Active Inference and Exploration for the Reinforcement Learning Problem[*]

Roland Dubb
Supervised by: Assoc. Prof. Jonathan Shock

March 2022

Abstract

Active inference agents can be considered as a type of reinforcement learning agent which select actions via the minimisation of free energy objective functionals. These functionals provide a flexible and probabilistic framework for modelling perceptual inference, learning parameters and, adaptively selecting actions to achieve the agent's goals. The active inference agent's objective function for action selection affords the agent a directed intrinsic exploratory motive which assists the agent in resolving the trade-off between exploration and exploitation which plagues reinforcement learning agents. This intrinsic exploratory drive motivates this study. This report begins by briefly introducing the reinforcement learning problem and some basic exploratory methods. Thereafter we endeavour to understand the intuitive and mathematical origins for the active inference agent's exploratory drive. This leads to a discussion of the free energy principle which motivates the construction of free energy-minimising agents. These *active inference* agents are briefly introduced, resulting in a discussion of their implementations in some benchmark reinforcement learning environments. The implementations provide evidence for the strong performance of the agent in sparse reward environments. This motivates an attempt to understand the mathematical origins of the intrinsic exploratory drive and the corresponding expected free energy objective function. Thereafter, a discussion of a taxonomic dichotomy between evidence and divergence objective functions is reviewed. Moreover, the taxonomy reveals a flexible scheme of heuristics for employing various variational objective functions for control. Lastly, we discuss a second taxonomy for exploratory methods in reinforcement learning. This provides additional perspective for the categorisation of the active inference agent's exploratory drive.

[*] This report serves as the written feedback submission for a reading module titled *Advanced Topics in Reinforcement Learning* in the department of Mathematics and Applied Mathematics at the University of Cape Town. The report seeks to study contemporary research in this field.

Contents

Notation

The notation below, follows [1]. We use this notation throughout all sections.

t, τ	Time index
s, s_t, s'	States of the environment
o, o_t	Observations
a, a_t	Actions
π, π'	A policy (sequence of actions, $a_1, a_2, a_3, \dots$)
π^*	The optimal policy
p	Known or unknown fixed probability density
q	An approximate probability density (variational density)
σ	Softmax function
D_{KL}	Kullback-Leibler divergence
$\mathcal{F}$	Variational free energy(VFE)
$\mathcal{G}$	Expected free energy (EFE)
$\mathcal{S}, \mathcal{R}, \mathcal{A}$	Sets of states, rewards and actions respectively.
r	Rewards
$\mathbb{E}$	Expectation function (mean)
C	Cumulative reward
γ	Discount factor hyperparameter (used for rewards)
V, Q or V_π, Q_π	State value function and state-action value function, respectively.
V^*, Q^*	Optimal state and state-action value functions.
$\mathcal{L}$	Objective (loss) function
$\tilde{s}, \tilde{a}, \tilde{r}$	Trajectories (for states, actions and rewards)
θ, ϕ, Ψ, ξ	Parameters (of neural networks)
$\tilde{p}$	Prior preference distribution (of the agent)
ϵ	Random noise term
α	Step-size hyperparameter
Ω	Binary variable representing optimality of a trajectory.
$\mathcal{H}$	Entropy function
$\tilde{\mathcal{F}}$	Free energy of expected future trajectories

1 Introduction and Overview

"Every good regulator of a system must be a model of that system" - Roger C. Conant and W. Ross Ashby [2]

The active inference [3] agent offers a general and flexible inference framework for addressing the reinforcement learning problem. This is the problem of inferring the optimal action sequence (the policy) that achieves the agent's goals in an uncertain environment. The active inference agent achieves this by minimising its variational free energy objective function. The agent, as derived from the free energy principle [4] of theoretical neuroscience, exhibits both reward-seeking and information seeking behaviour. This provides a means at addressing the trade-off between exploration and exploitation [5, 6]. Interestingly, the active inference agent's exploratory drive is one that exhibits a directed, intrinsic exploratory drive [1]. This drive provides our motivation for the study of exploratory methods in the context of the reinforcement learning problem problem.

The active inference agent is, specifically, one that provides a normative framework for decision-making in an uncertain environment, in order to achieve some goal. The goal corresponds to the agent's notion of preference [7]. As this problem corresponds to the reinforcement learning problem, the active inference agent can be considered a reinforcement learning agent. Perhaps one might argue that the generality and flexibility of the inference scheme afforded by the active inference agent (which makes inference as to the optimal action sequence to achieve the agent's goals) provides a more general framework than the reinforcement learning agent. Specifically, the active inference agent provides a theory of reinforcement learning that is grounded in variational inference and can therefore be straightforwardly extended to partially observable Markov decision processes (POMDPs), in a way that is non-trivial for reinforcement learning agents [1]. Perception, learning and action are unified in one Bayesian inference framework. This accommodates various types of data and hierarchies of distributions [7]. The modelling framework also offers the ability to represent preferences instead of rewards. This allows flexibility in learning non-monotonic, non-stationary and highly uncertain reward functions thus providing a Bayesian framework for handling uncertainty in the reward distributions [1]. Lastly the active inference agent's central tenet (based on the free energy principle) is to minimise its variational free energy objective function. As the expression of its objective function provides a unification of all components of the inference scheme, this affords, in principle, the training of all hyperparameters involved via the direct gradient descent on the variational free energy objective function, [1]. This offers the possibility to remove the requirement of the (computationally expensive) grid search for optimal hyperparameter tuning.

Conversely, the active inference framework, presented here, makes some design choices which are not required in reinforcement learning. These include that the agent's policy is a softmax distribution of the agent's expected free energy path integral (its 'value function') which effectively introduces Thompson sampling which is not necessary in contemporary reinforcement learning. Secondly, the computation of the agent's information gain term in its objective function requires the presence of a model of the state transition dynamics of the world. While this can be done in a model-free way [8], it is not required in traditional reinforcement learning either. However, it is precisely this information gain term which motivates our own exploration into the active inference agent in the context of the exploration-exploitation dilemma in the field of reinforcement learning (RL).

Specifically, we seek to understand the mathematically principled origins of this intrinsic exploratory objective. We also seek to understand the contexts in which it is beneficial. We note that this work directly follows an account of active inference and reinforcement learning which was reviewed for an honours project that studied the use of deep RL methods for scaling active inference [9]. Hence this paper looks to briefly (re)-introduce the RL problem and the exploration vs exploitation dilemma. In this context some common, basic methods of embedding exploratory behaviour in the RL agent are discussed. This provides the context in which we address the exploratory drive of the active inference agent. It also provides some exploratory methods for comparison.

As we seek to understand the principled mathematical origins of the active inference agent's exploratory drive and, since this is a mathematical paper, we change direction from the RL problem and review the origins of active inference. Here we introduce the free energy principle of contemporary neuroscience [4, 10, 11, 1]. This asks the question: can one characterise the necessary behaviour of any dynamical system, at a non-equilibrium steady state, that maintains a statistical separation from its environment, via a Markov Blanket? The application of the principle (which is conditioned upon several contentious assumptions[1]) means that one can characterise such a system as performing approximate, variational Bayesian inference. This grounds our understanding of the principled origins of active inference agents which, interestingly, motivate for biological plausibility [12]. The free energy principle affords us the ability to construct a decision-making agent (in an uncertain environment) that performs approximate Bayesian inference to achieve its preferences. This is our active inference agent.

Subsequent to the introduction of the free energy principle we provide a lightning refresher to the means through which active inference agents operate. This will be explained in terms of the inference schemes of perception, action and learning which are united via the minimisation of variational free energy (VFE). Via a decomposition of the agent's objective functions, we uncover the agent's directed intrinsic exploratory drive. As an empirical motivation for the study of this exploratory objective, three case studies of implementations of deep active inference agents are reviewed [8, 13, 14]. These provide some empirical evidence for the possible benefits of the directed exploratory motive (and also the active inference agent, especially as compared to benchmark RL agents). In order to study the implementations, the three algorithms are briefly revised (from the thorough review of our previous work [9]). Thereafter we discuss their empirical performance in 'proof-of-concept' trials in benchmark environments from OpenAI gym [15]. The three algorithms include a model-free algorithm [8], a model-based algorithm [13] and, a third algorithm which is a hybridisation of the previous two [14]. As the active inference agent's exploratory motive is *directed* it provides an efficiency gain when compared to more random methods of exploration within sparse reward environments [13].

After reviewing the performance of the deep active inference agents we are motivated to turn away from the empirical studies in order to study the mathematically principled origins of the exploratory gain term [1]. As will be discussed, we first see the exploratory gain term arise in the active inference agent's *expected free energy* objective function. However the study of [16] identifies this as a design choice. Moreover, a second objective function, the *free energy of the future* offers a principled application for the replacement of the expected free energy objective within the active inference framework. This objective can be seen to appear in the context of the *control as inference* framework [17] (which we revise from our introduction in [9, 18]). As the free energy of the future does not exhibit an information gain term, for exploration, this leads to the question of the mathematically principled origin of such a term.

The study of [19] proposes a dichotomy between two kinds of objectives for variational agents. These are the *evidence* and *divergence* objectives. It is precisely this divergence objective which yields the intrinsic information gain term for exploration (that is seen in the expected free energy objective). Furthermore, the identified dichotomy between evidence and divergence objectives prompts the formulation of a taxonomy for variational objectives for the control problem (and therefore for RL) [1].

The evidence vs divergence comparison forms a single dimension of this (three dimensional) taxonomy, with a second dimension discussing how value is encoded in the model (in an exogenous or endogenous way). The third dimension discusses design choices for the generative model embedded in the variational objective functions. This taxonomy affords us a classification scheme for some of the objectives discussed in the context of the active inference agents. Moreover, the taxonomy offers some mathematically principled heuristics for design choices of variational objective functions.

Finally, a second taxonomy, resulting from a survey of exploratory methods in RL is discussed [20]. This offers us a further means of categorisation and comparison for the active inference agent's intrinsic informa-

[1] ... depending on the use-case.

tion gain exploratory objective. Overall, this report seeks to try understand this exploratory objective. We try understand this in terms of both empirical benefit and in terms of mathematical and intuitive principles. Furthermore we seek to understand the principled origins of such a term. The discussion therefore offers a perspective on how such a term can benefit exploration within RL and moreover, offers a view of other unexplored variational objectives [1]. Furthermore, some interesting methods for exploration in RL are reviewed along the way.

2 RL and the Exploration-Exploitation Dilemma

This section provides the primary context for our study of the active inference agent's exploratory drive; the reinforcement learning (RL) problem and exploration-exploitation dilemma. This section contains only a lightning introduction to the RL problem where we briefly discuss some of its solution methods. This provides a refresher to the experienced reader of the basic concepts of the RL problem and does not comprise a sufficient introduction to the topic.[2] Subsequently to the study of the RL problem, the exploration and exploitation dilemma is reviewed. This leads to a discussion of several basic methods of exploration for RL agents. We also introduce a taxonomy from [20] which will aid in later discussion (about how the exploratory drive of the active inference objective function can be classified).

2.1 The Reinforcement Learning Problem

The RL problem asks how an agent can optimally select a sequence of actions in an uncertain environment so as to maximise a numerical reward signal [6]. RL is a subset of machine learning that can be considered an evaluative method. The two key characteristics of the RL agent are: (i) that it engages in *trial-and-error* learning and, (ii) that it engages in *reinforcement* learning via the feedback provided by the numerical reward signal. The RL problem is often framed in terms of a Markov decision process (MDP). This can be defined as the tuple $(\mathcal{S}, \mathcal{A}, r, p, \gamma)$ [22] which consists of states of the environment, $s \in \mathcal{S}$, actions available to the agent in a given state, $a \in \mathcal{A}$, a reward function, $r(s, a)$, transition probabilities between states (given actions and previous states) p and, a discount factor, γ. The processes of trial-and-error learning and reinforcement learning via feedback are as follows. As the agent finds itself in an uncertain and changing environment it must sample actions and states in order to ascertain which action sequences will obtain the greatest amount of cumulative reward (the central tenet of the agent). Thus it engages in trial-and-error learning by sampling states and actions and receiving feedback via the numerical reward signal provided by the environment. An example Markov decision process from [6] appears in figure 1.

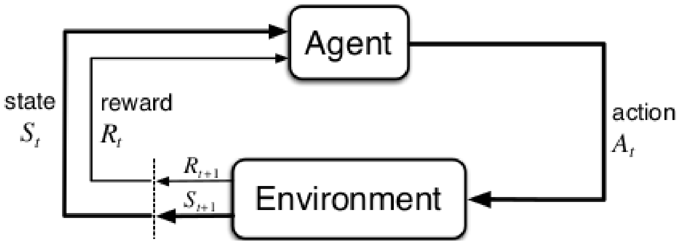

Figure 1: An example MDP from [6]. This illustrates what is referred to as the agent-environment interface whereby the agent selects actions which may change the state of the environment and which provide some amount of reward, emitted by the environment. The three signals transmitted in the RL problem are, therefore, the state, reward and action signals.

The RL problem has a whole zoo of associated solution algorithms. There is also a slew of taxonomic

[2] A sufficient introduction can be found via the textbook by Sutton and Barto [6] and via the lecture series by David Silver [21].

categories for these. Some of these categorisations look at the difference between model-free vs model-based, off-policy vs on-policy, temporal difference vs Monte Carlo, value-based vs policy-based, offline vs online, single-agent vs multi-agent, fully vs partially observable [6] and, iterative inference vs amortised inference [22]. Understanding these taxonomic categories is important for understanding the various solution algorithms for the RL problem and, when to apply them. For our purposes we highlight the difference between model-free and model-based RL, between temporal difference and Monte Carlo solution methods and, between value-based and policy based algorithms.

Solving the RL problem has made use of techniques of approximate dynamic programming [6]. Here we refer to the use of the recursive Bellman equations and the use of temporal difference (TD) learning. The Bellman equations offer a recursive bootstrapped expression for the values of states. The recursion occurs via the inclusion of the value of successor states. Defining state values and using this recursive expression allows for an ease of 'backups' of state values. This process makes use of the experience of the agent's trajectories. More specifically, TD learning uses these recursive statements to perform updates of state values. For the case of the *Q-learning* algorithm, state-action value functions, $Q(s,a)$, can be updated using the sample reward received by the agent along its experienced trajectory. This is added to the state value of its successor state via the expression below [6]:

$$Q\left(s_t, a_t\right) \leftarrow Q\left(s_t, a_t\right)+\alpha\left[r_{t+1}+\gamma \max_a Q\left(s_{t+1}, a\right)-Q\left(s_t, a_t\right)\right] \tag{1}$$

Hence, to select a policy (when given a state value function), the agent need only select actions such that they are *greedy* with respect to the estimated value function. This occurs while remembering that the central goal of the RL agent is to maximise its notion of cumulative reward. The corresponding policy can be denoted as [1]:

$$\pi\left(a_t|s_t;\theta\right)=\delta\left(a_t-\max_a Q_\pi(s_t,a_t)\right)=\begin{cases}1 & \text{if } a_t=\underset{a\in\mathcal{A}(s_t)}{\operatorname{argmax}}\ Q_\pi\left(s_t,a_t\right)\\ 0 & \text{otherwise}\end{cases} \tag{2}$$

The iteration between estimating values of states and selecting the corresponding greedy policies is called *generalised policy iteration* [6]. This makes use of the policy improvement theorem to suggest that this iteration will eventually converge on the optimal policy π^* and, the corresponding optimal value function V^*. The policy improvement theorem [6] states that if π and π' are any pair of policies such that the state-action value function $Q_\pi\left(s,\pi'(s)\right)\geq V_\pi(s), \forall s\in\mathcal{S}$ then this implies that the policy π' must be better than or equivalent to π.

The above Q-learning algorithm has been extended in the context of deep reinforcement learning with deep Q networks (DQN) [23, 24]. In our RL taxonomy, this is a model-free and value-based method. This highlights another key technique in the RL solution method toolkit, the use of Monte Carlo (MC) methods [6]. These can be contrasted with temporal difference methods in terms of the bias and variance trade-off. Here, the bootstrapped expression for value functions in temporal difference methods yield bias. MC methods refer to the agent's collection of sample trajectories to estimate expectations needed for both state value estimation and, policy function approximation. Hence, MC methods can yield high variance estimates. An example of the use of MC methods appears in the following *policy gradient method*.

The above Q-learning algorithm refers to a state-action value function. The derivation of the policy (equation 2) from this, refers to a value-based method for solving the RL problem [6]. Value-based methods (which occur in the model-free RL solution algorithm setting) can be contrasted with policy-based methods. Policy-based methods look to optimise a policy function, $\pi(a_t|s_t;\theta)$, that maps states to actions for the agent. The policy function is often approximated using a neural network with parameters, θ. The optimisation of the parameters, θ, can be achieved via gradient ascent on the RL agent's expected cumulative reward function. Hence, if the agent has policy π then, the expected cumulative reward expected under a trajectory, $\tilde{\tau}$, of states, actions and rewards is [6]:

$$J(\theta) = \mathrm{E}_{\tilde{\tau}\sim\pi}\left[\sum_t r\left(s_t, a_t\right)\right] \tag{3}$$

The *policy gradient theorem* [25] then says that, to perform gradient ascent on the above objective function, the gradient can be calculated as below. This says that the gradient of the objective function (with respect to the policy parameters) is the average gradient of the policy multiplied by the rewards received.

$$\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta \int d\tilde{\tau} r(\tilde{\tau})\pi_\theta(\tilde{\tau}) \\
&= \int d\tilde{\tau} r(\tilde{\tau})\nabla_\theta \pi_\theta(\tilde{\tau}) \\
&= \int d\tilde{\tau} r(\tilde{\tau})\frac{\nabla_\theta \pi_\theta(\tilde{\tau})}{\pi_\theta(\tilde{\tau})}\pi_\theta(\tilde{\tau}) \\
&= \int d\tilde{\tau} r(\tilde{\tau})\nabla_\theta \ln \pi_\theta(\tilde{\tau})\pi_\theta(\tilde{\tau}) \\
&= \mathbb{E}_{\pi_\theta(\tilde{\tau})}\left[r(\tilde{\tau})\nabla_\theta \ln \pi_\theta(\tilde{\tau})\right]
\end{aligned} \tag{4}$$

Hence, given a trajectory of states, actions and rewards; the agent has a means of estimating expectations of reward and performing gradient ascent to improve its policy. After collecting some MC sample trajectories from a policy, the gradient of the objective function with respect to the policy parameters is computed via stochastic gradient ascent. This is used for the purpose of updating the policy parameters with some step size hyperparameter α [6]:

$$\theta \leftarrow \theta + \alpha\nabla_\theta J\left(\theta\right)) \tag{5}$$

To contrast the MC approaches to temporal difference methods, one can replace the value of cumulative reward $r(\tilde{\tau})$ in the policy gradient (equation 4), with a bootstrapped expression. The above policy gradient method gives a vague sketch as to the algorithm REINFORCE [25]. Other popular contemporary algorithms include the value based method DQN [23, 24] (mentioned above) which scales up Q-learning with deep neural networks. The actor-critic architecture [25, 26] offers a third model-free algorithm paradigm. This modifies the policy-based methods with a baseline that reduce the fluctuation sizes of sample rewards. The baseline appears in the advantage function, A_t. This is defined in the policy gradient below, a state value function V is used as the baseline:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta(\tilde{\tau})}\left[\underbrace{\left(r_t - V\left(s_t\right)\right)}_{A_t}\nabla_\theta \ln \pi_\theta(\tilde{\tau})\right] \tag{6}$$

We classify the aforementioned algorithms as model-free RL algorithms. Some other noteworthy contemporary model-free algorithms include (in the policy optimisation category) *proximal policy optimisation* (PPO) [27] which updates the agent's policy by maximising a surrogate objective function. This gives a conservative estimate of the change in the $J(\theta)$ objective, with respect to the change in θ [28]. In the actor-critic model-free category there are the *Deep Deterministic Policy Gradients* (DDPG) [29] and *Soft Actor-Critics* (SAC) [30] algorithms. DDPG learns a deterministic policy alongside a Q-function [28]. SAC is an off-policy algorithm that uses maximum entropy RL (which will be discussed as an exploratory method) to train a stochastic policy and stabilise learning [28]. These model-free algorithms present solution methods to the RL problem that involve collecting sample trajectories of states, actions and reward via the agent's own experience of the environment. The sample averages of rewards can then be used to estimate values of states or approximate expectations used in the calculation of the agent's policies. The model-free approach is therefore useful in environments where it is easy and *cost effective* to sample states-action-reward trajectories.

Model-based methods, in contrast, offer sample efficiency [6, 31]. These can learn from fewer data points than model-free methods. These data are used to learn a model of the agent's environment, the central characteristic of model-based methods. The agent can then sample from its model to improve its policy via model-based *planning*. Planning refers to the use of the model to simulate trajectories to improve the agent's policy (or value function). In other words planning can take a model as input, evaluate trajectories using a value function and, output an optimised policy [6]. One example is presented by the cross-entropy method (CEM) planner [32]. However model-based planning relies on the accuracy of the model to provide an effective policy. Hence model-based methods present an issue of bias to solution methods for the RL problem [28]. Furthermore model-free methods have achieved greater asymptotic performance than model-based RL [31]. To combine the advantage of both model-free and model-based RL, hybrids of the two can be used. One such hybrid architecture is called *Dyna* [6, 33]. The performance of another hybrid which combines model-free and model-based deep active inference [14] will be discussed later in terms of the performance of the active inference agent on some benchmark tasks to highlight this agent's exploratory drive.

The aforementioned algorithms have participated in the success of RL methods at solving challenging and complex tasks. Some examples include an RL agent dominating the world champion of Go [34], having success at the Atari games [23] and, having success in the domain of robotics [35]. One complexity that arises in the context of the RL problem is that of the trade-off between exploration and exploitation. This is discussed next and, forms the primary context for this report.

2.2 The Exploration and Exploitation Dilemma

The exploration exploitation dilemma [5, 6] refers to the trade-off between the agent exploring new regions of its search-space and exploiting its knowledge about which trajectories of the space yield the highest rewards. Because of the difficulty of searching a large, unknown and dynamic space, learning an optimal policy presents a challenge. While exploring the extent of the space can be useful for locating a global maximum, this may incur an opportunity cost associated with the potential gain of the known optimal reward. On the other hand if the agent were to exploit its best possible knowledge of an optimal way to act then it incurs a recurring opportunity cost proportional to the distance between its local optimal reward and the global optimum [1]. This presents the dilemma between exploration and exploitation and a trade-off for the RL agent.

Exploration in RL presents a challenge for RL solution methods. Other than the inherent trade-off with exploitation, exploration poses difficulty in environments with the following characteristics. Large state-action spaces mean that exploring the extent of the space is exponentially more challenging than for small state-spaces. Sparse (and delayed) reward environments mean that exploratory strategies struggle to obtain informative feedback or discover valuable states [20]. Additionally, highly uncertain environmental dynamics, either in the reward or in state transitions, cause further uncertainty in estimates, increasing the difficulty of the RL problem.

2.2.1 Basic Methods of Exploration

To resolve the exploration-exploitation dilemma a number of exploratory methods have been devised to amend implementations of RL algorithms. Since we are interested in the active inference agent's *directed, intrinsic* exploratory drive (which will be discussed in the next section), it is useful to discuss a taxonomy of exploratory methods in RL. Two new surveys of exploration in RL offer the opportunity to review some of these methods and where they fit in. We will study the survey in [20], however we point to the taxonomy offered by the survey of [36] as useful for contrast. The taxonomy of exploration methods appears in figure 2.

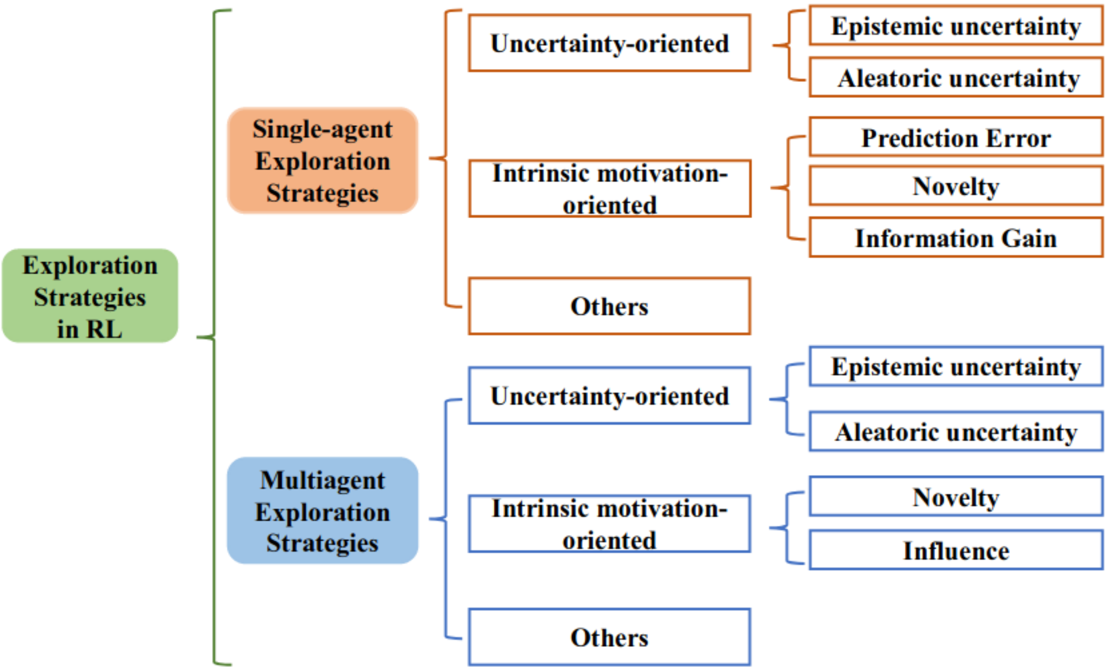

Figure 2: A taxonomy of exploration methods for deep RL and multi-agent RL as found in [20].

While the taxonomy in figure 2 classifies exploratory methods for both single and multi-agent RL, we focus only on the single agent case since the multi-agent case involves additional complexities which are not discussed in this work. In the context of single-agent RL the taxonomy of figure 2 outlines two major categories for exploratory methods. The first of these is the uncertainty oriented exploratory category. The second category is the intrinsic motivation-oriented category. Before discussing these categories, we turn to a review of some basic exploratory methods in RL. Subsequent to our discussion of the active inference agent, its implementations and its exploratory objective, we will discuss the above taxonomy in more detail, in section 6.

ϵ-Greedy

Arguably the most basic approach to exploration is called the ϵ-greedy method [6]. In this approach, the agent assigns a probability of $1 - \epsilon$ to selecting the greedy action. Complementing this, it assigns a probability of ϵ to selecting a random action. Typically a small probability is assigned for ϵ. A range of $0.02 - 0.05$ is suggested in [1]. While this approach is popular [20], the random selection of actions, necessarily, incurs an opportunity cost as the selected action may not benefit the exploratory effort of the agent. To resolve this (inefficiency), many approaches decay the value of ϵ with time to encourage convergence to the optimal policy.

Boltzmann Exploration

A second method of exploration is called Boltzmann exploration [37]. This applies to discrete state-action spaces where actions are selected with a probability drawn from a softmax function of the state-action value function $Q(s, a)$. This softmax function is:

$$q(a|s) = \frac{\beta \exp(-Q(a|s))}{\sum_a \exp(-Q(a|s))} \tag{7}$$

The β hyperparameter controls the extent of random exploration. A lower value for β yields a less peaked distribution and a more random selection of actions. This approach is used in implementations of discrete state-space active inference, with the state-action value function being replaced by an *expected free energy* path integral [7, 38, 39].

Upper Confidence Bound

The upper confidence bound (UCB) method arises from multi-armed bandit problems [20, 40]. While $\epsilon-$greedy methods only select actions on a random basis for exploration, UCB methods also employ greedy action selection but, amend the choice of action selection to select actions greedily that yield the highest upper confidence bound on the Q values. The upper bound calculates the potential of each action to generate a large reward. This yields an optimistic form of exploration. The upper confidence bound on the Q value appears below, in the expression for action selection [6]:

$$a_t = \arg\max_{a_t} \left[Q_t(a) + c\sqrt{\frac{\ln t}{N_t(a)}} \right] \tag{8}$$

This particular upper confidence bound incorporates a count, $N_t(a)$, of how many times the action a has been selected. The higher the count, the narrower the confidence bound. If $N_t(a)$ is 0 then a is considered to be the selected action [6]. Lastly, c is a hyperparameter that controls the degree of exploration.

Gaussian process UCB generalises the above method in a Bayesian manner [20]. If some function $f(x)$ provides a quantity of interest (such as a Q value) then if this is modelled as a Gaussian process with mean μ and variance σ then actions can be selected based on the upper confidence bound given by [20]:

$$\mu_{t-1}(x) + \beta_t^{1/2} \sigma_{t-1}(x) \tag{9}$$

If this applies to the Q value function, then actions are selected to maximise the expected reward added to one weighted standard deviation. The β_t hyperparameter can be tuned to reduce the influence of exploration over time.

Entropy Regularisation

Also known as maximum entropy RL [18], entropy regularisation [20] contrasts to the previous two methods. The previous two methods focused on value-based exploration which selects actions based on the Q values and adds randomness in that area. In contrast, entropy regularisation focuses on the policy gradient approach. As the policy function is trained on some objective function (previously equation 3), an entropy term can be added, to control the extent of stochasticity of the policy function, $q(a|s)$. This term measures the entropy of the stochastic policy distribution and acts as a regularisation term on an overall objective function for training the policy. This may appear as [18]:

$$= \mathbb{E}_{q(\tilde{a}|\tilde{s})p(\tilde{s})} \left[\sum_t^T r(s_t, a_t) \right] + \mathcal{H}[q(\tilde{a} \mid \tilde{s})] \tag{10}$$

The influence of the entropy term can be decayed with time to encourage convergence to the optimal policy.

Noise Perturbation

Noise perturbation offers a simple technique for deterministic policies to add exploration [20]. Suppose we have a deterministic policy $\pi(s)$ then a noise perturbation exploratory policy $\pi'(s)$ can be constructed by the addition of some stochastic process (noise) term N. This results in a random exploratory policy $\pi'(s) = \pi(s) + N$. The noise term can be chosen to suit the application. As with ϵ-greedy methods, this method necessarily incurs an efficiency loss.

Thompson Sampling

Thompson sampling [20, 41] is also known as posterior sampling. Given a random objective function $f(x)$, and some estimated posterior for this, $\text{Posterior}(f)$, then Thompson sampling samples a function $f(x)$ from this posterior and selects actions greedily with respect to this objective. For example if f represents a Q value function function then Thompson sampling samples from its posterior distribution of the Q functions and selects the action with the highest expected reward (from the sampled Q function). This method encapsulates uncertainty via the posterior distribution and thus, as the posterior estimates are improved, a more realistic sample will be chosen, thus improving efficiency as compared to random methods. In other words, when the posterior is peaked then it is likely that a true optimum has been found [1]. Conversely, a posterior that is approximately uniform will yield extra exploration. For Thompson sampling, actions are selected as follows [20]:

$$a_t = \arg\max_{a_t} f(x), \text{ s.t. } f \sim \text{Posterior}(f) \tag{11}$$

Expected Improvement

For the case of Q values expected improvement selects actions so as to maximise the expected improvement in Q values (for instance) relative to the best respective value in history, Q^+ [20]. This method can also be considered to be a Bayesian approach as the expectation is taken with respect to the posterior distribution for Q. Actions are selected so as to maximise the following [20]:

$$\mathbb{E}_{Q(s,a)\sim\mathcal{N}\left(\mu_{t-1},\sigma^2_{t-1}\right)}\left[\max\left(Q - Q^+_{t-1}, 0\right)\right] \tag{12}$$

Hence we have reviewed a number of basic exploratory methods for the RL problem. These methods provide us with some ideas for exploration in RL. They also provide us with some alternatives to compare the active inference agent's exploratory motive to. As mentioned, the active inference agent offers a means at exploration that is *directed* and intrinsic. As we will see, this offers a means of exploration that is effective in sparse reward environments [13]. In particular this ameliorates the inefficiencies of some of the algorithms above which employ random exploration. With the notable exception of maximum entropy exploration (which can be derived from a variational inference approach to the RL problem [18]), many other methods of exploration can be considered ad-hoc [1]. In contrast, as we will see, the active inference exploratory motive is explicitly derived from the minimisation of expected free energy. Hence we will later inquire as to where this (expected free energy) objective arises. As a primer to this question, we turn to the origins of the active inference agent, the free energy principle. It is from this principle, where active inference agents are explicitly derived [3]. Hence, before elaborating about this agent we digress to discussing the free energy principle. This will hopefully illuminate some of the origins of the active inference agent's objective functions which will (in turn) guide our questioning as to the origins of the expected free energy objective function and, the intrinsic exploratory drive. Furthermore the (mathematical) discussion of the free energy principle, provides us with theoretical foundations for constructing a decision-making agent which explores uncertain environments, aiming to achieve its goals.

3 The Free Energy Principle

The free energy principle (FEP) [4] is a formalisation of the principle written about by Ashby and Conant which says that “every good regulator of a system must be a model of that system” [1, 2]. It asks the question: *what characterises a self-organising, non-equilibrium, dynamical system that maintains a statistical separation from its environment*? Choosing to apply the principle would mean that one characterises such a system as performing approximate Bayesian Inference (also called variational inference). The principle relies on the existence of the self-organising system (into the non-equilibrium steady state). It also relies on the assumption that the system maintains a *Markov blanket* (a state structure with particular statistical independencies). Having satisfied these preconditions the principle provides the interpretation that the dynamics of the system perform variational Bayesian inference. This means that the internal states of the system perform inference about the external states (where internal and external states are defined by the Markov blanket structure). Since it is precisely this process, of minimisation of free energy, that maintains the system at a non-equilibrium steady state and since this process provides a Bayesian model of the system, then the FEP provides a regulator that models the system (as per Ashby and Conant).

From the neuroscientific perspective the FEP offers a homeostatic theory for how organisms are able to self-organise to resist the tendency of disorder, which occurs via the second law of thermodynamics [42]. As will be shown, via the minimisation of free energy, the organism conducts a process of *self-evidencing* [38]. Self-evidencing can be viewed as a homeostatic drive to remain in states of prior preference [42]. This characteristic relates to the origin of the active inference framework from the FEP. Thinking in terms of a decision-making agent that has some prior preferences, the preferences can be treated as a non-equilibrium steady-state (NESS) towards which the agent self-organises. Given this a-priori (and biased) treatment, the agent then selects actions in order to minimise its variational free energy and preserve its NESS. Thus the active inference agent fits into the FEP framework and, the FEP finds itself at the intersection of neuroscience and machine learning.

Hence, active inference modelling offers a useful framework in the context of neuroscience. The active inference process theory has, in fact, been motivated to offer biological plausibility [12]. It has also been used for modelling eye movements [43] and psychiatric disorders in the field of *computational psychiatry* [44]. While the neuroscientific aspects motivate our study of active inference, the focus of this paper is one of the RL problem and exploration-exploitation dilemma. What follows, in this section, is a formulation of the FEP to motivate for the origins of the active inference agent. In particular we follow the formulation presented by Beren Millidge in his chapter 2 of *Applications of the Free Energy Principle to Machine Learning and Neuroscience* [1] which, in turn, follows the arguments of Karl Friston in *A free energy principle for a particular physics* [10]. Hence the argument studied here is amended with the insights presented in [10] and [11].

Stochastic Differential Equation Dynamics and the NESS

Consider the following Langevin stochastic differential equation (SDE) where x can be a vector of states, $f(x)$ is a differentiable function and, ϵ is Normally distributed white noise s.t. $\epsilon \sim N(0, 2\Gamma)$ (Γ is half the variance of the noise):

$$\frac{dx}{dt} = f(x) + \epsilon \tag{13}$$

While the Langevin differential equation describes the dynamics of states, the Fokker-Planck equation describes the dynamics of a probability distribution over states, $p(x, t)$. Following the notation in [1] we have the Fokker-Planck equation denoted as:

$$\frac{dp(x,t)}{dt} = -\nabla_x f(x,t)p(x,t) + \nabla_x \Gamma \nabla_x p(x,t) \tag{14}$$

We assume that the dynamics of equation 13 tend toward a NESS density, p^*, i.e. we assume that $\lim_{t\to\infty} p(x,t) \to p^*(x)$ where a steady state is defined with $\frac{dp^*(x)}{dt} = 0$. In terms of a biological narrative, we note that biological systems that are self-organising are considered to be at a NESS whereby the NESS is maintained by steady input of energy over time (for otherwise the organism may dissipate into an equilibrium steady state). The steady input is often referred to as a *solenoidal flow.* Given that we are at a NESS, we take interest in the dynamics present at the NESS. We make use of the Helmholtz decomposition as per [11]. This decomposition (also known as the Fundamental Theorem of Vector Calculus) states that a vector field can be decomposed into divergence-free (solenoidal) and, curl-free (irrotational) components. For a vector field which we denote as F, following [11], we have that:

$$F = \Gamma(x) \cdot \nabla N + \nabla \times W \tag{15}$$

Here, N and W are a scalar and a vector potential respectively. $\Gamma(x) \cdot \nabla N$ is the curl-free, dissipative, component of the flow. $\nabla \times W$ is the solenoidal component of the flow. Following [11], we introduce $-Y(x)$, an anti-symmetric matrix ($-Y = Y^T$) and formulate the solenoidal component in terms of this such that[3]:

$$\begin{aligned} \nabla \times W &= -Y \nabla N \\ \implies F &= (\Gamma - Y)\nabla N \end{aligned} \tag{16}$$

The Helmholtz (Ao) decomposition rewrites the dynamics of the SDE, at the NESS, into a form that depicts the dissipative and solenoidal components to answer the question - how does the solenoidal flow prevent the system from relaxing into an equilibrium steady state [1] ? Γ gives the dissipative effects. Γ is the amplitude of the random fluctuations which try to increase the entropy and smooth out the NESS density. In contrast to the dissipative effects of Γ, the solenoidal component, Y, effectively opposes the dissipation arising from Γ. This happens despite the solenoidal component being orthogonal to the dissipative component. The implication is that the NESS is maintained at steady state due to the presence of flow (no detailed balance). One can see this in the following equation. The state dynamics are rewritten as $\frac{dx}{dt}$ and we choose $N = \ln p^*(x)$ (which we will see later is called negative surprise and which satisfies the Fokker-Planck equation). This gives us the following:

$$\frac{dx}{dt} = (\Gamma - Y)\nabla_x \ln p^*(x) \tag{17}$$

This satisifes the Fokker-Planck equation (for $f(x,t) = \frac{dx}{dt}$) to give the steady-state $\frac{dp^*(x)}{dt} = 0$, below. The last line results from the solenoidal flow being orthogonal to the gradient of the density and hence, the gradient of the solenoidal flow with respect to the gradient of the log density is zero [1]:

$$\begin{aligned} \frac{dp^*(x)}{dt} &= -\nabla_x \left[(\Gamma - Y)\nabla_x \ln p^*(x)\right] p^*(x) + \Gamma \nabla_x^2 p^*(x) \\ &= -\nabla_x \left[(\Gamma - Y)\frac{\nabla_x p^*(x)}{p^*(x)}\right] p^*(x) + \Gamma \nabla_x^2 p^*(x) \\ &= -\nabla_x \left[(\Gamma - Y)\nabla_x p^*(x)\right] + \Gamma \nabla_x^2 p^*(x) \\ &= -\Gamma \nabla_x^2 p^*(x) + \nabla_x Y \nabla_x p^*(x) + \Gamma \nabla_x^2 p^*(x) \\ &= \nabla_x Y \nabla_x p^*(x) = 0 \end{aligned} \tag{18}$$

[3] Note that in our formulation we remove the dependency on the states such that $\Gamma(x) = \Gamma; Y(x) = Y$, by form of assumption. We also assume that the random noise, Γ, contains no correlations between states.

We now have a SDE for our states x. This is in terms of a dissipative, Γ, and solenoidal, Y, flows which provides the insight that the solenoidal flow is the means through which the system is maintained at a NESS density [1].

Markov Blankets

To expand our insights into the dynamics of the states x we model their statistical properties. We start by forming a disjoint partition of the states $x = \{s, \mu, b\}$. s represent external states of the environment, μ represent the internal states of the agent and b represent the blanket states which separate the internal states from their environment. These will help us uncover the insight of the FEP, that maintaining a NESS against environmental perturbations requires that internal states model the external states via variational Bayesian inference. We note that all influence between internal and external states occurs through the blanket states. This can be viewed via the *Markov blanket* [45] property which says that internal and external states are conditionally independent, given the blanket states [1]:

$$p^*(s, \mu, b) = p^*(s|b)p^*(\mu|b)p^*(b) \tag{19}$$

This informs the dynamical flow of the system which can be viewed in figure 3. The figure depicts a further decomposition of the set of blanket states into sensory states, o, and active states, a. A causal loop is depicted whereby sensory states are the *causal children* of the external states. Hence the environment acts on the sensory states which can in turn influence internal states. The internal states can influence the active states which can influence the environment, hence defining a loop.

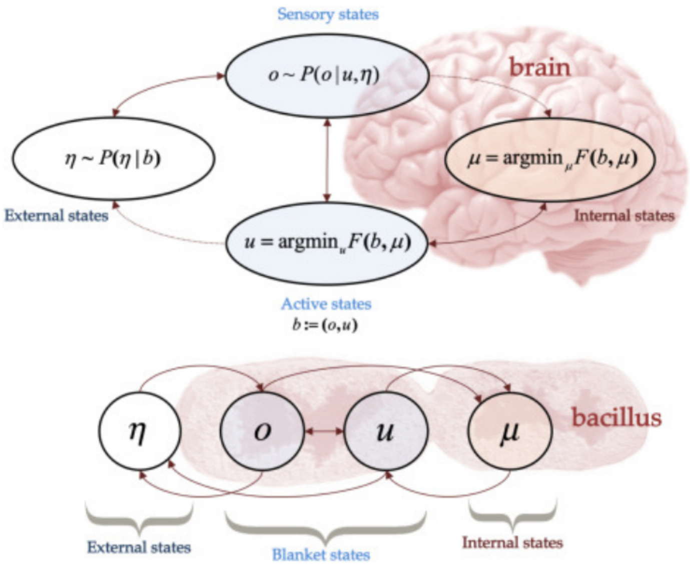

Figure 3: Markov Blanket loop. The internal states μ are separated from the external states (here as η, instead of s), via a Markov Blanket which consists of sensory states (here as u instead of o), and active states, a. The external states influence the sensory states which influence the internal states. The internal states can then influence the active states which can influence the environment. The diagram depicts the minimisation of free energy of the internal states (corresponding to perception) and the minimisation of expected free energy of the active states (corresponding to action). This figure appears in [1] and is adapted from [10] where the brain and bacteria (bacillus) are referred to.

Marginal Flow Lemma

To proceed with understanding the state dynamics under the condition of the Markov blanket and with our corresponding SDE in hand, we utilise the marginal flow lemma. The lemma connects the sparse influences mediated by the Langevin dynamics, to the conditional independencies among subsets of states mediated by the Markov blanket [10]. Marginal flow refers to the flow of certain subsets of states, averaged (marginalised) over the other states. The lemma takes the standard form for flow in equation 17 and generalises this to a partition of states that contain a Markov blanket. The Marginal flow lemma, as per [10], says that:

Lemma (marginal flow): *For any weakly mixing random dynamical system at a NESS, the marginal flow $f_\mu(s)$ of any subset of states $\mu \in X$, averaged under the complement of another subset $s \in X$ can be expressed in terms of gradients of the logarithm of the corresponding marginal density (where $\bar{s}$ denotes the complement of $s \in X$) such that:*

$$f_\mu(s) = \mathbb{E}_{p(\bar{s}|s)}\left[\frac{d\mu(x)}{dt}\right] = (\Gamma_{\mu\mu} - Y_{\mu\mu)}\nabla_\mu \ln p^*(s) + Y_{\mu\bar{\mu}}\nabla_{\bar{\mu}} \ln p^*(s) \tag{20}$$

Returning our attention to [1] and using the marginal flow lemma, we see that equation 20 describes the flow of internal states μ averaged under the complement $\bar{s}$ of particular states s. The marginal flow lemma allows us to express the flow of a subset of states, averaged under their complements, in terms of the SDE we derived earlier from the Helmholtz decomposition on the marginal NESS density. This is an expression of the conditional independencies implicit in the Markov Blanket structure in terms of the flows of the SDEs. This provides insights into the information-theoretic properties of the Markov Blanket loop and into the influences of one set of states over another.

Corollary (conditional independence): *If the flow of one subset of states does not depend on another, then it becomes the flow expected under the second subset [10].*

The corollary gives us the dynamics of the conditional independencies implicit in figure 3, [10]:

$$\begin{bmatrix} f_s(x) \\ f_o(x) \\ f_\mu(x) \\ f_a(x) \end{bmatrix} = \begin{bmatrix} f_s(s,b) \\ f_o(o,b) \\ f_\mu(\mu,b) \\ f_a(\mu,b) \end{bmatrix} = \begin{bmatrix} (Y_{ss} - \Gamma_{ss})\,\nabla_s \ln p^*(s,b) \\ (Y_{oo} - \Gamma_{oo})\,\nabla_o \ln p^*(s,b) + Y_{oa}\nabla_a \ln p^*(s,b) \\ (Y_{\mu\mu} - \Gamma_{\mu\mu})\,\nabla_\mu \ln p^*(\mu,b) \\ (Y_{aa} - \Gamma_{aa})\,\nabla_a \ln p^*(\mu,b) + Y_{ao}\nabla_o \ln p^*(\mu,b) \end{bmatrix} \tag{21}$$

Furthermore, a special case of the marginal flow lemma when $s = \mu$ and $Y_{s\bar{s}} = 0$ tells us that the expected flow of any subset of states, averaged over all other states will behave in the same way as all states considered together. It will perform a gradient descent on its marginal NESS density [10].

Corollary (expected flow): *The marginal flow of any subset $s \subset X$ averaged over all other states depends only on the gradients of its marginal density, provided there is no solenoidal coupling with its complement [10]:*

$$f_s(s) = (\Gamma_{ss} - Y_{ss})\,\nabla_s \ln p(s) \tag{22}$$

Studying equation system 21, demonstrates the conditional independencies of the Markov Blanket loop in figure 3 [10]. It reveals the information theoretic properties of interactions between the different sets of states. For instance, consider the marginal flow of the set of *autonomous* states (the active and internal states) $\alpha = \{a, \mu\}$, expected under the particular states, π. By the marginal flow lemma we have:

$$
\begin{aligned}
f_\alpha(\pi) &= (\Gamma_{\alpha\alpha} - \Upsilon_{\alpha\alpha})\, \nabla_\alpha \ln p^*(\pi) \\
\alpha &= \{a, \mu\} \\
\pi &= \{o, \alpha\}
\end{aligned}
\tag{23}
$$

We see that the set of states, α, follows a gradient descent on the marginal NESS density of the internal, sensory and active states. They attempt to minimise their surprise, $-\ln p^*$. To gain further insight into what this means we expand the surprise in terms of its interaction with external states beyond the blanket [1]:

$$
\begin{aligned}
-\ln p^*(\pi) &= \mathbb{E}_{p^*(s|\pi)}\left[-\ln p^*(\pi)\right] \\
&= \mathbb{E}_{p^*(s|\pi)}\left[\ln p^*(s|\pi) - \ln p^*(s,\pi)\right] \\
&= \mathbb{E}_{p^*(\mathfrak{s}|\pi)}\left[\ln p^*(s|\pi) - \ln p^*(\pi|s) - \ln p^*(s)\right] \\
&= \underbrace{\mathbb{E}_{p^*(s|\pi)}\left[-\ln p^*(\pi|s)\right]}_{\text{Inaccuracy}} + \underbrace{D_{\mathrm{KL}}\left[p^*(s|\pi) \| p(s)\right]}_{\text{Complexity}}
\end{aligned}
\tag{24}
$$

Here, we see that the flow of the autonomous states minimises the inaccuracy and complexity of the external states with respect to the particular states of the system in question [1]. In other words the minimisation of the inaccuracy term corresponds to maximising the likelihood of the internal states given the external states (self-evidencing). The complexity term corresponds with a minimisation of the divergence between the prior and posterior of the external states. This provides (an incomplete) clue of the key insight of the FEP - that maintaining a NESS against environmental perturbations requires that internal states model external states via Bayesian inference.

Variational Inference

In summary and so far, a model has been established of a partitioned set of states of the system. The model describes the conditional independencies in terms of the Markov Blanket and described the dynamics in terms of the SDEs formulated from the Fokker-Planck equation, Langevin equation and Helmholtz decomposition. The conditional independencies given by the Markov Blanket have been expressed in terms of how the marginal flow of a subset of states depends on other states. Lastly the model provides a clue as to how maintaining a NESS against environmental perturbations implies that internal states model the external state via variational Bayesian inference. Next, we provide an introduction to variational Bayesian inference. As has been expressed above, we can view self-organising systems that maintain themselves at a NESS to be performing approximate Bayesian inference. By this we mean that they are implicitly performing variational Bayesian inference and therefore implicitly minimising variational free energy (VFE). As a reminder, we define VFE as a tractable bound on the KL divergence between a desired true posterior and a variational posterior. This can be defined as follows [1]:

$$
\begin{aligned}
\mathcal{F}(o,\theta) &= D_{\mathrm{KL}}[q(s|o;\theta) \| p(s,o)] \\
&= D_{\mathrm{KL}}[q(s|o;\theta) \| p(s|o)] - \ln p(o) \\
&\geq D_{\mathrm{KL}}[q(s|o;\theta) \| p(s|o)]
\end{aligned}
\tag{25}
$$

By minimising the tractable upper bound on the divergence between the posterior and approximate posterior the agent can improve upon its estimate of the true posterior. VFE can also be posed as an upper bound on surprise and hence, when an agent minimises VFE it is implicitly minimising the upper bound on surprise [1]:

$$
\begin{aligned}
-\ln p(o) &= -D_{\mathrm{KL}}[q(s|o;\theta)\|p(s,o)] + \mathcal{F}(o,\theta) \\
&\leq \mathcal{F}(o,\theta)
\end{aligned}
\tag{26}
$$

The following (decomposition of VFE) provides insight into what happens during minimisation (where the entropy of a distribution is defined as $\mathcal{H}[q(s|o;\theta)] = \int dq(s|o;\theta)\ln q(s|Do;\theta))$ [1]. A second decomposition (in terms of accuracy and complexity) will be discussed in the next section where active inference agents are (briefly) defined. The decomposition in the expression, below, highlights that minimising VFE minimises the energy term which implies a maximisation of the joint probability of the generative model and, a maximisation of entropy. This means that the variational density should maximise the joint probability of the generative model (energy) while remaining as uncertain as possible (entropy). The second term acts as a regularisation on the Bayesian update.

$$
\mathcal{F}(o,\theta) = \underbrace{\mathbb{E}_{q(s|o;\theta)}[\ln p(s,o)]}_{\text{Energy}} - \underbrace{\mathcal{H}[q(s|o;\theta)]}_{\text{Entropy}}
\tag{27}
$$

Information Geometry

With the tools of variational inference, we turn to the problem of how we can interpret the dynamics of a system at a NESS as performing approximate Bayesian inference about its environment. Information geometry provides a tool through which we are able to link the concepts of dynamical motion in space and variational inference on parameters of distributions [1]. We want to understand the relationship between internal and external states. We will find that there exists a mapping. This is that the most likely internal states (given a specific blanket state) can be mapped to a *distribution* over external states. We define the most likely internal state and most likely external state, given a specific blanket state (assuming injectivity between most likely internal/external states and blanket states) as:

$$
\begin{aligned}
\boldsymbol{s}(b) &= \underset{s}{\operatorname{argmax}}\ p(s|b) \\
\boldsymbol{\mu}(b) &= \underset{\mu}{\operatorname{argmax}}\ p(\mu|b)
\end{aligned}
\tag{28}
$$

We assume that there exists a smooth and differentiable function ψ between the likeliest internal states and the most likely external states (given a blanket state) [1]:

$$
\boldsymbol{s}(b) = \psi(\boldsymbol{\mu}(b))
\tag{29}
$$

The most likely external states, $\boldsymbol{s}(b)$ is interpreted as parameterising a distribution of external states. More formally, if $q(s)$ represents a distribution of external states, then this is parameterised by $\boldsymbol{s}(b)$ such that $q(s;\boldsymbol{s}(b)) = q(s;\psi(\boldsymbol{\mu}(b)))$. Thus the space of internal states parameterises a space of distributions of external states [1]. We note that the space of distributions of external states is curved and non-euclidean and we measure the KL divergence to assess the distance between distributions. Hence to assess the change of a distribution with respect to a change in parameters we define [1]:

$$
\frac{\partial p(x;\theta)}{\partial \theta} = \lim_{\delta\theta \to 0} D_{\mathrm{KL}}[p(x;\theta)\|p(x;\theta+\delta\theta)]
\tag{30}
$$

For instance in the special case that we are dealing with parameters of distributions in the exponential family then information geometry tells us that the space of these parameters is a non-euclidean space with a Fisher Information metric, $\mathbb{F}$. If we define $\theta' = \theta + \delta\theta$ and we Taylor expand around $\theta' = \theta$, this yields [1, 10]:[4]

$$\begin{aligned}
D_{\mathrm{KL}}\left[p(x;\theta)\|p\left(x;\theta'\right)\right] \approx & \underbrace{D_{\mathrm{KL}}[p(x;\theta)\|p(x;\theta)]}_{=0} + \underbrace{\left.\frac{\partial D_{\mathrm{KL}}\left[p(x;\theta)\|p\left(x;\theta'\right)\right]}{\partial\theta}\right|_{\theta=\theta'}\left(\theta-\theta'\right)}_{=0} \\
& + \left.\frac{\partial^2 D_{\mathrm{KL}}\left[p(x;\theta)\|p\left(x;\theta'\right)\right]}{\partial\theta^2}\right|_{\theta=\theta'}\left(\theta-\theta'\right)^2 \\
& = \int p(x;\theta)\frac{\partial \ln p(x;\theta)}{\partial\theta}\frac{\partial \ln p(x;\theta)}{\partial\theta} \\
& = \mathbb{F}
\end{aligned} \tag{31}$$

Returning to the specifics of our Markov Blanket, we have that since the internal states parameterize distributions over external states, they lie on an information geometric manifold with a Fisher information metric (recall $q(s;\boldsymbol{s}(b)) = q(s;\psi(\boldsymbol{\mu}(b))))$ [1]. We refer to this as the *extrinsic geometry.* We define the *metric tensor*, m, over the space of external state densities, parameterized by internal states as [1, 10]:

$$m(\boldsymbol{s}) = \left.\frac{\partial^2 D_{\mathrm{KL}}[q(s;\boldsymbol{s})\|q(s;\boldsymbol{s}+\delta\boldsymbol{s})]}{\partial\boldsymbol{s}^2}\right|_{\boldsymbol{s}+\delta\boldsymbol{s}=\boldsymbol{s}} \tag{32}$$

Since the internal states also parameterize a second distribution over internal states $(q(\mu;\boldsymbol{\mu}))$ we also have another information geometry which we call the *intrinsic geometry.* The metric concerning the intrinsic information geometry is:

$$m(\boldsymbol{\mu}) = \left.\frac{\partial^2 D_{\mathrm{KL}}[q(\mu;\boldsymbol{\mu})\|q(\mu;\boldsymbol{\mu}+\delta\boldsymbol{\mu})]}{\partial\boldsymbol{\mu}^2}\right|_{\boldsymbol{\mu}+\delta\boldsymbol{\mu}=\boldsymbol{\mu}} \tag{33}$$

These information geometries enable us to mathematically link dynamical motion in space with variational inference as we see the motion of internal states as movement on the intrinsic and extrinsic statistical manifolds since $\boldsymbol{s} = \psi(\boldsymbol{\mu})$.[5]

The Free Energy Lemma

The free energy lemma says that the internal state dynamics can be seen to minimise a free energy functional over external states (therefore performing approximate Bayesian inference) [10]. To illustrate this lemma we refer back to equation 23 which illustrated the flow of the autonomous states in terms of the dissipative and solenoidal components of the gradient descent on $\ln p^*(\pi)$.:

$$\begin{aligned}
f_\alpha(\pi) &= (\Gamma_{\alpha\alpha} - Y_{\alpha\alpha})\,\nabla_\alpha \ln p^*(\pi) \\
\alpha &= \{a,\mu\} \\
\pi &= \{o,\alpha\}
\end{aligned} \tag{34}$$

[4] Here we have used that the Fisher information is the variance of the score function: $\mathbb{F}(\theta) = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2|\theta\right] = \int_{\mathbb{R}}\left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right)^2 f(x;\theta)dx$ and that the KL divergence is as usual $D_{\mathrm{KL}}(P\|Q) = \int_{-\infty}^{\infty} p(x)\log\left(\frac{p(x)}{q(x)}\right)dx$.

[5] This embeds our understanding of the inference procedure of internal states, as one that is adaptively active in changing external states.

If we assume that the variational density of the external states approximates the true posterior, $q(s|\pi) \approx p^*(s|\pi)$ then we can rewrite the dynamics of the autonomous states below [1, 10]:

$$\begin{aligned} \mathcal{F} &= D_{\mathrm{KL}}\left[q(s|\pi) \| p^*(s,\pi)\right] \\ &= \ln p^*(\pi) + D_{\mathrm{KL}}\left[q(s|\pi) \| p^*(s|\pi)\right] \\ &\approx \ln p^*(\pi) \end{aligned} \tag{35}$$

$$\implies f_\alpha(\pi) = (\Gamma - Y)\nabla_\alpha \mathcal{F}(\pi)$$

This demonstrates that the dynamics of systems (such as that in equation 34) which maintain a Markov Blanket at a NESS (despite external dissipative perturbations) can be interpreted as performing variational Bayesian inference to infer and optimise the posterior distribution over external states, parameterized by internal states [1]. This is the key statement of the FEP.

Maximum-a-Posteriori Modes

A maximum-a-posteriori distribution is the posterior distribution of the mode of a distribution. We can derive the flow of the external mode with respect to the flow of the internal mode using the expressions for the modes of the internal and external states, given the blanket states, $\boldsymbol{s}(b) = \psi(\boldsymbol{\mu}(b))$. Given the chain rule, this is[1]:

$$f_{\boldsymbol{s}}(b) = \frac{\partial \psi(\boldsymbol{\mu}(b))}{\partial \boldsymbol{\mu}(b)} f_{\boldsymbol{\mu}}(b) \tag{36}$$

and vice versa (assuming invertability):

$$f_{\boldsymbol{\mu}}(b) = \frac{\partial \psi(\boldsymbol{\mu}(b))^{-1}}{\partial \boldsymbol{\mu}(b)} f_{\boldsymbol{s}}(b) \tag{37}$$

We define the *synchronisation manifold* as the mapping between the two densities, of the external mode and the internal mode [1]. This is defined because of the above mapping, ψ and despite the separation of the Markov blanket. The synchronisation manifold is defined as:

$$\frac{\partial \ln p(\boldsymbol{s}(b)|b)}{\partial \mu} = \frac{\partial \ln p(\boldsymbol{s}(b)|b)}{\partial \boldsymbol{s}(b)} \frac{\partial \psi(\boldsymbol{\mu}(b))}{\partial \mu} \tag{38}$$

We combine equations 36, 37 and 38 along with the marginal flow lemma for the external mode, $f_{\boldsymbol{s}}(b) = (\Gamma_s - Y_s)\nabla_s \ln p(s(b)|b)$ to express the flow of the internal state mode [1]:

$$\begin{aligned} f_{\boldsymbol{\mu}}(b) &= \frac{\partial \psi(\boldsymbol{\mu}(b))}{\partial \boldsymbol{\mu}(b)}^{-1} \frac{d\boldsymbol{s}(b)}{dt} \\ &= \frac{\partial \psi(\boldsymbol{\mu}(b))^{-1}}{\partial \boldsymbol{\mu}(b)} (\Gamma_s - Y_s)\nabla_s \ln p(\boldsymbol{s}(b)|b) \\ &= \frac{\partial \psi(\boldsymbol{\mu}(b))^{-1}}{\partial \boldsymbol{\mu}(b)} \quad (\Gamma_s - Y_s) \frac{\partial \psi(\boldsymbol{\mu}(b))^{-1}}{\partial \boldsymbol{\mu}(b)} \quad \frac{\partial \psi(\boldsymbol{\mu}(b))}{\partial \boldsymbol{\mu}(b)} \nabla_s \ln p(\boldsymbol{s}(b)|b) \\ &= (\Gamma_\psi - Y_\psi)\nabla_\mu \ln p(\psi(\boldsymbol{\mu}(b))) \end{aligned} \tag{39}$$

,where $(\Gamma_\psi - Y_\psi) = \frac{\partial \psi(\boldsymbol{\mu}(b))}{\partial \boldsymbol{u}(b)}^{-1} (\Gamma_s - Y_s) \frac{\partial \psi(\boldsymbol{\mu}(b))}{\partial \boldsymbol{u}(b)}^{-1}$

Unpacking the above equation, we see that we have expressed the flow of the mode of the internal states as a gradient descent on the NESS density of the mode of the external states [1]. Furthermore, the mode of the external states is expressed as a function of the mode of internal states, given the blanket states. Additionally we have expressed the above in the form of the Helmholtz decomposition where Γ_ψ and Y_ψ are the curl-free and divergence-free components with respect to the internal states, modulated by the inverse of the mapping ψ. We have therefore expressed a transformation. We have moved from the coordinates of the flow of the external states to the coordinates of the flow of the mode of the external states as a function of the internal states [1]. To tie this to VFE we introduce the variational density of the mode of the external states, given the blanket states and parameterized by the mode of the internal states. We apply the Central Limit Theorem and Laplace approximation by assuming that the variational density is a Normal Distribution [1, 10]:

$$\begin{aligned} &q(\boldsymbol{s}|b;\boldsymbol{\mu}) = \mathcal{N}(\boldsymbol{s};\boldsymbol{\mu},\Sigma(\boldsymbol{\mu})) \\ &\text{where } \Sigma(\boldsymbol{\mu}) = \frac{\partial^2 \psi(\boldsymbol{\mu})^{-1}}{\partial \psi^2} \end{aligned} \tag{40}$$

We substitute this Gaussian variational density into the expression for VFE and we drop the constants (in terms of $\boldsymbol{\mu}$) to obtain the following, as found in [1]:

$$\begin{aligned} &\mathcal{F} = \ln p(\boldsymbol{\mu}, b) + \frac{1}{2}\operatorname{tr}(\Sigma(\boldsymbol{\mu}))\frac{\partial^2 \psi(\boldsymbol{\mu})}{\partial \psi^2} + \ln|\Sigma(\boldsymbol{\mu})| \\ &\implies \frac{\partial \mathcal{F}}{\partial \mu} = \frac{\partial \ln p(\boldsymbol{\mu}, b)}{\partial \mu} \end{aligned} \tag{41}$$

Combining this with the flow in 39 we obtain [1]:

$$f_{\boldsymbol{\mu}}(b) = (\Gamma_\psi - Y_\psi)\,\nabla_\mu \mathcal{F} \tag{42}$$

This demonstrates once again the crucial insight of the FEP. This time we have seen that with our Gaussian variational density, we have that the mode of the internal states infers the mode of the external states via the dynamics of gradient descent on the VFE. We are therefore equipped with the interpretation of the flow of the internal states as performing variational inference on the external states.

Active States and Active Inference

The story thus far can be seen as primarily concerning *perceptual inference.* This refers to the inference about the external states performed by the flow of the internal states. This inference occurred conditional upon the information from the blanket states and, specifically, the sensory states. The internal states were drawing inference as to the cause of changes to the sensory states. Now, we turn to the other blanket states, the active states. We apply the marginal flow lemma on active states, a, and similarly to 35 we write this as an approximate gradient descent of the VFE using the Helmholtz decomposition [1]:

$$\begin{aligned} &f_a(\pi) \approx (\Gamma_{aa} - Y_{aa})\,\nabla_a \mathcal{F}(\pi) \\ &\pi = \{o, \alpha\} \end{aligned} \tag{43}$$

If one turns to the inaccuracy and complexity decomposition of VFE below we see that the only term that depends on active states is the inaccuracy term [1]:

$$\begin{aligned}\mathcal{F}(\pi) &= D_{\mathrm{KL}}\left[q(s \mid \pi ; \boldsymbol{\mu}) \| p^*(s, \pi)\right] \\ &= -\underbrace{\mathbb{E}_{q(\pi \mid \boldsymbol{\mu})}\left[-\ln p^*(\pi \mid s)\right]}_{\text{Inaccuracy}}+\underbrace{D_{\mathrm{KL}}\left[q(s \mid \pi ; \boldsymbol{\mu}) \| p^*(s)\right]}_{\text{Complexity}}\end{aligned} \tag{44}$$

Hence, the inaccuracy term can be substituted into the marginal flow lemma for active states [1]:

$$f_a(x) \approx \left(\Gamma_{aa}-Y_{aa}\right) \nabla_a\left(-\mathbb{E}_{q(s \mid \boldsymbol{\mu})}\left[-\ln p^*(\pi \mid s)\right]\right) \tag{45}$$

We see here that the flow of the active states, at the NESS, maximises accuracy. Since active states can only influence external states, to maximise the accuracy about the beliefs of external states, the active states influence the external states to bring them in line with the internal and blanket states' beliefs about the external states [1]. This process is called *active inference*. To describe the active states of the system when the system is moving toward self-organisation we require the *expected free energy* (EFE) which forms an upper bound on surprise of the system throughout the process of self-organisation. It is this EFE term which brings about the directed, intrinsic exploratory behaviour of the active inference agent and, which we later interrogate.

Let $p\left(s_t, \mu_t, o_t, a_t \mid s_0, \mu_0, o_0, a_0\right)$ denote the probability density over the variables of the system at time t and given the set of variables at $t=0$. Let $\pi_0=\{\mu_0, o_0, a_0\}$. (We average over the initial external state, s_0, to simplify.) The expected free energy (EFE) is then defined as [1]:

$$\begin{aligned}G(\pi) &= \mathbb{E}_{p\left(s_t, \pi_t \mid \pi_0\right)}\left[\ln p\left(s_t \mid \pi_t, \pi_0\right)-\ln p^*(s, \pi)\right] \\ &= \underbrace{\mathbb{E}_{p\left(s_t, \pi_t \mid \pi_0\right)}\left[-\ln p^*(\pi \mid s)\right]}_{\text{Ambiguity}}+\underbrace{D_{\mathrm{KL}}\left[p\left(s_t \mid \pi_t, \pi_0\right) \| p^*(s)\right]}_{\text{Risk}}\end{aligned} \tag{46}$$

The expected free energy estimates the distance between the posterior at time t and the NESS density, with an expectation being taken using the predicted density at time t. We also note that the ambiguity and risk decomposition affords us the interpretation that minimising EFE minimises ambiguity (uncertainty) and, (risk) minimises the divergence between the current state density and the NESS density. The minimisation of this objective therefore encourages *self-organisation* toward the NESS. In fact, at convergence (to the NESS), the EFE forms an upper bound on surprise [1, 10]. The equation, below, reveals that when the system reaches the NESS (and self-organisation) that the EFE is equal to the surprise (when the KL divergence is zero) [1]:

$$\begin{aligned}& D_{\mathrm{KL}}\left[p\left(s_t, \pi_t \mid \pi_0\right) \| p^*(s, \pi)\right] \geq 0 \\ & \Longrightarrow \mathcal{G}\left(\pi_t\right)+\mathbb{E}_{p\left(s_t, \pi_t \mid \pi_t\right)} \ln p\left(\pi_t \mid \pi_0\right)] \geq 0 \\ & \Longrightarrow \mathcal{G}\left(\pi_t\right) \geq-\mathbb{E}_{p\left(s_t, \pi_t \mid \pi_t\right)}\left[\ln p\left(\pi_t \mid \pi_0\right)\right]\end{aligned} \tag{47}$$

Furthermore, this reveals that the EFE provides a Lyapunov value function for the policy of the system, near the NESS [11]. EFE at time t measures the distance between the current posterior at time t and the NESS density. Self-organising dynamics under a Markov blanket can, therefore, be interpreted as minimising EFE over time to maintain their NESS densities (by the FEP). Moreover, the system changes the active states to accomplish the NESS.

Conversely, instead of using the FEP to provide an interpretation of the dynamics of the NESS (under a Markov blanket), we can instead construct a decision-making agent, that aims to achieve some specific goals [1]. In this case, the agent's goals can correspond to the NESS density p^*,[6] to which it self-organises. Moreover, the agent can be defined such that it abides by the statistical independencies of the Markov blanket. Hence, by (applying) the FEP this agent can be interpreted as a free energy-minimising agent. This insight guides our research into active inference agents and, therefore, helps us define decision-making agents that can make use of inference to perform actions to achieve goals. The agent determines its actions via a minimisation of EFE (perhaps via gradient descent or model-based planning). We review the active inference agent in the next section, where we draw specific attention to the exploratory drive that is embedded in the EFE objective.[7]

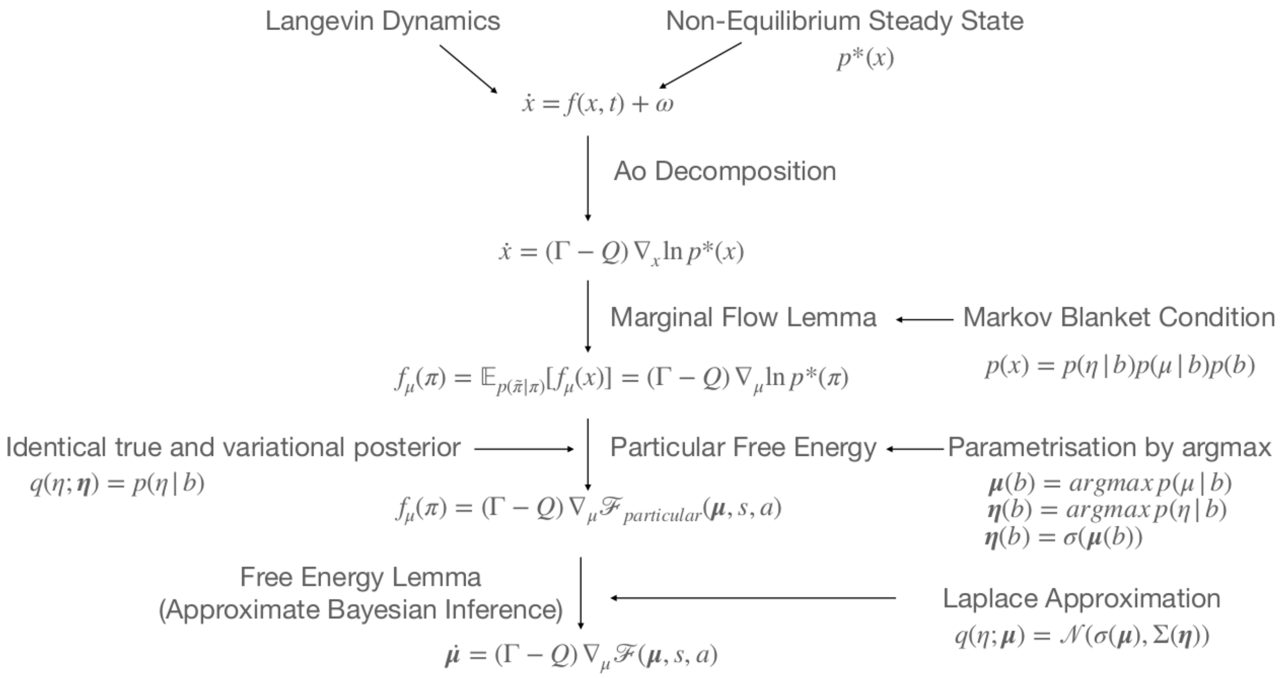
Langevin Dynamics
Non-Equilibrium Steady State
$p^*(x)$
$\dot{x} = f(x,t) + \omega$
Ao Decomposition
$\dot{x} = (\Gamma - Q)\nabla_x \ln p^*(x)$
Marginal Flow Lemma
Markov Blanket Condition
$p(x) = p(\eta \mid b)p(\mu \mid b)p(b)$
$f_\mu(\pi) = \mathbb{E}_{p(\tilde{\pi}|\pi)}[f_\mu(x)] = (\Gamma - Q)\nabla_\mu \ln p^*(\pi)$
Identical true and variational posterior
$q(\eta;\boldsymbol{\eta}) = p(\eta \mid b)$
Particular Free Energy
Parametrisation by argmax
$\boldsymbol{\mu}(b) = argmax\, p(\mu \mid b)$
$\boldsymbol{\eta}(b) = argmax\, p(\eta \mid b)$
$\boldsymbol{\eta}(b) = \sigma(\boldsymbol{\mu}(b))$
$f_\mu(\pi) = (\Gamma - Q)\nabla_\mu \mathscr{F}_{particular}(\boldsymbol{\mu}, s, a)$
Free Energy Lemma
(Approximate Bayesian Inference)
Laplace Approximation
$q(\eta;\boldsymbol{\mu}) = \mathscr{N}(\sigma(\boldsymbol{\mu}), \Sigma(\boldsymbol{\eta}))$
$\dot{\boldsymbol{\mu}} = (\Gamma - Q)\nabla_\mu \mathscr{F}(\boldsymbol{\mu}, s, a)$

Figure 4: A summary of the FEP formulation as it appears in [1]. Starting from the premise of the existence of a Langevin SDE which has a NESS we apply the Helmholtz (Ao) decomposition which interprets the dynamics as a gradient descent on surprise. We presuppose the existence of a Markov blanket and then use the marginal flow lemma to express the flows of subsets of states in terms of their own particular marginal flows. Then, interpreting the internal states as parameterizing a variational density over the external states and, applying the free energy lemma, we interpret the marginal flow of internal states as inferring the external states via the minimisation of VFE (using the Laplace approximation). The main text here omits the assumption of equality between the true and variational posterior which expresses the marginal flow lemma in terms of particular free energy as this is not necessary for the derivation. For an extended discussion of the FEP formulation, its assumptions and, its philosophical status see [1] and [10].

4 Exploration and Active Inference Agents

4.1 A Lightning Introduction to Active Inference Agents

The active inference agent, as derived from the FEP, provides a normative framework for decision-making in an uncertain environment to achieve its goals [7]. In our context, we therefore, focus on an RL agent, which is specifically defined to incorporate a generative model $p(o, s, a)$ of its environment and which selects

[6]... which are often represented as a Boltzmann distribution of environmental rewards such that $p^*(s, \pi) \propto \exp(-r(s))$ [1].

[7]... and, to the question of how this exploratory drive fairs in benchmark RL environments.

actions so as to minimise a variational free energy (VFE) objective. The VFE objective furnishes the agent with an optimal policy which includes an exploratory temperature hyperparameter, similarly to Boltzmann exploration. Furthermore, a second objective function called expected free energy (EFE), through which the agent selects actions, also furnishes the agent with a directed intrinsic exploratory drive which is not necessarily present for an RL agent. This directed intrinsic exploratory drive motivates the study of active inference agents in this paper. This section contains a refresher as to the decision-making framework of active inference agents (which was discussed in further detail in our previous work [9]).

The active inference agent is defined in a partially observable Markov decision process (POMDP) environment [1, 7, 38]. A POMDP can be defined as the tuple $(\mathcal{S}, \mathcal{A}, P, \mathcal{R}, \mathcal{O}, P_o)$ of states $s \in \mathcal{S}$, actions $a \in \mathcal{A}$, a state transition probability function $P = p(s_t|s_{t-1}, a_{t-1})$, rewards $r \in \mathcal{R}$, observations $o \in \mathcal{O}$, and an observation probability function $P_o = p(o|s)$. As a reminder to the reader this expresses a Markov decision process in which the agent cannot, necessarily, directly see the state s of the environment but instead receives observations which are generated by the state of the environment via $p(o|s)$.

Within the POMDP environment, the agent unites the processes of perception, action and learning via its minimisation of VFE (and therefore, the preservation of its non-equilibrium steady-state, under a Markov blanket). First, perception presents inference as to the agent's state of the environment. Perceptual inference can be viewed as the following inferential minimisation [1]:

$$\underset{q(s|o)}{\operatorname{argmin}}\, \mathcal{F}(o) \tag{48}$$

The agent's model learning involves the minimisation of VFE to obtain an optimal posterior for model parameters θ [1]:

$$\underset{q(\theta|s,o)}{\operatorname{argmin}}\, \mathcal{F}(o) \tag{49}$$

Putting these together, the joint posterior for states and parameters is $q(s, \theta|o) = q(s|o)q(\theta|s, o)$. Furthermore, minimising VFE can be decomposed for interpretation. The first term gives us that, minimising VFE increases model accuracy in terms of the observations received. However, the second term acts to regularise this procedure by keeping the posterior close to the prior (ensuring a Bayesian inference process).

$$\mathcal{F}(o, \theta) = -\underbrace{\mathbb{E}_{q(s|o;\theta)}[\ln p(o|s)]}_{\text{Accuracy}} + \underbrace{D_{\mathrm{KL}}[q(s|o;\theta)\|p(s)]}_{\text{Complexity}} \tag{50}$$

The agent's third inferential scheme, action selection inverts the usual question asked for RL agents. Instead of asking which actions to select to best achieve its goals, the agent assumes that it will achieve its goals and given this assumption it asks which actions it will most likely pursue. This involves defining the agent's biased, a-priori preferences. This is often defined in terms of a distribution over observations such that the agent prefers greater amounts of reward (although this can be defined more generally) [8]:[8]

$$\tilde{p}\left(o_{1:T}\right) = \prod_{t=1}^{T} \sigma\left(r\left(o_t\right)\right) \tag{51}$$

This preference distribution forms a component of the NESS density, toward which the agent self-organises. Given that the active inference agent meets the requirements to apply the FEP and, that it is therefore a

[8] σ represents a softmax distribution.

VFE-minimising agent that self-organises toward its NESS density, then the agent will necessarily minimise VFE in the future. Hence expected free energy (EFE) is defined as [1, 38]:

$$\begin{aligned}\mathcal{G}_{\pi}(o,s) &= \mathrm{E}_{q(o,s|\pi)}[\ln q(s|\pi) - \ln \tilde{p}(o,s|\pi)] \\ &= \underbrace{\mathrm{D}_{\mathrm{KL}}(q(o|\pi)\|\tilde{p}(o))}_{\text{Risk}} + \underbrace{\mathrm{E}_{q(s,o|\pi)}[\ln p(o|s)]}_{\text{Ambiguity}} \\ &= -\underbrace{\mathbb{E}_{q(o,s|\pi)}[\ln \tilde{p}(o)]}_{\text{Extrinsic Value}} - \underbrace{\mathbb{E}_{q(o|\pi)} D_{\mathrm{KL}}[q(s|o,\pi)\|q(s|\pi)]}_{\text{Intrinsic Value}} + \underbrace{\mathbb{E}_{q(o|\pi)} D_{\mathrm{KL}}[q(s|o,\pi)\|p(s|o,\pi)]}_{\text{Posterior Divergence}}\end{aligned} \tag{52}$$

The agent minimises this quantity which affords us the interpretation that the agent maximises extrinsic value, and therefore its preferences (and its amount of reward). Secondly, the agent maximises its intrinsic value which drives the expected variational posterior away from its variational prior, thus presenting an information gain term and, crucially, an exploratory drive. The exploratory drive is directed such that when the agent acts, it maximally updates its beliefs about the world in order to gain information which resolves uncertainty. Appropriately, the exploratory drive is incorporated into the mechanism of the agent's action selection. The action selection of the agent occurs via the following softmax distribution. The agent's (stochastic) policy is proportional to the exponentiated negative EFE value [1, 38]:

$$q(a|s) = \sigma(-\beta\mathcal{G}(o,s)) \tag{53}$$

The β hyperparameter is often referred to as precision. In a sense, this represents the given accuracy of the EFE estimate. This offers a second means of exploration whereby a smaller value of β will yield a less peaked softmax distribution for action selection, yielding greater amounts of exploration. The exploration here is similar to Boltzmann exploration.

Lastly, it is important to mention that the evaluation of EFE is computed via the ergodic path integral below [1]:

$$\mathcal{G}_{\pi}(o,s) = \sum_{t} \mathcal{G}_{\pi}(o_t, s_t) \tag{54}$$

4.2 Implementations of Deep Active Inference Agents

In our previous paper on scaling active inference methods using deep reinforcement learning [9], three active inference algorithms were reviewed [8, 13, 14]. These addressed the issue of scaling active inference methods to larger state-action spaces. In that work it was mentioned that 'proof-of-concept' implementations were run. In this section we review these implementations and study their performance on benchmark tasks in OpenAI Gym [15]. Crucially, as both RL and active inference address the problem of agent-based decision-making to achieve some goal, in uncertain environments, these algorithms take the insights of both RL and active inference and combine them. While the active inference agent benefits from the successes of deep RL, it is not clear how the RL agent benefits. One key potential benefit is the directed, intrinsic exploratory motive which is especially appropriate for sparse reward environments. This forms the focus of our discussion here, where we review the proof-of-concept implementations of deep active inference agents (in benchmark RL environments) and, compare these to benchmark RL agents.

4.2.1 Variational Policy Gradients: A Model-Free RL Approach to Active Inference

The first active inference agent algorithm implements an actor-critic, model-free architecture [8]. This algorithm is called Deep Active Inference. Here, a stochastic policy network $q(a|s)$ is trained as the actor. This is trained against the critic which is an EFE value network. The expression for the estimate of the EFE appears below as $\hat{\mathcal{G}}$. This appears as a recursive (Bellman) statement that incorporates a second frozen value network, G_ϕ, similarly to DQN [1, 8]:

$$\hat{\mathcal{G}}\left(o_{t:T}s_{t:T}\right) = \mathbb{E}_{q(o_t,s_t)}\left[r\left(o_t\right)\right] - \mathbb{E}_{q(o_t)}D_{\mathrm{KL}}\left[q\left(s_t|o_t\right)\|q\left(s_t\right)\right] + \mathbb{E}_{q(s_{t+1},o_{t+1}|s_t,a_t)q(a_t|s_t)}\left[\mathcal{G}_\phi\left(o_{t+1:T},s_{t+1:T}\right)\right] \tag{55}$$

This is trained on a squared error loss function:

$$|\hat{G}\left(o_t,s_t\right) - G_\phi\left(o_t,s_t\right)|^2 \tag{56}$$

The above value function is placed into the policy prior for the agent (where σ represents a softmax function and γ is a Boltzmann exploration hyperparameter):

$$p\left(a_t|s_t\right) = \sigma\left(\gamma\mathcal{G}_{a_{t:T}}\left(s_{t:T},o_{t:T}\right)\right) \tag{57}$$

The agent's policy $q(a|s)$ is trained against this action prior. Instead of being trained on cumulative reward, the agent is trained on the following VFE objective function [8]:

$$\begin{aligned}\mathcal{F}\left(o_{1:T}\right) &= \sum_{t=0}^{T}\mathcal{F}_t\left(o_t\right) = \sum_{t=0}^{T}D_{\mathrm{KL}}\left[q\left(a_t,s_t|o_t\right)\|p\left(s_t,a_t,o_t\right)\right] \\ &= \sum_{t=0}^{T} - \underbrace{\mathbb{E}_{q(a_t,s_t|o_t)}\left[\ln p\left(o_t|s_t\right)\right]}_{\text{Accuracy}} + \underbrace{\mathbb{E}_{q(a_t|s_t)}D_{\mathrm{KL}}\left[q\left(s_t|o_t\right)\|p\left(s_t|s_{t-1},a_{t-1}\right)\right]}_{\text{Complexity}} \\ &+ \underbrace{D_{\mathrm{KL}}\left[q\left(a_t|s_t\right)\|p\left(a_t|s_t\right)\right]}_{\text{Action Divergence}}\end{aligned} \tag{58}$$

This decomposition reveals the accuracy and complexity terms which were discussed in equation 50. The third term, action divergence, demonstrates that as the agent minimises VFE, it minimises the KL divergence between the policy posterior $q(a|s)$ and the policy prior. This embeds the exploratory behaviour of EFE, seen in equation 52 in the agent's policy. This also embeds the exploitatory behaviour via the EFE's extrinsic value term. This term attempts to maximise reward due to reward being encoded in the agent's preference distribution in equation 51 [8].

The other distributions appearing in equation 58 are the transition dynamics, $p(s_t|s_{t-1},a_{t-1})$, the posterior for states $q(s_t|o_t)$ and, the likelihood function $p(o_t|s_t)$. These are obtained from neural networks, trained on the agent's sampled experience from the environment (which is collected into a random experience replay buffer).

Implementation Results

Next we turn to the empirical results of the Deep Active Inference paper [8]. The authors compare the active inference agent to three other RL agents. They implement a DQN agent that uses Boltzmann exploration, they implement a vanilla policy gradient agent and, they implement an actor-critic agent. The actor-critic agent has a critic that is trained using Q-learning and an actor that is trained on a policy gradient objective that uses the critic's Q-values (similarly to equation 6). The form of the actor-critic policy used is a softmax function, thus implementing Boltzmann exploration. The policy network and value network of the actor-critic were trained using the same architecture and hyperparameters as the active inference agent for a fairer comparison.

The agents were compared on three environments from OpenAI gym [15]. The three environments used are called Cartpole, Acrobat and Lunar Lander. Cartpole involves the agent balancing a pole on a cart. The agent can move the cart left or right. The state space is continuous and includes the position and velocity of the cart and, the position and angular velocity of the pole. A reward of +1 is given for every time step (up to 200) that the episode does not end [1]. An episode ends if the cart moves off of the screen (corresponding to 2.4 distance units from the centre). An episode can also end if the angle of the pole is more than 15 degrees off of vertical.

The Acrobat environment possesses a double-jointed pendulum. The goal of the agent is to swing the pendulum up to a given height [15]. A six dimensional state space encapsulates the angles and velocities of the joints. The agent has a three dimensional action space corresponding to the forces it can apply on the pendulum [8]. The reward is 0 if the arm of the Acrobat is above the horizontal and -1 otherwise. This environment poses a challenging problem for exploration as random actions are unlikely to succeed.

Lastly, the Lunar Lander environment demands that the agent land a spacecraft between two markers on the moon. This has an eight dimensional state-space and a four dimensional action space [1]. The actions correspond to firing the agent's engine left, right or upwards and, to extending docking legs. The reward given is +100 for docking in the target area, +10 for each standing leg and, -0.3 for every time-step that the engine is firing [1].

The results of the implementations in the three environments appear below (figures 5, 6 and, 7). For each environment, each agent was run for 15000 episodes.

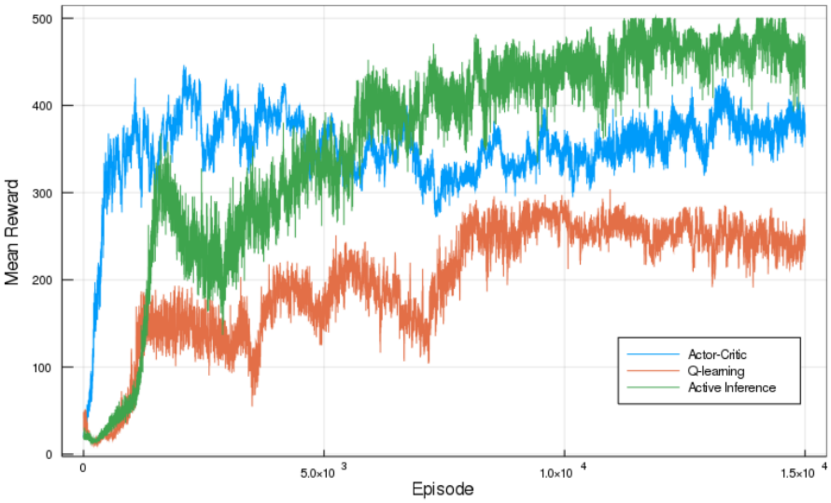

Figure 5: Cartpole environment [15] with a comparison of mean rewards (from 20 random seeds) achieved by the deep active inference agent as compared to the DQN and actor-critic agents. This image is from [1] while the study was done in [8]. The maximum possible reward here is 500 hence the active inference agent achieves a stronger performance over more episodes, while the actor-critic agent has a strong initial learning curve. Both the actor-critic and deep active inference agent outperform the Q-network.

Next the authors of [8] generate an ablation study to try understand how the components of the active inference agent affect its performance performance. The full active inference agent is compared to two ablated versions. The first version removes the entropy term which arises from the action divergence term of the VFE objective function in equation 58. The action divergence can be decomposed into an entropy

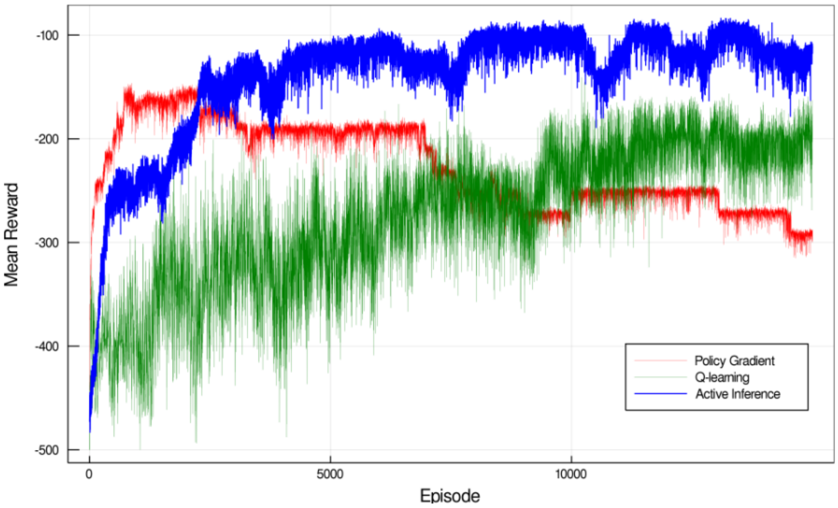

Figure 6: Acrobat environment [15] comparison of mean rewards (of 20 random seeds) achieved by the deep active inference agent as compared to the DQN and a policy gradient agent. This image is from [1] while the study was done in [8]. Here the maximum possible reward is 0 so none of the agents achieve optimality. The active inference agent in this instance outperforms the other two over many episodes of training.

term and an energy term below [8]:

$$D_{KL}[q(a|s)\|p(a|s)] = \int q(a|s)\ln p(a|s) + \mathcal{H}(q(a|s)) \tag{59}$$

While the above entropy term is removed from one ablated version of the active inference agent, the whole intrinsic value term is removed from the second ablated version of the active inference agent. This intrinsic value term appears in the EFE objective in equation 52. The results of the ablation study appear in the plots below [1] (figures 8 and 9).

The results of [8] highlight that the active inference agent can achieve similar performances to contemporary deep RL algorithms. It even outperforms these implementations in the Cartpole and Acrobat environments, while in the Lunar Lander environments it is beaten by a vanilla policy gradient agent. As the active inference agent achieves better performance than the actor-critic and DQN agents on Lunar Lander the authors believe that the weaker performance of the value-based methods is because of the bias and inaccuracy of the value functions (and the EFE estimates for the active inference agent). As the policy gradient agent uses unbiased Monte Carlo samples of the reward, instead of a bootstrapped value estimate, it is not affected by this bias, perhaps yielding a stronger performance in this particular task.

The ablation study highlights that, in the Cartpole environment, the entropy term provides a significant performance boost to the agent. The removal of the intrinsic value term leaves the active inference agent having a similar performance to DQN and actor-critic agents (figure 9). The reasons, hypothesised by the authors, for the lack of performance gain of the intrinsic value term are that: (i) the reward environment was dense and could be learnt by standard contemporary RL agents which use random exploration instead of an intrinsic information gain method and, (ii) the intrinsic value term (between the prior and posterior for states) is effectively a measure of the predictive success of the transition model [1]. [8] found that the transition model trained very quickly and converged rapidly as compared to the policy or value networks. Therefore in this MDP environment, the intrinsic exploration term had a negligible effect. Hence, in the next implementation (one inspired by model-based RL), which is a sparse reward environment implementation of an active inference agent, we see the importance of the information gain term.

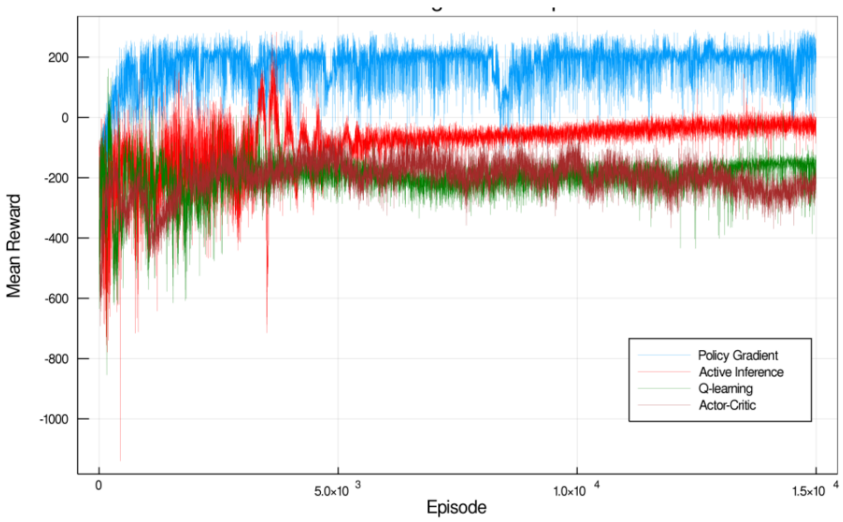

Figure 7: Lunar Lander environment [15] comparison of mean rewards (of 20 random seeds) achieved by the deep active inference agent as compared to the DQN, vanilla policy gradient and actor-critic agents. This image is from [1] while the study was done in [8]. In this environment a maximum reward of 200 is optimal. In this instance the policy gradient (blue) outperforms the active inference agent (red), although, the active inference agent outperforms the DQN (green) and actor-critic (maroon) agents.

4.2.2 A Model-Based RL Approach to Active Inference

The second algorithm discussed in the corresponding honours paper [9] was called the *free energy of the expected future* (FEEF) algorithm [13]. This was inspired by a model-based RL approach. While in that work the algorithm was discussed in detail, in this work we focus our attention on the implementation and empirical results. To briefly revise the algorithm we take note of the following.

The algorithm uses a new free energy objective function called FEEF [13]. From this, the optimal policy is derived to be the softmax distribution, which takes as its argument, a second free energy function, $\mathcal{F}_\pi$. A derivation of the optimal policy in [13] yields that a minimised FEEF objective function, $\tilde{\mathcal{F}}$, (with KL divergence 0) gives:

$$\tilde{\mathcal{F}} = 0 \Rightarrow D_{\mathrm{KL}}\left(q(\pi) \| \left(e^{-\tilde{\mathcal{F}}_\pi}\right)\right) = 0 \tag{60}$$

where (given the policy, π):

$$\tilde{\mathcal{F}}_\pi = D_{\mathrm{KL}}\left[q\left(o_t, s_t, \theta \mid \pi\right) \| \tilde{p}\left(o_t, s_t, \theta\right)\right] \tag{61}$$

Therefore the optimal policy, that minimises the FEEF, takes the softmax form $q(\pi) = \sigma\left(-\tilde{\mathcal{F}}_\pi\right)$. In this form policies which minimise $\tilde{\mathcal{F}}$ are more likely. The interpretation of $\tilde{\mathcal{F}}$ follows from the following decomposition [13]:

$$\begin{aligned}
\tilde{\mathcal{F}}_\pi &= D_{\mathrm{KL}}\left[q\left(o_t, s_t, \theta \mid \pi\right) \| \tilde{p}\left(o_t, s_t, \theta\right)\right] \\
&\approx \underbrace{\mathbb{E}_{q(s_t \mid \pi)} D_{\mathrm{KL}}\left[q\left(o_t \mid s_t\right) \| \tilde{p}\left(o_t\right)\right]}_{\text{Extrinsic Value}} - \underbrace{\mathbb{E}_{q(o_t;\theta)} D_{\mathrm{KL}}\left[q\left(s_t \mid o_t, \theta\right) \| q\left(s_t\right)\right]}_{\text{State Information Gain}} \\
&- \underbrace{D_{\mathrm{KL}}\left[q\left(\theta \mid s_t\right) \| q(\theta)\right]}_{\text{Parameter Information Gain}}
\end{aligned} \tag{62}$$

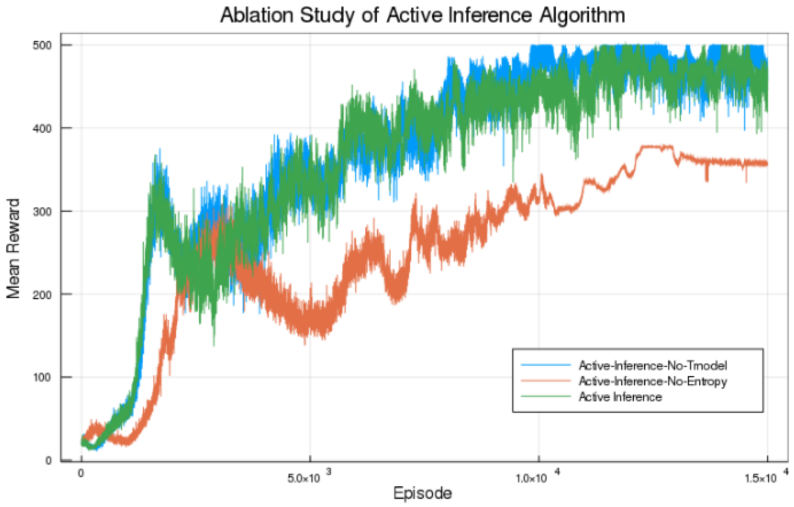

Figure 8: Ablation study of the active inference agent from [8]. The full active inference agent (green) is compared to an ablated version with no intrinsic value term (blue). These are compared to a second ablated version with no entropy term in the VFE objective function (orange). This image is drawn from [1]. The performance was compared in the Cartpole environment, over 15000 episodes.

The extrinsic value term represents the divergence between the prior preferences of the agent and its expected observations (under the variational belief distribution for observations). This encodes the reward seeking behaviour of the agent (if the agent's preferences are, once again, encoded as a preference for greater amounts of reward). The second and third terms represent the information-seeking behaviour of the agent (for its state posterior and its parameter posterior). In this implementation the model parameters are treated in a Bayesian way with a Bayesian prior $p(\theta)$ being used [13].

While the transition model, $p(s_t|s_{t-1}, a_{t-1})$, and the likelihood function, $p(o|s)$, are approximated by deep neural networks, the variational posteriors for states and observations are evaluated via a model-based planning approach for belief updating [13]:

$$\begin{aligned} q\left(o_t \mid s_t, \theta, \pi\right) &= \mathbb{E}_{q(s_t|\theta,\pi)}\left[p\left(o_t \mid s_t\right)\right] \\ q\left(s_t \mid s_{t-1}, \theta, \pi\right) &= \mathbb{E}_{q(s_{t-1}|\theta,\pi)}\left[p\left(s_t \mid s_{t-1}, \theta, \pi\right)\right] \end{aligned} \tag{63}$$

Using these distributions a planning horizon, H, is used for the evaluation of the $\tilde{\mathcal{F}}_\pi$ objective, such that the ergodicity assumption is applied $\tilde{\mathcal{F}}_\pi = \sum_t^{t+H} \tilde{\mathcal{F}}_{\pi_t}$. Then for a sampled trajectory of actions from an initial posterior $q(\pi)$ and a sampled selection of states for these actions, the objective function is calculated and the policy is updated via a softmax expression for the $\tilde{\mathcal{F}}_\pi$ value functions.

Implementation Results

While the previous (active inference agent) algorithm looked at dense reward environments, this algorithm studies four environments with the following three characteristics. These are, that the reward signal should be sparse, that the reward signal should be well-shaped or, that there should be no reward signal at all. The first environment used was the mountain-car environment from Open AI Gym [15]. This met the sparse reward characteristic.

In mountain-car the agent is required to steer a car up a one dimensional hill toward a goal. The environment presents a challenge since the agent must first learn to gather momentum, by driving up a hill opposite to the target hill. This environment presents a reward of +1 when the agent gets to the top of its target hill and 0 otherwise [13]. The agent has a one dimensional continuous action space, corresponding to the direction and power of its movement (within a limit). It has a two dimensional state space corresponding

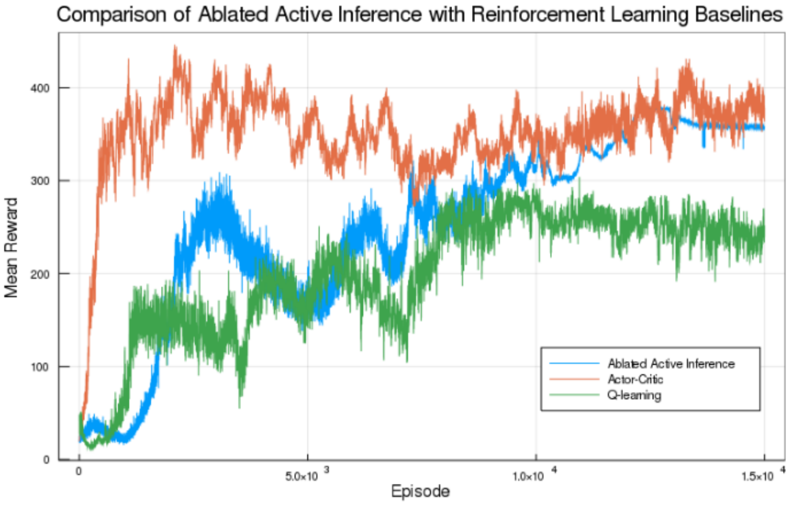

Figure 9: Ablation study of the active inference agent from [8]. In this image a fully ablated active inference agent, without both the intrinsic value term and the entropy term (blue), is compared to an actor-critic agent (orange) and a DQN agent (green). This image is drawn from [1]. The performance was compared in the Cartpole environment, over 15000 episodes.

to its position and velocity [15].

The second (of the four) environments studied, in [13], is the cup catch environment. This also presents a sparse reward environment as the agent receives a reward of $+1$ when it achieves its goal of catching a ball in a cup. The agent must control a cup, in which it must catch a ball, attached to the bottom of the cup. The cup catch environment has an eight dimensional state space with a two dimensional action space [13].

Thirdly, the half-cheetah environment from OpenAI gym [15], has the agent manifest as a cheetah. The agent must take control of the limbs of the cheetah which either runs on a plane or, in a second task, tries to flip itself over. The goal of the agent is to maximise the forward velocity of the cheetah, by learning to run or for the case of the flipping task, maximise the angular velocity of the cheetah [13]. The half-cheetah environment offers a chance to test our active inference agent on a well-shaped reward signal. The reward signal for the case of the running task is $v-0.1||a||^2$ and for the flipping task this is $\epsilon-0.1||a||^2$, where v and ϵ are the planar velocity and angular velocity respectively. The environment has the following dimensionality, specified by [13], $\left(\mathcal{S} \subseteq \mathbb{R}^{17}\mathcal{A} \subseteq \mathbb{R}^{6}\right)$.

Lastly, the ant-maze environment, used in [13], offers an environment where no reward is given. This was implemented to test the exploratory capability of the agent. The agent manifests itself as an ant and it explores as much of a maze as possible, whilst receiving no reward. This has the following dimensionality of the state-action space: $\left(\mathcal{S} \subseteq \mathbb{R}^{29}\mathcal{A} \subseteq \mathbb{R}^{8}\right)$ [13].

The implementations that were run in sparse reward environments tested the agent against two benchmark agents [13]. The first is a reward-seeking agent which only selects actions based on the extrinsic value term of equation 62. The second benchmark agent, for the sparse reward tasks, is a variance agent which seeks out highly uncertain state transitions by acting to maximise output variance of the state transition model [13]. The variance agent also holds the extrinsic value term to enable reward-seeking behaviour. Figure 10 **A** demonstrates that the FEEF agent is able to solve the mountain-car problem in only a single episode. This contrasts to the random agent and, the reward-seeking agent which struggle in this sparse reward environment. The strong performance of the FEEF agent is due to it being able to rapidly explore the state space. This is demonstrated by figure 11 **A** and **B** where the extent of exploration in the state space is studied [13]. Figure 11 **A** depicts the FEEF agent which explores rapidly and figure 11 **B** depicts the minimal extent of exploration with only the reward objective.

The cup-catch environment has the FEEF agent perform comparably to the other two agents in figure

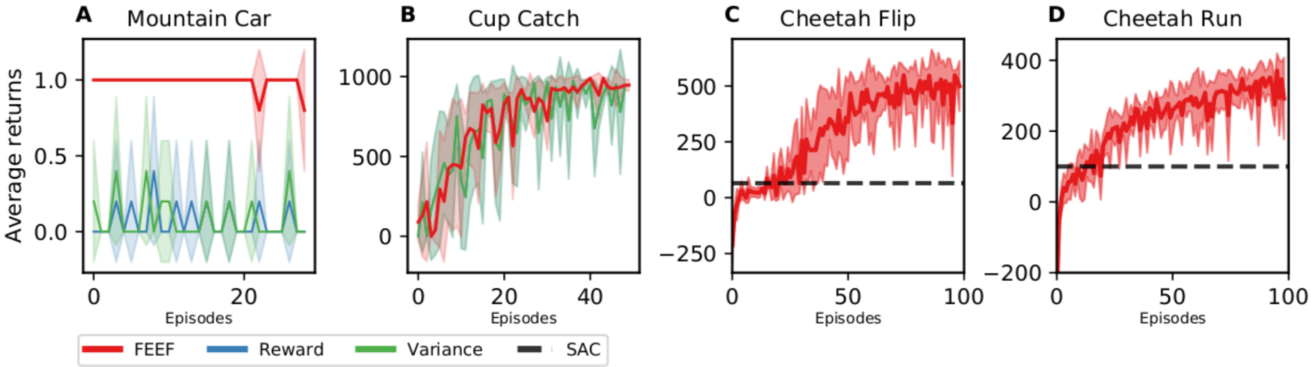

Figure 10: Results from an empirical study in [13]. **A** shows the Mountain-car environment (a sparse reward environment) performance (measured with the average reward received for each episode of 25 episodes) of the active inference agent against a reward-seeking agent and a variance agent. Similarly, **B** shows the cup-catch environment with the corresponding performance of the FEEF agent, the reward-seeking agent and the variance agent. **C** and **D** depict the average reward after each episode obtained by the FEEF agent as compared to a soft-actor-critic agent (SAC) [30] on the half cheetah environments (which provide well-shaped reward functions). The SAC agent's results are depicted after 100 episodes of learning.

10. Despite this being a sparse reward environment all agents achieve asymptotic performance after approximately 20 episodes [1]. The authors of [13] suggest that this is because (despite reward sparsity) it is simple to achieve success with random actions in this environment, meaning that the benefits of the directed exploratory drive are mitigated.

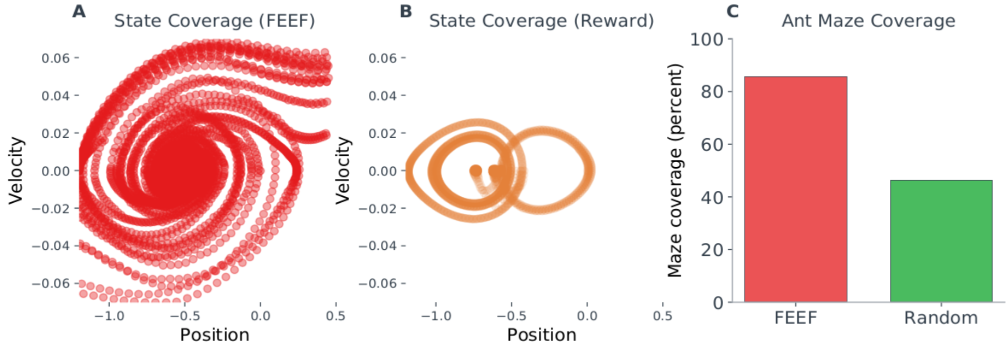

Figure 11: Exploratory results depicted in [13]. **A**: Exploration of the state space for the FEEF agent in the mountain-car environment (a sparse reward environment). **B**: Exploration of the state space of the mountain-car environment for the reward-seeking agent. **C**: Extent of exploration of the maze for the FEEF agent and random agent, in the ant-maze environment, after 35 episodes.

The results of the FEEF agent in the well-shaped reward environments, cheetah flip and cheetah run are depicted in figure 10 **C** and **D**. In these environments the FEEF agent was compared to a soft-actor-critic [30] which is employs a maximum entropy RL agent with a stochastic policy. Both of figure 10 **C** and **D** demonstrate the that the FEEF agent is able to outperform the soft-actor-critic agent in these well-shaped reward function environments. The authors of [13] also note that the FEEF algorithm provides sample efficiency as compared to the model-free SAC algorithm.

Lastly, the results, in [13], of the no-reward environment called ant-maze, are depicted in figure 11 **C**. The exploration of the FEEF agent is compared here to a random agent which conducts actions at random [13]. The results demonstrate that the intrinsic directed exploration of the FEEF agent is able to explore substantially more of the maze than the random agent. The result here, along with the results on the other 'proof-of-concept' tasks demonstrate that the FEEF agent is able to benefit from its intrinsic, directed exploration in reward landscapes that are specified as either sparse or well shaped and also, in environments with no reward signal at all. The FEEF agent therefore achieves a balance between exploration and exploitation in a variety of reward landscapes. Moreover, this is specifically contributed to by the information gain term of its objective function.

4.2.3 Control as Hybrid Inference

A third algorithm [14], which addressed the problem of scaling up active inference in the honours work in [9], designed a hybridisation of model-free and model-based RL. The hybrid attempts to combine the strong asymptotic performance of model-free RL with the sample efficiency of model-based RL [31]. It also seeks to combine the advantage of iterative and amortised inference [22]. The authors of [14] desire the rapid inference ability of amortised inference, with the flexibility of iterative inference.

The algorithm, called *control as hybrid inference* [14], as discussed in [9], operates by first training a model-free stochastic policy via the minimisation of VFE. The model-free policy is trained as a SAC architecture [30], using the objective function defined by a *control as inference* approach (a maximum entropy RL approach) [18]:

$$\mathcal{L}=\mathbb{E}_{q(\tau)}\left[\sum_{t=1}^{T} r\left(s_t, a_t\right)\right]+\mathcal{H}\left[q_\theta\left(a_{1:T} \mid s_{1:T}\right)\right] \tag{64}$$

This model-free policy is then used in conjunction with a trained transition model, $p_\lambda(s_{t+1}|s_t, a_t)$, to obtain an initial action distribution for a model-based planner. The model-based planner then iteratively updates a policy which takes a soft-actor-critic form [30]. The iterative updates occur via mirror descent [14]. This provides a generalisation of a CEM of planning [32], according to work by [46]. In fact [46] suggest that this iterative update provides a Bayesian generalisation for several stochastic optimisation methods used in model-based planning. The mirror descent update appears below, where $\mathcal{W}$ is the expected cumulative reward such that $\mathcal{W}(a_{t:T}) = \mathbb{E}_{q(s_{t:T}|a_{t:T},s_t;\theta)}[C_T(\{r(s_t, a_t)\}_{t=t}^{T})]$ [14]:

$$q^{(i+1)}\left(a_{t:T};\theta\right) \leftarrow \frac{q^{(i)}\left(a_{t:T};\theta\right)\cdot \mathcal{W}\left(a_{t:T}\right)\cdot q^{(i)}\left(a_{t:T};\theta\right)}{\mathbb{E}_{q^{(i)}\left(a_{t:T};\theta\right)}\left[\mathcal{W}\left(a_{t:T}\right)\cdot q^{(i)}\left(a_{t:T};\theta\right)\right]} \tag{65}$$

The algorithm returns the policy posterior after a fixed number of updates, giving more weight (in the posterior distribution) to policies that achieve greater amounts of cumulative reward.

Implementation Results

Once again, the Half-Cheetah 'run' environment was used to compare the performance of the (hybrid) agent against some benchmark agents, in [14]. As the agent is a hybridisation of a model-free and model-based component, it is compared against both a model-free SAC agent [30] and a model-predictive control algorithm that uses a CEM planner. Figure 12 depicts the superior performance of the hybrid as compared to the model-free and model-based benchmarks [14]. The authors note that the algorithm achieves the intended sample efficiency of model-based methods and the intended asymptotic efficiency of model-free methods. The 'proof-of-concept' results here, further contribute to our understanding of the benefits of VFE-minimising agents.

Summary thus far

Having introduced the RL problem with some basic methods of exploration and, the active inference agent (via the free energy principle) with a directed intrinsic exploratory drive, we have gone on to discuss some

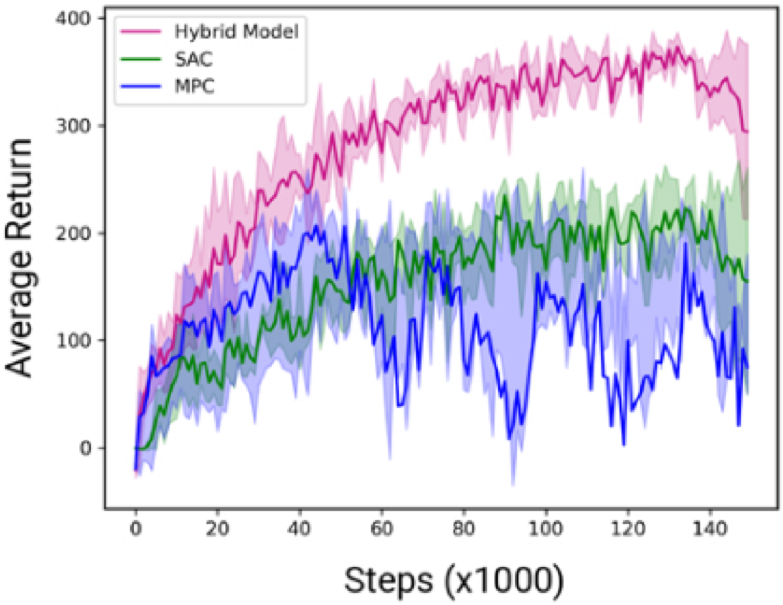

Figure 12: Average reward generated by the hybrid inference agent on the cheetah-run task as compared to its constituent components, a soft-actor-critic agent and a model-predictive control agent that uses a cross-entropy method planner. These results are from [14].

proof-of-concept empirical trials of the implementations of active inference agents. At a basic level these demonstrate the compelling effectiveness of the performance of active inference agents as compared to other contemporary RL agents. For the purpose of our intended study however, these empirical results demonstrate, via figure 10 in particular, that the intrinsic directed exploratory drive (of the active inference agent) offers benefit in environments with sparse or no rewards (and in environments with well-shaped reward functions) [13]. Interestingly, this term did not provide gains in the dense reward environment, as demonstrated by the ablation study of figure 8 [8]. In this instance, the entropy maximisation term provided a means of resolving the exploration-exploitation dilemma. In the next section we discuss an extension of the insights of these empirical studies with regard to resolving the exploration-exploitation dilemma. We turn to studying mathematical insights, which provide an abstract means of studying how the exploratory and intrinsic motives that characterise the active inference agent arise [1].

5 Mathematical Origins of Exploration

In the context of the exploration-exploitation problem in RL and the various means of exploration that we have discussed, it appears that there exist benefits to the directed information-seeking exploration as compared to random exploration. This has emerged both through a dialectical discussion about the wastefulness of random exploration (especially in sparse reward environments). And, this has emerged through some empirical proof-of-concept studies which serve to illustrate this property but, not to rigorously test it. There are two choices for the next step, either a rigorous empirical study of many methods of exploration is necessary in a variety of environments or, we can continue with a dialectical argument by exploring the mathematical origins of exploration. As this is a mathematical paper the latter is the approach adopted for this section. Specifically, we seek to understand the origins of the information gain term and, the origins of the expected free energy objective, from which is arises.

Information-seeking refers to reducing uncertainty about either the environment or the agent's model of the environment. We have noted that the opportunity cost of exploration would mean that random exploratory methods have an implicit inefficiency as compared to information-seeking methods. Furthermore we have noted that when information-seeking objective functions are combined in a single objective function with reward maximising objectives, that a goal-oriented exploration results. This is where the agent is only driven to explore contingencies which are also likely to lead to high reward due to the objectives being optimised in tandem. As seen in the empirical studies of deep active inference methods, these goal-oriented exploratory methods have performed well in RL tasks in sparse-reward environments in particular. These environments often pose a challenge to standard contemporary RL methods. As a result of these subjective insights, it is important to derive and study the mathematical origin and meaning of goal-directed exploration. We look to answer the questions about the nature of these objectives and how they may relate to other objectives. In practical contexts, exploratory objectives are often implanted in an ad-hoc manner. While these often offer sufficient amounts of desired behaviour it is our aim to understand the theory underlying the the choice of exploratory objective. Once again, this section follows Millidge in chapter 5 of his 2021 PhD work [1].

The first (sub)section studies the properties and mathematical origins of the EFE objective function. This is used ubiquitously in discrete active inference implementations [1, 7, 38, 39] with the exception of generalised free energy [1, 47]. For this reason we review an argument [16] that the EFE is not necessarily the only way to extend VFE into the future. For this purpose, we discuss an alternative VFE objective function, the free energy of the future (FEF) and compare this to EFE in terms of its properties and mathematical origins. Following from this, we discuss two differences [17] between the active inference framework and the control as inference framework from [18]. The control as inference framework is argued to use the FEF objective function [17]. Lastly, the Free Energy of Expected Future (FEEF) objective function is studied [13]. The FEEF objective offers both of an intuitively grounded starting point and, an information gain term (as with the EFE objective function).

A subsequent (sub)section takes an abstraction step away from the specifics of the three previously discussed objective functions (EFE, FEF and FEEF). This (sub)section reviews a general (mathematical) framework for understanding the origin of information-seeking exploration terms in objective functions [1, 19]. This framework identifies a dichotomy between evidence and divergence objectives. Here, evidence objectives refer to maximising the likelihood of a distribution of desires whereas divergence objectives try to minimise the KL divergence between predictive distributions and distributions of preferences [1]. This dichotomy leads to the discussion of a variety of variational objective functions for control [1]. This sheds light on some heuristics for design choices for assembling an objective function for an agent and, for a choice of exploratory objectives.

5.1 Origins of the Expected Free Energy Objective

In the usual active inference framework [1, 38] the expected free energy (EFE) is advocated for via the following reductio-ad-absurdum [1, 10]. As the free energy principle can be applied to the active inference agent (i.e. it self-organises to a non-equilibrium steady-state (NESS) that holds a Markov blanket), it therefore abides by the minimisation of VFE to maintain its NESS and perform approximate Bayesian inference about its environment. Since the agent is a free-energy-minimising agent and, since we have assumed that the agent will maintain its NESS distribution (which incorporates its prior preferences, $\tilde{p}$), it therefore will maintain this NESS into the future (via the minimisation of free energy). Extending this, the FEP says that the agent will perform actions which a-priori maintain its NESS. To perform actions, it thus asks the question: “given the assumption that I will achieve my preferences, which action sequence will I most likely pursue?” [1]. Hence the agent is assumed to minimise VFE into the future which facilitates its decision-making. This is because, if it did not minimise VFE into the future, it would not be a free-energy-minimising agent. Extending this argument, many formulations of active inference suggest that EFE is the resulting extension of VFE into the uncertain future [1].

However, contrasting to this claim, the authors of [16] suggest that, the natural extension of VFE into the future must have the following properties. Firstly and similarly to VFE, the extension of VFE into the future must consist of a KL-divergence between a variational posterior and a generative model of the environmental dynamics. Secondly, the objective function must form a variational upper bound on surprise, $-\ln(p(o))$ (i.e. a bound on the negative natural logarithm of the evidence of observations given a model). These two characteristics maintain the properties of VFE explicitly. In other words the first characteristic implicitly ensures that the variational posterior will be a good approximation of the true posterior (as per the methods of variational inference). Conversely, the second property affords us an explicit measure of how good the model is (due to the Bayesian model-comparison metric of log model evidence) [1].

These properties therefore define an alternative to EFE (which does not meet these properties). The alternative is called the *free energy of the future* (FEF) [16]. More precisely, this is defined in the same mathematical context as the active inference agent, namely a POMDP where the agent maintains a variational distribution over states and actions and a generative model over states s, observations o and policies $\pi = \{a_1, a_2 \ldots, a_T\}$. The FEF objective is defined below to meet the first property above, namely that it is a KL divergence between a variational posterior and a generative model (expected over future observations o_t). The FEF is defined as [1]:

$$\begin{aligned} \mathbb{FEF}_t(\pi) &= \mathbb{E}_{q(o_t, s_t|\pi)}\left[\ln q\left(s_t \mid o_t\right) - \ln \tilde{p}\left(o_t, s_t\right)\right] \\ &= \mathbb{E}_{q(o_t)} D_{\mathrm{KL}}\left[q\left(s_t \mid o_t\right) \| \tilde{p}\left(o_t, s_t\right)\right] \end{aligned} \tag{66}$$

The FEF objective function also meets the second condition [1]:

$$\begin{aligned} -\mathbb{E}_{q(o_t|\pi)[\ln \tilde{p}(o_t)]} &= -\mathbb{E}_{q(o_t|\pi)}\left[\ln \int ds_t \tilde{p}\left(o_t, s_t\right)\right] \\ &= -\mathbb{E}_{q(o_t|\pi)}\left[\ln \int ds_t \tilde{p}\left(o_t, s_t\right) \frac{q\left(s_t \mid o_t\right)}{q\left(s_t \mid o_t\right)}\right] \\ &\leq -\mathbb{E}_{q(o_t|\pi)} \int ds_t q\left(s_t \mid o_t\right)\left[\ln \frac{\tilde{p}\left(o_t, s_t\right)}{q\left(s_t \mid o_t\right)}\right] \\ &\leq -\mathbb{E}_{q(o_t, s_t|\pi)}\left[\ln \frac{\tilde{p}\left(o_t, s_t\right)}{q\left(s_t \mid o_t\right)}\right] \\ &\leq \mathbb{E}_{q(o_t, s_t|\pi)}\left[\ln \frac{q\left(s_t \mid o_t\right)}{\tilde{p}\left(o_t, s_t\right)}\right] \\ &\leq \mathbb{E}_{q(o_t|\pi)} \mathbb{D}_{KL}\left[q\left(s_t \mid o_t\right) \| \tilde{p}\left(o_t, s_t \mid \pi\right)\right] = \mathbb{FEF}(\pi) \end{aligned} \tag{67}$$

Thus the FEF objective displays the same properties (and benefits) as VFE. Viewing the two objectives (FEF and EFE) alongside one another provides additional contrast [1, 16]:

$$\begin{aligned} &\mathbb{FEF} = \mathbb{E}_{q(o_t, s_t|\pi)}\left[\ln q\left(s_t \mid o_t\right) - \ln \tilde{p}\left(o_t, s_t\right)\right] \\ &\mathbb{FEF} = -\underbrace{\mathbb{E}_{q(o_t, s_t|\pi)}\left[\ln \tilde{p}\left(o_t \mid s_t\right)\right]}_{\text{Extrinsic Value}} + \underbrace{\mathbb{E}_{q(o_t|\pi)} \mathbb{D}_{KL}\left[q\left(s_t \mid o_t\right) \| q\left(s_t \mid \pi\right)\right]}_{\text{Intrinsic Value}} \\ &\mathbb{EFE} = \mathbb{E}_{q(o_t, s_t|\pi)}\left[\ln q\left(s_t \mid \pi\right) - \ln \tilde{p}\left(o_t, s_t\right)\right] \\ &\mathbb{EFE} = \underbrace{-\mathbb{E}_{q(o_t, s_t|\pi)}\left[\ln \tilde{p}\left(o_t\right)\right]}_{\text{Extrinsic Value}} - \underbrace{\mathbb{E}_{q(o_t|\pi)} \mathbb{D}_{KL}\left[q\left(s_t \mid o_t\right) \| q\left(s_t \mid \pi\right)\right]}_{\text{Intrinsic Value}} \end{aligned} \tag{68}$$

While the respective definitions of EFE and FEF are similar, the decompositions of these into extrinsic and intrinsic value terms provides a clear contrast. While the EFE objective maximises the intrinsic value term, the FEF objective minimises it. This corresponds to an information gain term in the EFE (as previously elaborated) and a complexity term in the FEF. This complexity term is similar to that appearing for VFE in equation 50. Hence while the EFE objective imposes an exploratory incentive in the agent, the FEF objective does not and, in fact, attempts to tries to maximise reward whilst learning as little about the environment as possible [1].[9] The following decomposition highlights a key step in the approximation of EFE, namely approximating the posterior $p(s|o)$ with a variational posterior, $q(s|o)$ [1]. The resulting *posterior approximation error* term is often assumed to be zero.

$$
\begin{aligned}
\mathbb{EFE} &= \mathbb{E}_{q(o_t,s_t|\pi)}\left[\ln q\left(s_t \mid \pi\right) - \ln \tilde{p}\left(o_t, s_t\right)\right] \\
&\approx \mathbb{E}_{q(o_t,s_t|\pi)}\left[\ln q\left(s_t \mid \pi\right) - \ln p\left(s_t \mid o_t\right) - \ln \tilde{p}\left(o_t\right) + \ln q\left(s_t \mid o_t\right) - \ln q\left(s_t \mid o_t\right)\right] \\
&\approx \underbrace{-\mathbb{E}_{q(o_t|\pi)}\left[\ln \tilde{p}\left(o_t\right)\right]}_{\text{Negative Expected Log Model Evidence}} + \underbrace{\mathbb{E}_{q(o_t|\pi)}\mathbb{D}_{KL}\left[q\left(s_t \mid o_t\right) \| p\left(s_t \mid o_t\right)\right] |}_{\text{Posterior Approximation Error}} - \underbrace{\mathbb{E}_{q(o_t|\pi)}\mathbb{D}_{KL}\left[q\left(s_t \mid o_t\right) \| q\left(s_t \mid \pi\right)\right] |}_{\text{Information Gain}}
\end{aligned}
\tag{69}
$$

Given a zero-valued posterior approximation error term, we can see that the EFE provides a lower bound on the negative, expected natural logarithm of the model evidence (the surprise). But, our desired property was for an upper bound on this quantity, as with the FEF and with VFE. The addition of the posterior approximation error term does, however, provide an interpretation of the likely time-evolution of the EFE objective [1]. This is that the EFE will likely cycle around the bound of the surprise through time, until it is (potentially) reached (as per the assumption of a NESS which we reviewed in the FEP section, section 3). The interpretation here is that state inference at the start of the agent's training is poor and, that the posterior approximation error term will be larger, at this time, than the information gain term. Hence the EFE will provide an upper bound on surprise and will maximise the model evidence. As the information gain term (between the prior and posterior) gets larger the minimisation of EFE will then drive the agent to explore, and drive the EFE away from the true (negative) log model evidence for the environment [1]. Lastly, once the true and approximate posteriors have no sources of divergence left and, the agent has a perfect model of the environment (so that there is no residual source of information gain left[10]), then both the posterior approximation error term and the information term will be zero [1]. Hence, in this instance, the EFE will converge to the surprise. This provides an adaptive behaviour for the active inference agent. The derivation work in [1] also provides the insight that the EFE is equal to the FEF minus an additional information gain term:

$$
\begin{aligned}
\mathbb{FEF}_t(\pi) - \mathbb{IG}_t &= \mathbb{E}_{q(o_t,s_t|\pi)} \ln\left(\frac{q\left(s_t \mid o_t\right)}{\tilde{p}\left(o_t, s_t\right)}\right) - \mathbb{E}_{q(o_t,s_t|\pi)} \ln\left(\frac{q\left(s_t \mid o_t\right)}{q\left(s_t \mid \pi\right)}\right) \\
&= \mathbb{E}_{q(o_t,s_t|\pi)} \ln\left(\frac{q\left(s_t \mid o_t\right) q\left(s_t \mid \pi\right)}{\tilde{p}\left(o_t, s_t\right) q\left(s_t \mid o_t\right)}\right) \\
&= \mathbb{E}_{q(o_t,s_t|\pi)} \ln\left(\frac{q\left(s_t \mid \pi\right)}{\tilde{p}\left(o_t, s_t\right)}\right) \\
&= \mathbb{EFE}(\pi)_t
\end{aligned}
\tag{70}
$$

Hence the work of [1, 16] has revealed the origins of the EFE, that it is equal to the FEF objective with an information gain term being used, by choice of construction. This reveals that the exploratory, drive embedded in the EFE, is ad-hoc in the same way as many other exploratory methods. This raises the question of whether it is possible to derive a mathematically or intuitively principled objective which maintains the same information-gain term as the EFE. The FEEF objective, proposed by [1, 13, 16] provides this objective. As seen section 4.2 (and in our previous work [9]), minimising the FEEF objective functional as defined in [13], yields a policy posterior to be of a softmax form. Here, the argument of the softmax function is the negative $\mathcal{F}_\pi$ function defined in equation 62. In this section we remove the random variable representing parameters but note that the same arguments follow if this is included. We also refer to $\mathcal{F}_\pi$

[9] The complexity term (as with VFE) acts as a regularisation term in the objective. The author of [1] suggests that this provides benefit in the case of offline RL. In such cases the failure to generalise and extrapolate can often result in poor results, hence the regularisation term is useful to prevent overfitting.

[10] ... and the environment has no source of aleatoric uncertainty.

as the FEEF objective, below. This is defined to be the KL divergence between the variational belief distribution, $q(o, s)$ and the desired distribution, $\tilde{p}(o, s)$ [1]:

$$\begin{aligned}\mathbb{FEEF}(\pi)_t &= \mathbb{D}_{KL}\left[q\left(o_t, s_t \mid \pi\right) \| \tilde{p}\left(o_t, s_t\right)\right] \\ &\approx \underbrace{\mathbb{E}_{q(s_t|\pi)} \mathbb{D}_{KL}\left[q\left(o_t \mid s_t\right) \| \tilde{p}\left(o_t\right)\right]}_{\text{Extrinsic Value}} - \underbrace{\mathbb{E}_{q(o_t|\pi)} \mathbb{D}_{KL}\left[q\left(s_t \mid o_t\right) \| q\left(s_t \mid \pi\right)\right]}_{\text{Intrinsic Value}}\end{aligned} \tag{71}$$

This objective provides a strong intuitively principled basis for the adaptive action of the agent since it defines the objective of the agent to be the minimisation of the divergence between its predicted distribution and its desires [1]. As the desired distribution is fixed, the agent minimises this objective by taking actions so as to move its predictive distribution closer to its desires. The decomposition into extrinsic and intrinsic value terms reveals that this agent (which minimises the FEEF objective) maximises the same information gain term as appears in the EFE objective. The difference occurs with the extrinsic value term, as is emphasised by a second decomposition of the FEEF objective [1]:

$$\begin{aligned}\mathbb{FEEF}(\pi)_t &= \mathbb{D}_{KL}\left[q\left(o_t, s_t\right) \| \tilde{p}\left(o_t, s_t\right)\right] \\ &= \underbrace{\mathbb{E}_{q(o_t,s_t)}\left[\ln q\left(o_t \mid s_t\right)\right]}_{\text{Observation Likelihood}} + \underbrace{\mathbb{E}_{q(o_t,s_t)}\left[\ln \tilde{p}\left(o_t \mid s_t\right)\right] - \mathbb{E}_{q(o_t|\pi)} \mathbb{D}_{KL}\left[q\left(s_t \mid o_t\right) \| q\left(s_t \mid \pi\right)\right]}_{\text{EFE}}\end{aligned} \tag{72}$$

Equations 71 and 72 highlight the difference between the extrinsic value terms of EFE and the FEEF objectives. The EFE attempts to maximise the likelihood of the distribution of preferences, while the FEEF attempts to minimise the divergence between its distribution of preferred observations and its variational distribution for observations. Equation 72, specifically demonstrates the the FEEF is simply equal to the EFE with an added observation likelihood entropy term. As the entropy term (of the observation likelihood) is negative, it is maximised when the FEEF is minimised. This provides an extra, random exploratory term for the FEEF objective and hence the FEEF optimises EFE, while keeping its observation likelihood function as random as possible [1]. The occurrence of a distinction between extrinsic value terms (for the FEEF and EFE objectives), raises the question of what this means? This will be addressed in terms of the difference between evidence and divergence objective in a later subsection. We next identify the relationship between the FEF and EFE objectives.

5.2 Active Inference vs Control as Inference

The relationship between the FEF and EFE is used in this section which studies a comparison between the objective functions of active inference and control as inference. The control as inference framework looks to define the RL problem as a problem of inference. This is proposed by the work of [18], and was discussed in our previous work [9]. As a reminder to the reader the following paragraphs briefly introduce the control as inference framework.

The control as inference framework looks at the RL problem as one of inferring the optimal policy posterior distribution over a full trajectory. This is the policy posterior which maximises the expected cumulative reward, $J(\theta)$ [22]. To reformulate this objective in terms of probabilistic inference, [18] introduces the binary variables, Ω to the MDP context of the RL agent. These are equal to one if the time-step was optimal. As this formulation desires optimality the 'one' is dropped from notation, such that $p\left(\Omega_t = 1\right) = p\left(\Omega_t\right)$.

As the agent wishes to obtain the optimal posterior, which is $p(a, s|o, \Omega)$ within a POMDP, variational inference can be used [18]. This introduces the approximate posterior $q(a, s|o, \Omega)$. Here, the reward signal can be incorporated as $p\left(\Omega_t = 1|a_t, s_t\right) \propto \exp\left(r\left(s_t, a_t\right)\right)$ [18]. Lastly the control as inference objective function is simply the variational free energy, which forms an upper bound on the KL divergence between

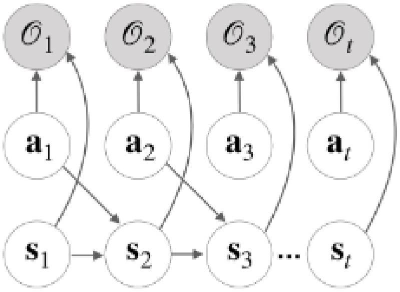

Figure 13: Graphical model of an MDP with added hidden optimality variable nodes, $\mathcal{O}$, from [18]. (These are Ω in the main text). The RL problem as inference infers the most probable action sequence, given that $\mathcal{O}$ is optimal (and equal to one).

the true and approximate posteriors. The free energy objective function for POMDPs can be denoted as [17]:

$$\begin{aligned}
\mathcal{L}(\phi)_{CAI} &= D_{\mathrm{KL}}\left(q_\phi\left(s_t, a_t\right) \| p\left(s_t, a_t, o_t, \Omega_t\right)\right) \\
&= \underbrace{-\mathbb{E}_{q_\phi\left(s_t, a_t\right)}\left[\ln p\left(\Omega | s_t, a_t\right)\right]}_{\text{Extrinsic Value}} + \underbrace{D_{\mathrm{KL}}\left(q\left(s_t\right) \| p\left(s_t \mid s_{t-1}, a_{t-1}\right)\right)}_{\text{State divergence}} \\
&+ \underbrace{\mathbb{E}_{q\left(s_t\right)}\left[D_{\mathrm{KL}}\left(q_\phi\left(a_t \mid s_t\right) \| p\left(a_t \mid s_t\right)\right)\right]}_{\text{Action Divergence}} - \underbrace{\mathbb{E}_{q_\phi\left(s_t, a_t\right)}\left[\ln p\left(o_t \mid s_t\right)\right]}_{\text{Observation Ambiguity}}
\end{aligned} \tag{73}$$

where this objective bounds the KL divergence between the true and approximate joint posteriors, $p(s, a|o, \Omega)$ and $q(s, a|o, \Omega)$ respectively [17]. With this objective in hand and, with our insights into the active inference objectives, we can review the difference between active inference and control as inference as explored by [17].

The first difference between the active inference and control as inference frameworks, that is noted by [17], is that active inference mainly infers action sequences while control as inference mainly infers single actions for a given time-step [1]. Of course, both active inference can be formulated to infer single actions and, control as inference can be reformulated to infer whole policies. For the case of control as inference the above objective function, in equation 73, can be reformulated to infer action sequences. This makes use of the structure of the POMDP [1, 17]:

$$\begin{aligned}
\mathcal{L}_{CAI} &= D_{\mathrm{KL}}\left(q\left(s_{t:T}, \pi\right) \| p\left(s_{t:T}, \pi, o_{t:T}, \Omega_{t:T}\right)\right) \\
&= D_{\mathrm{KL}}\left(q(\pi) \| p(\pi) \exp\left(-\sum_t^T \mathcal{L}_t(\pi)\right)\right) \\
&\implies q^*(\pi) = \sigma\left(p(\pi) - \sum_t^T \mathcal{L}_t(\pi)\right)
\end{aligned} \tag{74}$$

As equation 74 shows, the expression of the (minimised) control as inference objective in terms of action sequences yields the optimal policy posterior $q(\pi)$ to be a softmax distribution, that takes as its argument a variational free energy path integral $\mathcal{L}_t(\pi)$, augmented with optimality nodes and extended into the future [1]. This presents a similarity with the active inference optimal posterior. If the prior $p(\pi)$ is set to be uniform such that $p(\pi) \propto 1$ then the policy posterior is just a softmax distribution of a free energy path integral, just as the active inference policy posterior is a softmax distribution of the EFE path integral. $\mathcal{L}_t(\pi)$ can be expanded for interpretation, as seen in [17]:

$$
\begin{aligned}
\mathcal{L}_t(\pi) &= \mathbb{E}_{q(s_t|\pi)}\left[\ln q\left(s_t \mid \pi\right)-\ln p\left(s_t, \pi, o_t, \Omega_t\right)\right] \\
&= \underbrace{-\mathbb{E}_{q(s_t|\pi)}\left[\ln p\left(\Omega_t \mid s_t, \pi\right)\right]}_{\text{Extrinsic Value}}+\underbrace{D_{\mathrm{KL}}\left(q\left(s_t \mid \pi\right) \| p\left(s_t \mid s_{t-1}, \pi\right)\right)}_{\text{State divergence}}-\underbrace{\mathbb{E}_{q(s_t|\pi)}\left[\ln p\left(o_t \mid s_t\right)\right]}_{\text{Observation Ambiguity}}
\end{aligned} \tag{75}
$$

This is equivalent to equation 73 except that since whole action sequences are inferred the action divergence term is missing. [17] defines the active inference EFE objective to infer individual action instead of action sequences as:

$$
\begin{aligned}
-\mathcal{F}_t(\phi) &= \mathbb{E}_{q(o_t, s_t, a_t)}\left[\ln q_\phi\left(a_t, s_t\right)-\ln \tilde{p}\left(s_t, o_t, a_t\right)\right] \\
&= \underbrace{-\mathbb{E}_{q(o_t|a_t)}\left[\ln \tilde{p}\left(o_t \mid a_t\right)\right]}_{\text{Extrinsic Value}}-\underbrace{\mathbb{E}_{q(o_t, a_t|s_t)}\left[D_{\mathrm{KL}}\left(q\left(s_t \mid o_t, a_t\right) \| q\left(s_t \mid a_t\right)\right)\right]}_{\text{Intrinsic Value}}+\underbrace{\mathbb{E}_{q(s_t)}\left[D_{\mathrm{KL}}\left(q_\phi\left(a_t \mid s_t\right) \| p\left(a_t \mid s_t\right)\right)\right]}_{\text{Action Divergence}}
\end{aligned} \tag{76}
$$

[17] notes inferring single actions requires an action divergence term which is present in the control as inference formulation too. Specifically relevant to our review, is the comparison of equations 73 and 76 which hold two differences. The first is the presence of an observation ambiguity term in the control as inference objective. Secondly, the intrinsic value term from the EFE objective differs from the state divergence term from the control as inference objective. As we know that the intrinsic value term arises by choice we can instead choose the FEF objective (for the active inference agent), which is formulated to infer single-action policies below [1]:

$$
\begin{aligned}
-\hat{\mathcal{F}}_t(\phi) &= \mathbb{E}_{q_\phi(s_t, o_t, a_t)}\left[\ln q_\phi\left(s_t, a_t\right)-\ln \tilde{p}\left(o_t, s_t, a_t\right)\right] \\
&= \underbrace{-\mathbb{E}_{q_\phi(s_t, a_t)}\left[\ln \tilde{p}\left(o_t \mid s_t\right)\right]}_{\text{Extrinsic Value}}+\underbrace{D_{\mathrm{KL}}\left(q\left(s_t\right) \| p\left(s_t \mid s_{t-1}, a_{t-1}\right)\right)}_{\text{State divergence}}+\underbrace{D_{\mathrm{KL}}\left(q_\phi\left(a_t \mid s_t\right) \| p\left(a_t \mid s_t\right)\right)}_{\text{Action Divergence}}
\end{aligned} \tag{77}
$$

The FEF provides the same objective as the control as inference objective, except that there is a missing observation ambiguity term. The insight here is that, control as inference and active inference (with a FEF objective) possess fundamentally the same objective functions, apart from the observation ambiguity term. This difference is due to the encoding of goals in either framework [17]. Active inference encodes the agents goals in its biased observational likelihood distribution $\tilde{p}(o)$. In contrast, control as inference holds both an optimality likelihood $p(\Omega|s, a)$ and a likelihood for perceptual inference, $p(o|s)$. The result of this is the additional term in the control as inference objective [17].

According to [1], the distinction (between the approaches and objectives) is embedded in the philosophical foundations of the two approaches. While active inference views it unnecessary to separate perceptual inference and action inference, control as inference separates these. Active inference views this separation as unnecessary since its aim is adaptive action, in which perception serves as an intertwined process. In contrast, [1] suggests that control as inference maintains the 'modularity thesis'. In this case, perception aims to build an accurate model of the environment and, separately this model is used for the inference as to optimal action sequences. Hence control as inference estimates actions by conditioning on observing states with high rewards, generated by its world model and, in contrast, active inference selects actions by simply maximising the likelihood of its biased generative model.

While this insight is interesting to say the least, the focus of this section is on the mathematical origins of the exploratory terms. We have seen that the control as inference objective function presented here does not include an intrinsic information gain term and neither does the FEF. Furthermore, we have seen that only the control as inference objective holds an observation ambiguity term. Lastly, the intrinsic exploratory term in the EFE function is added only by choice of construction. Hence [19] asks whether we can determine a mathematically principled origin for this information gain term (and for the observation ambiguity term). We turn to this in the next (sub)section.

5.3 Evidence vs Divergence Objectives

In this subsection we review the paper titled *Understanding the origin of information-seeking exploration in probabilistic objectives for control* [19]. This proposes a dichotomy between evidence objectives and divergence objectives for exploration. First we define these as per [19]. The evidence objective is defined as a maximisation of the expectation of the likelihood of the desire distribution, with the expectation being taken over future observations:[11]

$$\begin{aligned}
\mathcal{L}_{\text{Evidence}} &= \underset{a_{t:T}}{\operatorname{argmax}} \mathbb{E}_{p(o_{t:T}|a_{t:T})}\left[\ln \tilde{p}\left(o_{t:T}\right)\right] \\
&= \underset{a_{t:T}}{\operatorname{argmax}} \mathbb{E}_{p(o_{t:T}|a_{t:T})}\left[\ln \tilde{p}\left(o_{t:T}\right) \frac{p\left(o_{t:T} \mid a_{t:T}\right)}{p\left(o_{t:T} \mid a_{t:T}\right)}\right] \\
&= \underset{a_{t:T}}{\operatorname{argmax}} - \underbrace{D_{\text{KL}}\left[p\left(o_{t:T} \mid a_{t:T}\right) \| p\left(o_{\widetilde{t:T}}\right)\right]}_{\text{Divergence}} - \underbrace{\mathbb{H}\left[p\left(o_{t:T} \mid a_{t:T}\right)\right]}_{\text{Expected Future Entropy}}
\end{aligned} \tag{78}$$

The divergence objective is defined as a minimisation of the KL divergence between the agent's predictive distribution of future observations and, its distribution of preferences for observations [19]:

$$\begin{aligned}
\mathcal{L}_{\text{Divergence}} &= \underset{a_{t:T}}{\operatorname{argmin}} D_{\text{KL}}\left[p\left(o_{t:T} \mid a_{t:T}\right) \| \tilde{p}\left(o_{t:T}\right)\right] \\
&= \underset{a_{t:T}}{\operatorname{argmin}} \underbrace{\mathbb{E}_{p(o_{t:T}|a_{t:T})}\left[\ln p\left(o_{t:T} \mid a_{t:T}\right)\right]}_{\text{Expected Future Entropy}} - \underbrace{\mathbb{E}_{p(o_{t:T}|a_{t:T})}\left[\ln \tilde{p}\left(o_{t:T}\right)\right]}_{\text{Evidence Objective}}
\end{aligned} \tag{79}$$

A distinction to draw between these two objectives is that the evidence objective is focused on trying to match the predictive distribution to the mode of the preference distribution whereas, the divergence objective seeks to, precisely, match the two distributions. The effect of this is illustrated by figure 14.

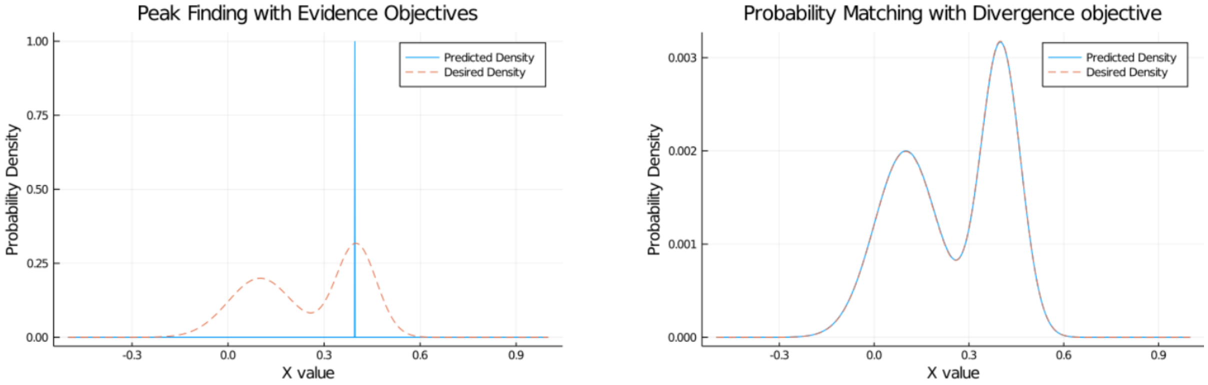

Figure 14: A numerical depiction of the evidence and divergence objectives from [19]. Both graphs depict predictive distributions being trained on their corresponding objective to match a multimodal preference distribution. (a) depicts the predictive distribution being matched to the mode of the preference distribution, while (b) depicts the predictive distribution precisely matching the preference distribution, through the use of the divergence objective.

[11] As usual, the natural logarithm of the desire distribution is taken, as this a monotonic increasing function so does not affect the optimum.

Figure 14 illustrates that the predictive distributions corresponding to the respective objectives (in blue) differ in their approaches to fitting the preference distribution (orange). For the case of the evidence objective, the agent forms a sharply peaked predictive distribution around the mode of the preference distribution. This starkly contrasts with an agent fitting a divergence objective, where in this case the agent tries to precisely match the full extent of the preference distribution. For the case of the agent with an evidence objective, it neglects all but the mode of the preference distribution. This has consequences for the exploratory behaviour of such an agent.

Equations 78 and 79 provide mathematical clarity on this distinction (between evidence and divergence objectives). The decomposition of the divergence objective provides us with the interpretation that, an agent using this objective seeks to maximise the expected likelihood of its prior preference distribution while maximising the entropy of its predictive distribution (and therefore keeping its future as broad as possible) [1]. In contrast, the interpretation provided by the decomposition of the evidence objective reveals that it attempts to balance the matching of its predicted and desired distributions for observations with the minimisation of the entropy of its predictive distribution. Hence, the behaviour depicted in figure 14 displays the evidence-maximising agent as resolving this trade-off by forming a low entropy distribution around the mode of the preference distribution.

The subtle differences in behaviour of these two objective reveals a difference in the underlying approach. The subtle differences are more apparent in scenarios where the distribution of preferences is broad and complex, since the divergence objective seeks to match this complexity, while the evidence objective neglects it to focus on the distribution mode [1]. It is this difference which gives rise to the information gain term (the intrinsic value term). This results from the inclusion of the (hidden) states into the divergence objective [1, 19]:[12]

$$\begin{aligned} D_{\mathrm{KL}}\left[p\left(o_{1:T} \mid a_{1:T}\right) \| \tilde{p}\left(o_{1:T}\right)\right] &= \mathbb{E}_{p(o_{1:T}|a_{1:T})}\left[\ln \frac{p\left(o_{1:T} \mid a_{1:T}\right)}{\tilde{p}\left(o_{1:T}\right)}\right] \\ &= \mathbb{E}_{p(o_{1:T}|a_{1:T})}\left[\ln \frac{p\left(o_{1:T}, s_{1:T} \mid a_{1:T}\right)}{\tilde{p}\left(o_{1:T}\right) p\left(s_{1:T} \mid o_{1:T}\right)}\right] \\ &= \underbrace{\mathbb{E}_{p(s_{1:T})} D_{\mathrm{KL}}\left[p\left(o_{1:T} \mid s_{1:T}\right) \| \tilde{p}\left(o_{1:T}\right)\right]}_{\text{Desire Divergence}} - \underbrace{\mathbb{E}_{p(o_{1:T}|a_{1:T})} D_{\mathrm{KL}}\left[p\left(s_{1:T} \mid o_{1:T}\right) \| p\left(s_{1:T}\right)\right]}_{\text{Information Gain}} \end{aligned} \tag{80}$$

An additional perspective demonstrates that the emergence of the information gain term results from the extension of the predictive distribution, to include the hidden states s. This can be seen from the perspective of the divergence objective's maximisation of the entropy of future observations [19]. The equation below shows that, to maximise the entropy of a distribution which holds dependence on a set of latent variables, the agent must both maximise the entropy of the observed variable given the latent variable, while also maximising the information gain about the latent variables.

$$\begin{aligned} \mathbb{H}\left[p\left(o_{1:T} \mid a_{1:T}\right)\right] &= \mathbb{E}_{p(o_{1:T}, s_{1:T}|a_{1:T})}\left[\ln p\left(o_{1:T} \mid a_{1:T}\right)\right] \\ &= \mathbb{E}_{p(o_{1:T}, s_{1:T}|a_{1:T})}\left[\ln \frac{p\left(o_{1:T}, s_{1:T} \mid a_{1:T}\right)}{p\left(s_{1:T} \mid o_{1:T}\right)}\right] \\ &= -\underbrace{E_{p(s_{1:T})} \mathbb{H}\left[p\left(o_{1:T} \mid s_{1:T}\right)\right]}_{\text{Likelihood Entropy}} - \underbrace{\mathbb{E}_{p(o_{1:T}|a_{1:T})} D_{\mathrm{KL}}\left[p\left(s_{1:T} \mid o_{1:T}\right) \| p\left(s_{1:T}\right)\right]}_{\text{Expected Information Gain}} \end{aligned} \tag{81}$$

The study of [19] has illuminated the differences between the evidence and divergence objectives. This has revealed that the two objectives hold a difference in philosophy. The evidence objective seeks a low-entropy future. This is a precise future, where the agent seeks to reach its goal whilst learning as little as it can

[12] As a reminder the hidden states are the states of the environment which generate observations o within the POMDP.

about its environment [1].[13] In contrast, an agent with a divergence objective seeks a broad future, which trades off the achievement of its goals with the maximisation of entropy of future observations. This affords it the ability to learn more complex distributions of desires. Furthermore, learning complex desires (and maximising the entropy of observations) requires modelling the relationship between the (hidden) states and the observations. Hence, it seeks to maximise the information gain about this relationship. Thus, we see that, it is the divergence objective which results in the embedding of an intrinsic directed exploratory drive in the agent, for information gain. We next turn to reviewing some examples of evidence and divergence objectives that appear in [19].

Examples of Evidence and Divergence Exploratory Objectives

The Control as Inference Objective

The control as inference objective, seen in [19] is one that seeks to obtain the variational posterior for sequences of actions, $p\left(a_{1:T} \mid \tilde{o}_{1:T}\right)$. In this case $\tilde{o}$ are a set of optimal actions which are conditioned upon. This yields the following control as inference objective function [19]:

$$\mathcal{L}_{CAI} = D_{\mathrm{KL}}\left[q\left(a_{1:T}\right) \| \tilde{p}\left(o_{1:T}, a_{1:T}\right)\right] \tag{82}$$

The derivation below demonstrates that this objective forms a lower bound on an evidence objective [19]. Hence, as per our usual variational inference argument, maximising the (negative) $\mathcal{L}_{CAI}$ objective (in terms of actions) will maximise the evidence objective (in terms of actions). As per the arguments of the previous section, this reveals why the control as inference objective does not yield an information gain term. Interestingly, this objective can yield a random action entropy maximising term for exploration [18]. As has been discussed this is efficient in dense reward environments but is increasingly inefficient in sparse reward environments, especially as compared to the exploration yielded by the intrinsic information gain term.

$$\begin{aligned}
\ln \tilde{p}\left(o_{1:T}\right) &= \ln \int dx \tilde{p}\left(o_{1:T} s_{1:T}\right) \\
&= \ln \int dx \frac{\tilde{p}\left(o_{1:T} s_{1:T}\right) q\left(a_{1:T}\right)}{q\left(a_{1:T}\right)} \\
&\geq \mathbb{E}_{q\left(a_{1:T}\right)}\left[\ln \frac{\tilde{p}\left(o_{1:T}, a_{1:T}\right)}{q\left(a_{1:T}\right)}\right] \\
&\geq -D_{\mathrm{KL}}\left[q\left(a_{1:T}\right) \| \tilde{p}\left(o_{1:T}, a_{1:T}\right)\right] = -\mathcal{L}_{CAI}
\end{aligned} \tag{83}$$

Expected Free Energy

As we have seen, the EFE objective of active inference does not bound the (negative) natural logarithm of model evidence. Rather it forms an upper bound when the posterior divergence term was greater than the information gain term and, a lower bound otherwise. If both the posterior divergence term and information gain term were zero the EFE converged to the log model evidence [1]. The relationship of EFE to the evidence objective (which is simply the log model evidence) is thus unclear [19]:

[13]... and as little as possible about the state-observation relationship as possible.

$$\begin{aligned}
\underbrace{\mathbb{E}_{q(o,s)}[\ln \tilde{p}(o)]}_{\text{Evidence Objective}} &= \underbrace{\mathbb{E}_{q(o,s)}[\ln q(s) - \ln \tilde{p}(o,s)]}_{\text{EFE}} + \underbrace{\mathbb{E}_{q(o)} D_{\mathrm{KL}}[q(s \mid o) \| q(s)]}_{\text{Information Gain}} - \underbrace{\mathbb{E}_{q(o)} D_{\mathrm{KL}}[q(s \mid o) \| p(s \mid o)]}_{\text{Posterior Divergence}} \\
&\implies \underbrace{\mathbb{E}_{q(o,s)}[\ln \tilde{p}(o)]}_{\text{Evidence Objective}} \geq \underbrace{\mathbb{E}_{q(o,s)}[\ln q(s) - \ln \tilde{p}(o,s)]}_{\text{EFE}} \\
&\text{If } \underbrace{\mathbb{E}_{q(o)} D_{\mathrm{KL}}[q(s \mid o) \| q(s)]}_{\text{Information Gain}} \geq \underbrace{\mathbb{E}_{q(o)} D_{\mathrm{KL}}[q(s \mid o) \| p(s \mid o)]}_{\text{Posterior Divergence}}
\end{aligned} \tag{84}$$

Similarly, the EFE objective does not pose a straightforward relationship with the divergence objective. The relationship between EFE and the divergence objective, seen below, interestingly includes the VFE. This reveals that the EFE provides a lower bound on the divergence objective if the information gain term is greater than VFE [19]:

$$\begin{aligned}
\underbrace{D_{\mathrm{KL}}[p(o) \| \tilde{p}(o)]}_{\text{Divergence Objective}} &= \mathbb{E}_{p(o)} \left[\ln \frac{\int ds p(o,s)}{\tilde{p}(o)} \right] \\
&= \mathbb{E}_{p(o)} \left[\ln \frac{\int ds p(o,s) q(s \mid o) q(o,s)}{\tilde{p}(o) q(s \mid o) q(o,s)} \right] \\
&\geq \mathbb{E}_{p(o)} \left[\ln \frac{\int ds p(o,s) q(o,s)}{\tilde{p}(o) q(s \mid o) q(o,s)} \right] \\
&\geq \mathbb{E}_{p(o)} \left[\ln \frac{\int ds p(o,s) q(o \mid s) q(s)}{\tilde{p}(o) q(s \mid o) q(s \mid o) q(o)} \right] \\
&\geq \underbrace{\mathbb{E}_{p(o) q(s|o)}[\ln q(s) - \ln q(s \mid o) - \ln \tilde{p}(o)]}_{\text{EFE}} - \underbrace{\mathbb{E}_{p(o)} D_{\mathrm{KL}}[q(s \mid o) \| p(o,s)]}_{\text{VFE}} + \underbrace{\mathbb{E}_{q(s|o)p(o)} D_{\mathrm{KL}}[q(s \mid o) \| q(s)]}_{\text{Information Gain}}
\end{aligned} \tag{85}$$

Similarly [19] expresses the EFE directly in terms of the divergence objective, below, where it is revealed that EFE provides an upper bound on the divergence objective under the condition that the information gain term is greater than the marginal entropy term. This yields the behaviour that when the information about the world that remains to be discovered is greater than the entropy of the agent's observations, that the EFE minimises a bound on the divergence objective, which thus yields greater amounts of exploratory behaviour in the active inference agent [19]. However, as the EFE does not provide a straightforward relationship with either the evidence or divergence objective (and instead oscillates between upper and lower bounds), its mathematical origins and behavioural origins remain unclear [1].

$$\begin{aligned}
\underbrace{D_{\mathrm{KL}}[p(o) \| \tilde{p}(o)]}_{\text{Divergence Objective}} &= \mathbb{E}_{q(s|o)p(o)} D_{\mathrm{KL}}[p(o) q(o,s) \| \tilde{p}(o) q(o,s)] \\
&= \mathbb{E}_{q(s|o)p(o)} D_{\mathrm{KL}}[p(o) q(o \mid s) q(s) \| \tilde{p}(o) q(s \mid o) q(o)] \\
&= \underbrace{\mathbb{E}_{q(s|o)p(o)}[\ln q(s) - \ln \tilde{p}(o) - \ln q(s \mid o)]}_{\text{EFE}} + \underbrace{\mathbb{E}_{p(o)} D_{\mathrm{KL}}[q(s \mid o) \| q(s)]}_{\text{Information Gain}} - \underbrace{\mathbb{H}[p(o)]}_{\text{Marginal Entropy}}
\end{aligned} \tag{86}$$

The Empowerment Objective for Exploration

The empowerment objective provides another means of exploration in RL. This objective as defined by [48] and, relies on the following definition of mutual information $\mathcal{I}(X;Y)$, between a 'sender' random variable X and a 'receiver' random variable Y:

$$\mathcal{I}(X;Y) = \sum_{\mathcal{X},\mathcal{Y}} p(y \mid x)p(x) \log_2 \frac{p(y \mid x)}{\sum_{\mathcal{X}} p(y \mid x)p(x)} \tag{87}$$

Mutual information is defined as the amount of information, measured in bits, that the received signal (for instance from observations) contains about the transmitted signal (for example from hidden states) [48]. The empowerment objective for exploration as seen in [19] is:

$$\mathcal{L}_{\text{Empowerment}} \ = \underset{a_{t:T}}{\operatorname{argmax}} \, \mathcal{I}\left[s_{t:T}, a_{t:T} \mid a_{1:t}, s_{1:t}\right] \tag{88}$$

An agent, with an empowerment objective, maximises the amount of information about the future that is contained in the present and past [19]. This objective encourages the agent to select actions such that future dynamics are predictable, given the information about the past. This also encourages the agent to allow for a broad future and thus can be seen to hold a relationship to the divergence objective (when this is extended to include actions and latent state variables). In fact the empowerment objective, as described by [19] below, provides one component of the divergence objective. The notation used by [19] in this decomposition of the divergence objective below, is simplified such that $p\left(o_{1:t}\right) = p\left(o_{<}\right)$and $p\left(o_{t:T}\right) = p\left(o_{>}\right)$:

$$\begin{aligned}
&D_{\mathrm{KL}}\left[p\left(o_{1:T}\right) \| \tilde{p}\left(o_{1:T}\right)\right] \\
=&D_{\mathrm{KL}}\left[p\left(o_{1:T}, s_{1:T}, a_{1:T}\right) \| p\left(s_{1:T}, a_{1:T} \mid o_{1:T}\right) \tilde{p}\left(o_{1:T}\right)\right] \\
=&D_{\mathrm{KL}}\left[p\left(o_{>} \mid s_{>}\right) p\left(s_{>}, a_{>} \mid s_{<}, a_{<}\right) p\left(s_{<}, a_{<} \mid o_{<}\right) p\left(o_{<}\right) \| p\left(s_{>}, a_{>} \mid o_{>}\right) p\left(s_{<} \mid o_{<}, s_{>}, a_{>}\right) \tilde{p}\left(o_{>}\right) \tilde{p}\left(o_{<}\right)\right] \\
=&\underbrace{\mathbb{E}_{p(s_{>}|a_{>})} D_{\mathrm{KL}}\left[p\left(o_{>} \mid s_{>}\right) \| \tilde{p}\left(o_{>}\right)\right]}_{\text{Future Divergence}} - \underbrace{\mathbb{E}_{p(s_{<},a_{<},o_{<})} D_{\mathrm{KL}}\left[p\left(s_{>}, a_{>} \mid o_{>}\right) \| p\left(s_{>}, a_{>} \mid s_{<}, a_{<}\right)\right]}_{\text{Generalised Empowerment}} \\
&- \underbrace{\mathbb{E}_{p(o_{>},s_{<},a_{<})} D_{\mathrm{KL}}\left[p\left(s_{<}, a_{<} \mid o_{<}, s_{>}, a_{>}\right) \| p\left(s_{<}, a_{<} \mid o_{<}\right)\right.}_{\text{Latent Filtering Information}} + \underbrace{D_{\mathrm{KL}}\left[p\left(o_{<}\right) \| \tilde{p}\left(o_{<}\right)\right]}_{\text{Past Divergence}}
\end{aligned} \tag{89}$$

A simplification of the (above) past divergence term involves noting that, the prior distribution at all times t is fixed such that $p\left(o_{<}\right) = \delta\left(o_{<} = \hat{o}_{<}\right)$, where $\hat{o}_{<}$ refers to the sequence of realised observations [19]. Hence the past divergence term becomes an entropy term of the past desire distribution and, since this is constant then the past divergence term vanishes completely. Elsewhere, the future divergence term supplies the reward-seeking behaviour (or extrinsic value term) of an RL agent, since this minimises the divergence between future preferences and future predicted observations. The generalised empowerment term[14] is maximised, to maximise the mutual information between future actions and states, given expected observations and, the 'prior' expectation of future actions and states. According to [19] this term underpins the exploration of state, action and parameter space as it encourages the agent to seek observations which allow for maximal predictability (and control) of the future. This affords the agent an *empowerment* as it takes actions which lead to a maximally controllable future environment but also to an environment that is different from the expected future states and actions, given the past states and actions [19].

Lastly, the *latent filtering information* term provides a complement of the empowerment objective, this tries to maximise the information about the past, given the future, meaning that it encourages the agent to explore the future so as to better understand the past. Therefore the agent selects actions which assist in its understanding of the past. Interestingly, [19] note that this term encourages the agent to 'discover future consequences of its past actions'. They note that this may encourage the agent to build a *causal* model of its actions such that the effects of past actions can be estimated, whereby future action choices can verify or disprove these estimates. This provides an interesting guide for the agent's exploration, when using a divergence objective.

[14] ... which makes use of the following definition of mutual information $\mathcal{I}(X;Y) = \mathbb{E}_Y\left[D_{\mathrm{KL}}\left(p(X \mid Y) \| p(X)\right)\right]$.

5.4 A Framework For Variational Objective Functionals for Control

To conclude this section (on the mathematical origins of exploration) we highlight a taxonomy for variational objective functions, which can be applied to a broad definition of the RL problem, namely the problem of an agent making decisions in an uncertain environment in order to achieve its goal. This section is developed by [1] and establishes a three dimensional taxonomy for the types of objective functional that we have been discussing. The three orthogonal dimensions include firstly, whether an evidence or divergence functional is used and secondly, whether value is encoded exogenously or endogenously. The last dimension classifies objective functionals depending on the type of generative model chosen for the objective.

As the last section elaborated on evidence and divergence objectives, here we elaborate on the other two dimensions of the taxonomy. In the comparison of the control as inference and active inference frameworks it emerged that these two frameworks differ not only due to the objectives used (FEF vs EFE) but also because of how value was encoded. Active inference encoded value in an *endogenous* way, via the biased generative model whereas, control as inference encoded value *exogenously*, via the addition of optimality nodes onto the (PO)MDP structure. Moreover, the last dimension of the taxonomy specifies the type of generative model used. By this we refer to whether latent states are encoded in the model or whether the model encodes various types of parameters [1].

Crucially, [1] mentions that all objectives appearing in this discussion are *mean-field* variational objective functionals . This refers to the objectives being able to be split into a number of independent objectives for each time-step of a trajectory, which can all be optimised independently. This assumption, of course, applies for the case of RL agents, for which it is a necessary precondition in order to make use of the Bellman equations. The taxonomy, provided by [1], affords mathematical heuristics to provide 'any given functional for variational control'. Furthermore, this provides an understanding of the objectives' decompositions and behaviour. This incorporates our understanding of the various exploratory behaviours that these functionals embed in the agent.

5.4.1 Exogenous vs Endogenous Value

[1] explores the design choices associated with encoding the agent's goals. Here, we discuss the first possible design choice (and first dimension of the taxonomy), whether to encode the agent's goals as exogenous variables which provide an additional variable in the process of inference or, whether to encode the goals endogenously by biasing one (or more) of the distributions of the agent's generative model.[15]

Maximum-Entropy RL and Exogenous Value

As we have discussed, the augmentation of the POMDP structure with exogenous binary variables representing optimality achieved a framework for framing the control problem as one of inference [18]. This resulted in four terms, as seen in equation 73. These appear again below for convenience, where we have also expanded the generative model in the second line for clarity [1]:

[15] The author of [1] also notes that a second design choice is available, whereby the goals can be encoded in the variational distribution of the agent instead of in its generative model. The author notes that the resulting algorithms provide an unexplored area in the literature.

$$
\begin{aligned}
\mathcal{L}_{CAI} &= D_{\mathrm{KL}}\left(q\left(s_t, a_t\right) \| p\left(s_t, a_t, o_t, \Omega_t\right)\right) \\
&= D_{\mathrm{KL}}\left(q\left(a_t \mid s_t\right) q\left(s_t \mid o_t\right) \| p\left(\Omega_t \mid s_t, a_t\right) p\left(o_t \mid s_t\right) p\left(a_t \mid s_t\right) p\left(s_t \mid s_{t-1}, a_{t-1}\right)\right) \\
&= \underbrace{-\mathbb{E}_{q\left(s_t, a_t\right)}\left[\ln p\left(\Omega \mid s_t, a_t\right)\right]}_{\text{Extrinsic Value}} + \underbrace{D_{\mathrm{KL}}\left(q\left(s_t\right) \| p\left(s_t \mid s_{t-1}, a_{t-1}\right)\right)}_{\text{State divergence}} \\
&+ \underbrace{\mathbb{E}_{q\left(s_t\right)}\left[D_{\mathrm{KL}}\left(q\left(a_t \mid s_t\right) \| p\left(a_t \mid s_t\right)\right)\right]}_{\text{Action Divergence}} - \underbrace{\mathbb{E}_{q\left(s_t, a_t\right)}\left[\ln p\left(o_t \mid s_t\right)\right]}_{\text{Observation Ambiguity}}
\end{aligned} \tag{90}
$$

As seen previously, the extrinsic value term can correspond to a reward maximisation if $\ln p\left(\Omega_t \mid s_t, a_t\right) = r\left(s_t, a_t\right)$. The observation ambiguity term (which arises in the POMDP setting) incentivises the agent to seek observations with a high expected likelihood, and low entropy. Crucially, this disincentivises the agent from exploration as the agent seeks well characterised regions of its search space (that minimise ambiguity) [1]. As previously discussed, the state divergence term acts as a regulariser, keeping the state posterior close to the prior states, which are expected under the generative model. If the transition model is learnt this again confines the agent to prioritise transitions with known dynamics [1]. Lastly, the action divergence term provides the agent with the objective to match its prior and posterior action distributions. However, the action prior is often set to be uniform this yields a maximum entropy term for the agent's variational policy. This can be seen below, where in the MDP setting the state divergence term vanishes [1]:

$$
\mathcal{F}_{\text{max ent}} = -\mathbb{E}_{q\left(s_t, a_t\right)}\left[\ln p\left(\Omega_t \mid s_t, a_t\right)\right] - \mathbb{E}_{q\left(s_t\right)}\left[\mathcal{H}\left[q\left(a_t \mid s_t\right)\right]\right] \tag{91}
$$

As discussed, this provides the agent with a motive for exploration, via random action selection. Interestingly, the addition of the observation ambiguity and state divergence terms, in the POMDP setting, incentivise the agent to exhibit behaviour that confines itself to regions of the search space that are as close to an MDP as possible [1].

Active Inference Agents and Endogenous Value

The difference between the addition of exogenous variables, to encode value, and an endogenous encoding can be viewed as follows [1]. The process of inference with the exogenous variables outputs the likely and unbiased trajectories for a given series of actions. Then, the trajectories are shifted to converge on the agent's goal by conditioning on the exogenous optimality variables. The agent then infers the action consistent with the most optimal trajectories. The process of inference for the endogenous variables and for active inference contrasts to this. Instead of outputting unbiased trajectories from the generative model, the active inference agent encodes value in a biased generative model, which outputs biased trajectories which can be used to infer actions [1]. The agent's variational free energy is optimised [1]:

$$
\begin{aligned}
\mathcal{F}_{\text{ActInf}} &= D_{\mathrm{KL}}\left(q\left(s_t, a_t \mid o_t\right) \| \tilde{p}\left(o_t, s_t, a_t\right)\right) \\
&= D_{\mathrm{KL}}\left(q\left(a_t \mid s_t\right) q\left(s_t \mid o_t\right) \| \tilde{p}\left(o_t \mid s_t\right) p\left(s_t \mid s_{t-1}, a_{t-1}\right) p\left(a_t \mid s_t\right)\right) \\
&= -\underbrace{\mathbb{E}_{q\left(s_t \mid o_t\right)}\left[\ln \tilde{p}\left(o_t \mid s_t\right)\right]}_{\text{Extrinsic Value}} + \underbrace{\mathbb{E}_{q\left(s_t \mid o_t\right)}\left[D_{\mathrm{KL}}\left(q\left(a_t \mid s_t\right) \| p\left(a_t \mid s_t\right)\right)\right]}_{\text{Action Divergence}} + \underbrace{D_{\mathrm{KL}}\left(q\left(s_t \mid o_t\right) \| p\left(s_t \mid s_{t-1}, a_{t-1}\right)\right)}_{\text{State Divergence}}
\end{aligned} \tag{92}
$$

As previously discussed, if reward is encoded in the biased prior preference distribution such that $\ln \tilde{p}(o|s) \propto r(s,a)$, then this is equivalent to the control as inference objective apart from the observation ambiguity term. As discussed, this term is missing since the encoding of preferences in terms of observations (instead of the addition of optimality variables) removes a degree of freedom from the active inference agent framework. Hence the active inference agents effectively bias the observation ambiguity term to favour its preferences

[1].[16]

The resulting advantage of the exogenous variable approach is that these types of agents are able to enforce a conservative bias which may be useful in environments where exploration is costly, environments where optimal policies are easy to find or, in offline RL, where venturing outside of the training distribution can have negative effects on the policy [1]. The author also notes the philosophical difference between the two approaches, suggesting that the exogenous approach maintains the modularity thesis where perception and action are kept separate. In contrast, an endogenous encoding of value selects actions through a process of biased perception. This highlights the difference in the philosophies behind RL and active inference according to [1]. While RL arises from a 'collectivist and representational tradition' in AI, which holds principled modularity in high esteem, active inference arises from a 'heavily embodied and enactivist' viewpoint which is influenced by the dynamical systems theory, included in the FEP, whereby systems are seen in terms of an action-perception loop, without modularity between subsystems [1].

5.4.2 Generative Model Choices

Finally, we discuss the third orthogonal dimension of the taxonomy for variational objectives for optimal control. The first two dimensions asked whether an evidence or divergence objective was used and whether value was encoded exogenously or endogenously. This third dimension concerns how the generative model is defined [1]. One aspect of this dimension is whether the model concerns an MDP or a POMDP and hence, whether latent (hidden) variables are included. For instance, the FEEF objective defined in terms of only an MDP is [1]:

$$\begin{aligned}\mathbb{FEEF}_{MDP} &= D_{\mathrm{KL}}\left(q\left(s_t, a_t\right) \| \tilde{p}\left(s_t, a_t\right)\right) \\ &= \underbrace{D_{\mathrm{KL}}\left(q\left(s_t\right) \| \tilde{p}\left(s_t\right)\right)}_{\text{Extrinsic Value}} + \underbrace{\mathbb{E}_{q(s_t)}\left[D_{\mathrm{KL}}\left(q\left(a_t \mid s_t\right) \| p\left(a_t \mid s_t\right)\right)\right]}_{\text{Action Divergence}}\end{aligned} \tag{93}$$

This omits the intrinsic value term as no information gain is possible as states are directly observed [1]. Hence exploration can occur through a random entropy term (from setting the prior action distribution to be uniform).

A second aspect of this taxonomic dimension is a choice of which variables are included into the generative model [1]. As we have seen the generative model can include parameters in addition to states, observations, actions and, binary optimality variables. Hence, it is possible to include hierarchies of (latent) variables in the generative model and generate corresponding variational objectives for control. The question is then: what additional behaviour do these inclusions generate in the agent and, what additional terms appear in the corresponding objectives? Take for example the inclusion of the parameters θ, which parameterise the state transition model $p(s_t|s_{t-1}, a_{t-1}; \theta)$ then, [1] denotes the variational free energy objective function which includes these parameters as:

$$\begin{aligned}\mathcal{F}_\theta &= D_{\mathrm{KL}}\left(q\left(s_t, a_t, \theta_t \mid o_t\right) \| \tilde{p}\left(o_t, s_t, a_t, \theta_t\right)\right) \\ &= D_{\mathrm{KL}}\left(q\left(a_t \mid s_t\right) q\left(\theta_t \mid s_t\right) q\left(s_t \mid o_t\right) \| \tilde{p}\left(o_t \mid s_t\right) p\left(a_t \mid s_t\right) p\left(s_t \mid s_{t-1}, a_{t-1}, \theta_t\right) p\left(\theta_t\right)\right) \\ &= -\underbrace{\mathbb{E}_{q(s_t|o_t)}\left[\ln \tilde{p}\left(o_t \mid s_t\right)\right]}_{\text{Extrinsic Value}} + \underbrace{\mathbb{E}_{q(s_t|o_t)}\left[D_{\mathrm{KL}}\left(q\left(a_t \mid s_t\right) \| p\left(a_t \mid s_t\right)\right)\right]}_{\text{Action Divergence}} \\ &\quad + \underbrace{\mathbb{E}_{q(\theta_t)}\left[D_{\mathrm{KL}}\left(q\left(s_t \mid o_t\right) \| p\left(s_t \mid s_{t-1}, a_{t-1}, \theta\right)\right)\right]}_{\text{State Divergence}} + \underbrace{D_{\mathrm{KL}}\left(q\left(\theta_t \mid s_t\right) \| p\left(\theta_t\right)\right)}_{\text{Parameter Divergence}}\end{aligned} \tag{94}$$

[16] The agent's preferences can also be introduced by biasing the state distribution $p(s|o)$ additionally or, the state prior $p(s)$ alone [1].

The above equation demonstrates that the inclusion of the parameters in the generative model embeds a parameter divergence term in the agent's objective function. As previously discussed, this is a regularisation term which acts to keep the parameter posterior close to the parameter prior, thus disincentivising exploration in the parameter space. The addition of the parameter divergence term provides insight as to how the addition of variables in the generative model affects the VFE objective function. In general, we see that this results in the inclusion of a divergence term which penalise the deviations between posteriors and priors [1].

Similarly, for the case of the FEEF objective function, including additional variables in the generative model results in additional information gain terms (along with the divergence terms). This can be seen, for the case of the inclusion of parameters, below [1]:

$$\begin{aligned}
\mathbb{FEEF}_{\theta} &= D_{\mathrm{KL}}\left(q\left(o_t, s_t, a_t, \theta_t\right) \| \tilde{p}\left(o_t, s_t, a_t, \theta_t\right)\right) \\
&= D_{\mathrm{KL}}\left(q\left(o_t \mid s_t\right) q\left(a_t \mid s_t\right) q\left(s_t \mid \theta\right) q\left(\theta_t\right) q\left(s_t \mid o_t\right) q\left(\theta_t \mid s_t\right)\right. \\
&\quad \| \tilde{p}\left(o_t \mid s_t\right) p\left(a_t \mid s_t\right) p\left(s_t \mid s_{t-1}, a_{t-1}, \theta_t\right) p\left(\theta_t\right) q\left(s_t \mid o_t\right) q\left(\theta_t \mid s_t\right)) \\
&= \underbrace{\mathbb{E}_{q\left(a_t, s_t\right)}\left[D_{\mathrm{KL}}\left(q\left(o_t \mid s_t\right) \| \tilde{p}\left(o_t \mid s_t\right)\right)\right]}_{\text{Extrinsic Value}} + \underbrace{\mathbb{E}_{q\left(o_t, s_t\right)}\left[D_{\mathrm{KL}}\left(q\left(a_t \mid s_t\right) \| p\left(a_t \mid s_t\right)\right)\right]}_{\text{Action Divergence}} - \underbrace{\mathbb{E}_{q\left(o_t\right)}\left[D_{\mathrm{KL}}\left(q\left(s_t \mid o_t\right) \| q\left(s_t\right)\right)\right]}_{\text{Expected Information Gain}} \\
&\quad + \underbrace{\mathbb{E}_{q\left(o_t\right) q\left(\theta_t\right)}\left[D_{\mathrm{KL}}\left(q\left(s_t \mid o_t\right) \| p\left(s_t \mid s_{t-1}, a_{t-1}, \theta_t\right)\right)\right]}_{\text{Expected Posterior Divergence}} - \underbrace{D_{\mathrm{KL}}\left(q\left(\theta_t \mid s_t\right) \| q\left(\theta_t\right)\right)}_{\text{Parameter Information Gain}} + \underbrace{D_{\mathrm{KL}}\left(q\left(\theta_t \mid s_t\right) \| p\left(\theta_t\right)\right)}_{\text{Parameter Divergence}}
\end{aligned} \tag{95}$$

And lastly, for the case of the posterior divergence objective, adding additional variables results in only additional information gain terms (without the regularising divergence term) [1]. This is seen below for the case of the entropy term embedded in the divergence objective:

$$\begin{aligned}
\mathbb{E}_{q\left(o_t\right)}\left[\ln q\left(o_t\right)\right] &= \mathbb{E}_{q\left(o_t, s_t, \theta_t\right)}\left[\ln q\left(o_t\right)\right] \\
&= \mathbb{E}_{q\left(o_t, s_t, \theta_t\right)}\left[\ln \frac{q\left(o_t, s_t, \theta_t\right)}{q\left(s_t, \theta_t \mid o_t\right)}\right] \\
&= \mathbb{E}_{q\left(o_t, s_t, \theta_t\right)}\left[\ln \frac{q\left(o_t \mid s_t\right) q\left(s_t \mid \theta_t\right) q\left(\theta_t\right)}{q\left(s_t \mid o_t\right) q\left(\theta_t \mid s_t\right)}\right] \\
&= -\underbrace{\mathcal{H}\left[q\left(o_t \mid s_t\right)\right]}_{\text{Likelihood Entropy}} - \underbrace{\mathbb{E}_{q\left(o_t\right) q\left(\theta_t\right)}\left[D_{\mathrm{KL}}\left(q\left(s_t \mid o_t\right) \| q\left(s_t \mid \theta_t\right)\right)\right]}_{\text{Expected State Information Gain}} - \underbrace{D_{\mathrm{KL}}\left(q\left(\theta_t \mid s_t\right) \| q\left(\theta_t\right)\right)}_{\text{Parameter Information Gain}}
\end{aligned} \tag{96}$$

Section Summary

In this section we have uncovered the mathematically principled origins for the intrinsic information gain motive, which appears in the EFE objective of active inference agents. This information gain term arises from divergence objectives which seek to match predicted and desired distributions. This involves a process of entropy maximisation of future observations, which induces the exploration in the latent variable space (in general for all latent variables added to the generative model). Therefore, agent's possessing a divergence objective can be interpreted as desiring a broad future, which requires exploration to update upon the resolvable uncertainty in the environment.

In contrast, we saw that the evidence objective was a preference-maximising objective. This induces an 'anti-exploratory' behaviour in the agent. As this agent desires a distribution with low entropy, with a narrow and precise future that is predictable and also maximally conforms to the agent's goals, it learns as little as possible in order to achieve these goals. Only the extent of uncertainty in the environment and in the reward function and also, the lack of controllability of the environment, prevent the agent from full maximisation which would result in a Dirac delta distribution around the future reward [1].[17]

The evidence objective revealed the reason for a lack of information gain term in the control as inference objective. Discussion here acted to illuminate the difference between the control as inference and active inference frameworks. This occurred not only through the difference in objective function (FEF vs EFE), but also through the encoding of value, in terms of exogenous optimality variables or, in an endogenous way, via the biasing of the active inference agent's generative model. This contrast, along with the evidence vs divergence contrast offered two dimensions of a taxonomy for variational objectives, with the last dimension involving understanding the effect of different definitions of the generative model. This taxonomy offers an area for future work, to investigate the impact of the design choices empirically, and specifically in challenging and sparse reward environments, as was briefly discussed with the active inference agent implementations of section 4.2 [1].

Lastly, we note that, the evidence objective (which the control as inference objective bounds) offers the insight that this reward-maximising agent cannot have an intrinsic exploratory drive since the nature of its objective compels it to minimise information gain [1]. However, [1] notes that it is possible to derive a form of information gain exploration from the reward maximisation, via the calculation of the value of information in terms of reward. Two heuristically driven approaches, thereof, are the upper-confidence bound [40] and Thompson sampling [41] methods of exploration.[18] In the context of the developed taxonomy we can now understand the use of any of the above variational objectives as heuristically driven design choices, although within the intuitively and mathematically principled description of the taxonomy.

6 A Taxonomy of Exploratory Methods

To conclude the report we return to a discussion of a second taxonomy for exploratory methods in RL which we saw in figure 2, in section 2.2.1. This taxonomy, arises from a survey of exploration in deep RL [20], which discusses both multi-agent and single-agent exploratory methods. In this section we confine our discussion to single-agent exploration and review the single-agent component of the taxonomy and some corresponding examples of exploratory methods. This can be seen in figure 15. The taxonomy identifies two categories of exploratory method. These concern uncertainty-oriented methods and, intrinsically motivated methods. This taxonomy offers us a second glance at a means of understanding where the active inference exploratory drive fits into a broader scheme of exploratory methods. This also offers a contrast to the previous taxonomy for variational objectives which took a more holistic approach as it included modelling choices for encoding value and design choices for the generative model.

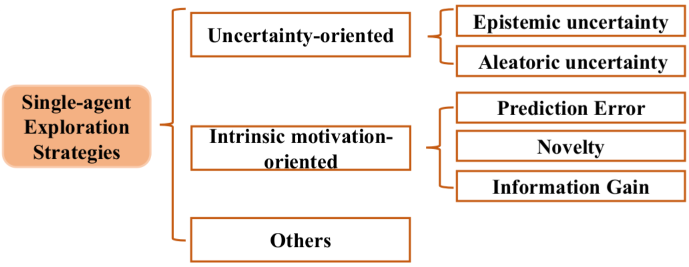

Figure 15: A taxonomy of exploration methods for deep single-agent RL as found in [20].

As the figure depicts, [20] discusses two categories of exploratory methods. The first of these, uncertainty-oriented methods, looks at optimistically biasing the objectives of the agent to explore its search space.

[17] ... as seen in figure 14.

[18] Upper confidence bound and Thompson sampling exploration was reviewed in section 2.2.1.

Uncertainty-oriented methods leverage the agent's quantification of both aleatoric and epistemic uncertainties to derive a means of efficient exploration. The objectives here, can help the agent explore areas with high epistemic uncertainty or even avoid areas with high aleatoric uncertainty (in addition) [20]. Here, the epistemic uncertainty is often modelled via a Bayesian posterior and, the aleatoric uncertainty via distributional value functions. This category includes, for instance, UCB and Thompson sampling.[19] Conversely, intrinsic motivation-oriented methods make use of a variety of heuristics and, often make use of reward-agnostic information to provide the agent with an *intrinsic* motive for exploration [20]. This category is subdivided into methods that focus on prediction error, novelty and information gain. This supplies us with a categorisation for the intrinsic value term within the EFE functional.[20]

6.1 Uncertainty-Oriented Methods

As mentioned, uncertainty-oriented methods leverage the agent's measurement of uncertainty to provide an exploratory drive [20]. There are two types of measurement, which [20] uses to categorise methods. These concern epistemic and aleatoric uncertainty.

First, epistemic uncertainty concerns the accuracy of the agent's knowledge of the environment. This often concerns the agent's model uncertainty. With epistemic uncertainty-oriented methods, the agent is often afforded an optimistic bias, whereby it is encouraged to explore areas of its search space with high epistemic uncertainty [20]. This often involves the approximation of a posterior distribution for the agent's value function, from which an uncertainty measurement can be drawn. The second type of uncertainty-oriented method, discussed by [20], concerns aleatoric uncertainty. This refers to the intrinsic uncertainty and randomness of the environment. Being optimistic with aleatoric uncertainty can damage the agent's performance since this implies favouring actions with inherently high variance. Hence, the agent can make use of an estimate of aleatoric uncertainty in order to avoid areas of the search-space which are noisy. [20] note that aleatoric uncertainty can be captured by a reward distribution. Next, we discuss the examples of UCB and Thompson sampling to illustrate the use of uncertainty-oriented methods for exploration [20].

Recall, that the upper-confidence bound (UCB) [40] method of exploration greedily selected actions based on an upper confidence bound on the Q value function. This invokes a principle of optimism as it implies an optimistic view of the extent of the uncertainty measurement in the Q value. The uncertainty measurement can be computed using the expression in section 2.2.1 or via a posterior distribution for the Q function. Viewing this as an uncertainty-oriented method gives the below [20]:

$$\begin{gathered} a_t=\arg\max_a Q^{+}\left(s_t, a\right) \text{ s.t. } \\ Q^{+}\left(s_t, a_t\right)=Q\left(s_t, a_t\right)+\text { Uncertainty }\left(s_t, a_t\right) \end{gathered} \tag{97}$$

Secondly, Thompson sampling [41] exploration incorporates the uncertainty via sampling from the Q value function's posterior distribution.[21] Recall Thompson sampling corresponds to [20]:

$$a_t=\arg\max_a Q_\theta\left(s_t, a\right), \quad Q_\theta \sim \text { Posterior }(Q) \tag{98}$$

[19] ... which were explored in section 2.2.1.

[20] We note that [20] also provide a comprehensive empirical study of the many exploratory methods, however this is not discussed here.

[21] [20] note that Thompson sampling offers a benefit as compared to UCB. This results from the use of the same Q value function throughout an episode, as opposed to performing optimistic action-selection in each time step. This affords the agent a temporal consistency of exploration and has advantages in tasks with a long time horizon.

6.2 Intrinsic Motivation-Oriented Methods

The second (broad) category of exploratory methods for single-agent RL, discussed by [20], concerns intrinsic motivation. This has been discussed at length in terms of the intrinsic value term which affords the active inference agent a means of *directed* exploration which facilitates its information-gain. We have also discussed that the origin of such a term mathematically arises from divergence objectives. In this (sub)section we study how the authors of [20] categorise such a method of exploration into their taxonomy.

As we have seen, intrinsic motivations for exploration often form an exploratory bonus (via the agent's objective function), which is analogous to assigning the agent an inherent reward for exploration. This exploratory bonus is usually agnostic to the measurement of the agent's reward or goals which defines why this method offers an *intrinsic* bonus. Three types of intrinsic motivation are defined in [20]. The first of these concerns methods that estimate prediction errors of the environmental dynamics. The second concerns an estimate of the novelty of visiting a state. Lastly, methods based on information gain are discussed.

Prediction Error Methods

Prediction error methods encourage the agent to explore areas of its search space which have high prediction errors [20]. Here, the intrinsic motivation is provided by the extent of the prediction error for the next state. This is defined by the following distance measure, where ϕ can be seen as a mapping from observations to latent states and, $\hat{f}$ predicts the next latent state, given the current (modelled) latent state and action [20]:

$$\text{Prediction Error}\left(s_t, s_{t+1}\right) = \text{dist}\left(\phi\left(s_{t+1}\right), \hat{f}\left(\phi\left(s_t\right), a_t\right)\right) \tag{99}$$

An example, provided by [20], of a prediction error intrinsic motivation-oriented method is exploration with mutual information (EMI) [49]. This method learns a mapping ϕ for both states and actions such that it holds $\phi_s(s)$ and $\phi_a(a)$. EMI combines these in a linear dynamics model, along with an error model which is a mapping from states and actions to the 'irreducible error' such that, $S_\gamma : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ [49]. The state and action latent spaces are trained via the maximisation of mutual information[22] $I\left(\left[\phi_s(s); \phi_a(a)\right]; \phi_s\left(s'\right)\right)$ and $I\left(\left[\phi_s(s); \phi_s\left(s'\right)\right]; \phi_a(a)\right)$. These are trained using a variational lower bound on MI [20]. Lastly, the intrinsic motivation term is computed from the prediction error of the linear dynamics model to present an intrinsic reward. This is, where s' represents the state following s [49]:

$$r_e\left(s_t, a_t, s_t'\right) = \left\|\phi_s\left(s_t\right) + \phi_a\left(a_t\right) + S\left(s_t, a_t\right) - \phi_s\left(s_t'\right)\right\|^2 \tag{100}$$

Novelty Methods

The second (sub)category under the intrinsic motivation-oriented exploration encourages agents to visits states that they have not been to frequently [20]. This involves assigning an agent an intrinsic bonus, which is inversely proportional to the number of visits to a state s [20]:

$$\text{Novelty}\left(s_t\right) = \frac{1}{N\left(s_t\right)} \tag{101}$$

This was previously implemented in the form of UCB given by [6] which we studied in section 2.2.1. As maintainting a count of the number of visits to states is difficult for large or continuous state spaces, another

[22] Mutual information was defined in section 5.3.

method for measuring novelty is to estimate the state novelty as a measure of distance between the current state s_t and, the states usually visited. This appears below where $\mathcal{B}$ is a distribution of recently visited states [20]:

$$\text{Novelty}\left(s_t\right)=\mathbb{E}_{s^{\prime} \in \mathcal{B}}\left[\operatorname{dist}\left(s_t ; s^{\prime}\right)\right] \tag{102}$$

Otherwise, a ‘pseudo-count’ can be defined to generalise a count of state-visits. This is proposed by [20] to use state probability models $p(s_t)$, where $p'(s_t)$ computes the probability of observing s_t after a new occurrence of s_t. The pseudo-count then appears below [20], where the intrinsic motivation appears as in equation 101:

$$\hat{N}\left(s_t\right)=\frac{p\left(s_t\right)\left(1-p^{\prime}\left(s_t\right)\right)}{p^{\prime}\left(s_t\right)-p\left(s_t\right)} \tag{103}$$

Information Gain Methods

Finally, the last (sub)category of intrinsic exploration, discussed by [20], is that of information gain which directs agents to explore areas of their search space, explicitly to resolve uncertainty. Information gain is defined by [20] as the reduction in uncertainty about environmental dynamics:

$$\text{Information Gain}\left(s_t, s_{t+k}\right)= \text{ Uncertainty }_{t+k}(\theta)- \text{ Uncertainty }_t(\theta) \tag{104}$$

This provides a category in the taxonomy of [20] for the exploration of the EFE and active inference agents.

Thus, the study of [20] has provided a second taxonomy of methods for exploration in RL, in which we can categorise the active inference agent’s exploratory drive. The taxonomy is, however, incomplete. As seen in figure 15, a number of exploratory methods cannot be categorised as either uncertainty-oriented or, as intrinsic motivation-oriented methods. This leads to a branch of their taxonomy with the label ‘other’. The authors of [20] identify other exploratory methods such as the empowerment objective [48], which was discussed in section 5.3, that they classify in this ‘other’ taxonomic category. Hence, this taxonomy offers to provide additional perspective on our classification of the information gain term. The authors of [20], additionally, provide a useful empirical (broad) survey of exploratory methods (which was not the focus of our discussion). Read in tandem with their taxonomy, they identify some key characteristics of the two taxonomic categories. These characteristics are based on an empirical study of the exploratory methods surveyed, in three benchmark RL tasks.

The authors of [20] identify that uncertainty-oriented methods struggle in sparse reward environments, as compared to intrinsic motivation-oriented methods. The authors note that this is because uncertainty-oriented methods rely on their value function estimation, which is hard to learn in environments with sparse feedback. However, despite the benefits of intrinsic-motivation methods in challenging tasks, the authors suggest that these may hinder the performance of the agent in some other tasks. They suggest this may be due to the trade-off in the objective functions of intrinsic motivation-oriented agents, as the intrinsic motivation may deviate the agents from optimal policies. However as we have seen, in our empirical studies of active inference agents, their particular information gain term serves to allow the agent to adaptively select actions, which, as per the empirical demonstrations in section 4.2, yields strong performance as compared to other contemporary RL algorithms (without hindering the agent).

7 Conclusions

Within the context of the RL problem and exploration-exploitation trade-off, the active inference agent holds an intrinsic, exploratory drive that is directed for information gain. It is this drive, in this context, which motivated our study. We sought to understand the mathematically principled origins of the intrinsic exploratory drive and, to understand the benefits of implementing such a motive. Lastly, we sought to categorise such a drive in a taxonomic scheme, for comparison to other methods of exploration in RL.

To achieve this, we began by establishing the RL problem and exploration-exploitation dilemma. A discussion of basic methods of exploration in RL provided some background for existing methods to resolve this dilemma. Some of these (basic methods) employed random exploration, which we discussed to be necessarily inefficient. Additionally, the employment of such methods was usually based on ad-hoc heuristics. In contrast, we noted that the active inference agent's exploratory drive displayed *direction*, to mitigate the inefficiency of random exploration. Moreover, this exploratory drive held a principled basis, resulting from the FEP.

Hence, to understand some component of the origins of the EFE objective, and the active inference agent, we studied a derivation of the FEP. This revealed the ability to construct a VFE-minimising agent. A lightning introduction to the active inference agent preceded a discussion of empirical 'proof-of-concept' implementations. These demonstrated the benefit of the directed, intrinsic exploration, especially in sparse reward environments. Moreover, the active inference agent demonstrated a strong performance as compared to some contemporary RL agents such as SAC [30] and DQN [23]. This motivated an inquiry as to the mathematically principled origins of the information-gain term, which yielded the strong performance in sparse reward environments.

The *mathematical origins of exploration* section uncovered that, the choice of the EFE objective remains unclear. Furthermore, the intrinsic exploratory drive, embedded in this term seems to appear by choice of construction. In contrast, an alternative objective called the FEF (for action selection of active inference agents) does not hold this information gain term. Thus, we asked the question of where such a term arises and found that, such a term arises from the entropy-maximising divergence objective.

The divergence objective posed a probability-matching method for the achievement of the agent's preferences. This is particularly useful for situations where the agent holds a complex distribution of preferences. The divergence objective desires a maximum-entropy, broad future. This is balanced with the achievement of the agent's goals. In order to allow for this broad future, and in order to account for the complex preferences of the agent, the divergence objective requires the information gain term, to direct the agent's learning about the environment. In contrast, evidence objectives seek to achieve the agent's goals, whilst learning as little as possible about the environment.

The contrast between evidence and divergence objectives adds to our perspective about types of variational objective functions. This lead to a taxonomic discussion about categorising such objectives. Other than comparing the evidence vs divergence objectives, the taxonomy holds two other orthogonal dimensions for categorisation. A second comparison is drawn between objectives which encode value via exogenous variables and, objectives which do this via an endogenous biasing of the generative model. This aligns to a comparison of the *control as inference* framework [18], which is compared to the active inference framework. Here we see that the control as inference framework displays exploration, via maximum entropy RL. Lastly, the taxonomy discusses the design choices for the generative model. This allows for the (flexible) inclusion of both POMDPs and MDPs and also, for the inclusion of various hierarchies of latent variables.

Finally, a review of a second taxonomy (for exploratory methods in RL) offers an additional means of categorising the information gain motive, which we had uncovered to arise from divergence objectives. Overall, this affords us an additional discussion of some other exploratory methods, along the way.

Hence, while this report reviews the active inference exploratory objective in detail, it leaves several interesting unanswered questions. Specifically, the taxonomy of variational objectives contains several objectives which have not been reviewed in the literature, according to [1]. This taxonomy offers mathematically principled heuristics for the selection of such objectives for applications. The flexibility and generality of such a scheme offers much material for empirical investigation. One aspect of flexibility is the ability to encode various complex preference distributions. This provides a challenge for RL agents. Hence, we note that while the empirical investigations, discussed in section 4.2, uncover some benefit for the active inference agent, a more exhaustive and rigorous empirical study is needed, to test the capabilities of the agent, using various choices of objective function. These tests require appropriate RL benchmark agents and RL benchmark environments. Perhaps the choices of environment, used in the survey of exploration in [20], provide a means for such a study.

Acknowledgements

Many thanks go to Dr Shock for his inspiring guidance in supervising this project. Dr Shock has afforded my classmates and I a freedom of exploration into topics of interest in RL. His class discussion groups have provided a stimulating environment for the sharing of ideas. Dr Shock's enthusiasm for learning and teaching has motivated me and my study. I am grateful for his support and for the learning opportunity he has given. I look forward to continuing our discussions in future.

References

[1] B. Millidge, "Applications of the Free Energy Principle to Machine Learning and Neuroscience," 2021. [Online]. Available: https://arxiv.org/abs/2107.00140v1

[2] R. C. Conant and W. R. Ashby, "Every good regulator of a system must be a model of that system," *Int. J. Systems Sci.*, vol. 1, no. 2, pp. 89–97, 1970.

[3] K. Friston, J. Daunizeau, and S. J. Kiebel, "Reinforcement learning or active inference?" *PloS one*, vol. 4, no. 7, p. e6421, July 2009. [Online]. Available: https://europepmc.org/articles/PMC2713351

[4] K. Friston, J. Kilner, and L. Harrison, "A free energy principle for the brain," *Journal of Physiology-Paris*, vol. 100, no. 1, pp. 70–87, 2006, theoretical and Computational Neuroscience: Understanding Brain Functions. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092842570600060X

[5] O. Berger-Tal, J. Nathan, E. Meron, and D. Saltz, "The Exploration-Exploitation Dilemma: A Multidisciplinary Framework," *PLOS ONE*, vol. 9, no. 4, pp. 1–8, 04 2014. [Online]. Available: https://doi.org/10.1371/journal.pone.0095693

[6] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. MIT Press, Cambridge, Massachusetts, 2018.

[7] L. D. Costa, N. Sajid, T. Parr, K. Friston, and R. Smith, "The relationship between dynamic programming and active inference: the discrete, finite-horizon case," 2020. [Online]. Available: https://arxiv.org/abs/2009.08111

[8] B. Millidge, "Deep Active Inference as Variational Policy Gradients," 2019. [Online]. Available: https://arxiv.org/abs/1907.03876

[9] R. Dubb, J. Shock, and M. Mavuso, "Deep Reinforcement Learning Methods For Scaling Active Inference," 2021, BSc. Honours thesis, University of Cape Town Department of Statistical Sciences.

[10] K. Friston, "A free energy principle for a particular physics," 2019. [Online]. Available: https://export.arxiv.org/abs/1906.10184

[11] K. Friston and P. Ao, "Free energy, value, and attractors," *Computational and mathematical methods in medicine*, vol. 2012, p. 937860, 2012. [Online]. Available: https://europepmc.org/articles/PMC3249597

[12] T. Parr and K. J. Friston, "The Anatomy of Inference: Generative Models and Brain Structure," *Frontiers in Computational Neuroscience*, vol. 12, p. 90, 2018. [Online]. Available: https://www.frontiersin.org/article/10.3389/fncom.2018.00090

[13] A. Tschantz, B. Millidge, A. K. Seth, and C. L. Buckley, "Reinforcement Learning through Active Inference," 2020. [Online]. Available: https://arxiv.org/abs/2002.12636

[14] ——, "Control as Hybrid Inference," 2020. [Online]. Available: https://arxiv.org/abs/2007.05838

[15] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "OpenAI Gym," 2016. [Online]. Available: https://arxiv.org/abs/1606.01540

[16] B. Millidge, A. Tschantz, and C. L. Buckley, "Whence the Expected Free Energy?" 2020. [Online]. Available: https://arxiv.org/abs/2004.08128

[17] B. Millidge, A. Tschantz, A. K. Seth, and C. L. Buckley, "On the Relationship Between Active Inference and Control as Inference," 2020. [Online]. Available: https://arxiv.org/abs/2006.12964

[18] S. Levine, "Reinforcement Learning and Control as Probabilistic Inference: Tutorial and Review," 2018. [Online]. Available: https://arxiv.org/abs/1805.00909

[19] B. Millidge, A. Tschantz, A. Seth, and C. Buckley, "Understanding the Origin of Information-Seeking Exploration in Probabilistic Objectives for Control," 2021. [Online]. Available: https://export.arxiv.org/abs/2103.06859

[20] T. Yang, H. Tang, C. Bai, J. Liu, J. Hao, Z. Meng, and P. Liu, "Exploration in Deep Reinforcement Learning: A Comprehensive Survey," 2021. [Online]. Available: https://arxiv.org/abs/2109.06668

[21] D. Silver, "Introduction to Reinforcement Learning with David Silver Lecture 7: Policy Gradient Methods," University Lecture, 2015. [Online]. Available: https://www.davidsilver.uk/wp-content/uploads/2020/03/pg.pdf

[22] B. Millidge, A. Tschantz, A. K. Seth, and C. L. Buckley, "Reinforcement learning as iterative and amortised inference," 2020. [Online]. Available: https://arxiv.org/abs/2006.10524v1#

[23] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing Atari with Deep Reinforcement Learning," 2013. [Online]. Available: https://arxiv.org/abs/1312.5602

[24] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529—533, February 2015. [Online]. Available: http://europepmc.org/article/MED/25719670

[25] R. S. Sutton, D. A. McAllester, S. Singh, and Y. Mansour, "Policy Gradient Methods for Reinforcement Learning with Function Approximation," in *NIPS*, 1999. [Online]. Available: https://papers.nips.cc/paper/1999/file/464d828b85b0bed98e80ade0a5c43b0f-Paper.pdf

[26] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous Methods for Deep Reinforcement Learning," 2016. [Online]. Available: https://arxiv.org/abs/1602.01783

[27] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," 2017. [Online]. Available: https://arxiv.org/abs/1707.06347

[28] J. Achiam, "Spinning Up in Deep Reinforcement Learning," 2018. [Online]. Available: https://spinningup.openai.com/en/latest/index.html

[29] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," 2019. [Online]. Available: https://arxiv.org/abs/1509.02971

[30] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," 2018. [Online]. Available: https://arxiv.org/abs/1801.01290

[31] K. Chua, R. Calandra, R. McAllister, and S. Levine, "Deep Reinforcement Learning in a Handful of Trials using Probabilistic Dynamics Models," 2018. [Online]. Available: https://arxiv.org/abs/1805.12114

[32] S. Levine, "Deep Reinforcement Learning Lecture 10: Optimal Control and Planning, CS285 at UC Berkeley," University Lecture, 2020. [Online]. Available: http://rail.eecs.berkeley.edu/deeprlcourse/static/slides/lec-10.pdf

[33] R. S. Sutton, "Dyna, an Integrated Architecture for Learning, Planning, and Reacting," *SIGART Bull.*, vol. 2, no. 4, p. 160–163, Jul. 1991. [Online]. Available: https://doi.org/10.1145/122344.122377

[34] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, pp. 484–, Jan. 2016. [Online]. Available: http://dx.doi.org/10.1038/nature16961

[35] G. Williams, P. Drews, B. Goldfain, J. M. Rehg, and E. A. Theodorou, "Aggressive driving with model predictive path integral control," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 2016, pp. 1433–1440.

[36] S. Amin, M. Gomrokchi, H. Satija, H. van Hoof, and D. Precup, "A Survey of Exploration Methods in Reinforcement Learning," 2021. [Online]. Available: https://arxiv.org/abs/2109.00157v1

[37] N. Cesa-Bianchi, C. Gentile, G. Lugosi, and G. Neu, "Boltzmann Exploration Done Right," 2017. [Online]. Available: https://arxiv.org/abs/1705.10257

[38] R. Smith, K. Friston, and C. Whyte, "A Step-by-Step Tutorial on Active Inference and its Application to Empirical Data," *PsyArXiv*, 2021. [Online]. Available: https://doi.org/10.31234/osf.io/b4jm6

[39] L. Da Costa, T. Parr, N. Sajid, S. Veselic, V. Neacsu, and K. Friston, "Active inference on discrete state-spaces: A synthesis," *Journal of Mathematical Psychology*, vol. 99, p. 102447, Dec 2020. [Online]. Available: http://dx.doi.org/10.1016/j.jmp.2020.102447

[40] T. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4–22, 1985. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0196885885900028

[41] D. Russo and B. V. Roy, "An information-theoretic analysis of thompson sampling," 2015. [Online]. Available: https://arxiv.org/abs/1403.5341

[42] K. Friston, "Life as we know it," *Journal of the Royal Society, Interface / the Royal Society*, vol. 10, p. 20130475, 06 2013. [Online]. Available: https://doi.org/10.1098/rsif.2013.0475

[43] M. B. Mirza, R. A. Adams, C. D. Mathys, and K. J. Friston, "Scene Construction, Visual Foraging, and Active Inference," *Frontiers in computational neuroscience*, vol. 10, p. 56, 2016. [Online]. Available: https://europepmc.org/articles/PMC4906014

[44] M. Cullen, B. Davey, K. J. Friston, and R. J. Moran, “Active Inference in OpenAI Gym: A Paradigm for Computational Investigations Into Psychiatric Illness,” *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, vol. 3, no. 9, pp. 809–818, 2018, computational Methods and Modeling in Psychiatry. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2451902218301617

[45] J. Pearl, “Chapter 3 - MARKOV AND BAYESIAN NETWORKS: Two Graphical Representations of Probabilistic Knowledge,” in *Probabilistic Reasoning in Intelligent Systems*, J. Pearl, Ed. San Francisco (CA): Morgan Kaufmann, 1988, pp. 77–141. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780080514895500096

[46] M. Okada, N. Kosaka, and T. Taniguchi, “PlaNet of the Bayesians: Reconsidering and Improving Deep Planning Network by Incorporating Bayesian Inference,” 2020. [Online]. Available: https://arxiv.org/abs/2003.00370

[47] K. Friston, F. Rigoli, D. Ognibene, C. Mathys, T. Fitzgerald, and G. Pezzulo, “Active inference and epistemic value,” *Cognitive Neuroscience*, vol. 6, no. 4, pp. 187–214, 2015, pMID: 25689102. [Online]. Available: https://doi.org/10.1080/17588928.2015.1020053

[48] A. Klyubin, D. Polani, and C. Nehaniv, “Empowerment: a universal agent-centric measure of control,” in *2005 IEEE Congress on Evolutionary Computation*, vol. 1, 2005, pp. 128–135 Vol.1. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/1554676

[49] H. Kim, J. Kim, Y. Jeong, S. Levine, and H. O. Song, “Emi: Exploration with mutual information,” 2019. [Online]. Available: https://arxiv.org/abs/1810.01176