Empowerment, Free Energy Principle and Maximum Occupancy Principle Compared

Rubén Moreno-Bote
Department of Communication and Information Technologies,
Serra Húnter Fellow Programme
Universitat Pompeu Fabra
Barcelona, Spain
ruben.moreno@upf.edu

Jorge Ramírez-Ruiz
Department of Communication and Information Technologies
Universitat Pompeu Fabra
Barcelona, Spain
jorge.ramirez@upf.edu

Abstract

While the objective of reward maximization in reinforcement learning has lead to impressive achievements in several games and artificial environments, animals seem to be driven by intrinsic signals that are not purely extrinsic, such as curiosity. Several reward-free approaches have emerged in the fields of cognitive neuroscience and artificial intelligence that primarily make use of signals different from extrinsic rewards to guide exploration and ultimately drive behavior, but a comparison between these approaches is lacking. Here we focus on two popular reward-free approaches, known as empowerment (MPOW) and free energy principle (FEP), and a recently developed one, called maximum occupancy principle (MOP), and compare them in sequential problems and fully-observable environments. We find that MPOW shows a preference for unstable fixed points of the dynamical system that defines the agent and environment. FEP is shown to be equivalent to reward maximization in certain cases. None of these two principles of behavior seem to consistently generate variable behavior: behavior collapses within a small repertoire of possible action-state trajectories or fixed points. Collapse to an optimal deterministic policy can be proved in specific, recent implementations of FEP, with the only exception of policy degeneracy due to ties. In contrast, MOP consistently generates variable action-state trajectories. In two simple environments, a balancing cartpole and a grid world, we find that both MPOW and FEP agents stick to a relatively small set of states and actions, while MOP agents generate short of exploratory and dancing-like motions.

1 Introduction

The objective of reward or utility maximization pervades the fields of reinforcement learning (RL) [1], economics [2] and psychology [3]. However, not all behaviors seem to be explainable in terms of reward maximization. For instance, spontaneous movements of babies [4] or curiosity in kids [5] are often taken as examples of intrinsically motivated behaviors that are driven independently of extrinsic rewards. Thus, several frameworks have emerged where intrinsic signals are major drives

Information-Theoretic Principles in Cognitive Systems Workshop at the 37th Conference on Neural Information Processing Systems (NeurIPS 2023).

of behavior: information seeking [6, 7], novelty seeking [8], empowerment [9, 10], minimizing free energy [11, 12], or occupying action-state space [13] are goals that per se are independent of reward maximization, and seem to capture certain aspects of the observed behavior of living creatures.

As each of these reward-free algorithms are designed to maximize a different objective, a comparison between them has proven difficult, and to our knowledge experimental comparison is lacking [14]. This difficulty does not arise in standard reward-maximization algorithms, as all of them are compared on the basis of the reward that each delivers in a common, agreed-upon benchmark [15]. One possible route to compare reward-free approaches is to generate behaviors (action-state trajectories) from them and study their differences. Which reward-free objective leads to behaviors that are more complex, rich and interesting?

In this paper, we focus on three reward-free approaches and compared their resulting behaviors in two different environments in a Markov Decision Process setting. We chose two popular sequential reward-free approaches, Empowerment (MPOW) [9, 10] and the Free Energy Principle (FEP)[11, 12, 16], and a recently developed one, Maximum Occupancy Principle (MOP) [13, 17]. These approaches have been used across several disciplines, and therefore they cover a wide spectrum of ideas and constrains. In MPOW, an agent seeks to maximize the mutual information between actions and the resulting state transitions at some future time; MPOW agent thus tend to move to regions of action-state space where there is a large repertoire of possible controllable transitions. In FEP, an agent seeks to minimize the KL divergence between the distribution of states and a target distribution that models desirable states; FEP agents thus tend to move to regions of action-state space where it is possible to keep a homeostatic balance. In MOP, an agent seeks to maximize a combination of future cumulative action and state path entropy; MOP agents thus tend to move to regions where it is possible to generate more actions and visit more states.

Although these are probably the most popular reward-free approaches, they do not fully cover the whole spectrum. Further, we purposely assume that the environments are fully observable and known, so that there is nothing to learn and to explore, and thus novelty and information seeking lose their relevance. Although this is indeed a limited setup, it is the one that has largely been employed in reward-free problems [10, 12, 16, 18, 13, 17]. By virtue of simplifying the environments, we also expect to discover some fundamental differences between the three approaches that could be obscured in more complex environments; these differences are expected to be robust to increasing complexity.

2 Describing and comparing MPOW, FEP and MOP

We are interested in modelling sequential processes, where actions are made and state transitions are observed in sequence. We assume that the agent follows a Markov Decision Process in discrete time. At time t the agent is at state $s_t \in \mathcal{S}$ and performs action $a_t \in \mathcal{A}$ with probability (density) $\pi(a_t|s_t)$, which defines the policy π. Then, a transition results to state s_{t+1} at the next time $t+1$ with probability (density) $p(s_{t+1}|s_t, a_t)$. The goals of the three reward-free approaches are:

Empowerment (MPOW): Our implementation is identical to the one used in [10]. Given current state s_t, the agent computes the empowerment of all of its successor states. The agent then moves to the state with the highest empowerment. Empowerment of a state s_t, denoted $\mathcal{E}(s_t)$, is defined as the maximum mutual information (channel capacity) between the n-step actions $a_t^n = (a_t, a_{t+1}, ..., a_{t+n-1}) \in \mathcal{A}^n$ and the resulting state s_{t+n} (actions are planned in an open loop manner, that is, without any feedback being provided about intermediate states), and it is written as

$$\mathcal{E}(s_t) = \max_{\tau(a_t^n|s_t)} \sum_{a_t^n, s_{t+1}} p(s_{t+n}|s_t, a_t^n)\tau(a_t^n|s_t) \log\left(\frac{p(s_{t+n}|s_t, a_t^n)}{\sum_{a_t^n} p(s_{t+n}|s_t, a_t^n)\tau(a_t^n|s_t)}\right), \quad (1)$$

where $\tau(a_t^n|s_t)$ is the probability distribution of n-step actions ("imagined" n-step policy, not actually used to generate actions) that mutual information is maximized over, and $p(s_{t+n}|s_t, a_t^n) = \tau(a_t^n|s_t) \prod_{\tau=t}^{t+n-1} p(s_{\tau+1}|s_\tau, a_\tau)$ is the conditional n-step state transition probability. The horizon n is chosen to be small (< 10), so that the integrals in Eq. (1) are numerically tractable. For continuous action and state spaces, the sums are replaced by integrals. The optimization of the n-step policy is done via the Blahut-Arimoto algorithm [19]

Biases from MPOW: The agent tends to visit regions of action-state space where there are many future accessible states but where actions can predictably lead to them – this maximizes mutual information between future states and the sequence of actions. In a sense, the agent looks for states where it is in control of what it will happen. In deterministic continuous dynamical systems, highly empowered states typically correspond to unstable fixed points, as in those states small actions lead to large predictable state changes.

Remarks: MPOW is based on maximizing mutual information, which is not additive over action-state paths [13]. This implies that no Bellman equation (recursion) exists for MPOW. One consequence of lacking a Bellman equation is that information that has been gained about the value (empowerment) of a state cannot be reused in the future, and thus empowerment needs to be fully recomputed at every step. One attempt of writing the objective in Eq. (1) as a Bellman equation is based on dividing the trajectory in one-step mutual information and maximize its discounted sum [20], but the sum does not in general equal the mutual information in the full trajectory [13].

Free Energy Principle (FEP): Here we follow the sophisticated active inference formalization of FEP, where an agent computes its expected free energy (EFE) assuming it will follow a policy that minimizes its EFE in the future [16]. The agent, starting at state s_t at time t and performing actions $a_t^{T-1} = (a_t, a_{t+1}, ..., a_{T-1})$, traverses the states $s_{t+1}^T = (s_{t+1}, s_{t+2}, ..., s_T)$ up to some final time T that defines the end of an episode. The agent is characterized by a target probability distribution $q(s_t^T)$ assigning large probability to a subset of states ("desired" states). Desired states define "homeostatic" configurations of the agent that allows its survival, and the probability distribution defines how big the set of those desired configurations is. Formally, the goal of a FEP agent is to find the 1-step policy π that minimizes the expected KL divergence between the generated and the target distribution, defined as

$$\begin{aligned} G_{\pi,t}(s_t) &= \mathbb{E}_{a_t^{T-1}\sim\pi}\mathrm{KL}\left(p(s_{t+1}^T|a_t^{T-1}, s_t)||q(s_{t+1}^T)\right) \\ &= \sum_{s_{t+1}^T, a_t^{T-1}} p_\pi(s_{t+1}^T, a_t^{T-1}|s_t)\log\frac{p(s_{t+1}^T|a_t^{T-1}, s_t)}{q(s_{t+1}^T)}\ , \end{aligned} \tag{2}$$

where $p_\pi(s_{t+1}^T, a_t^{T-1}|s_t) = \prod_{\tau=t}^{T-1}\pi(a_\tau|s_\tau)p(s_{\tau+1}|s_\tau, a_\tau)$ and $p(s_{t+1}^T|a_t^{T-1}, s_t) = \prod_{\tau=t}^{T-1} p(s_{\tau+1}|s_\tau, a_\tau)$ are the joint probabilities of future states and actions, and their conditional, respectively, given an initial state. We assume that the target probability $q(s_{t+1}^T)$ factorizes as $q(s_{t+1}^T) = \prod_{\tau=t}^{T-1} q(s_{\tau+1})$, where $q(s)$ is a time-independent probability describing the "desired" states of the agent, modelling the idea that desired states are time-independent.

Note that because the time horizon is finite, here we need to consider time-dependent policies, so $\pi(a_t|s_t)$ is understood as the probability of selecting action a_t at time t given that the state at time t is s_t. Time-independent policies will be suboptimal in general in finite horizon MDPs.

By virtue of the Markov property, the objective in Eq. (2) can be recursively written as

$$G_{\pi,t}(s_t) = \sum_{s_{t+1}, a_t} \pi(a_t|s_t)p(s_{t+1}|s_t, a_t)\left[\log\frac{p(s_{t+1}|s_t, a_t)}{q(s_{t+1})} + G_{\pi,t+1}(s_{t+1})\right] \tag{3}$$

for $t < T-1$ and the terminal cost is

$$G_{\pi,T-1}(s_{T-1}) = \sum_{s_T, a_{T-1}} \pi(a_{T-1}|s_{T-1})p(s_T|s_{T-1}, a_{T-1})\log\frac{p(s_T|s_{T-1}, a_{T-1})}{q(s_T)}, \tag{4}$$

as at time T the episode terminates. Note that the above formalization slightly generalizes EFE [16] by allowing the possibility that the optimal policy is stochastic.

In Sec. (A.1) we show (i) that the optimal policy is deterministic and (ii) that in deterministic environments the FEP agent is identical to a reward-maximizing agent with reward $\log q(s)$.

Biases from FEP: An FEF agent minimizes (2), and therefore its goal is to minimize the KL divergence between the future and desired state distributions by selecting a suitable policy. This amounts to finding a policy that makes the desired states highly probable under the environment dynamics.

Maximum Occupancy Principle (MOP): We follow the formalization of MOP provided in [13]. The goal of a MOP agent is to maximize the cumulative future (weighted) action and state transition entropy. Starting at $t = 0$ in state s_0, an agent performing a sequence of actions and experiencing

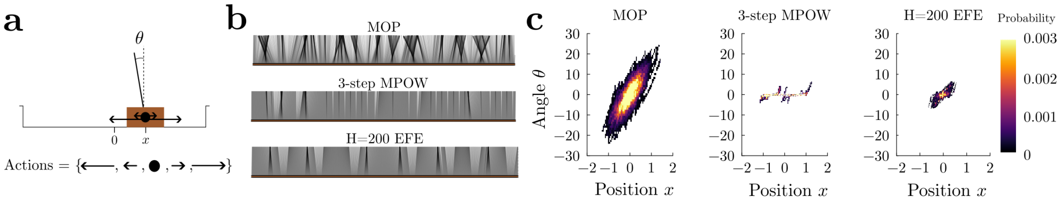

Figure 1: Dancing in a MOP cartpole, but not in MPOW and FEP (EFE implementation) cartpoles. (a) The cart has a pole attached. (b) Time-shifted snapshots of the pole in the reference frame of the cart as a function of time. (c) Position and angle occupation for for the three agents.

state transitions traverses the path $\tau \equiv (s_0, a_0, s_1, ..., a_t, s_{t+1}, ...)$, and gets a return (intrinsic reward) equal to

$$R(\tau) = -\sum_{t=0}^{\infty} \gamma^t \ln\left(\pi^{\alpha}(a_t|s_t) p^{\beta}(s_{t+1}|s_t, a_t)\right) , \tag{5}$$

with weights $\alpha > 0$ and $\beta \geq 0$ and discount factor $0 < \gamma < 1$. A large return is obtained when a low-probability action is followed by a low-probability transition. This captures the idea that a MOP agent tends to "occupy" actions and states that have a low visit probability. Action and state entropy is a consistent way of quantifying the intuitive notion of action-state occupancy [13].

The agent is assumed to follow the policy π that maximizes the state-value $V_\pi(s)$, defined as

$$\begin{aligned} V_\pi(s) &\equiv \mathbb{E}_{a_t \sim \pi(a_t|s_t), s_{t+1} \sim p(s_{t+1}|s_t, a_t)}[R(\tau)|s_0 = s] \\ &= \mathbb{E}_{a_t \sim \pi(a_t|s_t), s_{t+1} \sim p(s_{t+1}|s_t, a_t)} \left[\sum_{t=0}^{\infty} \gamma^t \left(\alpha \mathcal{H}(A|s_t) + \beta \mathcal{H}(S'|s_t, a_t) \right) \Big| s_0 = s \right] , \end{aligned} \tag{6}$$

which is the expected return given the initial condition $s_0 = s$ and following policy π, where we have defined the entropies $\mathcal{H}(A|s) = -\sum_a \pi(a|s) \ln \pi(a|s)$ and $\mathcal{H}(S'|s, a) = -\sum_{s'} p(s'|s, a) \ln p(s'|s, a)$. An optimality Bellman equation exists for Eq. (6) (Sec. A.2). The advantage of the existence of a Bellman equation is algorithmic efficiency, as the values that have been learnt for some states can be reused in the future if during planning the agent expects to visit those states.

Biases from MOP: Maximizing the value function in Eq. (6) over policies π means that a MOP agent will tend to look for actions that lead to future states with many potential actions and state transitions. Terminal states are avoided to keep obtaining future action and state entropy. States that provide energy for the agent are also looked for because they guarantee further future motion and thus further future action-state entropy.

Comparison between MPOW, FEP and MOP. Comparing MPOW, FEP and MOP requires making choices about model parameters beyond environment specifications. MPOW has only the number of look-ahead actions n as free parameter. FEP requires defining the horizon $T = H$ (a receding horizon in our simulations), but more importantly the target distribution $q(s)$. As we are interested in generating rich behavior, we do not restrict the repertoire of desired states: we take $q(s)$ to be approximately uniform, except for the terminal states, which need to be avoided. Specifically, terminal states are given low target probabilities, while all other accessible states have similarly high target probability. MOP has two free parameters, the ratio β/α and the discount factor γ. In deterministic environments state transition entropy is equivalent to action entropy, and thus we arbitrarily set $\beta = 0$ and $\alpha = 1$.

We have tested the three agents in a cartpole environment (see Sec. A.4.2 for details). A cart can move laterally within some boundaries, and a pole is attached to it (Fig. 1a). A terminal state is reached if the angle $|\theta| > 36\,\mathrm{deg}$ or the borders are hit. The three agents thus try to avoid the terminal states while maximizing their respective objectives. We find that MPOW and FEP agents tend to prefer upright pole positions, while the MOP agent generates some sort of dancing behavior (Fig. 1b). The upright pole position is the configuration with highest empowerment, and thus it is selected by the MPOW agent. The FEP agent in the EFE implementation acts as a reward maximizer, and thus it looks for the safest state: the upright position is the one furthest from the terminal states. In contrast, the MOP agent generates random, but guided, actions to occupy action state while avoiding the terminal states (Fig. 1c, see Video 1). Analogous differences are observed in a 4-room grid world with food sources, where agents can die due to starvation (Appendix Sec. A.3; see Video 2).

3 Conclusions

We have contrasted three reward-free approaches (MPOW, FEP and MOP agents) in two different environments. While MPOW and FEP prefers unstable fixed points and low-risk states, respectively, MOP prefers continual motion.

Acknowledgments and Disclosure of Funding

This project was supported by grants from the Howard Hughes Medical Institute (HHMI; Ref: 55008742), The Bial Foundation (Ref: 106/2022), Ministerio de Ciencia e Innovación (Ref: PID2020-114196GB-I00/AEI) and ICREA Academia (2022) to R.M.B; and MINECO/ESF (Spain; PRE2018-084757) to J.R.-R.

References

[1] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[2] John Von Neumann and Oskar Morgenstern. *Theory of games and economic behavior*. Princeton university press, 2007.

[3] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific, 2013.

[4] Karen E Adolph and Sarah E Berger. Motor development. *Handbook of child psychology*, 2, 2007.

[5] Celeste Kidd and Benjamin Y Hayden. The psychology and neuroscience of curiosity. *Neuron*, 88(3):449–460, 2015.

[6] Susanne Still and Doina Precup. An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, 131(3):139–148, 2012.

[7] Jacqueline Gottlieb, Pierre-Yves Oudeyer, Manuel Lopes, and Adrien Baranes. Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends in cognitive sciences*, 17(11):585–593, 2013.

[8] Joel Lehman and Kenneth O Stanley. Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary computation*, 19(2):189–223, 2011.

[9] Alexander S Klyubin, Daniel Polani, and Chrystopher L Nehaniv. Empowerment: A universal agent-centric measure of control. In *2005 ieee congress on evolutionary computation*, volume 1, pages 128–135. IEEE, 2005.

[10] Tobias Jung, Daniel Polani, and Peter Stone. Empowerment for continuous agent—environment systems. *Adaptive Behavior*, 19(1):16–39, 2011.

[11] Karl Friston, Philipp Schwartenbeck, Thomas FitzGerald, Michael Moutoussis, Timothy Behrens, and Raymond J Dolan. The anatomy of choice: active inference and agency. *Frontiers in human neuroscience*, 7:598, 2013.

[12] Karl J Friston, Jean Daunizeau, and Stefan J Kiebel. Reinforcement learning or active inference? *PloS one*, 4(7):e6421, 2009.

[13] Jorge Ramírez-Ruiz, Dmytro Grytskyy, and Rubén Moreno-Bote. Seeking entropy: complex behavior from intrinsic motivation to occupy action-state path space. *arXiv preprint arXiv:2205.10316*, 2022.

[14] Danijar Hafner, Pedro A Ortega, Jimmy Ba, Thomas Parr, Karl Friston, and Nicolas Heess. Action and perception as divergence minimization. *arXiv preprint arXiv:2009.01791*, 2020.

[15] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018.

[16] Lancelot Da Costa, Noor Sajid, Thomas Parr, Karl Friston, and Ryan Smith. Reward maximization through discrete active inference. *Neural Computation*, 35(5):807–852, 2023.

[17] Dmytro Grytskyy, Jorge Ramírez-Ruiz, and Rubén Moreno-Bote. A general markov decision process formalism for action-state entropy-regularized reward maximization. *arXiv preprint arXiv:2302.01098*, 2023.

[18] Alexander Tschantz, Beren Millidge, Anil K Seth, and Christopher L Buckley. Reinforcement learning through active inference. *arXiv preprint arXiv:2002.12636*, 2020.

[19] R. Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory*, 18(4):460–473, July 1972. ISSN 1557-9654. doi: 10.1109/TIT.1972.1054855. Conference Name: IEEE Transactions on Information Theory.

[20] Felix Leibfried, Sergio Pascual-Diaz, and Jordi Grau-Moya. A unified bellman optimality principle combining reward maximization and empowerment. *Advances in Neural Information Processing Systems*, 32, 2019.

[21] Donald E Kirk. *Optimal control theory: an introduction*. Courier Corporation, 2004.

[22] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

A Appendix

A.1 Optimal policy for EFE

To find the optimal policy in Eqs. (3,4), we proceed backwards in time [21]. At time $T-1$ the optimal policy is deterministic because Eq. (4) is linear in the policy. The only exception is that there could be ties between several actions having the same value of the objective, in which case one arbitrary action can be always chosen, or any action can randomly be chosen. Therefore, the optimal action is

$$a^*_{T-1}(s_{T-1}) = \arg\min_a \sum_{s_T} p(s_T|s_{T-1},a) \log \frac{p(s_T|s_{T-1},a)}{q(s_T)} \ . \tag{7}$$

The optimal return at time $T-1$ becomes

$$G^*_{T-1}(s_{T-1}) = \sum_{s_T} p(s_T|s_{T-1},a^*_{T-1}(s_{T-1})) \log \frac{p(s_T|s_{T-1},a^*_{T-1}(s_{T-1}))}{q(s_T)} . \tag{8}$$

Proceeding backwards, with $t = T-2, T-3, ...$, we find that again for all times the optimal policy is deterministic, and that the optimal action is

$$a^*_t(s_t) = \arg\min_a \sum_{s_{t+1}} p(s_{t+1}|s_t,a) \left[\log \frac{p(s_{t+1}|s_t,a)}{q(s_{t+1})} + G^*_{t+1}(s_{t+1}) \right] \ , \tag{9}$$

where the optimal return is recursively computed as

$$G^*_t(s_t) = \sum_{s_{t+1}} p(s_{t+1}|s_t,a^*_t(s_t)) \left[\log \frac{p(s_{t+1}|s_t,a^*_t(s_t))}{q(s_{t+1})} + G^*_{t+1}(s_{t+1}) \right] \ . \tag{10}$$

In deterministic environments, the transition probability gives probability one just for a single successor state $s' = s'(s,a)$. In this case, the first sum in Eqs. (9,10) reduces to a single term, equal to $-\log q(s'(s,a))$, which can be readily interpreted as the immediate reward $R(s,a)$ [16] when the equations are mapped into classical dynamical programming of reward-maximization.

A.2 Bellman equation for MOP

The state-value $V_\pi(s)$ in Eq. (6) can be recursively written using the values of successor states through the standard Bellman equation

$$\begin{aligned} V_\pi(s) &= \alpha \mathcal{H}(A|s) + \beta \sum_a \pi(a|s) \mathcal{H}(S'|s,a) + \gamma \sum_{a,s'} \pi(a|s) p(s'|s,a) V_\pi(s') \\ &= \sum_{a,s'} \pi(a|s) p(s'|s,a) \left(-\alpha \ln \pi(a|s) - \beta \ln p(s'|s,a) + \gamma V_\pi(s') \right), \end{aligned} \tag{11}$$

where the sum is over the available actions a from state s and over the successor states s' given the performed action at state s. The optimal policy π^* that maximizes the state-value is defined as $\pi^* = \arg\max_\pi V_\pi$ and the optimal state-value is

$$V^*(s) = \max_\pi V_\pi(s), \tag{12}$$

where the maximization is with respect to the $\{\pi(\cdot|\cdot)\}$ for all actions and states.

A.3 Comparing MPOW, FEP and MOP in a grid world

We find that both MPOW and FEP agents visit states in a quite restricted manner compared to the MOP agent (Fig. 2). This is expected, as the MPOW agent tries to find states with large accessibility of states, which happens to be close to the food sources. Similarly, the FEP agent in this deterministic environment acts as a reward maximizer in terms of survival time, and thus it sticks close to a food source as well. In contrast, the MOP strikes a balance between visiting the food sources to get just enough energy so that it can generate actions in the future (see Video 2).

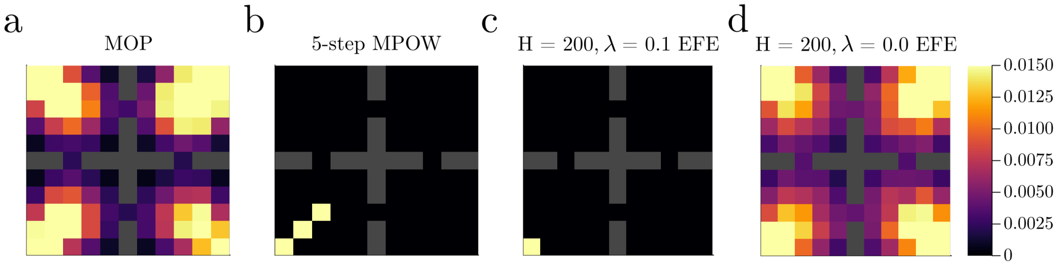

Figure 2: Large coverage for MOP agent in a gridworld. MPOW shows preference between standing away from the food, and gathering food (food source in the corners of arena). EFE agent with a small reward for the food stays on it forever ($\lambda > 0$ in Eq. A.4.3), and only a degenerate EFE agent ($\lambda = 0$) that maximizes survival shows large coverage.

A.4 Experiments

Both the grid world and the cartpole experiments have been described elsewhere [13], so we limit the description to the implementation of the algorithms.

A.4.1 Four-room grid world

The arena is composed of four rooms. At each of these rooms, there is a food source located in the corner furthest from the openings. The states are the Cartesian product between location (x, y) and internal state u. The internal state u is simply a scalar value between a minimum of 0 and a maximum capacity of 100, representing the agent's current energy level. All $u = 0$ states are absorbing states, independently of the location (x, y). The agent has a maximum of 9 actions: `up, down, left, right, up left, up right, down left, down right`, and `nothing`. Whenever the agent is next to a wall, the number of available actions decreases such that the agent cannot choose to go into walls. Finally, whenever the agent is in an absorbing state, only `nothing` is available. At any transition, there is a reduction in u of 1 unit for being alive. On the other hand, whenever the agent is located at a food source, there is an increase in u that we fix to 10. The initial condition for all agents is $s_0 = (x_0, y_0, u_0) = (3, 3, 30)$, and they all have a capacity of $u_{\text{max}} = 100$.

A.4.2 Cartpole

A cart is placed in a one-dimensional track with boundaries at $|x| = 1.8$. It has a pole attached to it, that rotates like an inverted pendulum with its pivot point on the cart. The system can be described by a four-dimensional external state (x, v, θ, ω), where x is the position of the cart, v is its linear velocity, θ is the angle of the pole with respect to the vertical which grows counterclockwise, and ω is its angular velocity. In this case, we model the internal state u simply with the binary variable `alive, dead`, where the agent enters the absorbing state `dead` if its position exceeds the boundaries, or if its angle exceeds 36 degrees. This amplitude of angles is larger than that typically assumed (12 degrees in [22]), and therefore our system is allowed to be more non-linear and unstable. Any time the agent is `alive`, it has 5 possible actions: forces of $\{-40, -10, 0, 10, 40\}$, where zero force is understood as `nothing`. If the agent is `dead`, then only `nothing` is allowed.

A.4.3 Agents

MOP agent

The objective function that this agent maximizes in general is Eq. (6). The α and β parameters control the weights of action and next-state entropies to the objective function, respectively. Unless indicated otherwise, we always use $\alpha = 1, \beta = 0$ for the experiments. Given that the environments are deterministic, then the next-state entropy is $\mathcal{H}(S'|s, a) = -\sum_{s'} p(s'|s, a) \ln p(s'|s, a) = 0$, and therefore β does not change the optimal policy.

We have implemented an iterative map to solve for the optimal value. It has been proven that this iterative map finds the unique optimal value regardless of the initial condition of the value function in the first orthant [13].

MPOW agent

Grid world: For the 4-room gridworld, we implemented empowerment in its original discrete formulation, in [9]. The agent looks ahead at all possible subsequent states s', computes their empowerment, and greedily chooses the action that corresponds to the successor state with highest empowerment (environment is deterministic). In our particular formulation, we allowed for a stochastic choice of action in case of empowerment ties between successor states. Given the nature of the arena, we implemented $5-$step empowerment, to give the agent enough lookahead to consider going into other rooms, while keeping the computations tractable, given the large amount of 1-step actions (9 for center cells).

Cartpole: For the case of the cartpole experiment, we implemented continuous-state empowerment, as developed in [10]. In order to have enough lookahead without needing high n, we used $3-$step empowerment with each action in the $3-$step action held constant for $k = 10$ time steps, in order for the computation of empowerment to be meaningfully different between states. The computation of empowerment is done similarly as in the gridworld, through a Blahut-Arimoto algorithm described in [10]. Similarly, the agent looks ahead at successor states, computes their empowerment and greedily chooses the action that corresponds to the state with the highest empowerment.

EFE agent

We constructed the preferred distribution using the reward scheme described in Ref. [16],

$$q_\lambda(s) = \frac{\exp(\lambda R(s))}{\sum_{s'} \exp(\lambda R(s'))},$$

where λ is an inverse temperature parameter that controls how reward-driven the agent is, and $R(s)$ is the reward of being in state s. This description leads to an EFE agent that is minimizing free energy as equivalent as maximizing reward.

In the case of infinite temperature, $\lambda = 0$, $q_0(s)$ is uniform over (living) state space, and thus the free energy is completely degenerate, rendering the EFE agent a survival maximizer. For small positive λ, $q_\lambda(s)$ is almost uniform, but has a little bump at around the reward location. Due to the $\min$ operator, even a tiny bump will create a massive preference of the reward state, as shown in Fig. (2c,d).

For any temperature, the preferred distribution $q_\lambda(s)$ only has support over living states, such that $q_\lambda(x, y, u = 0) = 0$. In other words, $R(\texttt{dead}) = -\infty$. In practice, in order for the information about absorbing states to propagate, we set $q_\lambda(x, y, u = 0) \ll 1$, such that $\log(q_\lambda(x, y, u = 0))$ is finite. The actual value is not relevant, as long as $\log(q_\lambda(x, y, u = 0))$ is less than zero, the reward for surviving, which we have verified empirically.

For the particular experiments, we chose the following reward functions,

Grid world:

$$R(x, y, u) = \begin{cases} 1, & \text{if } u > 0 \text{, and } (x, y) \text{ corresponds to a food source} \\ 0, & \text{if } u > 0 \text{, and } (x, y) \text{ is not a food source} \\ -\infty & \text{if } u = 0 \end{cases}$$

Cartpole:

$$R(x, y, u) = \begin{cases} 0, & \text{if } u = \texttt{alive} \\ -\infty & \text{if } u = \texttt{dead} \end{cases}$$

Additionally, we implemented a receding horizon for the EFE agent, which means that the optimal expected free energy was computed for a given horizon H, and the action at each time step was drawn from using the initial free energy at the given state for horizon H. The choice of horizon for the grid world is irrelevant for horizons that are longer than the diameter of a room ($H > 5$). For the cartpole, we chose the longest horizon at which MOP and EFE live about the same. Coverage results are robust for choice of horizon, for sufficiently long horizons.