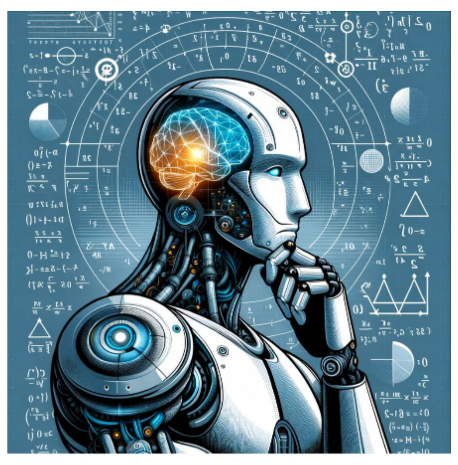


Path Divergence Objective

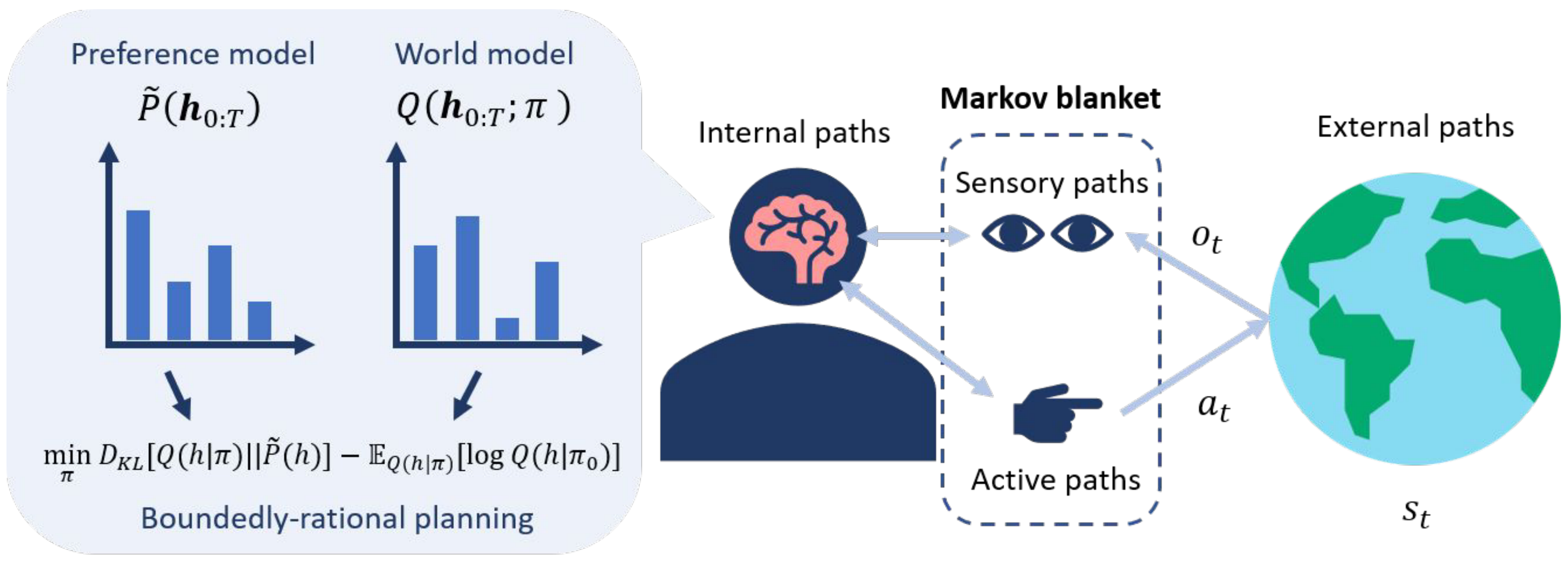
Boundedly Rational Decision-Making in Partially-Observable Environments

Tomáš Gavenčiak* David Hyland* Lancelot Da Costa Michael J. Wooldridge Jan Kulveit

Motivation and Summary

How can we improve our models of decision-making of realistic agents in real-world scenarios?

1. **Partially-observable, stochastic environments**
2. **Cost of information-processing**
3. **Applicable to both biological and artificial agents**



We propose a novel class of objectives to model 1—3 using **information-theoretic bounded rationality**, and show how this leads to phenomena like **curiosity** and **cognitive dissonance minimising behaviour**.

Environment and Agent Models

Partially-Observable MDPs (POMDPs) with:

Utility $\mathcal{U} : \mathbb{H} \to \mathbb{R}$ ***with*** $\tilde{P}(\mathbf{h}_{0:T}) := \dfrac{\exp(\beta\mathcal{U}(\mathbf{h}_{0:T}))}{\sum_{\mathbf{h}'_{0:T}} \exp(\beta\mathcal{U}(\mathbf{h}'_{0:T}))}$

Policies $\pi : (o_1, o_2, ..., o_t) \to \Delta(A), t \in \{0, ..., T\}$

Path Divergence Objective (PDO)

- We assume and generalise **Information-Theoretic Bounded Rationality** [Ortega et al., 2015] model, which is related to **Rational Inattention** [Sims 2003] and **Capacity-limited Bayesian RL** [Arumugam et al., 2024]
- ***Agent pays a cost for policy (and belief) updates*** from a prior belief about trajectories to a posterior belief, measured using the KL divergence:

$$\max_{\pi^*} \underbrace{\mathbb{E}_{Q(\mathbf{h}_{0:T};\pi^*)}[\mathcal{U}(\mathbf{h}_{0:T})]}_{\text{Expected utility}} - \underbrace{\frac{1}{\beta}\mathsf{D}_{\mathsf{KL}}[Q(\mathbf{h}_{0:T};\pi^*) \,||\, Q(\mathbf{h}_{0:T};\pi_0)]}_{\text{Cost of information processing}}$$

$$|||$$

$$\min_{\pi^*} G(\pi;\pi_0) := \underbrace{\underbrace{\mathsf{D}_{\mathsf{KL}}\left[Q(\mathbf{h}_{0:T};\pi^*) \,||\, \tilde{P}(\mathbf{h}_{0:T})\right]}_{\text{Divergence}} - \underbrace{\mathbb{E}_{Q(\mathbf{h}_{0:T};\pi^*)}[\log Q(\mathbf{h}_{0:T};\pi_0)]}_{\text{Cross-Entropy}}}_{\text{Path Divergence Objective}}$$

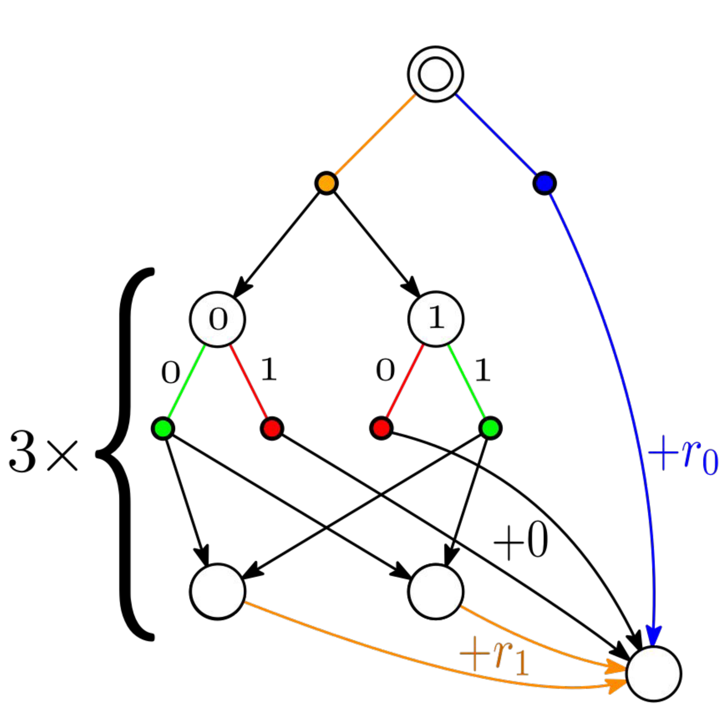
- **Intuition:** β is the **rationality level** or **information processing efficiency**. The divergence term has 3 parts:

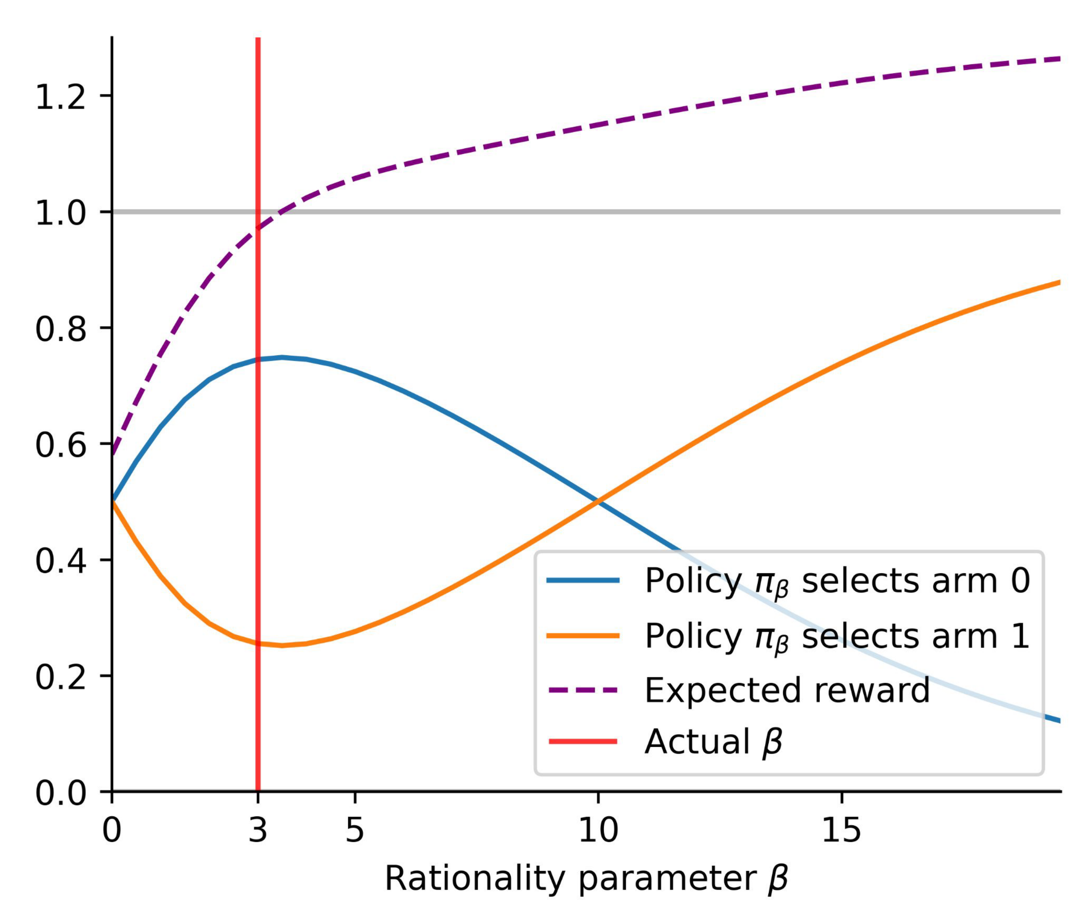
$$\begin{aligned}\mathsf{D}_{\mathsf{KL}}\left[Q(\mathbf{h}_{0:T};\pi) \,||\, \tilde{P}(\mathbf{h}_{0:T})\right] = &- \underbrace{\mathbb{E}_{Q(o_{0:T},a_{0:T};\pi)}[\mathsf{D}_{\mathsf{KL}}[Q(s_{0:T}|o_{0:T},a_{0:T}) \,||\, Q(s_{0:T}|a_{0:T})]]}_{\text{Epistemic Value}} \\ &+ \underbrace{\mathbb{E}_{Q(s_{0:T},a_{0:T};\pi)}\left[\mathsf{D}_{\mathsf{KL}}\left[Q(o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

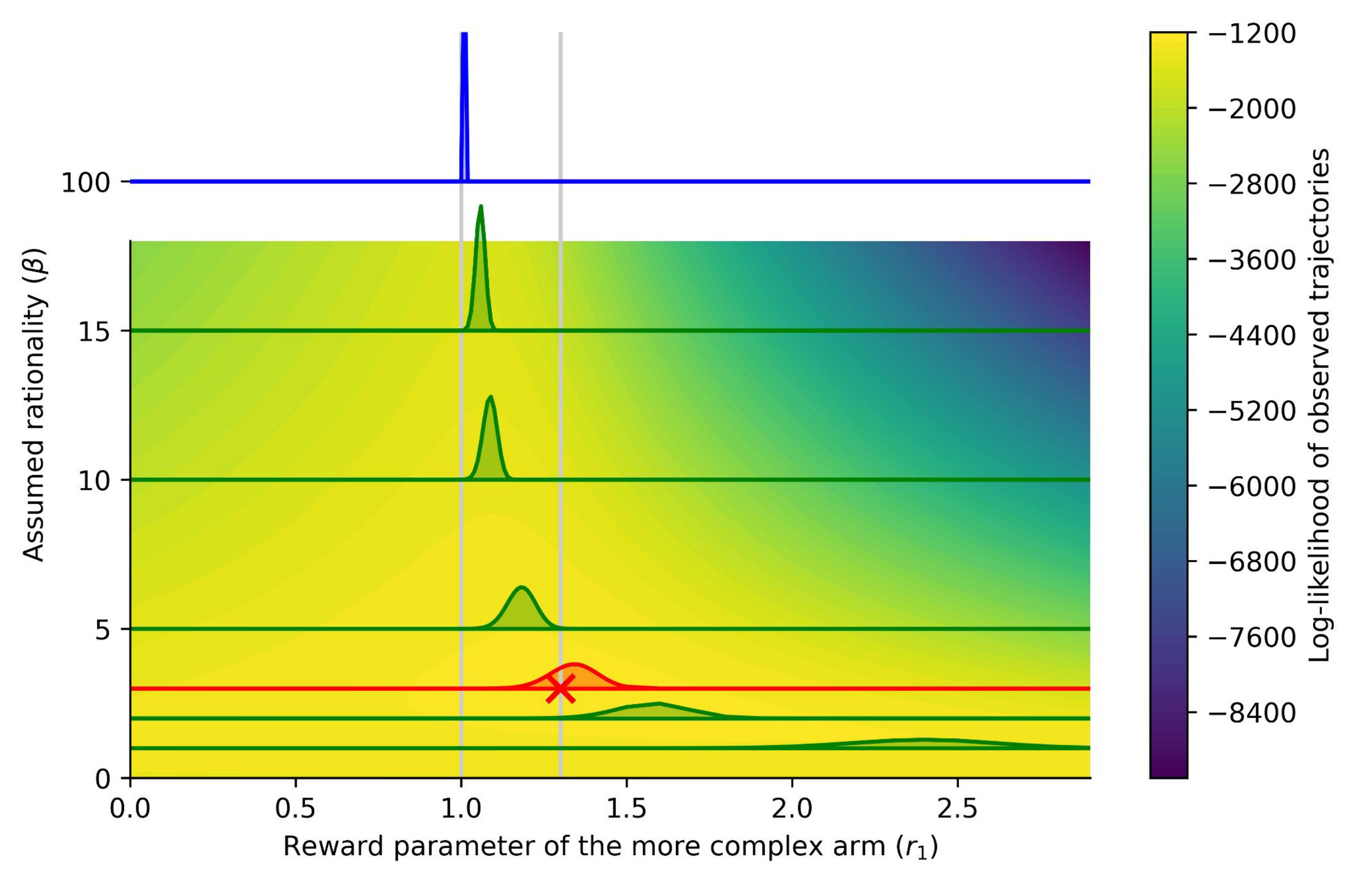
Demonstration: Value inference under incorrect rationality assumptions

Inference of preferences generally fails without an adequate model of agent's rationality.

Skill-based bandit: A two-armed bandit where one of the arms requires the agent to correctly input 3 bits based on a 3-bit observation in order to get reward r_1. The observer knows that $r_0=1$ (direct reward), and r_1 is inferred from observing 300 trajectories.







Key Applications

AI alignment for PDO-minimising agents to include varying rationality levels, biased world models, and information-seeking behavior, e.g., modeling human-AI interactions; "bounded assistance games" *(CIRL)*

Game theory with PDO-minimising agents - **free-energy equilibria** as a normative *and* descriptive solution concept.

Mechanism design for boundedly-rational agents to develop incentive structures accounting for limited rationality.

Questions? Interested in collaboration?
gavento@acsresearch.org,
david.hyland@cs.ox.ac.uk

ACS
ACSresearch.org

Charles University

UNIVERSITY OF OXFORD