4(7). p.e6421
- Al-Shedivat, M., Bansal, T., Burda, Y., Sutskever, I., Mordatch, I., & Abbeel, P. (2018) Continuous Adaptation via Meta-Learning in Nonstationary and Competitive Environments. *CoRR, arXiv:1710.03641*.
- Gershman, S., & Niv, Y. (2010) Learning latent structure: carving nature at its joints. *Current opinion in neurobiology*, 20(2), 251-256
- Tervo, D., Tenenbaum, J., & Gershman, S. (2016) Toward the neural implementation of structure learning. *Current opinion in neurobiology*, 37, 99-105
- Gershman, S., & Beck, J. (2017) Complex probabilistic inference. *Computational Models of Brain and Behavior*, 453
- Beck, J., Pouget, A., & Heller, K. (2012) Complex inference in neural circuits with probabilistic population codes and topic models. In: *Advances in Neural Information Processing Systems*, 305930
- Tijmsma, A.D., Drugan, M.M., & Wiering, M.A. (2016) Comparing exploration strategies for Q-learning in random stochastic mazes. In: *IEEE Symposium Series on Computational Intelligence (SSCI)* , 19
- Vermorel, J., & Mohri, M. (2005) Multi-armed bandit algorithms and empirical evaluation. In: *European conference on machine learning*. Springer, 437-448
- Sutton, R.S. (1990) Integrated architectures for learning, planning, and re-acting based on approximating dynamic programming. In: *Proceedings of the seventh international conference on machine learning*., 216224
- Wiering, M. & Schmidhuber, J. (1998) Efficient model-based exploration. In: *Proceedings of the Fifth International Conference on Simulation of Adaptive Behavior (SAB98)*, 223228.
- Watkins, C. & Dayan, P. (1992) Q-learning. *Kluwer Academic Publishers, Boston. Manufactured in The Netherlands.*, 8 279-292.
- Poupart, P. (2018) Lecture slides on Bayesian Reinforcement Learning from CS885: <https://cs.uwaterloo.ca/~ppoupart/teaching/cs885-spring18/slides/cs885-lecture10.pdf>
- Camacho, E. & Bordons, C. (2013) Model predictive control. *Springer-Verlag London*

- Watkins, C.J.C.H (1989) Learning from delayed rewards. *PhD thesis, University of Cambridge*, Cambridge, England
- Ghavamzadeh, M., Mannor, S., Pineau, J. & Tamar, A., (2015) Bayesian reinforcement learning: A survey.Q-learning. *Foundations and Trends in Machine Learning*, 8(5-6) 359-483.
- Ng, A. Y. (2003) Shaping and policy search in reinforcement learning. *PhD thesis, University of California, Berkeley.*, USA
- Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, John., Mané, D., (2016) Concrete Problems in AI Safety. *CoRR, arXiv:1606.06565*.
- 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*, 369:20130481.
- Da Costa, L., Parr, T., Sajid, N., Veselic, S., Neacsu, V. & Friston, K.J. (2020) Active inference on discrete state-spaces: a synthesis. *arXiv, 2001.07203*.
- Schmidhuber, J. (2006) Developmental robotics, optimal artificial curiosity, creativity, music, and the fine arts. *Connection Science*, 18(2) 173-187.