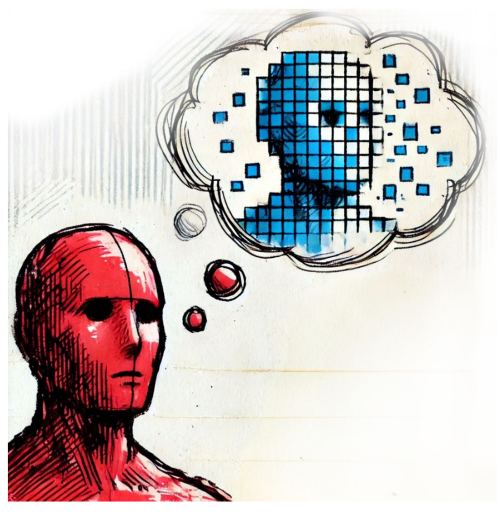
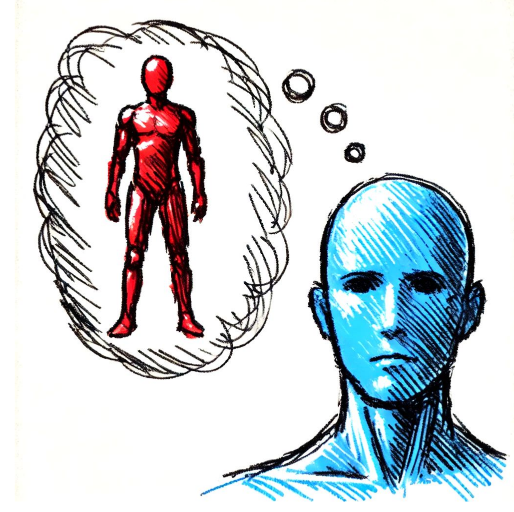


Free-Energy Equilibria

Toward a Game-Theoretic Foundation of Multi-Agent Active Inference



David Hyland* Tomáš Gavenčiak* Lancelot Da Costa Conor Heins
Vojtěch Kovařík Julian Gutierrez Michael J. Wooldridge Jan Kulveit

Motivation and goals

***Understand* and *shape* realistic strategic agent interactions in complex systems**

- Develop a **unified framework** for modelling **boundedly-rational agents** in **stochastic, partially observable** environments
  - Real-world agents have limited information and cognitive capacity
  - Traditional game theory often assumes perfect rationality
- Develop tools for modelling **human-AI interactions** and **AI alignment problems**
  - Account for agents with *different levels of rationality* and *biased beliefs*
- Bridge *game theory*, *bounded rationality*, *information theory*, and *active inference*

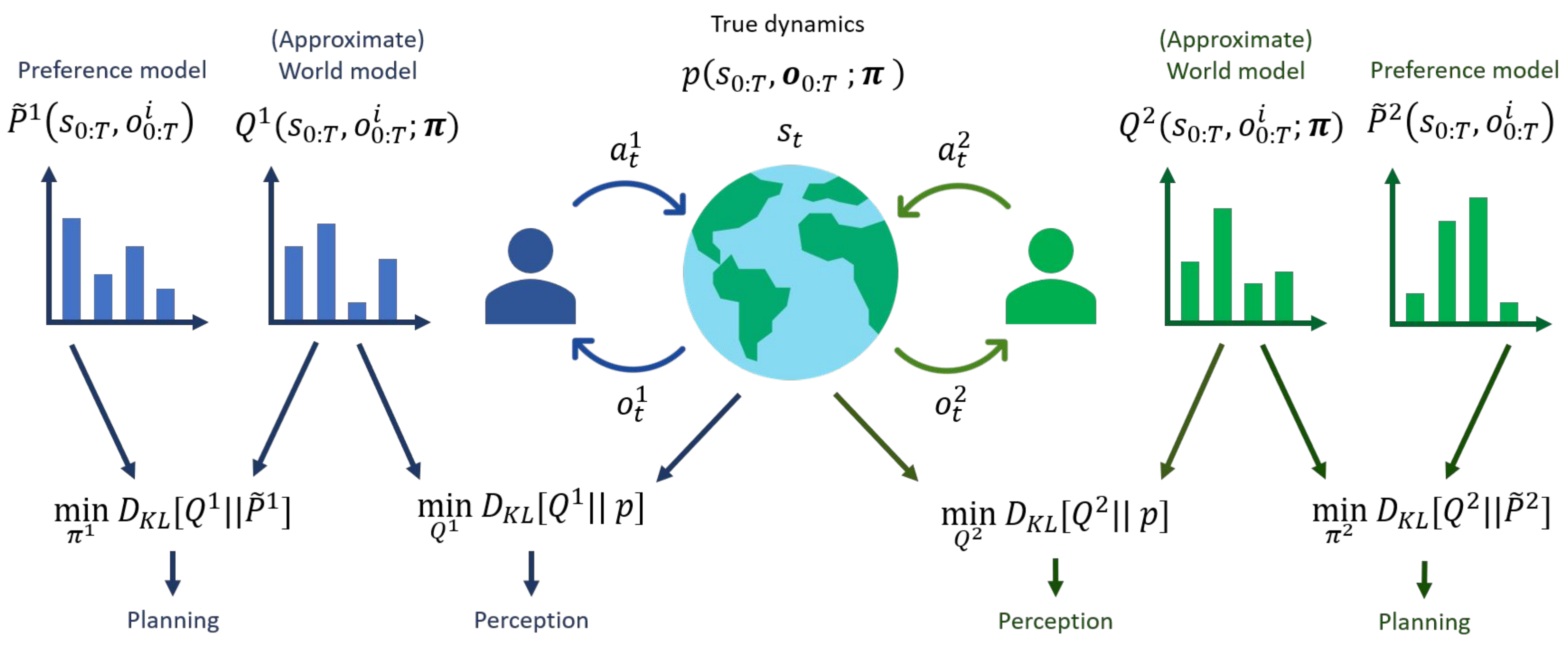
The setting

Partially-observable stochastic games (POSGs)

- POSGs generalize POMDPs to multi-agent settings
- A *strategy:* $\pi^i : (o^i_{0,\dots,}\, o^i_t) \to \Delta\left(A^i\right),\ t \in \{0, \dots, T\}$
- A *strategy profile:* $\boldsymbol{\pi} = \left(\pi^1, \dots, \pi^N\right)$

Bounded rationality model

- We assume and generalize the *Information-Theoretic Bounded Rationality* [Ortega, Braun 2015] model, equivalent to *Rational Inattention* [Sims 2003]
- Each agent has a cost of policy (and belief) updates from a prior policy π^i_0, and minimizes $G^i(\pi) := \underbrace{\mathbb{E}_{Q^i(\mathbf{h}_{0:T};\pi)}\left[U^i(\mathbf{h}_{0:T})\right]}_{\text{expected utility}} - \underbrace{\frac{1}{\beta}\mathrm{D}_{\mathrm{KL}}\left[Q^i(\mathbf{h}_{0:T};\pi)\;||\;Q^i(\mathbf{h}_{0:T};\pi^0)\right]}_{\text{cost of information processing}}$
- *Intuition:* β represents the *level of rationality* or *efficiency of information processing*: $\beta \cong 0$ prevents any update from a prior policy, $\beta \cong \infty$ means a perfectly rational agent

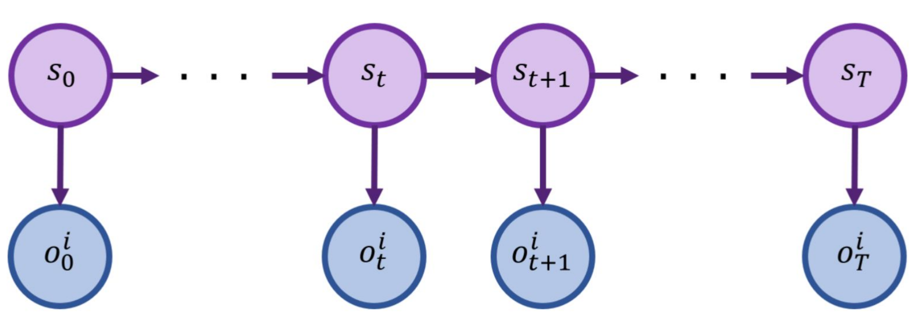


Modelling Agents

Agents minimize divergence between *predictive* and *preferential* distributions

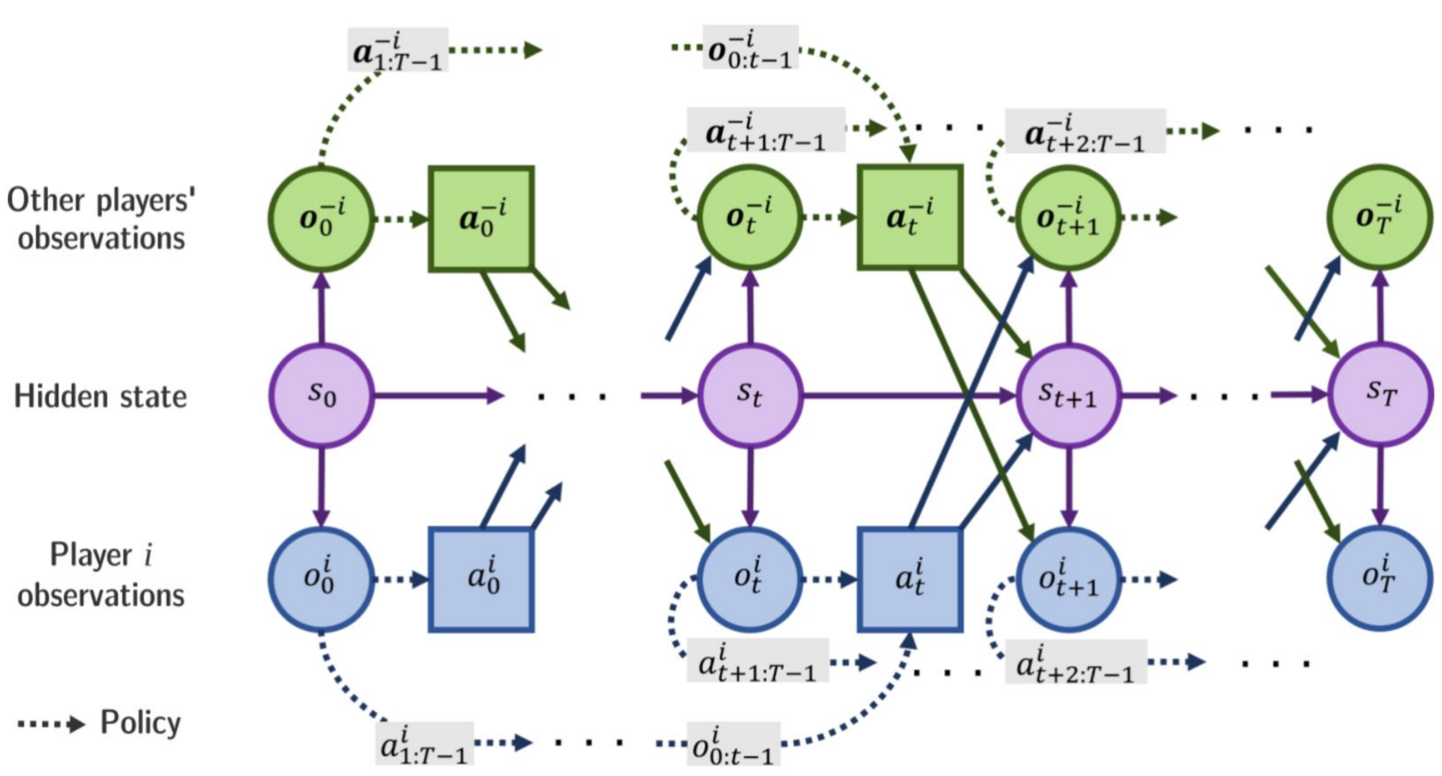
Preference model $\tilde{P}^i\left(s_{0:t}, o_{0:t}\right)$

- The model captures any finite utility function over states, joint observations and joint actions, including e.g., non-Markovian reward functions:



World model $Q^i\left(s_{0:t}, \vec{o}_{0:t}, \vec{a}_{0:t};\ \mu\right)$

- By definition, every player estimates the actions and observations of the other players, though other, local formulations of Q are possible.

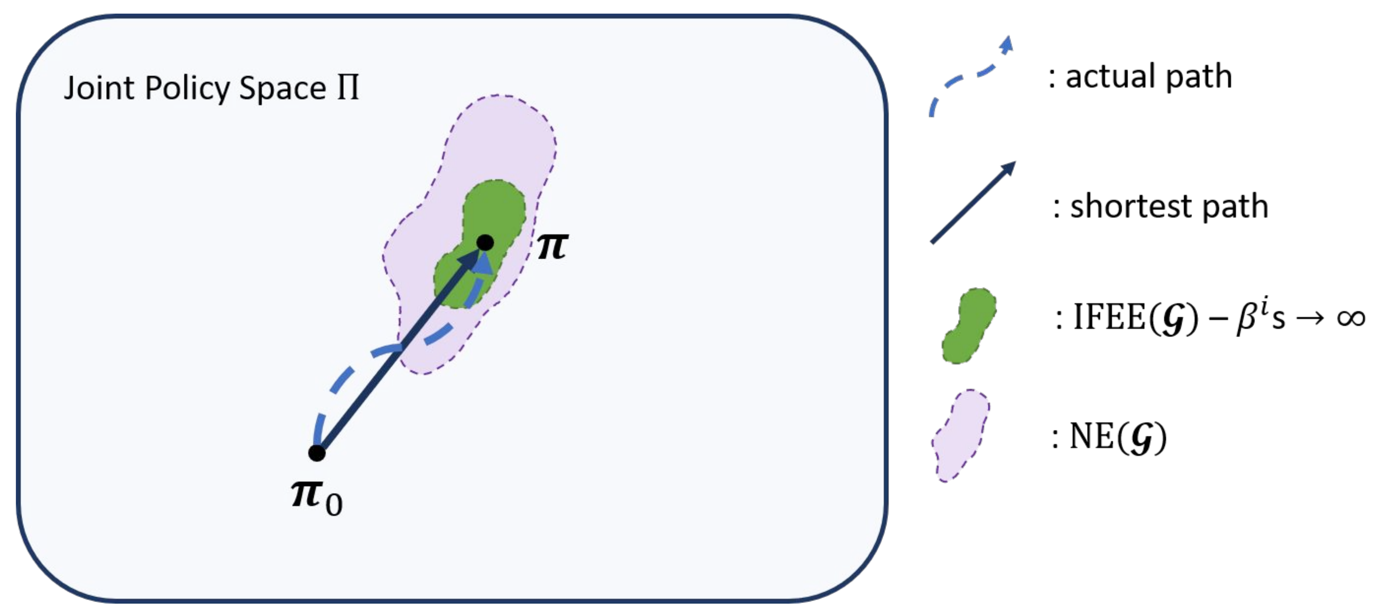


Free-Energy Equilibria (FEE)

- **Definition:** $\forall \hat{\pi}^i : G^i\left(\left(\hat{\pi}^i, \boldsymbol{\pi}^{-i}\right)\right) \geq G^i(\boldsymbol{\pi})$ for every player *i*, where G^i is a *free energy functional* of player *i*. That is, no player can decrease their subjective free energy by *unilaterally* changing their strategy
- This coincides with *Nash equilibrium* for $G^i(\boldsymbol{\pi}) = -V^i(\boldsymbol{\pi})$
- A similar FEE definition and correspondence for coarse correlated equilibria

Path divergence objective (PDO)

- Free energy functional generalizing the inf. theor. bounded rationality
- $$G^i(\pi) = \underbrace{\mathrm{D}_{\mathrm{KL}}\left[Q^i(\mathbf{h}_{0:T};\pi)\;||\;\tilde{P}^i(\mathbf{h}_{0:T})\right]}_{\text{Divergence from preferences}} + \underbrace{\mathbb{E}_{Q^i(\mathbf{h}_{0:T};\pi)}\left[-\log Q^i(\mathbf{h}_{0:T};\pi_0)\right]}_{\text{Cross entropy from prior}}$$
$$= \underbrace{\mathbb{E}_{Q^i(\mathbf{h}_{0:T};\pi)}\left[-\log \tilde{P}^i(\mathbf{h}_{0:T})\right]}_{\text{-Value (Energy)}} + \underbrace{\mathrm{D}_{\mathrm{KL}}\left[Q^i(\mathbf{h}_{0:T};\pi)\;||\;Q^i(\mathbf{h}_{0:T};\pi_0)\right]}_{\text{Divergence from prior}}$$
$$\geq \underbrace{\mathbb{E}_{Q^i(\mathbf{h}_{0:T};\pi)}\left[-\log \tilde{P}^i(\mathbf{h}_{0:T})\right]}_{\text{-Value (Energy)}} - \underbrace{H\left(Q^i(\mathbf{h}_{0:T};\pi)\right)}_{\text{Entropy}}$$
where $\mathbf{h}_{0:T} = \left(s_0, \vec{o}_0, \vec{a}_0, s_1, \dots, \vec{a}_T\right)$
- PDO is a lower bound on any real-world ("inf. processing cost" - "reward")



Generalization of Nash and Coarse correlated equilibria

- When $\beta \to \infty$ all PDO FEEs converge to Nash equilibria; similarly for CCE

Applications and Research Directions

Free energy formulations of AI alignment proposals to enhance realism by incorporating varying rationality, biased world models, and information-seeking behavior. Examples include: modeling of human-AI interactions (large rationality difference); bounded rationality formulations of the *Assistance game (CIRL).*

Joint vs individual free energy as a **measure of cooperation**, indicating collaboration or conflict levels and potentially quantifying collective agency.

Learning and non-equilibrium dynamics, algorithms and their convergence. Learning the generative model formulated as hidden state discovery. Identify potential convergent policy-learning algorithms. Generalize from maximum entropy over states to maximum caliber over trajectories.

Models of agents' internal cognition include graphical models of *P* and *Q*, hierarchical architectures, and metacognition, and could integrate perception, learning, belief updating, planning, and action selection.

Mechanism design for boundedly-rational agents involves developing incentive structures accounting for limited rationality and optimizing information provision. This could lead to more effective and fair system designs.

FEE-based multi-agent systems as models of collective decision-making, social norm formation, and emergent communication protocols.

Theoretical extensions should focus on linking FEE with other equilibrium concepts, developing microfoundations for PDO-based FEE, and mapping the space of FEEs over various active inference and other functionals.

ICML MFHAIA paper

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

ACS
ACSresearch.org

Charles University

UNIVERSITY OF OXFORD