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

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce the Path Divergence Objective (PDO), a novel model of boundedly-rational decision-making in stochastic, partially-observable environments. The PDO is derived from fundamental physical principles, including embodiment and the inherent costs of information processing. This framework enables us to model key features observed in real-world agent behavior, such as curiosity-driven exploration, novelty-seeking, and the intention-behavior 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 modeling 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 modeling 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 the machine learning methods and capabilities have advanced significantly, progress in modeling 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 approach to modeling bounded rationality in Partially Observable Markov Decision Processes (POMDPs).

The concept of bounded rationality, originally developed to model human decision-making (50; 51), has broader applications in modeling 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 (13; 15; 18; 30; 56).

Our propsoed framework builds upon and generalizes an information-theoretic model of bounded rationality (41; 42), focusing on the computational cost of finding a good policy. Our framework offers a principled approach to modeling decision-making in complex, uncertain environments, naturally capturing trade-offs between exploitation and exploration. We anticipate its applicability to a wide range of agents with varying internal structures and intelligence levels, from individual neurons to advanced AI systems, providing a unified framework for understanding decision-making across different scales of complexity.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

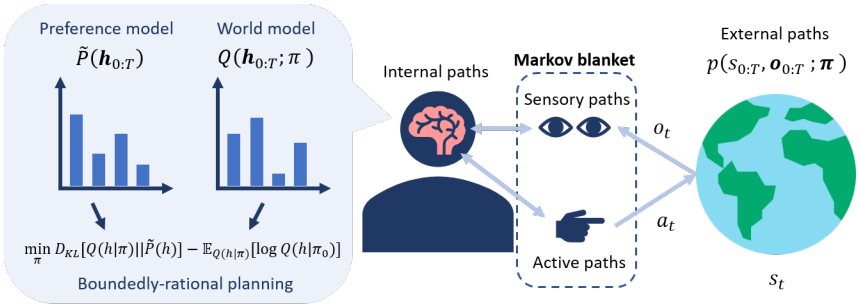

Figure 1: Illustration of the framework. The agent possesses an internal model, which is decomposed into a world model and a preference model, minimising the discrepancy (PDO) between them.

Bounded Rationality and Rational Inattention: The PDO formalizes and extends Simon's original concept of bounded rationality (50) in three key ways: 1) It introduces partial observability to information-theoretic models of bounded rationality (5; 41; 42), maintaining a spectrum of rationality levels; 2) It completes the bridge to active inference models (11; 45), demonstrating features like information-seeking behaviour which are a common feature of such models; and 3) It generalizes rational inattention (36; 37; 52) to dynamic, sequential decision-making, modelling how agents balance information costs with rewards over time.

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 (11; 16; 17; 22; 45). Indeed, one view of this work is a derivation of a broader class of active inference or divergence objectives from the starting point of bounded rationality (38), which includes several existing objectives such as the Expected Free Energy (7; 20), the Free Energy of the Expected Future (39), and Action Perception as Divergence minimisation (24).

Reinforcement Learning and Control Theory: While traditional Reinforcement Learning (RL) focuses on maximising expected rewards (53), recent work has explored information-theoretic objectives in RL and control theory (4; 29; 33; 55). 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 way 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 PDO in certain environments; and 4) A comparative analysis of the PDO with expected utility maximisation and Expected Free Energy, illustrating novel insights and predictions provided by our approach.

2 Preliminaries

POMDPs, Policies, and World Models: A ***Partially Observable Markov Decision Process*** (1; 2) 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 (53) and non-Markovian rewards generated by, e.g., a reward machine (27). In order to compute expected rewards in $\mathcal{M}$, we define the *reach probability* of a history $\mathbf{h}$ under a policy π as $p(\mathbf{h}; \pi)$.[1] 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.

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 $V(\pi) := \mathbb{E}_{Q(\mathbf{h}_{0:T};\pi)} \left[\mathcal{U}(\mathbf{h}_{0:T})\right]$; 2) The agent has a *prior policy* π_0, which represents their *a priori* 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 (44); 3) Observing that information processing incurs a cost (31), 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 (42; 44). Thus, the optimisation problem that the agent appears to be solving is

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

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[2]:

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.

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

[2] All proofs are deferred to the appendices.

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:*

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

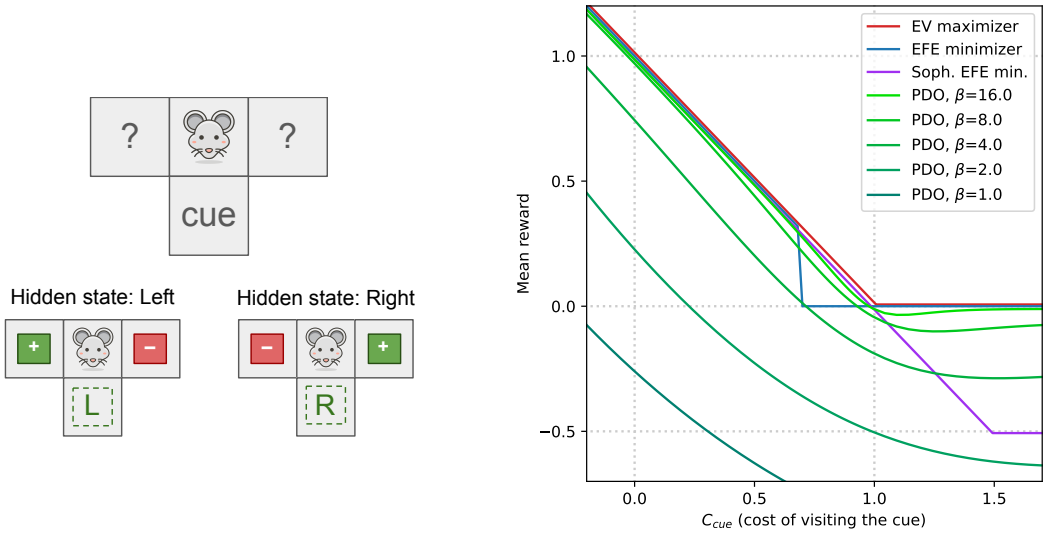

Figure 2: **Left:** Schematic representation of the T-Maze environment. The maze consists of a start position from which two goal arms (left and right) extend, along with a third *cue* arm (bottom). The maze randomly starts in one of two states: a reward in the left arm and a punishment in the right 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, and the cost of visiting the cue to C_{cue}. **Right:** 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 both play suboptimally in different ranges: EFE stops visiting the cue when the C_{cue} is larger than its information gain, and Sophisticated EFE over-values the information of the cue (for $1.0 < C_{cue} < 1.5$) and then over-values the information gained by visiting a random arm and inferring the cue from there, correcting on the second action. Note that neither variant of EFE exhibits bounded rationality outside of those ranges of C_{cue}. (Curves are ofsetted to avoid overlaps.)

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 (11; 12; 45; 48). Firstly, interpreting (negative) expected utility as an energy, the PDO is an upper bound on expected energy minus entropy:

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

Furthermore, decomposing the divergence term in the PDO reveals a natural decomposition in terms of *epistemic value*, *pragmatic value*, and an *intention-behaviour gap*,[3] all of which have been robustly empirically observed in human behaviour (6; 8; 10).

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

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:*

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

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 (19; 40). This a simple and commonly-used environment for studying cognition, information-seeking, and decision-making under uncertainty. See Figure 2 (left) for a description of the environment.

Figure 2 (right) compares the expected reward of π under various models of decision-making: thse 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) (49) and the Sophisticated Expected Free Energy (17). Note that the primary goal here is not to try to maximise the expected value, but rather to study the qualitative differences between the models.

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 (54).

Future research directions include connecting the PDO with applications in 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 biological 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.

References

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

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

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

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

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

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

[7] 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).

[8] 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.

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

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

[11] 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.

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

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

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

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

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

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

[18] 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.

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

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

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

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

[23] FRISTON, K. J., AND FRITH, C. D. Active inference, communication and hermeneutics. *cortex 68* (2015), 129–143.

[24] 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).

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

[26] 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.

[27] 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.

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

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

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

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

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

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

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

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

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

[37] 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.

[38] 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).

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

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

[41] 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.

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

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

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

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

[46] 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.

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

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

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

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

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

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

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

[54] 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.

[55] 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).

[56] 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

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

□

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:*

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

Proof. We can write the divergence term $\mathrm{D}_{\mathrm{KL}}\left[Q(\mathbf{h}_{0:T};\pi) \mid\mid \tilde{P}(\mathbf{h}_{0:T})\right]$ 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:

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

□

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 (47), 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 (23; 43; 46), 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 (9; 34; 35).

2. The *pragmatic value* similarly scores the expected divergence between the agent's predictions about their own observations and their preferences over the same (21). We can additionally decompose the pragmatic value term further as $\mathbb{E}_{Q(s_{0:T},a_{0:T};\pi)}\Big[-H[Q(o_{0:T}|s_{0:T},a_{0:T})] - \mathbb{E}_{Q(o_{0:T}|s_{0:T},a_{0:T})}\Big[\tilde{P}(o_{0:T}|a_{0:T})\Big]\Big]$. The first term can be interpreted as an entropy-regulariser which motivates the agent to seek out diverse or novel experiences (3; 32), 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 (10). Such a gap is one contributor towards the experience of cognitive dissonance (14; 25) or predictive dissonance (28), 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. *The Path Divergence Objective can be expressed in the following recursive forms:*

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

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

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

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

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 PDO as

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

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

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

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*

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

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

$$\tag{15}$$

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

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

$$G(\pi;\pi_0) = \mathbb{E}_{Q(h_{0:T};\pi)} \sum_{t=0}^{T} \Big[\log \frac{Q(h_{0:t};\pi)}{Q(h_{0:t-1};\pi)} - \log \frac{Q(h_{0:t};\pi_0)}{Q(h_{0:t-1};\pi_0)} \Big] - \mathbb{E}_{Q(h_{0:T};\pi)} \log \tilde{P}(h_{0:T}) \tag{17}$$

$$= \mathbb{E}_{Q(h_{0:T};\pi)} \sum_{t=0}^{T} \Big[\log \frac{Q(h_{0:t};\pi)}{Q(h_{0:t-1};\pi)} - \log \frac{\tilde{P}(h_{0:t})}{\tilde{P}(h_{0:t-1})} - \log \frac{Q(h_{0:t};\pi_0)}{Q(h_{0:t-1};\pi_0)} \Big] \tag{18}$$

$$= \sum_{t=0}^{T} \mathbb{E}_{Q(h_{0:t};\pi)} \Big[\log \frac{Q(h_{0:t};\pi)}{Q(h_{0:t-1};\pi)} - \log \frac{\tilde{P}(h_{0:t})}{\tilde{P}(h_{0:t-1})} - \log \frac{Q(h_{0:t};\pi_0)}{Q(h_{0:t-1};\pi_0)} \Big] \tag{19}$$

$$= \sum_{t=0}^{T} \mathbb{E}_{Q(h_{0:t};\pi)} \Big[\log \frac{Q(h_{0:t};\pi)}{Q(h_{0:t-1};\pi)} - \log \frac{Q(h_{0:t};\pi_0)}{Q(h_{0:t-1};\pi_0)} \Big] - \mathbb{E}_{Q(h_{0:T};\pi)} \tilde{P}(h_{0:T}) \tag{20}$$

$$\tag{21}$$

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 $\tilde{P}_t$. Recall that

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

Variant c) is derived analoguously 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 thereis an efficient algorithm for finding the $\hat{\pi}$ minimizing $G(\pi;\pi_0)$ for any given perfect recall envirionment, any such $\tilde{P}_t$, and any given π_0.*

The algorithm runs 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 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 $\tilde{P}_t$ resp Q.

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.

$$G(\pi;\pi_0) = \mathbb{E}_{Q(o_0)} G_0^{\mathrm{M}'}(\pi|o_0, S_0 = Q(S_0|o_0);\pi_0), \text{ where} \tag{22}$$

$$G_t^{\mathrm{M}'}(\pi|o_{0:t}, S_t;\pi_0) = \mathbb{E}_{\pi(a_t|o_{0:t})}\Big[\log\pi(a_t|o_{0:t}) - \log\pi_0(a_t|o_{0:t}) + \mathbb{E}_{Q(S_{t+1}|S_t,a_t)Q(o_{t+1}|S_{t+1})}\big[$$
$$- \mathbb{E}_{s_{t+1}\sim S_{t+1}}\log\tilde{P}(a_t, s_{t+1}, o_{t+1}|o_{0:t}) + G_{t+1}^{\mathrm{M}'}(\pi|o_{0:t+1}, S_{t+1};\pi_0)\big]\Big] \tag{23}$$

$$= \mathbb{E}_{\pi(a_t|o_{0:t})}\Big[\log\pi(a_t|o_{0:t}) + F_t(a_t, o_{0:t}, S_t, \pi_0)\Big] \tag{24}$$

$$G_T^{\mathrm{M}'}(\pi|o_{0:T}, S_T;\pi_0) = 0 \tag{25}$$

Here $F_t(a_t, o_{0:t}, S_t, \pi_0)$ merely collects all the terms of the outer expectation in (23) except the first. Notably, it does not depend on π and can be evaluated for every individual a_t independently.

The algotihm 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 rahter 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:

$$\hat{\pi}(a_t|o_{0:t}) = \frac{e^{-F_t(a_t, o_{0:t}, S_t, \pi_0)}}{Z_t(o_{0:t}, S_t, \pi_0)}, \tag{26}$$

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 (26), 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 $o_{0:t} \in O_{0:<T}$ 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 (26), adding our own implementation of the PDO and an expectation-maximizing algorithm into the framework. We will publish our work after the publication of this work.

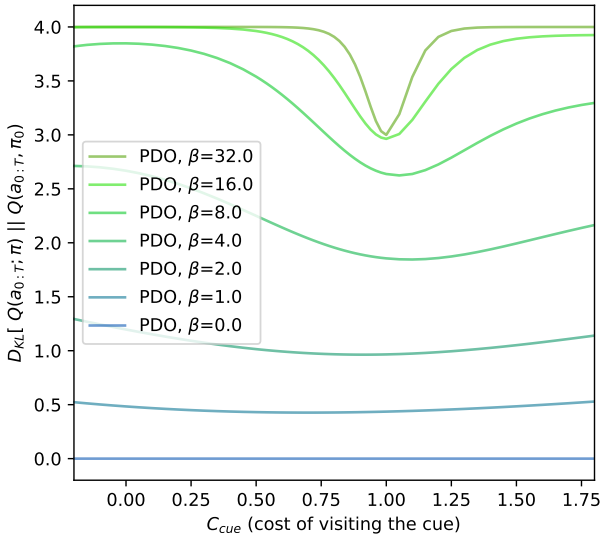

Figure 3: The divergence of action distribution under Q when playing according to π vs according to π_0. With an observation-agnostic prior policy π_0, this can be seen as the expected divergence $\sum_{t=0}^{T} \mathbb{E}_{Q(o_{0:t;\pi})} \mathrm{D}_{\mathrm{KL}} \left[\pi(o_{0:t}) \ ||\ \pi_0(o_{t:0})\right]$ 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.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The paper's content is summarized by the introduction and abstract.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The Conclusion and other parts of the paper list some of the limitation, and frame the work as ongoing work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The proofs, algorithm outlines and technical details are provided in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The details of the experiment are described in detail, as well as the baseline implementation used (PyMDP). Our code is not attached but will be published after publication.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Our code will be published (probably as a part of PyMDP) after publication and de-anonymization. We are happy to provide it at any point upon request.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper includes experiment details, and does not rely on any dataset.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper does not make claims of statistical significance of the results, the experiments are illustrative.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Information not provided, no special hardware needed (i.e. no GPU or large amount of compute).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

Justification: The work conforms with the code of ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: The paper proposes a theoretical model. Future applications will need to include societal consideratons.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: NA

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: In references.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: No new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No human subjects involved.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: No human subjects involved.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.