Path Divergence Objective: Boundedly-Rational Decision Making in Partially Observable Environments

Tomáš Gavenčiak*
Alignment of Complex Systems
Charles University
Prague, Czech Republic

David Hyland*
Department of Computer Science
University of Oxford
Oxford, United Kingdom

Lancelot Da Costa
VERSES AI Research Lab
Los Angeles, USA

Michael Wooldridge
Department of Computer Science
University of Oxford
Oxford, United Kingdom

Jan Kulveit
Alignment of Complex Systems
Charles University
Prague, Czech Republic

Abstract

We introduce the Path Divergence Objective (PDO), an information-theoretic model of boundedly-rational decision-making in stochastic, partially-observable environments. The PDO is derived from physical principles, modelling the inherent costs of information processing in model-based planning for embodied agents. This framework enables us to model key features observed in real-world agent behaviour, such as curiosity-driven exploration, novelty-seeking, and the intention-behaviour gap. By adjusting a single parameter, the PDO can describe a continuous spectrum of decision-making strategies, ranging from highly irrational to perfectly rational. This flexibility makes the PDO applicable to a wide range of scenarios, including modelling biological organisms, simulating interactions between agents with varying degrees of bounded rationality, addressing AI alignment challenges, and designing AI systems that interact more effectively with humans.

1 Introduction

Accurately predicting and modelling the decisions of real-world agents—from humans to AI systems—remains a fundamental challenge across cognitive science, neuroscience, and artificial intelligence, including the alignment of AI systems to human preferences. While machine learning methods and capabilities have advanced significantly, progress in modelling real-world decision-making under cognitive and informational constraints has been comparatively slower. To address this gap, we propose the Path Divergence Objective (PDO), a novel objective for modelling and building agents with bounded rationality in Partially Observable Markov Decision Processes (POMDPs).

*Equal contribution. Corresponding authors: gavento@acsresearch.org, david.hyland@cs.ox.ac.uk

NeuroAI Workshop at the 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

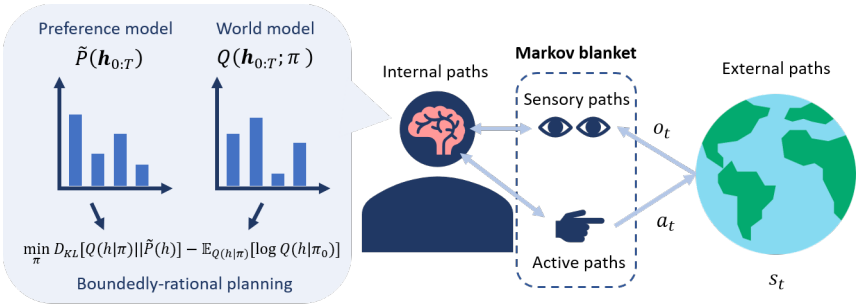

Figure 1: Illustration of the framework. The agent possesses an internal model, which is decomposed into a world model and a preference model, minimising the discrepancy (PDO) between them. The agent's interface to the external world is known as its *Markov blanket* [38, 24, 58], which is comprised of active and sensory paths, which are state trajectories that mediate the interactions between internal and external paths.

The concept of bounded rationality, originally developed to more realistically model (the limits of) human decision-making [64, 65], has broader applications in modelling any teleological physical system. This includes not only humans but also AI systems and other biological entities. The universality of this approach stems from the fact that all physical systems operate within thermodynamic constraints, converting available energy into useful work [19, 21, 24, 40, 71].

Our proposed framework builds upon and generalises an information-theoretic model of bounded rationality [53, 54], focusing on the computational cost of finding a good policy. This model offers a principled approach to modelling decision-making in complex, uncertain environments, naturally capturing trade-offs between information processing and goal attainment. We anticipate its applicability to a wide range of agents with varying internal structures and levels of intelligence, from individual neurons to advanced AI systems, providing a unified framework for understanding decision-making across different scales of complexity.

Bounded Rationality and Rational Inattention: The PDO formalizes and extends Simon's original concept of bounded rationality [64] in three key ways: 1) It introduces partial observability to information-theoretic models of bounded rationality [9, 53, 54], which models agents across a spectrum of rationality levels; 2) It completes the bridge to active inference models [16, 57], demonstrating features like information-seeking behaviour which are a common feature of such models; and 3) It expands the applicability of models like rational inattention [48, 49, 66] to dynamic, sequential decision-making under partial-observability, modelling how agents balance information costs with rewards over time.

Resource and Computational Rationality: Our model can also be viewed through the lens of resource-rationality [6, 45], which is centered on the idea of viewing boundedly-rational agents as making "rational use of limited cognitive resources" for decision-making, and has been found to predict various facets of human decision-making better than existing models [7, 35]. Closely related is the notion of computational rationality [44, 30], who recognise the importance of modelling the tradeoffs that exist between computation and expected utility.

Active Inference and Divergence Objectives: The PDO shares conceptual foundations with active inference, a framework for modelling perception and action based on free-energy minimisation [16, 22, 23, 29, 57]. Indeed, one view of this work is as a derivation of a broader class of active inference or divergence objectives from the starting point of bounded rationality [50], which includes several existing objectives such as the Expected Free Energy [11, 26], the Free Energy of the Expected Future [51], and Action Perception as Divergence minimization [32].

Reinforcement Learning and Control Theory: The field of Reinforcement Learning (RL) [67] has long benefitted from concepts in information theory and statistical mechanics, including entropy/KL-divergence/mutual information regularisation [8, 31, 39, 43, 70, 2].

KL-divergence or mutual information regularisation has seen increasingly widespread adoption in reinforcement learning as a practical approach to updating an agent's policy [1, 2, 13, 59]. Another complementary perspective is to start from an information-theoretic objective, such as path entropy [5] or empowerment [68], and incorporate reward-seeking biases as constraints or regularisers. In both approaches, one arrives at a family of objectives that encode the inherent tradeoffs between goal attainment and information processing/acquisition.

Similarly, the PDO offers a principled framework for incorporating these ideas into partially observable settings, which, to our knowledge, has not been studied as an RL objective. Our approach may provide a theoretical foundation for understanding how RL agents might balance exploration and exploitation in a more sample-efficient manner that more closely mimics human decision-making, potentially leading to more robust and adaptive AI systems.

Main contributions: 1) The derivation and introduction of the Path Divergence Objective, a novel framework for modelling bounded rationality in partially observable environments; 2) An analysis of the PDO through various decompositions to understand the decision-making trade-offs underlying PDO-minimisation; 3) An efficient algorithm to compute the PDO in certain environments; and 4) A comparative analysis of the PDO with expected utility maximisation and the (Sophisticated) Expected Free Energy [16, 23], illustrating novel insights and predictions provided by our approach.

2 Preliminaries

POMDPs, Policies, and World Models: A ***Partially Observable Markov Decision Process*** [3, 4] is a tuple $\mathcal{M} = (S, A, \Omega, O, T, p, I, \mathcal{U})$, where: 1) S is a finite set of *states*; 2) A is a finite set of *actions*; 3) Ω is a finite set of *observations*; 4) $O : A \times S \rightharpoonup \Delta(\Omega)$ is the partial *observation likelihood function*; 5) $T \in \mathbb{Z}^+$ is a finite *time horizon*; 6) $p : S \times A \to \Delta(S)$ is the *probabilistic transition function*; 7) $I \in \Delta(S)$ is the *initial state distribution*; 8) $\mathcal{U} : \mathbb{H} \to \mathbb{R}$ is the *history utility function* which models the agent's preferences, where $\mathbb{H}$ is the set of all histories $\mathbf{h}_{0:t} = s_0 o_0 a_0 s_1 \dots s_t o_t, t \in \{1, \dots, T\}$. We similarly use the notation $s_{0:t}$, $o_{0:t}$, and $a_{0:t}$ to denote state, observation, and action trajectories respectively. A *policy function* $\pi : \mathbb{O} \to \Delta(A)$ maps each observation history of the agent to a probability distribution over their actions. We let Π denote the set of all policies.

In a POMDP, an agent does not have direct access to the true state of the environment. Instead, it receives observations that provide partial information about the state. The agent's goal is to maximise the expected utility of its trajectories. The goal, as defined here, accounts for a wide range of special cases commonly encountered in reinforcement learning, such as the expected sum of discounted rewards [67] and non-Markovian rewards generated by, e.g., a reward machine [36]. In order to compute expected rewards in $\mathcal{M}$, we define the *reach probability* of a history $\mathbf{h}$ under a policy π as $p(\mathbf{h}; \pi)$.[2] Additionally, we assume that agents possess a probabilistic world model $Q(\mathbf{h}_{0:t}; \pi)$, which captures their beliefs about the past and future.

Value functions and Solution Concepts: Perhaps the central concept of interest in control theory and RL is the (objective) *value function*, which measures the expected reward/utility to-go for the agent from a given time point t until the end of the episode, under the policy π. Formally, the value function of the agent is given by $\mathcal{V}(o_{0:t}; \pi) = \mathbb{E}_{p(\mathbf{h}_{0:T} \mid o_{0:t}; \pi)} \left[\mathcal{U}(\mathbf{h}_{0:T})\right]$, where $\mathbf{h}_{0:T} = \mathbf{h}_{0:t}\, a_t s_{t+1} \dots s_T o_T$. By $\mathcal{V}(\pi) = \mathcal{V}(\varnothing; \pi)$, we denote the total expected utility under π.

[2] Please refer to Appendix B.2 for formal mathematical definitions.

3 Path Divergence Objective

Here, we introduce the PDO, outlining its derivation and discussing some of its properties. In this framework, we make three additional assumptions: 1) The agent has a sufficiently accurate world model such that the objective value function $\mathcal{V}(\pi)$ can be replaced by a subjective value function $V(\pi) := \mathbb{E}_{Q(\mathbf{h}_{0:T};\pi)}[\mathcal{U}(\mathbf{h}_{0:T})]$; 2) The agent has a *prior policy* π_0, which represents their *a priori* current best guess at what a good policy might be. The prior policy can be thought of as an agent's default or habitual policy when they do not devote any time to planning. Hence, this 'cognitive effort' can be read as the mental exertion required to overcome one's habitual or instinctual behaviour [56]; 3) Observing that information processing incurs a cost [41], and that agents expend effort when computing a posterior policy to improve the value function, we assume that this expenditure reduces the agent's utility linearly, and that the cost incurred can be measured by the Kullback-Leibler (KL) divergence [54, 56]. Thus, the optimisation problem that the agent appears to be solving is

$$\max_{\pi \in \Pi} \mathbb{E}_{Q(\mathbf{h}_{0:T};\pi)}[\mathcal{U}(\mathbf{h}_{0:T})] - \frac{1}{\beta} \mathrm{D}_{\mathrm{KL}}\left[Q(\mathbf{h}_{0:T};\pi) \,||\, Q(\mathbf{h}_{0:T};\pi_0)\right], \tag{1}$$

for some $\beta > 0$. Now, suppose that we define a probability distribution $\tilde{P}(\mathbf{h}_{0:T}) := \frac{\exp(\beta \mathcal{U}(\mathbf{h}_{0:T}))}{Z(\beta;\mathcal{U})}$, where we let $Z(\beta;\mathcal{U}) := \sum_{\mathbf{h}'_{0:T} \in \mathbb{H}_T} \exp\left(\beta \mathcal{U}(\mathbf{h}'_{0:T})\right)$. We call this distribution the *preference model*, as it is another way of representing the agent's preferences in the form of a probability distribution. Then, re-writing the problem using the preference model, we have the following result[3]:

Lemma 1. *The optimisation problem in* (1) *is equivalent to the following optimisation problem:*

$$\min_{\pi \in \Pi} D_{KL}\left[Q(\mathbf{h}_{0:T};\pi) \,||\, \tilde{P}(\mathbf{h}_{0:T})\right] - \mathbb{E}_{Q(\mathbf{h}_{0:T};\pi)}\left[\log Q(\mathbf{h}_{0:T};\pi_0)\right]. \tag{2}$$

Thus, we see that the planning objective for a boundedly-rational agent can be viewed as finding a policy that minimises the KL divergence between its prediction model and a preference model $\tilde{P}$, with an additional cross entropy term that acts as a penalty for large differences between π and π_0. In other words, one can think of the KL divergence term as the expected excess surprise when the agent wishfully believes that trajectories are distributed according to $\tilde{P}$, when its actual belief is Q.

Definition 2. *The **Path Divergence Objective** (PDO) for an agent i in a POMDP $\mathcal{M}$ given a prior policy π_0 and a posterior policy π is given by:*

$$G(\pi;\pi_0) := D_{KL}\left[Q(\mathbf{h}_{0:T};\pi) \,||\, \tilde{P}(\mathbf{h}_{0:T})\right] - \mathbb{E}_{Q(\mathbf{h}_{0:T};\pi)}\left[\log Q(\mathbf{h}_{0:T};\pi_0)\right]. \tag{3}$$

Decomposition of the PDO. The PDO can be decomposed in several ways, which sheds light on its connections to active inference and intrinsic motivation in reinforcement learning [16, 18, 57, 62]. Firstly, interpreting (negative) expected utility as an energy, the PDO is an upper bound on expected energy minus entropy:

$$\begin{aligned} G(\pi;\pi_0) &= \underbrace{\mathbb{E}_{Q(\mathbf{h}_{0:T};\pi)}\left[-\log \tilde{P}(\mathbf{h}_{0:T})\right]}_{\text{-Value (Energy)}} + \underbrace{\mathrm{D}_{\mathrm{KL}}\left[Q(\mathbf{h}_{0:T};\pi) \,||\, Q(\mathbf{h}_{0:T};\pi_0)\big)\right]}_{\text{Divergence from prior}} \\ &\geqslant \underbrace{\mathbb{E}_{Q(\mathbf{h}_{0:T};\pi)}\left[-\log \tilde{P}(\mathbf{h}_{0:T})\right]}_{\text{-Value (Energy)}} - \underbrace{H\big(\,Q(\mathbf{h}_{0:T};\pi)\,\big)}_{\text{Entropy}}. \end{aligned}$$

One way of thinking of the PDO is therefore as a free energy *relative* to some knowledge in the form of a prior distribution.

Furthermore, decomposing the divergence term in the PDO reveals a natural decomposition in terms of *epistemic value*, *pragmatic value*, and an *intention-behaviour gap*,[4] all of which have been robustly empirically observed in human behaviour [10, 12, 15].

[3] All proofs are deferred to the appendices.

[4] For a more detailed discussion of the decomposition, please refer to Appendix A.4

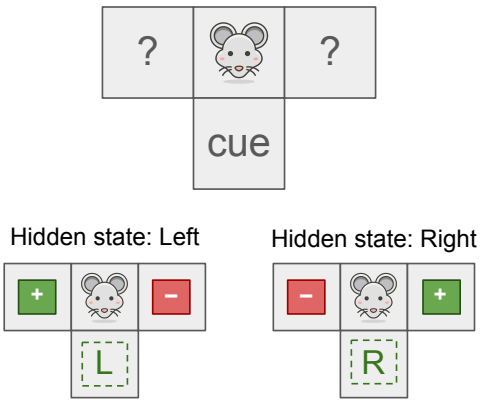

Figure 2: Schematic representation of the T-Maze environment. The maze consists of a start position from which two goal arms (left and right) extend, along with a third *cue* arm (bottom). The maze randomly starts in one of two states: a reward in the left arm and a punishment in the right arm, or vice versa. This state is initially hidden from the agent, but the information about the hidden state is positioned in the cue arm. The agent can visit two locations in a single experiment. We set the reward to +1 and the punishment to -4 in order to disincentivize visiting random arms, and the cost of visiting the cue to C_{cue}.

Theorem 3. *If $\tilde{P}(s_{0:T}|o_{0:T}, a_{0:T}) = Q(s_{0:T}|o_{0:T}, a_{0:T})$, then the divergence term in the PDO can be decomposed as:*

$$D_{KL}\left[Q(\mathbf{h}_{0:T};\pi) \,||\, \tilde{P}(\mathbf{h}_{0:T})\right] = \underbrace{-\mathbb{E}_{Q(o_{0:T},a_{0:T};\pi)}\left[D_{KL}\left[Q(s_{0:T}|o_{0:T},a_{0:T}) \,||\, Q(s_{0:T}|a_{0:T})\right]\right]}_{\textit{Epistemic Value}}$$
$$\underbrace{+\mathbb{E}_{Q(s_{0:T},a_{0:T};\pi)}\left[D_{KL}\left[Q(o_{0:T}|s_{0:T},a_{0:T}) \,||\, \tilde{P}(o_{0:T}|a_{0:T})\right]\right]}_{\textit{Pragmatic Value}} + \underbrace{D_{KL}\left[Q(a_{0:T};\pi) \,||\, \tilde{P}(a_{0:T})\right]}_{\textit{Intention-Behaviour Gap}}.$$

4 Algorithmic and experimental results

Optimal policy search. We propose and implement an efficient algorithm to compute a PDO-minimising policy under the following assumtions: 1) an environment with perfect recall of actions, i.e. every *reachable* observation sequence $o_{0:t}$ uniquely determines the sequence of actions $a_{0:t-1}$ that has led to it. Secondly, a decomposition of $\tilde{P}$ into temporal factors $\tilde{P}_t$ such that we have $\tilde{P}(h_{0:T}) = \prod_{t=0}^{T-1} \tilde{P}_t(a_t, s_{t+1}, o_{t+1}|a_{0:t-1}, o_{0:t}, s_t)$.

The algorithm computes the optimal policy π minimising $G(\pi;\pi_0)$ for any such environment, any given π_0, and any $\tilde{P}_t$ as above, in time $\mathcal{O}(|O_{0:<T}||S|(T_{\tilde{P}_t} + T_Q))$, where S is the set of all states, $O_{0:<T}$ is the set of all prefixes of reachable sequences of observations, and $T_{\tilde{P}_t}$ and T_Q are the times required to evaluate $\tilde{P}_t$ resp Q. See Appendix A.6 for details.

Experimental demonstration of PDO. We study properties of the PDO on a standard T-Maze environment with a cue [17, 25, 52]. This a minimally effective and commonly-used task for studying cognition, information-seeking, and decision-making under uncertainty. See Figure 2 for a description of the environment.

Figure 3 compares the expected reward of an agent's policy π under various models of decision-making: the PDO for various values of β, the expected-value-maximising policy, and two other models of agency and information-seeking under uncertainty: the Expected Free Energy (EFE) [63] and the Sophisticated Expected Free Energy [23]. Note that the primary goal here is not to try to maximise the expected value, but rather to demonstrate some of the qualitative differences between the models in the context of bounded rationality.

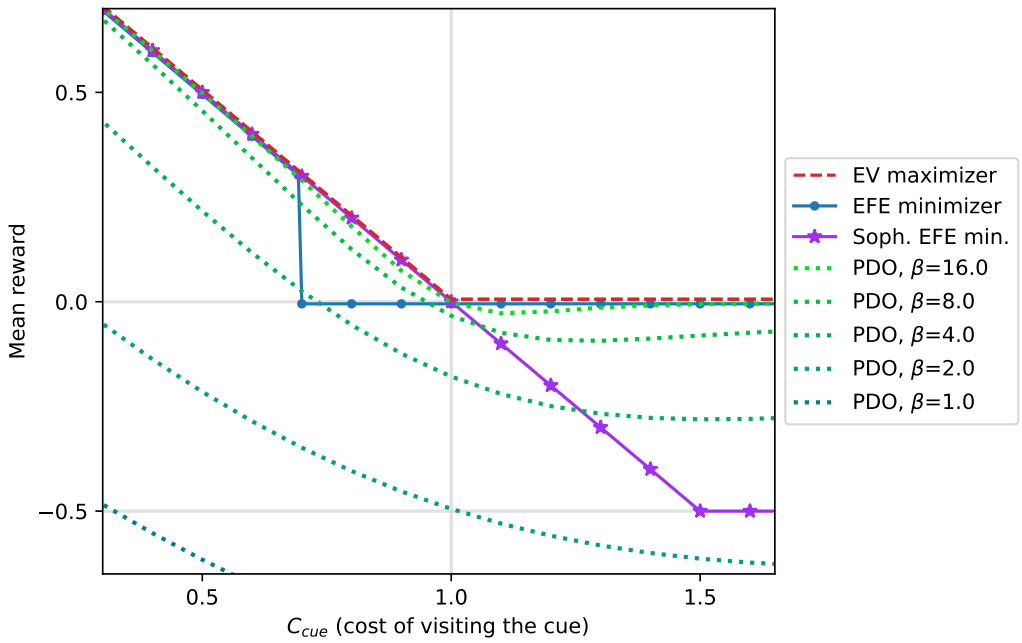

Figure 3: A plot of the mean reward obtained by several decision-making models depending on C_{cue}. The EV maximiser plays the optimal strategy: when the cost of visiting the cue is over 1.0, it is optimal to do nothing. The PDO models various degrees of rationality (β) and smoothly approaches this optimum for $\beta \to \infty$; for $\beta \to 0$, this would correspond to playing π_0 (here a uniform policy). The EFE and Sophisticated EFE exhibit different types of bounded rationality and decide suboptimally in different ranges: EFE stops visiting the cue when the C_{cue} is larger than its information gain term. Sophisticated EFE visits the cue even for $1.0 < C_{cue} < 1.5$, over-estimating the value of the cue information, and then for $C_{cue} > 1.5$ it still learns the value of the cue *indirectly* by visiting a random arm and inferring the cue from there, correcting the action on the second turn to get mean reward -0.5. (Note: the plot lines are slightly ofset to minimize visual overlap.)

5 Conclusion

In this paper, we have introduced the Path Divergence Objective, a novel objective for modelling boundedly-rational model-based planning in partially observable environments. Derived from an information-theoretic model of bounded rationality, the PDO balances reward-seeking behavior with information processing constraints, parameterised by a single "rationality" parameter β. We have then demonstrated how to naturally decompose the PDO into epistemic value, pragmatic value, and intention-behaviour gap, and derived an efficient algorithm for computing PDO-optimal policies in perfect recall environments. Importantly, the PDO converges to expected value maximisation as β approaches infinity, establishing a clear link to classical decision theory [69].

Future research directions include applying the PDO to behavioural modelling, incentive design, AI alignment, and game theory. We aim to develop more scalable algorithms using MCTS-like approaches and function approximators, and empirically compare the PDO's behaviour against existing RL and POMDP algorithms. This flexible, theoretically-grounded framework opens up new possibilities for developing robust AI systems and advancing our understanding of the foundational principles of cognition. We also plan to investigate learning dynamics under the PDO and develop more detailed models incorporating additional cognitive structures, potentially inspiring novel directions in AI research and cognitive modelling. Current limitations include, e.g., not accounting for the cost of learning world model parameters, inferring posteriors, and imperfect plan execution. We hope that further development of the PDO will lead to a versatile toolset for analysing and designing decision-makers to accommodate a wide range of cognitive constraints and real-world scenarios.

Acknowledgments and Disclosure of Funding

We would like to thank Conor Heins, Ran Wei, and Karl Friston for fruitful discussions and technical advice.

Tomáš Gavenčiak and Jan Kulveit are supported by a Charles University grant PRIMUS 22/HUM/020 (Complex Risks in Complex Systems). Michael Wooldridge was supported by a UKRI AI World Leading Researcher Fellowship (grant number EP/W002949/1).

References

[1] ABDOLMALEKI, A., LIOUTIKOV, R., PETERS, J. R., LAU, N., PUALO REIS, L., AND NEUMANN, G. Model-based relative entropy stochastic search. *Advances in Neural Information Processing Systems 28* (2015).

[2] ARUMUGAM, D., HO, M. K., GOODMAN, N. D., AND VAN ROY, B. Bayesian reinforcement learning with limited cognitive load. *Open Mind 8* (04 2024), 395–438.

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

[4] BARTO, A., AND SUTTON, R. *Reinforcement Learning: An Introduction*. 1992.

[5] BERRUETA, T. A., PINOSKY, A., AND MURPHEY, T. D. Maximum diffusion reinforcement learning. *Nature Machine Intelligence* (2024), 1–11.

[6] BHUI, R., LAI, L., AND GERSHMAN, S. J. Resource-rational decision making. *Current Opinion in Behavioral Sciences 41* (2021), 15–21.

[7] BINZ, M., AND SCHULZ, E. Modeling human exploration through resource-rational reinforcement learning. *Advances in neural information processing systems 35* (2022), 31755–31768.

[8] BOTVINICK, M., AND TOUSSAINT, M. Planning as inference. *Trends in cognitive sciences 16*, 10 (2012), 485–488.

[9] BRAUN, D. A., AND ORTEGA, P. A. Information-theoretic bounded rationality and ε-optimality. *Entropy 16*, 8 (2014), 4662–4676.

[10] BROMBERG-MARTIN, E. S., AND MONOSOV, I. E. Neural circuitry of information seeking. *Current Opinion in Behavioral Sciences 35* (2020), 62–70.

[11] CHAMPION, T., BOWMAN, H., MARKOVIĆ, D., AND GRZEŚ, M. Reframing the expected free energy: Four formulations and a unification. *arXiv preprint arXiv:2402.14460* (2024).

[12] CHARPENTIER, C. J., AND COGLIATI DEZZA, I. Information-seeking in the brain. *The Drive for Knowledge: The Science of Human Information Seeking* (2022), 195–216.

[13] CHOW, Y., CUI, B., RYU, M., AND GHAVAMZADEH, M. Variational model-based policy optimization. *arXiv preprint arXiv:2006.05443* (2020).

[14] COHN, D., ATLAS, L., AND LADNER, R. Improving generalization with active learning. *Machine learning 15* (1994), 201–221.

[15] CONNER, M., AND NORMAN, P. Understanding the intention-behavior gap: The role of intention strength. *Frontiers in Psychology 13* (2022), 923464.

[16] DA COSTA, L., PARR, T., SAJID, N., VESELIC, S., NEACSU, V., AND FRISTON, K. Active inference on discrete state-spaces: A synthesis. *Journal of Mathematical Psychology 99* (2020), 102447.

[17] DEACON, R. M., AND RAWLINS, J. N. P. T-maze alternation in the rodent. *Nature protocols 1*, 1 (2006), 7–12.

[18] DECI, E., AND RYAN, R. M. *Intrinsic Motivation and Self-Determination in Human Behavior*. Perspectives in Social Psychology. Springer US, New York, 1985.

[19] DELI, E., PETERS, J., AND KISVÁRDAY, Z. The thermodynamics of cognition: a mathematical treatment. *Computational and Structural Biotechnology Journal 19* (2021), 784–793.

[20] FESTINGER, L. A theory of cognitive dissonance. *Stanford University Press* (1957).

[21] FIELDS, C., GOLDSTEIN, A., AND SANDVED-SMITH, L. Making the thermodynamic cost of active inference explicit. *Entropy 26*, 8 (2024), 622.

[22] FRISTON, K. Life as we know it. *Journal of the Royal Society Interface 10*, 86 (2013), 20130475.

[23] FRISTON, K., DA COSTA, L., HAFNER, D., HESP, C., AND PARR, T. Sophisticated inference. *Neural Computation 33*, 3 (2021), 713–763.

[24] FRISTON, K., DA COSTA, L., SAJID, N., HEINS, C., UELTZHÖFFER, K., PAVLIOTIS, G. A., AND PARR, T. The free energy principle made simpler but not too simple. *Physics Reports 1024* (June 2023), 1–29.

[25] FRISTON, K., FITZGERALD, T., RIGOLI, F., SCHWARTENBECK, P., AND PEZZULO, G. Active Inference: A Process Theory. *Neural Computation 29*, 1 (Jan. 2017), 1–49.

[26] FRISTON, K., FITZGERALD, T., RIGOLI, F., SCHWARTENBECK, P., PEZZULO, G., ET AL. Active inference and learning. *Neuroscience & Biobehavioral Reviews 68* (2016), 862–879.

[27] FRISTON, K., FITZGERALD, T., RIGOLI, F., SCHWARTENBECK, P., PEZZULO, G., ET AL. Active inference and learning. *Neuroscience & Biobehavioral Reviews 68* (2016), 862–879.

[28] FRISTON, K., RIGOLI, F., OGNIBENE, D., MATHYS, C., FITZGERALD, T., AND PEZZULO, G. Active inference and epistemic value. *Cognitive neuroscience 6*, 4 (2015), 187–214.

[29] FRISTON, K. J., DAUNIZEAU, J., KILNER, J., AND KIEBEL, S. J. Action and behavior: A free-energy formulation. 227–260.

[30] GERSHMAN, S. J., HORVITZ, E. J., AND TENENBAUM, J. B. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science 349*, 6245 (2015), 273–278.

[31] HAARNOJA, T., ZHOU, A., ABBEEL, P., AND LEVINE, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning* (2018), PMLR, pp. 1861–1870.

[32] HAFNER, D., ORTEGA, P. A., BA, J., PARR, T., FRISTON, K., AND HEESS, N. Action and perception as divergence minimization. *arXiv preprint arXiv:2009.01791* (2020).

[33] HARMON-JONES, E., AND MILLS, J. An introduction to cognitive dissonance theory and an overview of current perspectives on the theory.

[34] HEINS, C., MILLIDGE, B., DEMEKAS, D., KLEIN, B., FRISTON, K., COUZIN, I. D., AND TSCHANTZ, A. pymdp: A python library for active inference in discrete state spaces. *Journal of Open Source Software 7*, 73 (2022), 4098.

[35] HO, M. K., ABEL, D., COHEN, J. D., LITTMAN, M. L., AND GRIFFITHS, T. L. The efficiency of human cognition reflects planned information processing. In *Proceedings of the 34th AAAI conference on artificial intelligence* (2020).

[36] ICARTE, R. T., KLASSEN, T., VALENZANO, R., AND MCILRAITH, S. Using reward machines for high-level task specification and decomposition in reinforcement learning. In *International Conference on Machine Learning* (2018), PMLR, pp. 2107–2116.

[37] KAARONEN, R. O. A theory of predictive dissonance: Predictive processing presents a new take on cognitive dissonance. *Frontiers in psychology 9* (2018), 2218.

[38] KIRCHHOFF, M., PARR, T., PALACIOS, E., FRISTON, K., AND KIVERSTEIN, J. The Markov blankets of life: Autonomy, active inference and the free energy principle. *Journal of The Royal Society Interface 15*, 138 (Jan. 2018), 20170792.

[39] KORBAK, T., PEREZ, E., AND BUCKLEY, C. L. Rl with kl penalties is better viewed as bayesian inference. *arXiv preprint arXiv:2205.11275* (2022).

[40] KRINGELBACH, M. L., PERL, Y. S., AND DECO, G. The thermodynamics of mind. *Trends in Cognitive Sciences* (2024).

[41] LANDAUER, R. Irreversibility and Heat Generation in the Computing Process. *IBM Journal of Research and Development 5*, 3 (July 1961), 183–191.

[42] LEHMAN, J., AND STANLEY, K. O. Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary computation 19*, 2 (2011), 189–223.

[43] LEVINE, S. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909* (2018).

[44] LEWIS, R. L., HOWES, A., AND SINGH, S. Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in cognitive science 6*, 2 (2014), 279–311.

[45] LIEDER, F., AND GRIFFITHS, T. L. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and brain sciences 43* (2020), e1.

[46] LINDLEY, D. V. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics 27*, 4 (1956), 986–1005.

[47] MACKAY, D. J. Information-based objective functions for active data selection. *Neural computation 4*, 4 (1992), 590–604.

[48] MAĆKOWIAK, B., MATĚJKA, F., AND WIEDERHOLT, M. Rational inattention: A review. *Journal of Economic Literature 61*, 1 (2023), 226–273.

[49] MATĚJKA, F., AND MCKAY, A. Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review 105*, 1 (2015), 272–298.

[50] MILLIDGE, B., SETH, A., AND BUCKLEY, C. Understanding the origin of information-seeking exploration in probabilistic objectives for control. *arXiv preprint arXiv:2103.06859* (2021).

[51] MILLIDGE, B., TSCHANTZ, A., AND BUCKLEY, C. L. Whence the expected free energy? *Neural Computation 33*, 2 (2021), 447–482.

[52] OLTON, D. S. Mazes, maps, and memory. *American psychologist 34*, 7 (1979), 583.

[53] ORTEGA, P. A., AND BRAUN, D. A. Thermodynamics as a theory of decision-making with information-processing costs. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 469*, 2153 (2013), 20120683.

[54] ORTEGA, P. A., BRAUN, D. A., DYER, J., KIM, K.-E., AND TISHBY, N. Information-theoretic bounded rationality. *arXiv preprint arXiv:1512.06789* (2015).

[55] PARR, T., FRISTON, K., AND ZEIDMAN, P. Active data selection and information seeking. *Algorithms 17*, 3 (2024), 118.

[56] PARR, T., HOLMES, E., FRISTON, K. J., AND PEZZULO, G. Cognitive effort and active inference. *Neuropsychologia 184* (2023), 108562.

[57] PARR, T., PEZZULO, G., AND FRISTON, K. J. *Active inference: the free energy principle in mind, brain, and behavior*. MIT Press, 2022.

[58] PEARL, J. *Causality*, 2nd edition ed. Cambridge University Press, Cambridge, U.K. ; New York, Sept. 2009.

[59] PETERS, J., MULLING, K., AND ALTUN, Y. Relative entropy policy search. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2010), vol. 24, pp. 1607–1612.

[60] SAJID, N., DA COSTA, L., PARR, T., AND FRISTON, K. Active inference, bayesian optimal design, and expected utility. *The Drive for Knowledge: The Science of Human Information Seeking* (2022), 124–146.

[61] SANDVED-SMITH, L., AND DA COSTA, L. Metacognitive particles, mental action and the sense of agency. *arXiv preprint arXiv:2405.12941* (2024).

[62] SCHMIDHUBER, J. Formal Theory of Creativity, Fun, and Intrinsic Motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development 2*, 3 (Sept. 2010), 230–247.

[63] SCHWARTENBECK, P., FITZGERALD, T., DOLAN, R. J., AND FRISTON, K. Exploration, novelty, surprise, and free energy minimization. 710.

[64] SIMON, H. A. A behavioral model of rational choice. *The quarterly journal of economics* (1955), 99–118.

[65] SIMON, H. A. Theories of bounded rationality. 161–176.

[66] SIMS, C. A. Implications of rational inattention. *Journal of monetary Economics 50*, 3 (2003), 665–690.

[67] SUTTON, R. S., AND BARTO, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

[68] VOLPI, N. C., AND POLANI, D. Goal-directed empowerment: combining intrinsic motivation and task-oriented behavior. *IEEE Transactions on Cognitive and Developmental Systems 15*, 2 (2020), 361–372.

[69] VON NEUMANN, J., AND MORGENSTERN, O. *Theory of Games and Economic Behavior*. Theory of Games and Economic Behavior. Princeton University Press, Princeton, NJ, US, 1944.

[70] WEI, R., LAMBERT, N., MCDONALD, A. D., GARCIA, A., AND CALANDRA, R. A unified view on solving objective mismatch in model-based reinforcement learning. *Transactions on Machine Learning Research* (2024).

[71] WOLPERT, D., KORBEL, J., LYNN, C., TASNIM, F., GROCHOW, J., KARDEŞ, G., AIMONE, J., BALASUBRAMANIAN, V., DE GIULI, E., DOTY, D., ET AL. Is stochastic thermodynamics the key to understanding the energy costs of computation? *arXiv preprint arXiv:2311.17166* (2023).

A Notation, Proofs and Technical Details

A.1 Notation

Let Y be a finite set. We let $\Delta(Y)$ be the set of probability distributions over Y. For a discrete random variable X with distribution P, we write $H[X]$ for the Shannon entropy of X. For two probability distributions P and Q defined over the same domain X, we write $\mathrm{D}_{\mathrm{KL}}\left[P(x) \mid\mid Q(x)\right]$ to denote the Kullback-Leibler (KL) divergence or relative entropy from Q to P.

A.2 Formal definitions in POMDPs

The reach probability of a history $\mathbf{h}_{0:T}$ under a policy π is defined as

$$p(\mathbf{h};\pi) := I(s_0) \cdot \left(\prod_{\tau=0}^{T-1} O(o_\tau|a_{\tau-1},s_\tau) \cdot \pi(a_\tau|o_{0:\tau}) \cdot p(s_{\tau+1}|s_\tau,a_\tau) \right) \cdot O(o_T|s_T),$$

where $\pi(a_\tau|o_{0:\tau}) := \prod_{i=1}^{n} \pi(a_\tau|o_{0:\tau})$.

The world model is formally defined as

$$Q(\mathbf{h}_{0:t}\,;\,\pi) := Q(s_0) \cdot \left(\prod_{\tau=0}^{t-1} Q(o_\tau|s_\tau) \cdot Q(s_{\tau+1}|s_\tau,a_\tau) \cdot \pi(a_\tau|o_{0:\tau}) \right) \cdot Q(o_t|s_t).$$

A.3 Derivation of the PDO

Lemma 1. *The optimisation problem in 1 is equivalent to the following optimisation problem:*

$$\min_{\pi\in\Pi} D_{KL}\left[Q(\mathbf{h}_{0:T};\pi) \mid\mid \tilde{P}(\mathbf{h}_{0:T}) \right] - \mathbb{E}_{Q(\mathbf{h}_{0:T};\pi)}\left[\log Q(\mathbf{h}_{0:T};\pi_0)\right]. \tag{4}$$

Proof. Recall assumptions 1-3:

$Q(\mathbf{h}_{0:T};\pi) \approx p(\mathbf{h}_{0:T};\pi)$;

The agent has a prior policy π_0;

The agent trades off between utility and information processing additively. Under these, the agent can be seen as optimising the following objective function:

$$\max_{\pi\in\Pi} \mathbb{E}_{Q(\mathbf{h}_{0:T};\pi)}\left[\mathcal{U}(\mathbf{h}_{0:T})\right] - \frac{1}{\beta}\mathrm{D}_{\mathrm{KL}}\left[Q(\mathbf{h}_{0:T};\pi) \mid\mid Q(\mathbf{h}_{0:T};\pi_0)\right]. \tag{5}$$

Now, defining the preference model as

$$\tilde{P}(\mathbf{h}_{0:T}) := \frac{\exp\left(\beta\mathcal{U}(\mathbf{h}_{0:T})\right)}{Z(\beta;\mathcal{U})},$$

we can rearrange this for $\mathcal{U}$, and we obtain

$$\mathcal{U}(\mathbf{h}_{0:T}) = \frac{1}{\beta} \cdot \log\left[\tilde{P}(\mathbf{h}_{0:T}) \cdot Z(\beta;\mathcal{U}) \right]. \tag{6}$$

Using this, we obtain the equivalent problem

$$\begin{aligned}
&\max_{\pi\in\Pi}\frac{1}{\beta}\mathbb{E}_{Q(\mathbf{h}_{0:T};\pi^i)}\left[\log\tilde{P}(\mathbf{h}_{0:T})+\log Z(\beta;\mathcal{U})\right]-\frac{1}{\beta}\mathrm{D}_{\mathrm{KL}}\left[Q(\mathbf{h}_{0:T};\pi)\;||\;Q(\mathbf{h}_{0:T};\pi_0)\right]\\
=&\max_{\pi\in\Pi}\mathbb{E}_{Q(\mathbf{h}_{0:T};\pi^i)}\left[\log\tilde{P}(\mathbf{h}_{0:T})+\log Z(\beta;\mathcal{U})-\log Q(\mathbf{h}_{0:T};\pi)+\log Q(\mathbf{h}_{0:T};\pi_0)\right]\\
=&\min_{\pi\in\Pi}\mathbb{E}_{Q(\mathbf{h}_{0:T};\pi)}\left[\log Q(\mathbf{h}_{0:T};\pi)-\log\tilde{P}(\mathbf{h}_{0:T})-\log Z(\beta;\mathcal{U})-\log Q(\mathbf{h}_{0:T};\pi_0)\right]\\
=&\min_{\pi\in\Pi}\mathrm{D}_{\mathrm{KL}}\left[Q(\mathbf{h}_{0:T};\pi)\;||\;\tilde{P}(\mathbf{h}_{0:T})\right]-\mathbb{E}_{Q(\mathbf{h}_{0:T};\pi)}\left[\log Q(\mathbf{h}_{0:T};\pi_0)\right].
\end{aligned}$$

□

A.4 Decomposition of the PDO

Theorem 3. *If $\tilde{P}(s_{0:T}|o_{0:T},a_{0:T})=Q(s_{0:T}|o_{0:T},a_{0:T})$, then the divergence term in the PDO can be decomposed as:*

$$\begin{aligned}
D_{KL}\left[Q(\mathbf{h}_{0:T};\pi)\;||\;\tilde{P}(\mathbf{h}_{0:T})\right]=&-\underbrace{\mathbb{E}_{Q(o_{0:T},a_{0:T};\pi)}\left[D_{KL}\left[Q(s_{0:T}|o_{0:T},a_{0:T})\;||\;Q(s_{0:T}|a_{0:T})\right]\right]}_{\textit{Epistemic Value}}\\
+\underbrace{\mathbb{E}_{Q(s_{0:T},a_{0:T};\pi)}\left[D_{KL}\left[Q(o_{0:T}|s_{0:T},a_{0:T})\;||\;\tilde{P}(o_{0:T}|a_{0:T})\right]\right]}_{\textit{Pragmatic Value}}&+\underbrace{D_{KL}\left[Q(a_{0:T};\pi)\;||\;\tilde{P}(a_{0:T})\right]}_{\textit{Intention-Behaviour Gap}}.
\end{aligned}$$

Proof. We can write the divergence term $\mathrm{D}_{\mathrm{KL}}\left[Q(\mathbf{h}_{0:T};\pi)\;||\;\tilde{P}(\mathbf{h}_{0:T})\right]$ under the assumption that $\tilde{P}(s_{0:T}|o_{0:T},a_{0:T})=Q(s_{0:T}|o_{0:T},a_{0:T})$ as follows:

$$\begin{aligned}
&\mathbb{E}_{Q(\mathbf{h}_{0:T};\pi)}\left[\log Q(s_{0:T}|a_{0:T})+\log Q(o_{0:T}|s_{0:T},a_{0:T})+\log Q(a_{0:T})\right.\\
&\qquad\left.-\log\tilde{P}(s_{0:T}|o_{0:T},a_{0:T})-\log\tilde{P}(o_{0:T}|a_{0:T})-\log\tilde{P}(a_{0:T})\right]\\
=&\mathbb{E}_{Q(\mathbf{h}_{0:T};\pi)}\left[\log Q(s_{0:T}|a_{0:T})+\log Q(o_{0:T}|s_{0:T},a_{0:T})+\log Q(a_{0:T})\right.\\
&\qquad\left.-\log Q(s_{0:T}|o_{0:T},a_{0:T})-\log\tilde{P}(o_{0:T}|a_{0:T})-\log\tilde{P}(a_{0:T})\right]\\
=&-\underbrace{\mathbb{E}_{Q(o_{0:T},a_{0:T};\pi)}\left[\mathrm{D}_{\mathrm{KL}}\left[Q(s_{0:T}|o_{0:T},a_{0:T})\;||\;Q(s_{0:T}|a_{0:T})\right]\right]}_{\text{Epistemic Value}}\\
&+\underbrace{\mathbb{E}_{Q(s_{0:T},a_{0:T};\pi)}\left[\mathrm{D}_{\mathrm{KL}}\left[Q(o_{0:T}|s_{0:T},a_{0:T})\;||\;\tilde{P}(o_{0:T}|a_{0:T})\right]\right]}_{\text{Pragmatic Value}}\\
&+\underbrace{\mathrm{D}_{\mathrm{KL}}\left[Q(a_{0:T};\pi)\;||\;\tilde{P}(a_{0:T})\right]}_{\text{Intention-Behaviour Gap}}.
\end{aligned}$$

□

The condition in Theorem 3 can be interpreted as the assumption that agents' preferences are only defined over components of their interface with the environment, i.e., their Markov blanket, and not directly over underlying states of the world. This represents what we might call *preference empiricism*, where the stance is taken that an agent's preferences can only be defined over parts of the world which are observable or controllable by them. In the case of metacognitive agents [61], preferences may not be restricted only to one's observations or actions, but could also be defined over one's own internal world model.

Unpacking this decomposition intuitively, we observe the following:

1. The *epistemic value*, also known as the expected information gain [28, 55, 60], scores the expected reduction in uncertainty about the state trajectory before and after knowing the observation trajectory. Notice that since the distributions which are being compared are conditional on the chosen action trajectory, the agent has a bias towards *active data sampling* to advance their understanding about the underlying state of the world [14, 46, 47].

2. The *pragmatic value* similarly scores the expected divergence between the agent's predictions about their own observations and their preferences over the same [27]. We can additionally decompose the pragmatic value term further as $\mathbb{E}_{Q(s_{0:T},a_{0:T};\pi)}\left[-H[Q(o_{0:T}|s_{0:T},a_{0:T})] - \mathbb{E}_{Q(o_{0:T}|s_{0:T},a_{0:T})}\left[\tilde{P}(o_{0:T}|a_{0:T})\right]\right]$. The first term can be interpreted as an entropy-regulariser which motivates the agent to seek out diverse or novel experiences [5, 42], while the second term can be interpreted as the expected utility.

3. The *intention-behaviour gap*, or value-action gap, can be interpreted as capturing the difference between an agent's preferences over their own actions and what their expectations over the same, given the posterior policy [15]. Such a gap is one contributor towards the experience of cognitive dissonance [20, 33] or predictive dissonance [37], which agents will attempt to minimise under this decomposition. The situation of this term amongst the epistemic and pragmatic value components may partially explain why individuals do not always act in a way consistent with their stated preferences, that is, the epistemic or pragmatic benefits of acting in a certain manner may outweigh the intention-behaviour gap induced by such behaviour.

A.5 Recursive formulations of the PDO

Theorem 4. *Under the assumption that the world model Q is fixed for the transition and observation likelihood functions, the Path Divergence Objective can be expressed in the following recursive forms:*

a) With $\tilde{P}$ over the full path:

$$G(\pi;\pi_0) = \mathbb{E}_{Q(s_0,o_0)} G_0^{\mathrm{P}}(\pi|s_0,o_0;\pi_0), \textit{ where} \tag{7}$$

$$G_t^{\mathrm{P}}(\pi|h_{0:t};\pi_0) = D_{KL}\left[\pi(a_t|o_{0:t}) \,||\, \pi_0(a_t|o_{0:t})\right] + \mathbb{E}_{Q(a_t,s_{t+1},o_{t+1}|h_{0:t};\pi)} G_{t+1}^{\mathrm{P}}(\pi|h_{0:t+1};\pi_0) \tag{8}$$

$$G_T^{\mathrm{P}}(\pi|h_{0:T};\pi_0) = -\log \tilde{P}(h_{0:T}). \tag{9}$$

b) With $\tilde{P}$ as conditionals: *For any decomposition of $\tilde{P}$ into a chain of conditional distributions of the form*

$$\tilde{P}(a_t, s_{t+1}, o_{t+1}|h_{0:t}) = \prod_{t=0}^{T-1} \tilde{P}_t(a_t, s_{t+1}, o_{t+1}|h_{0:t})$$

we can express the PDO as

$$G(\pi;\pi_0) = \mathbb{E}_{Q(s_0,o_0)} G_0^{\mathrm{C}}(\pi|s_0,o_0;\pi_0), \textit{ where} \tag{10}$$

$$\begin{aligned} G_t^{\mathrm{C}}(\pi|h_{0:t};\pi_0) &= D_{KL}\left[\pi(a_t|o_{0:t}) \,||\, \pi_0(a_t|o_{0:t})\right] + \\ &\quad + \mathbb{E}_{Q(a_t,s_{t+1},o_{t+1}|h_{0:t};\pi)}\Big[G_{t+1}^{\mathrm{C}}(\pi|h_{0:t+1};\pi_0) - \log \tilde{P}_t(a_t, s_{t+1}, o_{t+1}|h_{0:t})\Big] \end{aligned} \tag{11}$$

$$G_T^{\mathrm{C}}(\pi|h_{0:T};\pi_0) = 0. \tag{12}$$

c) Markovian preferential distribution: *Assuming that $\tilde{P}_t$ from b) only depends on the previous state and the observation history, i.e. $\tilde{P}_t(a_t, s_{t+1}, o_{t+1}|h_{0:t}) = \tilde{P}_t(a_t, s_{t+1}, o_{t+1}|o_{0:t}, s_t)$, we have*

$$G(\pi;\pi_0) = \mathbb{E}_{Q(s_0)Q(o_0|s_0)} G_0^{\mathrm{M}}(\pi|o_0, s_0; \pi_0), \text{ where} \tag{13}$$

$$G_t^{\mathrm{M}}(\pi|o_{0:t}, s_t; \pi_0) = D_{KL}\left[\pi(a_t|o_{0:t}) \;||\; \pi_0(a_t|o_{0:t})\right] + \mathbb{E}_{\pi(a_t|o_{0:t})Q(s_{t+1}|s_t,a_t)Q(o_{t+1}|s_{t+1})}\Big[\\ - \log \tilde{P}_t(a_t, s_{t+1}, o_{t+1}|o_{0:t}, s_t) + G_{t+1}^{\mathrm{M}}(\pi|o_{0:t+1}, s_{t+1}; \pi_0)\Big] \tag{14}$$

$$\tag{15}$$

$$G_T^{\mathrm{M}}(\pi|o_{0:T}, s_T; \pi_0) = 0. \tag{16}$$

Proof. All of the variants are shown by expanding the KL-divergence in the Definition 2, and then introducing a telescopic products over $\tilde{P}$ and Q.

$$G(\pi;\pi_0) = \mathbb{E}_{Q(h_{0:T};\pi)}\Big[\log Q(h_{0:t};\pi) - \log Q(h_{0:t};\pi_0)\Big] - \mathbb{E}_{Q(h_{0:T};\pi)} \log \tilde{P}(h_{0:T}) \tag{17}$$

$$= \mathbb{E}_{Q(h_{0:t};\pi)}\Big[\sum_{t=0}^{T} \log Q(a_t, s_{t+1}, o_{t+1}|h_{0:t};\pi) - \log Q(a_t, s_{t+1}, o_{t+1}|h_{0:t};\pi_0)\Big] \\ - \mathbb{E}_{Q(\mathbf{h}_{0:T};\pi}\Big[\log \tilde{P}(h_{0:T})\Big] \tag{18}$$

$$= \mathbb{E}_{Q(h_{0:t};\pi)}\Big[\sum_{t=0}^{T} \log \pi(a_t|o_{0:t}) - \log \pi_0(a_t|o_{0:t})\Big] - \mathbb{E}_{Q(\mathbf{h}_{0:T};\pi}\Big[\log \tilde{P}(h_{0:T})\Big], \tag{19}$$

where the last equality follows from the fact that

$$Q(a_t, s_{t+1}, o_{t+1}|h_{0:t};\pi) = \pi(a_t|o_{0:t})Q(s_{t+1}|s_t, a_t)Q(o_{t+1}|s_{t+1}),$$

so all non-policy terms cancel out under the assumption that these are fixed.

Now we rearrange the components of the sum into a tree of T levels by matching prefixes of $h_{0:t}$, decomposing the expectation $\mathbb{E}_{Q(h_{0:T};\pi)}$ into a chain of expectations $\prod_{t=0}^{T-1} \mathbb{E}_{Q(a_t,s_{t+1},o_{t+1}|h_{0:t};\pi)}$. This can be then directly rewritten in the recursive form of a) by leaving $\tilde{P}$ intact as in (17), or b) by decomposing $\tilde{P}$ into factors $\tilde{P}_t$. Variant c) is derived analogously to b) using the stated assumptions and subsequently removing irrelevant variables (i.e. a_t and all but the last s_t) from the parameters of G. □

A.6 Algorithm computing the PDO

A *perfect recall environment* is one where the agent observes and remembers not just all its observations but also all its actions, i.e. any reachable sequence $o_{0:t}$ uniquely determines the only sequence of $a_{0:t-1}$ that may have lead to it. Each action sequence may lead to multiple observation sequences (non-determinism), there may be unreachable observation sequences.

Theorem 5. *Assume that conditions of Theorem 4.c hold, that is $\tilde{P}$ can be decomposed into factors $\tilde{P}_t$ such that $\tilde{P}(a_t, s_{t+1}, o_{t+1}|h_{0:t}) = \prod_{t=0}^{T-1} \tilde{P}_t(a_t, s_{t+1}, o_{t+1}|h_{0:t})$ and $\tilde{P}_t$ only depends on the previous state and the observation history, i.e. $\tilde{P}_t(a_t, s_{t+1}, o_{t+1}|h_{0:t}) = \tilde{P}_t(a_t, s_{t+1}, o_{t+1}|o_{0:t}, s_t)$. Then there is an efficient algorithm for finding the $\hat{\pi}$ minimizing $G(\pi;\pi_0)$ for any given perfect recall environment, any such $\tilde{P}_t$, and any given π_0.*

The algorithm runs in time $\mathcal{O}(|\mathbb{O}||S|(T_{\tilde{P}_t} + T_Q))$, where S is the set of all states, $\mathbb{O}$ is the set of all reachable sequences of observations (within the time horizon) and their prefixes, and $T_{\tilde{P}_t}$ and T_Q are the times required to evaluate $\tilde{P}_t$ and Q, respectively.

Note that this algorithm can also work for a "full path" formulation similar to Theorem 4.a if $\tilde{P}$ only depends on the observation sequence and the last state (i.e. $\tilde{P}(h_{0:T}) = \tilde{P}(a_{0:T-1}, o_{0:T}, s_T)$,

as $\tilde{P}_t$ can be assumed to be trivial (e.g. uniform) for all $t < T$, and only have nontrivial $\tilde{P}_T(a_{T-1}, s_T, o_T|o_{0:T-1}) = \tilde{P}(a_{0:T-1}, o_{0:T}, s_T)$ (note that due to the perfect recall assumption, past actions are implied by the past observations).

Proof. First, define $G^{\mathrm{M}'}$, a variant of G^{M} where the conditioning is not on the last state but rather on a distribution (belief) of the last state, S_t.

$$G(\pi; \pi_0) = \mathbb{E}_{Q(o_0)} G_0^{\mathrm{M}'}(\pi|o_0, S_0 = Q(S_0|o_0); \pi_0), \text{ where} \tag{20}$$

$$G_t^{\mathrm{M}'}(\pi|o_{0:t}, S_t; \pi_0) = \mathbb{E}_{\pi(a_t|o_{0:t})}\Big[\log \pi(a_t|o_{0:t}) - \log \pi_0(a_t|o_{0:t}) + \mathbb{E}_{Q(S_{t+1}|S_t,a_t)Q(o_{t+1}|S_{t+1})}\big[- \mathbb{E}_{s_{t+1}\sim S_{t+1}} \log \tilde{P}(a_t, s_{t+1}, o_{t+1}|o_{0:t}) + G_{t+1}^{\mathrm{M}'}(\pi|o_{0:t+1}, S_{t+1}; \pi_0)\big]\Big] \tag{21}$$

$$= \mathbb{E}_{\pi(a_t|o_{0:t})}\Big[\log \pi(a_t|o_{0:t}) + F_t(a_t, o_{0:t}, S_t, \pi_0)\Big] \tag{22}$$

$$G_T^{\mathrm{M}'}(\pi|o_{0:T}, S_T; \pi_0) = 0 \tag{23}$$

Here $F_t(a_t, o_{0:t}, S_t, \pi_0)$ merely collects all the terms of the outer expectation in (21) except the first. Notably, it does not depend on π and can be evaluated for every individual a_t independently.

The algorithm to find $\hat{\pi}$ proceeds as if evaluating $G_0^{\mathrm{M}'}(\pi|o_0, S_0)$ by expanding it recursively, finding the optimal $\hat{\pi}$ along the way and returning it, along with the final value of G. We start with several observations before stating the algorithm.

Observe that in the evaluation tree, $G_t^{\mathrm{M}'}$ is only evaluated once for any given $o_{0:t}$, and $\pi(a_t|o_{0:t})$ only appears in that one evaluation, and moreover $\pi(a_t|o_{0:t})$ can be chosen independently from π for all other observations. Further observe that $G^{\mathrm{M}'}$ can in fact be minimised by minimizing each $G^{\mathrm{M}'}(o_{0:t}, S_t)$ independently, since $G^{\mathrm{M}'}$ only appears as a positive term in other $G^{\mathrm{M}'}(\dots)$'s, and the value of S_{t+1} passed down the recursion does *not* depend on π but rather is conditioned on a single action a_t.

Therefore, $\pi(a_t|o_{0:t})$ can be optimised locally after first evaluating all $G_{t+1}^{\mathrm{M}'}(\pi|o_{0:t+1}, S_{t+1}; \pi_0)$. The $\pi(a_t|o_{0:t})$ minimizing $\mathbb{E}_{\pi(a_t|o_{0:t})}\Big[\log \pi(a_t|o_{0:t}) + F_t(a_t, o_{0:t}, S_t, \pi_0)\Big]$ is the Boltzmann distribution where F plays the role of the expected energy of the action:

$$\hat{\pi}(a_t|o_{0:t}) = \frac{e^{-F_t(a_t, o_{0:t}, S_t, \pi_0)}}{Z_t(o_{0:t}, S_t, \pi_0)}, \tag{24}$$

where $Z_t(o_{0:t}, S_t, \pi_0)$ is a distribution-normalization constant.

The algorithm is then as follows: Traverse the tree of evaluating $G^{\mathrm{M}'}$ recursively. While evaluating the tree node $G_t^{\mathrm{M}'}(\pi|o_{0:t}, S_t; \pi_0)$, first evaluate $G_{t+1}^{\mathrm{M}'}(\pi|o_{0:t+1}, S_{t+1}; \pi_0)$ for all a_t and o_{t+1} recursively, combining the returned partial policies $\hat{\pi}$. Then set $\hat{\pi}(a_t|o_{0:t})$ according to equation (24), and return the updated policy along with the (directly computed) value of $G_t^{\mathrm{M}'}(\pi|o_{0:t}, S_t; \pi_0)$.

The runtime follows from visiting each $o_{0:t} \in \mathbb{O}$ only once, and each evaluation does $\mathcal{O}(|S|(T_{\tilde{P}_t} + T_Q))$ work. The algorithm is efficient since every algorithm without further assumptions on $\tilde{P}_t$ and Q needs to evaluate them on all observation sequences, otherwise we can engineer $\tilde{P}_t$ and Q that would encode an exceedingly high reward in the omitted branch. □

B Supplementary Materials for Experiments

Here we include the omitted materials regarding our experimental results and design.

The experiments were carried out with the PyMDP library [34], adding our own implementation of the PDO-minimizing and an expectation-maximizing algorithms into the framework. Our code can be found at `https://github.com/gavento/pymdp/tree/neurips2024-v1` with the implementation of PDO and expected value minimization under `pymdp/pdo_agents/` and the code to execute our experiments and generate the plots under `notebooks/`. We hope our algorithm will be eventually merged into the PyMDP library.

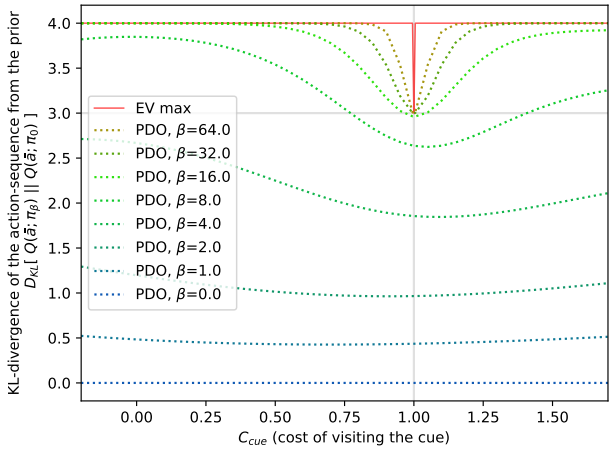

Figure 4: The divergence of action distribution under Q when playing according to π vs according to π_0. With an observation-agnostic prior policy π_0, this can be seen as the expected divergence $\sum_{t=0}^{T} \mathbb{E}_{Q(o_{0:t};\pi)} \mathrm{D}_{\mathrm{KL}}\left[\pi(o_{0:t}) \,||\, \pi_0(o_{0:t})\right]$ of the policies in $\pi(o_{0:t})$ from the prior policy $\pi_0(o_{0:t})$, where the expectation is over observations seen by an agent acting according to π. Note that $\beta = 0$ implies playing π_0 (here a uniform policy), perfect control requires 4 bits (2 for each round) and higher values of β mostly require the same level of control regardless of C_{cue} with the exception of a region around $C_{cue} = 1.0$ where there are multiple almost-optimal courses of action.