REINFORCEMENT LEARNING THROUGH ACTIVE INFERENCE

Alexander Tschantz
Sackler Centre for Consciousness Science
Evolutionary & Adaptive Systems Research Group
University of Sussex
Brighton, UK
tschantz.alec@gmail.com

Beren Millidge
University of Edinburgh
Edinburgh, UK
s1686853@sms.ed.ac.uk

Anil K. Seth
Sackler Centre for Consciousness Science
Evolutionary & Adaptive Systems Research Group
University of Sussex
Brighton, UK
Canadian Institute for Advanced Research

Christopher L. Buckley
Evolutionary & Adaptive Systems Research Group
University of Sussex
Brighton, UK

ABSTRACT

The central tenet of reinforcement learning (RL) is that agents seek to maximize the sum of cumulative rewards. In contrast, active inference, an emerging framework within cognitive and computational neuroscience, proposes that agents act to maximize the evidence for a biased generative model. Here, we illustrate how ideas from active inference can augment traditional RL approaches by (i) furnishing an inherent balance of exploration and exploitation, and (ii) providing a more flexible conceptualization of reward. Inspired by active inference, we develop and implement a novel objective for decision making, which we term the *free energy of the expected future*. We demonstrate that the resulting algorithm successfully balances exploration and exploitation, simultaneously achieving robust performance on several challenging RL benchmarks with sparse, well-shaped, and no rewards.

1 INTRODUCTION

Both biological and artificial agents must learn to make adaptive decisions in unknown environments. In the field of reinforcement learning (RL), agents aim to learn a policy that maximises the sum of expected rewards (Sutton et al., 1998). In contrast, the framework of active inference (Friston et al., 2009; Friston, 2019a) suggests that agents aim to maximise the evidence for a biased model of the world. This framework extends influential theories of Bayesian perception and learning (Knill & Pouget, 2004; L Griffiths et al., 2008) to incorporate action, and comes equipped with a biologically plausible process theory (Friston et al., 2017a) that enjoys considerable empirical support (Walsh et al., 2020).

In the current paper, we explore whether insights from active inference can inform the development of novel RL algorithms. We focus on two aspects of active inference which are of particular relevance for RL. First, active inference suggests that agents embody a generative model of their preferred environment and seek to maximise the evidence for this model. In this context, the concept of reward is replaced with prior probabilities over observations. This allows for greater flexibility when specifying an agent's goals (Friston et al., 2012), provides a principled (i.e. Bayesian) method for learning preferences (Sajid et al., 2019), and is consistent with neurophysiological data demonstrating the distributional nature of reward representations (Dabney et al., 2020). Second, casting adaptive behavior in terms of model evidence naturally encompasses both exploration and exploitation under a single objective, obviating the need for adding ad-hoc exploratory terms.

Translating these conceptual insights into practical benefits for RL has proven challenging. Current implementations of active inference have generally been confined to discrete state spaces and toy problems (Friston et al., 2015; 2017b;c), although see (Tschantz et al., 2019a; Millidge, 2019; Catal et al., 2019; Ueltzhffer, 2018). Therefore, it has not yet been possible to evaluate the effectiveness of active inference in challenging environments; as a result, active inference has not yet been widely taken up within the RL community.

In this paper, we consider active inference in the context of decision making[1]. We propose and implement a novel objective function motivated by active inference - the *free energy of the expected future* - and show that this quantity provides a tractable bound on established RL objectives. We evaluate the performance of this algorithm on a selection of challenging continuous control tasks, and show strong performance on environments with sparse, well-shaped, and no rewards, demonstrating our algorithm's ability to effectively balance exploration and exploitation.

2 Active Inference

Both active inference and RL can be formulated in the context of a partially observed Markov decision process POMDPs (Murphy, 1982). The goal of RL is to learn a policy that maximises the expected sum of rewards $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t \mathrm{r}_t]$ (Sutton et al., 1998). In contrast, the goal of active inference is to maximise the Bayesian model evidence for an agent's generative model $p^{\Phi}(\mathrm{o}, \mathrm{s}, \theta)$, where $\mathrm{o} \in \mathcal{O}$ denote observations, $\mathrm{s} \in \mathcal{S}$ denote latent or hidden states, and $\theta \in \Theta$ denote model parameters. Crucially, active inference suggests that an agent's generative model is biased towards favourable states of affairs (Friston, 2019b), such that the model assigns probability to events that are both likely and beneficial for an agent's success. We use the notation $p^{\Phi}(\cdot)$ to represent a distribution biased by an agent's preferences.

Given a generative model, agents can perform approximate Bayesian inference by encoding an arbitrary distribution $q(\mathrm{s}, \theta)$ and minimising *variational free energy* $\mathcal{F} = D_{\mathrm{KL}}\big(q(\mathrm{s}, \theta) \| p^{\Phi}(\mathrm{o}, \mathrm{s}, \theta)\big)$. From observations, $\mathcal{F}$ can be minimized through standard variational methods (Bishop, 2006; Buckley et al., 2017), causing $q(\mathrm{s}, \theta)$ to tend towards the true posterior $p(\mathrm{s}, \theta | \mathrm{o})$. Note that treating model parameters θ as random variables casts learning as a process of inference (Blundell et al., 2015).

In the current context, agents additionally maintain beliefs over policies $\pi = \{a_t, ..., a_T\}$, which are themselves random variables. Policy selection is then implemented by identifying the $q(\pi)$ which minimizes $\mathcal{F}$, thus casting policy selection as a process of approximate inference (Friston et al., 2015). While the standard free energy functional $\mathcal{F}$ is generally defined for a single time point t, π refers to a temporal sequence of variables. Therefore, we augment the free energy functional $\mathcal{F}$ to encompass future variables, leading to the *free energy of the expected future* $\tilde{\mathcal{F}}$. This quantity measures the KL-divergence between a sequence of beliefs about future variables and an agent's biased generative model [2].

2.1 Free energy of the expected future

Let $\mathrm{x}_{t:T}$ denote a sequence of variables through time, $\mathrm{x}_{t:T} = \{\mathrm{x}_t, ..., \mathrm{x}_T\}$. We wish to minimize the free energy of the expected future $\tilde{\mathcal{F}}$, which is defined as:

$$\tilde{\mathcal{F}} = D_{\mathrm{KL}}\Big(q(\mathrm{o}_{0:T}, \mathrm{s}_{0:T}, \theta, \pi) \| p^{\Phi}(\mathrm{o}_{0:T}, \mathrm{s}_{0:T}, \theta)\Big) \tag{1}$$

where $q(\mathrm{o}_{t:T}, \mathrm{s}_{t:T}, \theta, \pi)$ represents an agent's beliefs about future variables, and $p^{\Phi}(\mathrm{o}_{t:T}, \mathrm{s}_{t:T}, \theta)$ represents an agent's biased generative model. Note that the beliefs about future variables include beliefs about future observations, $\mathrm{o}_{t:T}$, which are unknown and thus treated as random variables.

In order to find $q(\pi)$ which minimizes $\tilde{\mathcal{F}}$ we note that (see Appendix F):

$$\begin{aligned} \tilde{\mathcal{F}} = 0 \Rightarrow D_{\mathrm{KL}}\Big(q(\pi) \; \| \big(- e^{-\tilde{\mathcal{F}}_{\pi}} \big)\Big) = 0 \\ \tilde{\mathcal{F}}_{\pi} = D_{\mathrm{KL}}\Big(q(\mathrm{o}_{0:T}, \mathrm{s}_{0:T}, \theta | \pi) \; \| \; p^{\Phi}(\mathrm{o}_{0:T}, \mathrm{s}_{0:T}, \theta)\Big) \end{aligned} \tag{2}$$

[1]A full treatment of active inference would consider inference and learning, see (Buckley et al., 2017) for an overview.

[2]For readers familiar with the active inference framework, we highlight that the *free energy of the expected future* differs subtly from *expected free energy* (Friston et al., 2015). We draw comparisons in Appendix E.

Thus, the free energy of the expected future is minimized when $q(\pi) = \sigma(-\tilde{\mathcal{F}}_\pi)$, or in other words, policies are more likely when they minimise $\tilde{\mathcal{F}}_\pi$.

In order to provide an intuition for what minimizing $\tilde{\mathcal{F}}_\pi$ entails, we factorize the agent's generative models as $p^\Phi(\mathrm{o}_{0:T}, \mathrm{s}_{0:T}, \theta) = p(\mathrm{s}_{0:T}, \theta | \mathrm{o}_{0:T}) p^\Phi(\mathrm{o}_{0:T})$, implying that the model is only biased in its beliefs over observations. To retain consistency with RL nomenclature, we treat 'rewards' r as a separate observation modality, such that $p^\Phi(\mathrm{o}_{t:T})$ specifies a distribution over preferred rewards and $q(\mathrm{o}_{t:T} | \mathrm{s}_{t:T}, \theta, \pi)$ specifies beliefs about future rewards, given a policy. We describe our implementation of $p^\Phi(\mathrm{o}_{t:T})$ in Appendix I. Given this factorization, it is straightforward to show that $-\tilde{\mathcal{F}}_\pi$ decomposes into an expected information gain term and an extrinsic term (see Appendix D)[3]:

$$
\begin{aligned}
-\tilde{\mathcal{F}}_\pi \approx & -\underbrace{\mathbb{E}_{q(\mathrm{o}_{0:T}|\pi)}\Big[D_{\mathrm{KL}}\Big(q(\mathrm{s}_{0:T}, \theta | \mathrm{o}_{0:T}, \pi) \| q(\mathrm{s}_{0:T}, \theta | \pi)\Big)\Big]}_{\text{Expected information gain}} \\
& + \underbrace{\mathbb{E}_{q(\mathrm{s}_{0:T}, \theta|\pi)}\Big[D_{\mathrm{KL}}\Big(q(\mathrm{o}_{0:T} | \mathrm{s}_{0:T}, \theta, \pi) \| p^\Phi(\mathrm{o}_{t:T})\Big)\Big]}_{\text{Extrinsic term}}
\end{aligned} \tag{3}
$$

Maximizing Eq.3 has two functional consequences. First, it maximises expected information gain, which quantifies the amount of information an agent expects to gain from executing some policy, both in terms of beliefs about parameters and states. Second, it minimizes the extrinsic term - which is the KL-divergence between an agent's (policy-conditioned) beliefs about future observations and their preferred observations. In the current context, it measures the KL-divergence between the rewards an agent expects from a policy and the rewards an agent desires. In summary, selecting policies to minimise $\tilde{\mathcal{F}}$ invokes a natural balance between exploration and exploitation.

Evaluating $\tilde{\mathcal{F}}_\pi$ To select actions, we optimise $q(\pi)$ at each time step, and execute the first action specified by the most likely policy. This requires (i) a method for evaluating beliefs about future variables $q(\mathrm{s}_{t:T}, \mathrm{o}_{t:T}, \theta | \pi)$, (ii) an efficient method for evaluating $\mathcal{F}_\pi$, and (iii) a method for optimising $q(\pi)$ such that $q(\pi) = \sigma(-\mathcal{F}_\pi)$. We list our design choices and describe the full implementation in Appendix B.

3 Experiments

To determine whether our algorithm can successfully balance exploration and exploitation, we investigate its performance in domains with (i) well-shaped rewards, (ii) extremely sparse rewards and (iii) a complete absence of rewards.

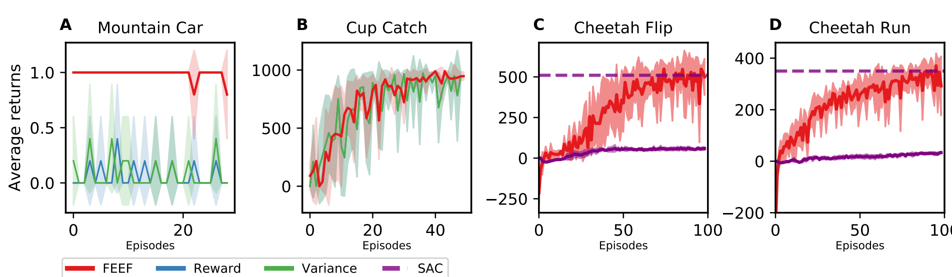

Figure 1: **(A & B) Mountain Car / Cup Catch:** Average return after each episode on the sparse-reward Mountain Car & Cup Catch tasks. **(C & D) Half Cheetah:** Average return after each episode on the well-shaped Half Cheetah environment, for the running and flipping tasks, respectively. Dotted line shows SAC performance after 2000 episodes. Each curve is the mean of 5 seeds and filled regions show +/- std.

[3]The approximation in Eq. 3 arises from the approximation $q(\mathrm{s}_{0:T}, \theta | \mathrm{o}_{0:T}, \pi) \approx p(\mathrm{s}_{0:T}, \theta | \mathrm{o}_{0:T}, \pi)$, which is justifiable given that $q(\cdot)$ represents a variational approximation of the true posterior (Friston et al., 2017a).

Sparse reward We use the **Mountain Car** and **Cup Catch** environments, where agents only receive reward when the goal is achieved (see Appendix K for details on all environments). We compare our algorithm to two baselines, (i) a **reward** algorithm which only selects policies based on the extrinsic term (i.e. ignores expected information gain), and (ii) a **variance** algorithm that seeks out uncertain transitions by acting to maximise the output variance of the transition model (see Appendix I). Note that the variance agent is augmented with the extrinsic term to enable comparison.

In the **Mountain Car** task (Fig. 1A), our algorithm is able to efficiently explore and solve the environment within a single trial. We visualize the exploration in Fig. 2, which plots the points in state space visited with and without exploration, and demonstrates that the free energy of the expected future objective allows the agent to explore the state space while still solving the task. Our algorithm performs comparably to the baselines in the Cup Catch environment, shown in Fig. 1B. We hypothesize that this is because, while the reward structure is technically sparse, it is simple enough to reach the goal with random actions, and thus the directed exploration afforded by our method provides little benefit.

Well-shaped reward For well-shaped rewards, we use the challenging **Half Cheetah** environment, using both the running and flipping tasks. We compare the performance of our algorithm to the soft-actor-critic (SAC) (Haarnoja et al., 2018), which encourages exploration by seeking to maximise the entropy of the policy distribution. Figure 1 C&D demonstrate that our algorithm exhibits robust performance in environments with well-shaped rewards and provides considerable improvements in sample-efficiency, relative to SAC.

No reward For domains without reward, we use the **Ant Maze** environment, where there are no rewards and success is measured by the percent of the maze covered. We compare our algorithm to a baseline which conducts actions at random. Fig. 2B shows that our algorithms rate of exploration is substantially higher, demonstrating that minimising the free energy of the expected future leads to directed exploration. Taken together, these results show that our proposed algorithm - which naturally balances exploration and exploitation - can successfully master challenging domains with a variety of reward structures.

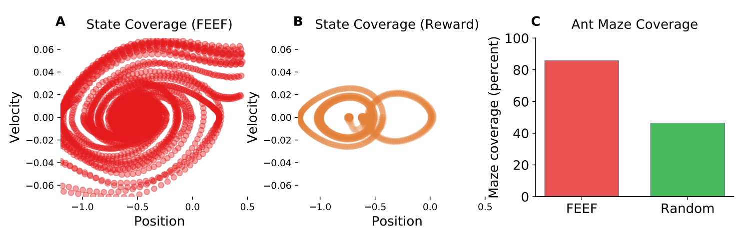

Figure 2: **(A & B) Mountain Car state space coverage**: State space plots from the FEEF and reward agent over 20 episodes. **(C) Ant Maze Coverage**: We plot the percentage of the maze covered after 35 episodes, comparing the FEEF agent to an agent acting randomly. These results are the average of 4 seeds.

4 Discussion

Despite originating from different intellectual traditions, both active inference and RL address fundamental questions about adaptive decision-making in unknown environments. Exploiting this conceptual overlap, we have applied an active inference perspective to the reward maximization objective of RL, recasting it as minimizing the divergence between desired and expected futures. We derived a novel objective that naturally balances exploration and exploitation and instantiated this objective within a model-based RL context. Our algorithm exhibits robust performance and flexibility in a variety of environments and reward structures known to be challenging for RL. By implementing active inference using tools from RL, such as amortising inference with neural networks, deep ensembles and sophisticated algorithms for planning, we have demonstrated that active inference can scale to high dimensional tasks with continuous state and action spaces.

Acknowledgements

AT is funded by a PhD studentship from the Dr. Mortimer and Theresa Sackler Foundation and the School of Engineering and Informatics at the University of Sussex. BM is supported by an EPSRC funded PhDS Studentship. CLB is supported by BBRSC grant number BB/P022197/1. AT and AKS are grateful to the Dr. Mortimer and Theresa Sackler Foundation, which supports the Sackler Centre for Consciousness Science. AKS is additionally grateful to the Canadian Institute for Advanced Research (Azrieli Programme on Brain, Mind, and Consciousness).

Author Contributions

A.T, B.M and C.L.B contributed to the conceptualization of this work. A.T and B.M contributed to the coding and generation of experimental results. A.T, B.M, C.L.B, A.K.S contributed to the writing of the manuscript.

Data

Code for the model can be found at https://github.com/alec-tschantz/rl-inference

References

Marc G. Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. 2016. URL http://arxiv.org/abs/1606.01868.

Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. 2015. URL http://arxiv.org/abs/1505.05424.

Christopher L Buckley, Chang Sub Kim, Simon McGregor, and Anil K Seth. The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology*, 81:55–79, 2017.

Ozan Catal, Johannes Nauta, Tim Verbelen, Pieter Simoens, and Bart Dhoedt. Bayesian policy selection using active inference. 2019. URL http://arxiv.org/abs/1904.08149.

Konstantinos Chatzilygeroudis, Vassilis Vassiliades, Freek Stulp, Sylvain Calinon, and Jean-Baptiste Mouret. A survey on policy search algorithms for learning robot controllers in a handful of trials. 2018. URL http://arxiv.org/abs/1807.02303.

Nuttapong Chentanez, Andrew G. Barto, and Satinder P. Singh. Intrinsically motivated reinforcement learning. In L. K. Saul, Y. Weiss, and L. Bottou (eds.), *Advances in Neural Information Processing Systems 17*, pp. 1281–1288. MIT Press, 2005. URL http://papers.nips.cc/paper/2552-intrinsically-motivated-reinforcement-learning.pdf.

Kashyap Chitta, Jose M. Alvarez, and Adam Lesnikowski. Deep probabilistic ensembles: Approximate variational inference through KL regularization. 2018. URL http://arxiv.org/abs/1811.02640.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In *Advances in Neural Information Processing Systems*, pp. 4754–4765, 2018a.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. 2018b. URL http://arxiv.org/abs/1805.12114.

Maell Cullen, Ben Davey, Karl J Friston, and Rosalyn J Moran. Active inference in openai gym: A paradigm for computational investigations into psychiatric illness. *Biological psychiatry: cognitive neuroscience and neuroimaging*, 3(9):809–818, 2018.

Will Dabney, Zeb Kurth-Nelson, Naoshige Uchida, Clara Kwon Starkweather, Demis Hassabis, Rémi Munos, and Matthew Botvinick. A distributional code for value in dopamine-based reinforcement learning. *Nature*, pp. 1–5, 2020.

Ildefons Magrans de Abril and Ryota Kanai. A unified strategy for implementing curiosity and empowerment driven reinforcement learning. *arXiv preprint arXiv:1806.06505*, 2018.

Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. 2019. URL http://arxiv.org/abs/1912.02757.

Karl Friston. A free energy principle for a particular physics. *arXiv preprint arXiv:1906.10184*, 2019a.

Karl Friston. A free energy principle for a particular physics. 2019b. URL https://arxiv.org/abs/1906.10184v1.

Karl Friston, Spyridon Samothrakis, and Read Montague. Active inference and agency: optimal control without cost functions. *Biological cybernetics*, 106(8-9):523–541, 2012.

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

Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, Giovanni Pezzulo, et al. Active inference and learning. *Neuroscience & Biobehavioral Reviews*, 68:862–879, 2016.

Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, and Giovanni Pezzulo. Active inference: a process theory. *Neural computation*, 29(1):1–49, 2017a.

Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, and Giovanni Pezzulo. Active inference: A process theory. 29(1):1–49, 2017b. ISSN 1530-888X. doi: 10.1162/NECO_a_00912.

Karl J Friston, Jean Daunizeau, and Stefan J Kiebel. Reinforcement learning or active inference? *PloS one*, 4(7), 2009.

Karl J. Friston, Marco Lin, Christopher D. Frith, Giovanni Pezzulo, J. Allan Hobson, and Sasha Ondobaka. Active inference, curiosity and insight. 29(10):2633–2683, 2017c. ISSN 0899-7667. doi: 10.1162/neco_a_00999. URL https://doi.org/10.1162/neco_a_00999.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, pp. 2450–2462, 2018.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*, 2018.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. *arXiv preprint arXiv:1811.04551*, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. Curiosity-driven exploration in deep reinforcement learning via bayesian neural networks. *arXiv preprint arXiv:1605.09674*, 2016a.

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. VIME: Variational information maximizing exploration. 2016b. URL http://arxiv.org/abs/1605.09674.

Lukasz Kaiser, Mohammad Babaeizadeh, Piotr Milos, Blazej Osinski, Roy H. Campbell, Konrad Czechowski, Dumitru Erhan, Chelsea Finn, Piotr Kozakowski, Sergey Levine, Afroz Mohiuddin, Ryan Sepassi, George Tucker, and Henryk Michalewski. Model-based reinforcement learning for atari. 2019. URL http://arxiv.org/abs/1903.00374.

Hilbert J Kappen, Vicenç Gómez, and Manfred Opper. Optimal control as a graphical model inference problem. *Machine learning*, 87(2):159–182, 2012.

Hyoungseok Kim, Jaekyeom Kim, Yeonwoo Jeong, Sergey Levine, and Hyun Oh Song. EMI: Exploration with mutual information. 2018. URL http://arxiv.org/abs/1810.01176.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. 2013. URL http://arxiv.org/abs/1312.6114.

David C Knill and Alexandre Pouget. The bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, 27(12):712–719, 2004.

Thomas L Griffiths, Charles Kemp, and Joshua B Tenenbaum. Bayesian models of cognition. 2008.

Lisa Lee, Benjamin Eysenbach, Emilio Parisotto, Eric Xing, Sergey Levine, and Ruslan Salakhutdinov. Efficient exploration via state marginal matching. *arXiv preprint arXiv:1906.05274*, 2019.

Felix Leibfried, Sergio Pascual-Diaz, and Jordi Grau-Moya. A unified bellman optimality principle combining reward maximization and empowerment. In *Advances in Neural Information Processing Systems*, pp. 7867–7878, 2019.

Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018.

D. V. Lindley. On a measure of the information provided by an experiment. 27(4):986–1005, 1956. ISSN 0003-4851, 2168-8990. doi: 10.1214/aoms/1177728069. URL https://projecteuclid.org/euclid.aoms/1177728069.

Beren Millidge. Deep active inference as variational policy gradients. 2019. URL http://arxiv.org/abs/1907.03876.

Atanas Mirchev, Baris Kayalibay, Maximilian Soelch, Patrick van der Smagt, and Justin Bayer. Approximate bayesian inference in spatial environments. 2018. URL http://arxiv.org/abs/1805.07206.

Shakir Mohamed and Danilo Jimenez Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. 2015. URL http://arxiv.org/abs/1509.08731.

KP Murphy. A survey of pomdp solution techniques: Theory. *Models, and algorithms, management science*, 28, 1982.

Anusha Nagabandi, Gregory Kahn, Ronald S Fearing, and Sergey Levine. Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7559–7566. IEEE, 2018.

Anusha Nagabandi, Kurt Konoglie, Sergey Levine, and Vikash Kumar. Deep dynamics models for learning dexterous manipulation. *arXiv preprint arXiv:1909.11652*, 2019.

Brendan O'Donoghue, Ian Osband, Remi Munos, and Volodymyr Mnih. The uncertainty bellman equation and exploration. *arXiv preprint arXiv:1709.05380*, 2017.

Masashi Okada and Tadahiro Taniguchi. Variational inference MPC for bayesian model-based reinforcement learning. 2019. URL http://arxiv.org/abs/1907.04202.

Pierre-Yves Oudeyer and Frederic Kaplan. What is intrinsic motivation? a typology of computational approaches. *Frontiers in neurorobotics*, 1:6, 2009.

Thomas Parr and Karl J Friston. The active construction of the visual world. *Neuropsychologia*, 104:92–101, 2017.

Thomas Parr, Dimitrije Markovic, Stefan J Kiebel, and Karl J Friston. Neuronal message passing using mean-field, bethe, and marginal approximations. *Scientific reports*, 9(1):1–18, 2019.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 16–17, 2017.

Giovanni Pezzulo, Emilio Cartoni, Francesco Rigoli, Léo Pio-Lopez, and Karl Friston. Active inference, epistemic value, and vicarious trial and error. *Learning & Memory*, 23(7):322–338, 2016.

Konrad Rawlik, Marc Toussaint, and Sethu Vijayakumar. On stochastic optimal control and reinforcement learning by approximate inference. In *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

Konrad Cyrus Rawlik. On probabilistic inference approaches to stochastic optimal control. 2013.

Reuven Y Rubinstein. Optimization of computer simulation models with rare events. *European Journal of Operational Research*, 99(1):89–112, 1997.

Noor Sajid, Philip J Ball, and Karl J Friston. Demystifying active inference. *arXiv preprint arXiv:1909.10863*, 2019.

Jürgen Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proc. of the international conference on simulation of adaptive behavior: From animals to animats*, pp. 222–227, 1991.

Jürgen Schmidhuber. Simple algorithmic principles of discovery, subjective beauty, selective attention, curiosity & creativity. In *International Conference on Discovery Science*, pp. 26–38. Springer, 2007.

Philipp Schwartenbeck, Johannes Passecker, Tobias U Hauser, Thomas HB FitzGerald, Martin Kronbichler, and Karl J Friston. Computational mechanisms of curiosity and goal-directed exploration. 8:e41703, 2019. ISSN 2050-084X. doi: 10.7554/eLife.41703. URL https://doi.org/10.7554/eLife.41703.

Pranav Shyam, Wojciech Jaśkowski, and Faustino Gomez. Model-based active exploration. *arXiv preprint arXiv:1810.12162*, 2018.

Pranav Shyam, Wojciech Jakowski, and Faustino Gomez. Model-based active exploration. In *International Conference on Machine Learning*, pp. 5779–5788, 2019. URL http://proceedings.mlr.press/v97/shyam19a.html.

Susanne Still and Doina Precup. An information-theoretic approach to curiosity-driven reinforcement learning. 131(3):139–148, 2012. ISSN 1611-7530. doi: 10.1007/s12064-011-0142-z.

Jan Storck, Sepp Hochreiter, and Jürgen Schmidhuber. Reinforcement driven information acquisition in non-deterministic environments. In *Proceedings of the international conference on artificial neural networks, Paris*, volume 2, pp. 159–164. Citeseer, 1995.

Yi Sun, Faustino Gomez, and Juergen Schmidhuber. Planning to be surprised: Optimal bayesian exploration in dynamic environments. 2011. URL http://arxiv.org/abs/1103.5708.

Richard S Sutton, Andrew G Barto, et al. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge, 1998.

Bjrn Ivar Teigen. An active learning perspective on exploration in reinforcement learning. 2018. URL https://www.duo.uio.no/handle/10852/62823.

Alexander Tschantz, Manuel Baltieri, Anil K. Seth, and Christopher L. Buckley. Scaling active inference. 2019a. URL http://arxiv.org/abs/1911.10601.

Alexander Tschantz, Anil K. Seth, and Christopher L. Buckley. Learning action-oriented models through active inference. pp. 764969, 2019b. doi: 10.1101/764969. URL https://www.biorxiv.org/content/10.1101/764969v1.

Sebastian Tschiatschek, Kai Arulkumaran, Jan Sthmer, and Katja Hofmann. Variational inference for data-efficient model learning in POMDPs. 2018. URL http://arxiv.org/abs/1805.09281.

Kai Ueltzhffer. Deep active inference. 112(6):547–573, 2018. ISSN 0340-1200, 1432-0770. doi: 10.1007/s00422-018-0785-7. URL http://arxiv.org/abs/1709.02341.

Kevin S Walsh, David P McGovern, Andy Clark, and Redmond G O'Connell. Evaluating the neurophysiological evidence for predictive processing as a model of perception. *Annals of the New York Academy of Sciences*, 2020.

Manuel Watter, Jost Tobias Springenberg, Joschka Boedecker, and Martin Riedmiller. Embed to control: A locally linear latent dynamics model for control from raw images. 2015. URL http://arxiv.org/abs/1506.07365.

Grady Williams, Paul Drews, Brian Goldfain, James M Rehg, and Evangelos A Theodorou. Aggressive driving with model predictive path integral control. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1433–1440. IEEE, 2016.

Yarin Gal, Rowan McAllister, and Carl Edward Rasmussen. Improving PILCO with bayesian neural network dynamics models. In *Data-Efficient Machine Learning workshop*, 2016.

A Relationship to probabilistic RL

In recent years, there have been several attempts to formalize RL in terms of probabilistic inference (Levine, 2018), such as KL-control (Rawlik, 2013), control-as-inference (Kappen et al., 2012), and state-marginal matching (Lee et al., 2019). In many of these approaches, the RL objective is broadly conceptualized as minimising $D_{\mathrm{KL}}\Big(p(\mathrm{o}_{0:T}|\pi)\|\ p^{\Phi}(o_{0:T})\Big)$[4]. In Appendix H, we demonstrate that the free energy of the expected future $\tilde{\mathcal{F}}$ provides a tractable bound on this objective:

$$\tilde{\mathcal{F}} \geq D_{\mathrm{KL}}\Big(p(\mathrm{o}_{t:T}|\pi)\|\ p^{\Phi}(o_{t:T})\Big) \tag{4}$$

These results suggest a deep homology between active inference and existing approaches to probabilistic RL.

B Implementation

In this section, we describe an efficient implementation of the proposed objective function in the context of model-based RL. To select actions, we optimise $q(\pi)$ at each time step, and execute the first action specified by the most likely policy. This requires (i) a method for evaluating beliefs about future variables $q(\mathrm{s}_{t:T}, \mathrm{o}_{t:T}, \theta|\pi)$, (ii) an efficient method for evaluating $\mathcal{F}_{\pi}$, and (iii) a method for optimising $q(\pi)$ such that $q(\pi) = \sigma(-\mathcal{F}_{\pi})$.

Evaluating beliefs about the future We factorize and evaluate the beliefs about the future as:

$$\begin{aligned} q(\mathrm{s}_{t:T}, \mathrm{o}_{t:T}, \theta|\pi) &= q(\theta) \prod_{t=\tau}^{T} q(\mathrm{o}_{\tau}|\mathrm{s}_{\tau}, \theta, \pi) q(\mathrm{s}_{\tau}|\mathrm{s}_{\tau-1}, \theta, \pi) \\ q(\mathrm{o}_{\tau}|\mathrm{s}_{\tau}, \theta, \pi) &= \mathbb{E}_{q(\mathrm{s}_{\tau}|\theta,\pi)}\big[p(\mathrm{o}_{\tau}|\mathrm{s}_{\tau})\big] \\ q(\mathrm{s}_{\tau}|\mathrm{s}_{\tau-1}, \theta, \pi) &= \mathbb{E}_{q(\mathrm{s}_{\tau-1}|\theta,\pi)}\big[p(\mathrm{s}_{\tau}|\mathrm{s}_{\tau-1}, \theta, \pi)\big] \end{aligned} \tag{5}$$

where we have here factorized the generative model as $p(\mathrm{o}_{\tau}, \mathrm{s}_{\tau}, \theta|\pi) = p(\mathrm{o}_{\tau}|\mathrm{s}_{\tau}, \pi)p(\mathrm{s}_{\tau}|\mathrm{s}_{\tau-1}, \theta, \pi)p(\theta)$. We describe the implementation and learning of the likelihood $p(\mathrm{o}_{\tau}|\mathrm{s}_{\tau}, \pi)$, transition model $p(\mathrm{s}_{\tau}|\mathrm{s}_{\tau-1}, \theta, \pi)$ and parameter prior $p(\theta)$ in Appendix I.

[4]We acknowledge that not all objectives follow this exact formulation.

Evaluating $\tilde{\mathcal{F}}_\pi$ Note that $-\tilde{\mathcal{F}}_\pi = \sum_{\tau=t}^{t+H} -\tilde{\mathcal{F}}_{\pi_\tau}$, where H is the planning horizon. Given beliefs about future variables, the free energy of the expected future for a single time point can be efficiently computed as (see Appendix G):

$$\begin{aligned} -\tilde{\mathcal{F}}_{\pi_\tau} \approx\ & E_{q(\mathrm{s}_\tau,\theta|\pi)}\Big[D_{\mathrm{KL}}\Big(q(\mathrm{o}_\tau|\mathrm{s}_\tau,\theta,\pi)\|p^\Phi(\mathrm{o}_\tau)\Big)\Big] \\ & + \underbrace{\mathbf{H}[q(\mathrm{o}_\tau|\pi)] - \mathbb{E}_{q(\mathrm{s}_\tau|\pi)}\Big[\mathbf{H}[q(\mathrm{o}_\tau|\mathrm{s}_\tau,\pi)]\Big]}_{\text{State information gain}} \\ & + \underbrace{\mathbf{H}[q(\mathrm{s}_\tau|\mathrm{s}_{\tau-1},\theta,\pi)] - \mathbb{E}_{q(\theta)}\Big[\mathbf{H}[q(\mathrm{s}_\tau|\mathrm{s}_{\tau-1},\pi,\theta)]\Big]}_{\text{Parameter information gain}} \end{aligned} \tag{6}$$

In the current paper, agents observe the true state of the environment s_t, such that the only partial observability is in rewards r_t. As as a result, the second term of equation 6 is redundant, as there is no uncertainty about states. The first (extrinsic) term can be calculated analytically (see Appendix I). We describe our approximation of the final term (parameter information gain) in Appendix G.

Optimising the policy distribution We choose to parametrize $q(\pi)$ as a diagonal Gaussian. We use the CEM algorithm (Rubinstein, 1997) to optimise the parameters of $q(\pi)$ such that $q(\pi) \propto -\mathcal{F}_\pi$. While this solution will fail to capture the exact shape of $-\mathcal{F}_\pi$, agents need only identify the peak of the landscape to enact the optimal policy.

Algorithm 1: Inference of $q(\pi)$

Input: Planning horizon H — Optimisation iterations I — Number of candidate policies J — Current state s_t — Likelihood $p(\mathrm{o}_\tau|\mathrm{s}_\tau)$ — Transition distribution $p(\mathrm{s}_\tau|\mathrm{s}_{\tau-1},\theta,\pi)$ — Parameter distribution $P(\theta)$ — Global prior $p^\Phi(\mathrm{o}_\tau)$

```
Initialize factorized belief over action sequences q(π) ← N(0, I).
for optimisation iteration i = 1...I do
    Sample J candidate policies from q(π)
    for candidate policy j = 1...J do
        π^(j) ~ q(π)
        -F̃^j_π = 0
        for τ = t...t + H do
            q(s_τ|s_{τ-1}, θ, π^(j)) = E_{q(s_{τ-1}|θ,π^(j))}[p(s_τ|s_{τ-1}, θ, π^(j))]
            q(o_τ|s_τ, θ, π^(j)) = E_{q(s_τ|θ,π^(j))}[p(o_τ|s_τ)]
            -F̃^j_π ← -F̃^j_π + E_{q(s_τ,θ|π^(j))}[D_KL(q(o_τ|s_τ, θ, π^(j))||p^Φ(o_τ))] +
            H[q(s_τ|s_{τ-1}, θ, π^(j))] - E_{q(θ)}[H[q(s_τ|s_{τ-1}, π^(j), θ)]]
        end
    end
    q(π) ← refit(-F̃^j_π)
end
return q(π)
```

C Related work

Active inference There is an extensive literature on active inference in discrete state-spaces, covering a wide variety of tasks, from epistemic foraging in saccades (Parr & Friston, 2017; Friston, 2019b; Schwartenbeck et al., 2019), exploring mazes (Friston et al., 2015; Pezzulo et al., 2016; Friston et al., 2016), to playing Atari games (Cullen et al., 2018). Active inference also comes equipped with a well-developed neural process theory (Friston et al., 2017a; Parr et al., 2019) which can account for a substantial range of neural dynamics. There have also been prior attempts to scale up active inference to continuous RL tasks (Tschantz et al., 2019a; Millidge, 2019; Ueltzhffer, 2018), which we build upon here.

Model based RL Model based reinforcement learning has been in a recent renaissance, with implementations vastly exceeding the sample efficiency of model-free methods, while also approaching their asymptotic performance (Ha & Schmidhuber, 2018; Nagabandi et al., 2018; Chua et al., 2018a; Hafner et al., 2018). There have been recent successes on challenging domains such as Atari (Kaiser et al., 2019), and high dimensional robot locomotion (Hafner et al., 2018; 2019) and manipulation (Nagabandi et al., 2019) tasks. Key advances include variational autoencoders (Kingma & Welling, 2013) to flexibly construct latent spaces in partially observed environments, Bayesian approaches such as Bayes by backprop (Houthooft et al., 2016a), deep ensembles (Shyam et al., 2018; Chua et al., 2018a), and other variational approaches (Okada & Taniguchi, 2019; Tschiatschek et al., 2018; Yarin Gal et al., 2016), which quantify uncertainty in the dynamics models, and enable the model to learn a latent space that is useful for action (Tschantz et al., 2019b; Watter et al., 2015). Finally, progress has been aided by powerful planning algorithms capable of online planning in continuous state and action spaces (Williams et al., 2016; Rubinstein, 1997).

Intrinsic Measures Using intrinsic measures to encourage exploration has a long history in RL (Schmidhuber, 1991; 2007; Storck et al., 1995; Oudeyer & Kaplan, 2009; Chentanez et al., 2005). Recent model-free and model based-intrinsic measures that have been proposed in the literature include policy entropy (Rawlik, 2013; Rawlik et al., 2013; Haarnoja et al., 2018), state entropy (Lee et al., 2019), information-gain (Houthooft et al., 2016b; Okada & Taniguchi, 2019; Kim et al., 2018; Shyam et al., 2019; Teigen, 2018), prediction error (Pathak et al., 2017), divergence of ensembles (Shyam et al., 2019; Chua et al., 2018b), uncertain state bonuses (Bellemare et al., 2016; O'Donoghue et al., 2017), and empowerment (de Abril & Kanai, 2018; Leibfried et al., 2019; Mohamed & Rezende, 2015). Information gain additionally has a substantial history outside the RL framework, going back to (Lindley, 1956; Still & Precup, 2012; Sun et al., 2011).

D DERIVATION FOR THE FREE ENERGY OF THE EXPECTED FUTURE

We begin with the full free energy of the expected future and decompose this into the free energy of the expected future given policies, and the negative policy entropy:

$$\begin{aligned} \tilde{\mathcal{F}} &= \mathbb{E}_{q(\mathrm{o},\mathrm{s},\theta,\pi)}[\log q(\mathrm{o},\mathrm{s},\theta,\pi) - \log p^{\Phi}(\mathrm{o},\mathrm{s},\theta)] \\ &= \mathbb{E}_{q(\pi)}[\tilde{\mathcal{F}}_{\pi}] - \mathbf{H}[q(\pi)] \end{aligned} \tag{7}$$

We now show the free energy of the expected future given policies can be decomposed into extrinsic and information gain terms:

$$\begin{aligned} \tilde{\mathcal{F}}_{\pi} &= \mathbb{E}_{q(\mathrm{o},\mathrm{s},\theta|\pi)}[\log q(\mathrm{o},\mathrm{s},\theta|\pi) - \log p^{\Phi}(\mathrm{o},\mathrm{s},\theta)] \\ &= \mathbb{E}_{q(\mathrm{o},\mathrm{s},\theta|\pi)}[\log q(\mathrm{s},\theta|\pi) + \log q(\mathrm{o}|\mathrm{s},\theta,\pi) - \log p(\mathrm{s},\theta|\mathrm{o}) - \log p^{\Phi}(\mathrm{o})] \\ &\approx \mathbb{E}_{q(\mathrm{o},\mathrm{s},\theta|\pi)}[\log q(\mathrm{s},\theta|\pi) + \log q(\mathrm{o}|\mathrm{s},\theta,\pi) - \log q(\mathrm{s},\theta|\mathrm{o},\theta) - \log p^{\Phi}(\mathrm{o})] \\ &= \mathbb{E}_{q(\mathrm{o},\mathrm{s},\theta|\pi)}[\log q(\mathrm{s},\theta|\pi) - \log q(\mathrm{s},\theta|\mathrm{o},\pi)] + \mathbb{E}_{q(\mathrm{o},\mathrm{s},\theta|\pi)}[\log q(\mathrm{o}|\mathrm{s},\theta,\pi) - \log p^{\Phi}(\mathrm{o})] \\ -\tilde{\mathcal{F}}_{\pi} &= \mathbb{E}_{q(\mathrm{o},\mathrm{s},\theta|\pi)}[\log q(\mathrm{s},\theta|\mathrm{o},\pi) - \log q(\mathrm{s},\theta|\pi)] + \mathbb{E}_{q(\mathrm{o},\mathrm{s},\theta|\pi)}[\log p^{\Phi}(\mathrm{o}) - \log q(\mathrm{o}|\mathrm{s},\theta,\pi)] \\ &= \underbrace{\mathbb{E}_{q(\mathrm{o}|\pi)}\Big[D_{\mathrm{KL}}\Big(q(\mathrm{s},\theta|\mathrm{o},\pi)\|q(\mathrm{s},\theta|\pi)\Big)\Big]}_{\text{Expected Information Gain}} - \underbrace{\mathbb{E}_{q(\mathrm{s},\theta|\pi)}\Big[D_{\mathrm{KL}}\Big(q(\mathrm{o}|\mathrm{s},\theta,\pi)\|p^{\Phi}(\mathrm{o})\Big)\Big]}_{\text{Extrinsic Value}} \end{aligned} \tag{8}$$

Where we have assumed that $p(\mathrm{s},\theta|\mathrm{o}) \approx q(\mathrm{s},\theta|\mathrm{o},\pi)$. We wish to minimize $\tilde{\mathcal{F}}_{\pi}$, and thus maximize $-\tilde{\mathcal{F}}_{\pi}$. This means we wish to maximize the information gain and minimize the KL-divergence between expected and preferred observations.

By noting that $q(\mathrm{s}, \theta|\mathrm{o}, \pi) \approx q(\mathrm{s}|\mathrm{o}, \pi)q(\theta|\mathrm{s})$, we can split the expected information gain term into state and parameter information gain terms:

$$\begin{aligned}
&\mathbb{E}_{q(\mathrm{o}|\pi)}\Big[D_{\mathrm{KL}}\Big(q(\mathrm{s}, \theta|\mathrm{o}, \pi)\|q(\mathrm{s}, \theta|\pi)\Big)\Big] \\
&= \mathbb{E}_{q(\mathrm{o}|\pi)q(\mathrm{s},\theta|\mathrm{o},\pi)}\big[\log q(\mathrm{s}, \theta|\mathrm{o}, \pi) - \log q(\mathrm{s}, \theta|\pi)\big] \\
&= \mathbb{E}_{q(\mathrm{o}|\pi)q(\mathrm{s},\theta|\mathrm{o},\pi)}\big[\log q(\mathrm{s}|\mathrm{o}, \pi) + \log q(\theta|\mathrm{s}) - \log q(\mathrm{s}|\theta, \pi) - \log q(\theta)\big] \\
&= \mathbb{E}_{q(\mathrm{o}|\pi)q(\mathrm{s},\theta|\mathrm{o},\pi)}\big[\log q(\mathrm{s}|\mathrm{o}, \pi) - \log q(\mathrm{s}|\theta, \pi)\big] + \mathbb{E}_{q(\mathrm{o}|\pi)q(\mathrm{s},\theta|\mathrm{o},\pi)}\big[\log q(\theta|\mathrm{s}) - \log q(\theta)\big] \\
&= \underbrace{\mathbb{E}_{q(\mathrm{o}|\pi)q(\theta)}\Big[D_{\mathrm{KL}}\big(q(\mathrm{s}|\mathrm{o}, \pi)\|q(\mathrm{s}|\theta)\big)\Big]}_{\text{Expected State Information Gain}} + \underbrace{\mathbb{E}_{q(\mathrm{s}|\theta)}\Big[D_{\mathrm{KL}}\big(q(\theta|\mathrm{s})\|q(\theta)\big)\Big]}_{\text{Expected Parameter Information Gain}}
\end{aligned} \tag{9}$$

E Relationship to expected free energy

In the active inference literature, policy selection minimises expected free energy $\mathcal{G}$, which is defined for a policy π as (Friston, 2019b):

$$\mathcal{G}(\pi) = \mathbb{E}_{q(\mathrm{o},\mathrm{s},\theta|\pi)}[\log q(\mathrm{s}, \theta|\pi) - \log p^{\Phi}(\mathrm{o}, \mathrm{s}, \theta)] \tag{10}$$

Expected free energy also decomposes into an expected information gain term and a (different) extrinsic term. We have proposed the free energy of the expected future as a viable alternative to approximating the free energy that will occur from some policy. While $\mathcal{G}$ and $\tilde{\mathcal{F}}$ look extremely similair, they have a number of different properties. We leave it to future work to explore these in full.

F Derivation of the optimal policy

We derive the distribution for $q(\pi)$ which minimizes $\tilde{\mathcal{F}}$:

$$\begin{aligned}
\tilde{\mathcal{F}} &= D_{\mathrm{KL}}\Big(q(\mathrm{o}, \mathrm{s}, \theta, \pi)\|p^{\Phi}(\mathrm{o}, \mathrm{s}, \theta)\Big) \\
&= \mathbb{E}_{q(\mathrm{o},\mathrm{s},\theta,\pi)}[\log q(\mathrm{o}, \mathrm{s}, \theta|\pi) + \log q(\pi) - \log p^{\Phi}(\mathrm{o}, \mathrm{s}, \theta, \pi)] \\
&= \mathbb{E}_{q(\pi)}\Big[\mathbb{E}_{q(\mathrm{o},\mathrm{s},\theta|\pi)}[\log q(\pi) - [\log p^{\Phi}(\mathrm{o}, \mathrm{s}, \theta) - \log q(\mathrm{o}, \mathrm{s}, \theta|\pi)]\Big] \\
&= \mathbb{E}_{q(\pi)}\Big[\log q(\pi) - \mathbb{E}_{q(\mathrm{o},\mathrm{s},\theta|\pi)}[\log p^{\Phi}(\mathrm{o}, \mathrm{s}, \theta) - \log q(\mathrm{o}, \mathrm{s}, \theta|\pi)]\Big] \\
&= \mathbb{E}_{q(\pi)}\Big[\log q(\pi) - \big[-\mathbb{E}_{q(\mathrm{o},\mathrm{s},\theta|\pi)}[\log q(\mathrm{o}, \mathrm{s}, \theta|\pi) - \log p^{\Phi}(\mathrm{o}, \mathrm{s}, \theta)]\big]\Big] \\
&= \mathbb{E}_{q(\pi)}\Big[\log q(\pi) - \log e^{-\big[-\mathbb{E}_{q(\mathrm{o},\mathrm{s},\theta|\pi)}[\log q(\mathrm{o},\mathrm{s},\theta|\pi) - \log p^{\Phi}(\mathrm{o},\mathrm{s},\theta)]\big]}\Big] \\
&= \mathbb{E}_{q(\pi)}\Big[\log q(\pi) - \log e^{-D_{\mathrm{KL}}\big(q(\mathrm{o},\mathrm{s},\theta|\pi)\|p^{\Phi}(\mathrm{o},\mathrm{s},\theta)\big)}\Big] \\
&= D_{\mathrm{KL}}\Big(q(\pi)\|e^{-D_{\mathrm{KL}}\big(q(\mathrm{o},\mathrm{s},\theta|\pi)\|p^{\Phi}(\mathrm{o},\mathrm{s},\theta)\big)}\Big) \\
&= D_{\mathrm{KL}}\Big(q(\pi)\ \|e^{-\tilde{\mathcal{F}}_{\pi}}\Big)
\end{aligned} \tag{11}$$

G Expected information gain

In Eq. 3, expected parameter information gain was presented in the form $\mathbb{E}_{q(\mathrm{s}|\theta)}D_{\mathrm{KL}}\big(q(\theta|\mathrm{s})\|q(\theta)\big)$. While this provides a nice intuition about the effect of the information gain term on behaviour, it cannot be computed directly, due to the intractability of identifying true posteriors over parameters. We here show that, through a simple application of Bayes' rule, it is straightforward to derive an equivalent expression for the expected information gain as the divergence between the state likelihood and marginal, given the parameters, which decomposes into an entropy of an average minus an average of entropies:

$$\begin{aligned}
&\mathbb{E}_{q(\mathrm{s}|\theta)} D_{\mathrm{KL}}\Big(q(\theta|\mathrm{s})\|q(\theta)\Big) \\
&= \mathbb{E}_{q(\mathrm{s}|\theta)q(\theta|\mathrm{s})}\big[\log q(\theta|\mathrm{s}) - \log q(\theta)\big] \\
&= \mathbb{E}_{q(\mathrm{s},\theta)}\big[\log q(\mathrm{s}|\theta) + \log q(\theta) - \log q(\mathrm{s}) - \log q(\theta)\big] \\
&= \mathbb{E}_{q(\mathrm{s},\theta)}\big[\log q(\mathrm{s}|\theta) - \log q(\mathrm{s})\big] \\
&= \mathbb{E}_{q(\theta)q(\mathrm{s}|\theta)}\big[\log q(\mathrm{s}|\theta)\big] - \mathbb{E}_{q(\theta)q(\mathrm{s}|\theta)}\big[\log \mathbb{E}_{q(\theta)} q(\mathrm{s}|\theta)\big] \\
&= -\mathbb{E}_{q(\theta)}\mathbf{H}\big[q(\mathrm{s}|\theta)\big] + \mathbf{H}\big[\mathbb{E}_{q(\theta)} q(\mathrm{s}|\theta)\big]
\end{aligned} \tag{12}$$

The first term is the (negative) average of the entropies. The average over the parameters θ is achieved simply by averaging over the dynamics models in the ensemble. The entropy of the likelihoods $\mathbf{H}\big[p(\mathrm{s}|\theta)\big]$ can be computed analytically since each network in the ensemble outputs a Gaussian distribution for which the entropy is a known analytical result. The second term is the entropy of the average $\mathbf{H}\big[\mathbb{E}_{p(\theta)} p(\mathrm{s}|\theta)\big]$. Unfortunately, this term does not have an analytical solution. However, it can be approximated numerically using a variety of techniques for entropy estimation. In our paper, we use the nearest neighbour entropy approximation (Mirchev et al., 2018).

H Derivation of RL bound

Here we show that the free energy of the expected future is a bound on the divergence between expected and desired observations. The proof proceeds straightforwardly by importance sampling on the approximate posterior and then applying Jensen's inequality:

$$\begin{aligned}
D_{\mathrm{KL}}\big(q(\mathrm{o}_{t:T}|\pi)\|\, p^{\Phi}(o_{t:T})\big) &= \mathbb{E}_{q(\mathrm{o}_{t:T}|\pi)}\big[\log q(\mathrm{o}_{t:T}|\pi) - \log p^{\Phi}(\mathrm{o})\big] \\
&= \mathbb{E}_{q(\mathrm{o}_{t:T}|\pi)}\Bigg[\log\Big(\int dx_{1:T}\int d\theta_{1:T}\frac{q(\mathrm{o}_{t:T},\mathrm{s}_{t:T},\theta_{t:T}|\pi)q(\mathrm{s}_{t:T},\theta_{t:T}|\mathrm{o}_{t:T})}{p^{\Phi}(\mathrm{o}_{t:T})q(\mathrm{s}_{t:T},\theta_{t:T}|\mathrm{o}_{t:T})}\Big)\Bigg] \\
&\leq \mathbb{E}_{q(\mathrm{o}_{t:T},\mathrm{s}_{t:T},\theta_{t:T}|\pi)}\Big[\log\Big(\frac{q(\mathrm{o}_{t:T},\mathrm{s}_{t:T},\theta_{t:T}|\pi)}{p^{\Phi}(\mathrm{o}_{t:T},\mathrm{s}_{t:T},\theta_{t:T})}\Big)\Big] \\
&\leq D_{\mathrm{KL}}\Big(q(\mathrm{o}_{t:T},\mathrm{s}_{t:T},\theta|\pi)\|p^{\Phi}(\mathrm{o}_{t:T},\mathrm{s}_{t:T},\theta)\Big) = \tilde{\mathcal{F}}
\end{aligned} \tag{13}$$

I Model details

In the current work, we implemented our probabilistic model using an ensemble-based approach (Chua et al., 2018a; Fort et al., 2019; Chitta et al., 2018). Here, an ensemble of point-estimate parameters $\theta = \{\theta_0, ..., \theta_B\}$ trained on different batches of the dataset $\mathcal{D}$ are maintained and treated as samples as from the posterior distribution $p(\theta|\mathcal{D})$. Besides consistency with the active inference framework, probabilistic models enable the active resolution of model uncertainty, capture both epistemic and aleatoric uncertainty, and help avoid over-fitting in low data regimes (Fort et al., 2019; Chitta et al., 2018; Chatzilygeroudis et al., 2018; Chua et al., 2018b).

This design choice means that we use a trajectory sampling method when evaluating beliefs about future variables (Chua et al., 2018a), as each pass through the transition model $p(\mathrm{s}_t|\mathrm{s}_{t-1},\theta,\pi)$ evokes B samples from s_t.

Transition model We implement the transition model as $p(\mathrm{s}_t|\mathrm{s}_{t-1},\theta,\pi)$ as $\mathcal{N}(\mathrm{s}_t; f_\theta(\mathrm{s}_{t-1}), f_\theta(\mathrm{s}_{t-1}))$, where $f_\theta(\cdot)$ are a set of function approximators $f_\theta(\cdot) = \{f_{\theta_0}(\cdot), ..., f_{\theta_B}(\cdot)\}$. In the current paper, $f_{\theta_i}(\mathrm{s}_{t-1})$ is a two-layer feed-forward network with 400 hidden units and swish activation function. Following previous work, we predict state deltas rather than the next states (Shyam et al., 2018).

Reward model We implement the reward model as $p(\mathrm{o}_\tau|\mathrm{s}_\tau,\theta,\pi) = \mathcal{N}(\mathrm{o}_\tau; f_\lambda(\mathrm{s}_\tau), \mathbf{1})$, where $f_\lambda(\mathrm{s}_\tau)$ is some arbitrary function approximator[5]. In the current paper, $f_\lambda(\mathrm{s}_\tau)$ is a two layer feed for-

[5]Formally, this is an observation model, but we retain RL terminology for clarity.

ward network with 400 hidden units and ReLU activation function. Learning a reward model offers several plausible benefits outside of the active inference framework, as it abolishes the requirement that rewards can be directly calculated from observations or states (Chua et al., 2018a).

Global prior We implement the global prior $p^{\Phi}(\mathrm{o})$ as a Gaussian with unit variance centred around the maximum reward for the respective environment. We leave it to future work to explore the effects of more intricate priors.

J Implementation details

For all tasks, we initialize a dataset $\mathcal{D}$ with a single episode of data collected from a random agent. For each episode, we train the ensemble transition model and reward model for 100 epochs, using the negative-log likelihood loss. We found cold-starting training at each episode to lead to more consistent behaviour. We then let the agent act in the environment based on Algorithm 1, and append the collected data to the dataset $\mathcal{D}$.

We list the full set of hyperparameters below:

Hyperparameters	
Hidden layer size	400
Learning rate	0.001
Training-epochs	100
Planning-horizon	30
N-candidates (CEM)	700
Top-candidates (CEM)	70
Optimisation-iterations (CEM)	7

K Environment details

The Mountain Car environment ($\mathcal{S} \subseteq \mathbb{R}^2 \mathcal{A} \subseteq \mathbb{R}^1$) requires an agent to drive up the side of a hill, where the car is underactuated requiring it first to gain momentum by driving up the opposing hill. A reward of one is generated when the agent reaches the goal, and zero otherwise. The Cup Catch environment ($\mathcal{S} \subseteq \mathbb{R}^8 \mathcal{A} \subseteq \mathbb{R}^2$) requires the agent to actuate a cup and catch a ball attached to its bottom. A reward of one is generated when the agent reaches the goal, and zero otherwise. The Half Cheetah environment ($\mathcal{S} \subseteq \mathbb{R}^1 7\mathcal{A} \subseteq \mathbb{R}^6$) describes a running planar biped. For the running task, a reward of $v - 0.1||a||^2$ is received, where v is the agent's velocity, and for the flipping task, a reward of $\epsilon - 0.1||a||^2$ is received, where ϵ is the angular velocity. The Ant Maze environment ($\mathcal{S} \subseteq \mathbb{R}^2 9\mathcal{A} \subseteq \mathbb{R}^8$) involves a quadruped agent exploring a rectangular maze.