Free-Energy Equilibria: Toward a Theory of Interactions Between Boundedly-Rational Agents

David Hyland [1 *] **Tomáš Gavenčiak** [2 *] **Lancelot Da Costa** [3 4 5] **Conor Heins** [3 6] **Vojtech Kovarik** [7]
Julian Gutierrez [8] **Michael J. Wooldridge** [1] **Jan Kulveit** [2]

Abstract

We propose a novel framework for modelling strategic interactions between boundedly-rational agents in complex, partially observable environments. Our approach introduces agents that minimize a free-energy functional, capturing the divergence between their beliefs about future trajectories and their preferences, which are represented by a biased probabilistic model. We extend this to multi-agent settings and introduce Free-Energy Equilibria, a new class of game-theoretic solution concepts. We begin by establishing the relationship between Free-Energy Equilibria and existing game-theoretic solution concepts. Then, we propose an approach to studying cooperation by contrasting Free-Energy Equilibria with joint free-energy minimization, extending key concepts from mechanism design. Our framework allows for modelling interactions between agents with varying levels of rationality and biased or incorrect world models, providing insights into human-AI interaction and AI alignment.

1. Introduction

Efforts to understand human behaviour, cooperation, and the alignment of AI systems with human values rely on formal models of human and AI agency. While simplifying assumptions can lead to useful insights, the problem of aligning AI systems with humans requires us to critically examine the realism of these models. Unrealistic assumptions about human agency or artificially intelligent systems, when used to align powerful AI systems, can lead to unethical and potentially catastrophic outcomes (Russell, 2019; Critch & Krueger, 2020).

A realistic model of both natural and artificial agency should at least capture (but is not limited to) (i) embodied agents who persist through time, model their world, and interact with it through an interface/Markov blanket (Friston et al., 2023a;b; Ramstead et al., 2023), (ii) strategic interactions between agents who may have different beliefs and preferences, (iii) partially observable, stochastic environments, and (iv) varying degrees of agent rationality. Traditional game theory provides a foundation for analyzing strategic interactions but assumes perfect rationality. It also typically omits explicit models of agents' beliefs and how they are revised over time, limiting its applicability to realistic scenarios with cognitively constrained agents (Simon, 1964; Kahneman, 2003). Free-energy-based frameworks, such as Active Inference (Friston et al., 2010) and Action-Perception Divergence (APD) minimization (Hafner et al., 2020), are a compelling class of models of perception and action based on (variational) free-energy minimization, but are currently lacking a general theory of multi-agent interactions.

We propose an application of game theory in studying interactions between free-energy minimizing agents to bridge the gap between idealized game-theoretic results and the behaviour of realistic agents. Our model studies boundedly-rational agents as those who minimize a free-energy functional, which captures the divergence between their beliefs (represented by a predictive generative model) and preferences (represented by a biased generative model) over future trajectories. This leads us to the concept of Free-Energy Equilibrium (FEE) and its generalizations.

Our framework is built upon Partially Observable Stochastic Games (POSGs) (Hansen et al., 2004b), extending the Partially Observable Markov Decision Process (POMDP) to multi-agent settings. Our framework allows one to model

[*]Equal contribution [1]Department of Computer Science, University of Oxford, United Kingdom [2]Alignment of Complex Systems, Charles University, Prague, Czech Republic [3]VERSES AI Research Lab, Los Angeles, CA 90016, USA [4]Department of Mathematics, Imperial College London, United Kingdom [5]Wellcome Centre for Human Neuroimaging, University College London, United Kingdom [6]Department of Collective Behaviour, Max Planck Institute of Animal Behavior, Konstanz D-78457, Germany [7]Foundations of Cooperative AI Lab, Carnegie Mellon University, Pittsburgh, USA [8]Department of Data Science & AI, Monash University, Australia. Correspondence to: David Hyland <david.hyland@cs.ox.ac.uk>, Tomáš Gavenčiak <tomas.gavenciak@acsresearch.org>.

ICML 2024 Workshop on Models of Human Feedback for AI Alignment, Vienna, Austria. 2024.

boundedly-rational agents through (1) approximate latent state estimation using variational inference, (2) sub-optimal policy selection due to bounded rationality, and (3) varying levels of belief updating and self-modelling. Preferences are modelled by assuming that agents desire to minimize the KL divergence between 'expected' and 'desired' trajectories, a common approach in free-energy methods, bounded rationality models, and variational inference (Friston et al., 2010; Hafner et al., 2020; Ortega & Braun, 2013).

The implications of our work extend beyond the introduction of free-energy equilibria and their theoretical analysis: We introduce the concept of joint free-energy minimization as a framework for studying cooperation, conflict, and alignment in multi-agent systems. We hope to lay the foundation for understanding the dynamics of human-AI interactions and addressing key challenges in AI alignment. We also illustrate the use of the framework by translating an existing formal alignment proposal, Cooperative Inverse Reinforcement Learning (Hadfield-Menell et al., 2016), to our model of bounded rationality.

2. Related Work

Formal models of human agency permeate many fields, ranging from psychology and cognitive science to machine learning and microeconomics. Due to space constraints, we briefly review only the most pertinent areas.

2.1. Economics and Game Theory

Economics and game theory are concerned with developing models of bounded rationality to capture real-world decision-makers' behaviors (Simon, 1964; Braun & Ortega, 2014). The objective we study extends information-theoretic models of bounded rationality, which assume that agents incur information-processing costs (Ortega & Braun, 2013; Braun & Ortega, 2014). It is also closely related to rational inattention, which models the costs of information acquisition in decision-making (Maćkowiak et al., 2023; Matějka & McKay, 2015; Sims, 2003). We study a more general class of dynamic decision-making problems and focus more on the costs of policy and belief updating rather than on information acquisition, although the latter can also be incorporated into our model. Bounded rationality has been applied to strategic interactions using statistical and information-theoretic models (Evans & Prokopenko, 2021; Gottwald & Braun, 2019; Rogers et al., 2009; McKelvey & Palfrey, 1995; 1998; Wolpert, 2006a;b). However, these models do not explicitly account for the internal models that agents use to plan and make decisions.

Mechanism design, a subfield of game theory, focuses on designing rules that lead to desirable outcomes in strategic settings. Foundational concepts were developed by Leonid Hurwicz (Hurwicz, 1973), while later contributions introduced key ideas such as incentive compatibility (Myerson, 1981) and equilibrium selection (Harsanyi & Selten, 1988). Social choice theory explores the aggregation of individual preferences to implement collective decision-making. This framework could be applied to explaining behavioural experiments, offering a theoretical framework for policy design, studying social behaviours, and potentially providing microfoundations for macroeconomic phenomena due to its applicability across scales (Yudkowsky, 2017; Ramstead et al., 2021b). In incentive design, it can suggest ways to combine 'utility-based' and 'information-based' incentives to shape behaviour (Ratliff et al., 2019).

2.2. Computer Science and Biology

Information theory and Bayesian inference form a bridge between modern AI and biologically-inspired theories of agency. For example, many machine learning problems can be viewed as KL divergence minimization (Hafner et al., 2020; Millidge et al., 2021a;b; Alemi, 2023) or Evidence Lower Bound maximization (Beal, 2003; Blei et al., 2017). Recent language modelling approaches using Reinforcement Learning (RL) with KL divergence penalties can be viewed as approximate Bayesian inference (Korbak et al., 2022). Given this, it is natural to investigate whether artificial agents based on neural networks and RL can be modelled as boundedly-rational in the ways we assume. With the rapid growth of access to computational power, sample-efficient model-based methods in (multi-agent) RL have been garnering interest (Eysenbach et al., 2021; Moerland et al., 2020) alongside more data-intensive model-free methods (Albrecht et al., 2024; Sutton & Barto, 2018). The model we propose here may provide a general framework for developing model-based (multi-agent) RL algorithms in partially observable settings by learning generative models of players and environments for planning (Ray et al., 2008; Chandrasekaran et al., 2017; Wang et al., 2022). Indeed, a closely related framework known as Maximum Diffusion RL has recently been proposed as a principled way of regularizing reward functions for RL agents to decorrelate their *experiences*, which leads to an objective very similar to the one we propose here (Berrueta et al., 2024), but arrived at from a different starting point.

Active Inference is a theoretical framework for modelling biological systems (Parr et al., 2022), which has more recently found useful applications in describing generally agentic systems in the formalism of POMDPs (Da Costa et al., 2020). Within the field of active inference, there has been much interest in studying interactions between multiple active inference agents, focusing on applications such as communication (Albarracin et al., 2022; 2024a; Friston & Frith, 2015a;b; Friston et al., 2023d; Vasil et al., 2020), competition (Demekas et al., 2023), coordination (Friedman

et al., 2021; Levchuk et al., 2019; Maisto et al., 2023; Pöppel et al., 2022), social cognition (Constant et al., 2019; Gallagher & Allen, 2018; Hipólito & van Es, 2022; Veissière et al., 2020; Yoshida et al., 2008), collective behaviour (Friston et al., 2015a; 2024; Heins et al., 2022; 2023; Kaufmann et al., 2021; Ramstead et al., 2021a), and hierarchical self-organisation (Friston et al., 2015a; Hesp et al., 2019; Kuchling et al., 2020; Palacios et al., 2020; Ramstead et al., 2019; Sims, 2021). Despite this, the field is currently lacking a general formalism and theory that can be used to model such multi-agent interactions, in the same way that POMDPs (Da Costa et al., 2020; 2023; 2024) and algorithms for (approximately) solving them (Fountas et al., 2020; Maisto et al., 2021) have been applied to model active inference in single-agent settings. We hope to take steps towards addressing this gap in the literature with our proposed model.

3. Preliminaries

We define some basic notation and conventions, with technical details deferred to the appendix. For a set X, let $\Delta(X)$ denote the set of probability distributions over X. Where X is a discrete random variable with distribution P, we write $H(X)$ to denote the Shannon entropy of X. For two probability distributions P and Q defined over the same domain X, we write $\mathrm{D}_{\mathrm{KL}}\left[P(x) \,||\, Q(x)\right]$ to denote the Kullback-Leibler (KL) divergence or relative entropy from Q to P. Finally, we use boldface to denote n-tuples, e.g., $\mathbf{x} = (x_1, \ldots, x_n)$. For $I \subseteq \{1, \ldots, n\}$, we let $\mathbf{x}^I$ denote the tuple $(x^i)_{i \in I}$ and $\mathbf{x}^{-I}$ to denote the tuple $(x^i)_{i \in \{1,\ldots,n\} \setminus I}$. When adding an additional time index $t \in \mathbb{Z}$, we denote $\mathbf{x}_t = (x_t^1, \ldots, x_t^n)$. For $s, t \in \mathbb{Z}$ such that $s \leqslant t$, we use $\mathbf{x}_{s:t}$ to denote the sequence $\mathbf{x}_s, \mathbf{x}_{s+1}, \ldots, \mathbf{x}_t$ and $x_{s:t}^i$ to denote $x_s^i, \ldots, x_t^i$.

3.1. Partially Observable Stochastic Games

We use the framework of Partially Observable Stochastic Games (POSGs) (Hansen et al., 2004a), which are a natural extension of Partially Observable Markov Decision Processes (POMDPs) to situations involving multiple agents acting concurrently. POMDPs are a widely used model for individual decision-making in dynamic stochastic environments where agents may be uncertain about the true state of the environment (Åström, 1965). They have been successfully applied to problems in reinforcement learning (Kimura et al., 1997), active inference (Parr et al., 2022), robotics (Kurniawati, 2022), and many other disciplines. The wide applicability of POMDPs thus makes their multi-agent extension an obvious first choice for the underlying model in which to describe situations involving multiple heterogeneous interacting active inference agents. POSGS are also a suitably general framework upon which we can build a model that satisfies the desiderata outlined in Section 1.

Definition 1. *A finite-horizon **Partially Observable Stochastic Game** (POSG) is a tuple*

$$\mathcal{G} = \left(N, S, (A^i)_{i \in N}, (\Omega^i)_{i \in N}, (O^i)_{i \in N}, T, P, I, (R^i)_{i \in N}\right)$$

*where: (i) N is a finite set of **agents**; (ii) S is a set of **states**; (iii) A^i is a set of **actions** for each $i \in N$. We write $A = \times_{i \in N} A^i$ for the set of **joint actions**; (iv) Ω^i is a **set of observations** for each $i \in N$. We write $\Omega = \times_{i \in N} \Omega^i$ for the set of **joint observations**; (v) $T \in \mathbb{Z}^+$ is the **time horizon**. We write $\mathbb{T} = \{0, \ldots, T\}$ for the set of time steps in the game; (vi) $O^i : A \times S \rightharpoonup \Delta(\Omega^i)$ is a partial **observation probability function** (or observation likelihood function) for each agent $i \in N$. For a joint observation $\mathbf{o} = (o^i)_{i \in N} \in \Omega$ and a time $t \in \mathbb{T}$, we write $O(\mathbf{o}_t | s_t, \mathbf{a}_{t-1}) = \times_{i \in N} O^i(o_t^i | s_t, \mathbf{a}_{t-1})$ for the joint probability that each agent $i \in N$ receives observations o_t^i, given that the state is s_t and the most recently played joint action was $\mathbf{a}_{t-1}$. (vii) $p : S \times A \to \Delta(S)$ is a Markovian **probabilistic transition function**, which can also be written as a conditional probability distribution $P(s_{t+1} | s_t, \mathbf{a}_t)$ for times $t \in \{0, \ldots, T-1\}$; (viii) $I \in \Delta(S)$ is the **initial state distribution**; (ix) $R^i : S \to \mathbb{R}$ is agent i's **state-reward function**. We write $R(s) = (R^1(s), \ldots, R^N(s))$.*

To capture the output of this process between any two timesteps t_0 and t, $0 \leqslant t_0 \leqslant t \leqslant T$, we use the notion of a *trajectory* $\mathbf{h}_{t_0:t} := s_{t_0} \mathbf{o}_{t_0} \mathbf{a}_{t_0} \ldots s_{t-1} \mathbf{o}_{t-1} \mathbf{a}_{t-1} s_t \mathbf{o}_t$ — that is, an alternating sequence of states, joint observations, and joint actions. We use the term *history* (or run) $\mathbf{h}_{0:t}$ to refer to trajectories that start at timestep $t_0 = 0$, and similarly for state histories $\mathbf{s}_{0:t}$, observation histories $\mathbf{o}_{0:t}$, etc. Let $\mathbb{O}_t$ be the set of all possible observation trajectories of length t and $\mathbb{O} := \bigcup_{t=1}^{T} \mathbb{O}_t$ as the set of all possible observation trajectories of any length. Similarly, we use $\mathbb{O}_t^i = (\Omega^i)^t$ and $\mathbb{O}^i = \bigcup_{t=1}^{T} (\Omega^i)^t$ to denote the sets of observation histories of a given player.

A *policy* $\pi^i : \mathbb{O}^i \to \Delta(A^i)$ of player i maps each observation history of i to a probability distribution over their actions. An independent *policy profile* (or a profile of *independent* policies) is a tuple $\boldsymbol{\pi} = (\pi^1, \ldots, \pi^n)$, where each π^i is a policy of player i. A policy (or a policy profile) is said to be *pure* when each distribution $\pi^i(\,\cdot\,|o_{0:t}^i)$ is deterministic. A *correlated policy profile* $\boldsymbol{\mu}$ is a probability distribution over pure policy profiles.[2] We use $\Pi_{\text{ind}} = \times_{i=1}^{n} \Pi^i$ and Π_{corr} to denote the sets of all independent, resp. correlated, policy profiles. For a correlated policy profile $\boldsymbol{\mu}$, we will use $\boldsymbol{\mu}^{-i}$ to denote the corresponding marginal distribution over $\times_{j \neq i} \Pi^j$.

In order to compute expected rewards in $\mathcal{G}$, it is useful to consider the auxiliary notion of the probability of reaching

[2] Note that the definitions relevant to correlated policies make sense even for probability distribution over profiles of non-deterministic policies.

a given history. We define the *reach probability* of a history $\mathbf{h} = \mathbf{h}_{0:t}$ under an independent policy profile $\boldsymbol{\pi}$ as

$$p(\mathbf{h};\boldsymbol{\pi}) := I(s_0)\cdot\Big(\prod_{\tau=0}^{T-1} O(\mathbf{o}_\tau|s_\tau,a_{\tau-1})\cdot\boldsymbol{\pi}(\mathbf{a}_\tau|\mathbf{o}_{0:\tau}) \cdot p(s_{\tau+1}|s_\tau,\mathbf{a}_\tau)\Big)\cdot O(\mathbf{o}_T|s_T), \quad (1)$$

where $\boldsymbol{\pi}(\mathbf{a}_\tau|\mathbf{o}_{0:\tau}) := \prod_{i=1}^n \pi^i(a^i_\tau|o^i_{0:\tau})$. Analogously, we define the reach probability of $\mathbf{h}$ under a *correlated* policy profile $\boldsymbol{\mu}$ as $p(\mathbf{h};\boldsymbol{\mu}) := \mathbb{E}_{\boldsymbol{\pi}\sim\boldsymbol{\mu}}\, p(\mathbf{h};\boldsymbol{\pi})$. For any (independent or correlated) policy profile $\boldsymbol{\pi}$, these notions can be straightforwardly extended to track other probability distributions related to the game (Kovařík et al., 2023), such as the conditional probabilities $p(\mathbf{h}_{0:t}|\mathbf{h}_{0:t_0};\boldsymbol{\pi})$ of reaching a given history $\mathbf{h}_{0:t}$ given that the current history is $\mathbf{h}_{0:t_0}$, $t_0 \leqslant t$.

3.2. Value functions and Solution Concepts

A central concept in control theory and reinforcement learning is known as the (subjective) *value function*, which is a measure of the expected reward to-go for a player $i \in N$ from a given time point t in a run until the end of the episode, under the policy profile $\boldsymbol{\pi}$.

Definition 2. *Given an (independent or correlated) policy profile $\boldsymbol{\pi}$ and a history $\mathbf{h}_{0:t}$, the **value function** of agent i is given by*

$$V^i(o^i_{0:t};\boldsymbol{\pi}) = \mathbb{E}_{p(\mathbf{h}_{0:T}\,|\,o^i_{0:t};\boldsymbol{\pi})}\left[\sum_{\tau=t+1}^{T} R^i(s_\tau)\right], \quad (2)$$

where $\mathbf{h}_{0:T} = \mathbf{h}_{0:t}\,\mathbf{a}_t s_{t+1}\ldots s_T\mathbf{o}_T$. By $V^i(\boldsymbol{\pi}) = V^i(\varnothing;\boldsymbol{\pi})$, we denote the total expected reward under $\boldsymbol{\pi}$.

This measures a player's expectation of the reward-to-go under a given joint policy $\boldsymbol{\pi}$ and the information that they have access to, i.e., $o^i_{0:t}$. The objective of a classical reward-maximizing agent is thus to select a policy which maximizes its initial value function, given information about the policies of the other players. Game theory typically assumes an equilibrium state where each player knows the policies of the others (Aumann & Brandenburger, 1995). However, this assumption of equilibrium can be relaxed, giving rise to the need to learn and possibly even shape opponent policies (Foerster et al., 2017; He et al., 2016; Yu et al., 2022).

The historically prominent solution concept in game theory is Nash equilibrium — a profile of independent policies that allows no profitable deviations for any individual player. However, in the context of this paper, we will be more interested in the notion of correlated equilibrium, where players may use an external source of randomness (e.g., a mediator or the outcome of some random process) to coordinate on which joint actions to play.

While the notion of a correlated equilibrium is relatively straightforward in single-step interactions (i.e., normal form games), there are several ways of extending the idea to sequential settings (Zhang et al., 2022). What all these variants have in common is that (i) the players use the correlated policy profile $\boldsymbol{\mu}$ to randomly select a pure policy profile $\boldsymbol{\pi} \sim \boldsymbol{\mu}$ to adopt, and (ii) $\boldsymbol{\mu}$ is said to be a correlated equilibrium if none of the players can benefit by unilaterally deviating from this plan. The variants differ in the types of deviations available to the players.

Definition 3. *A **Nash Equilibrium** (NE) is an* independent *policy profile $\boldsymbol{\pi}$ s.t. for all $i \in N$ and $\tilde{\pi}^i \in \Pi^i$, $V^i(\boldsymbol{\pi}) \geqslant V^i((\tilde{\pi}^i, \boldsymbol{\pi}^{-i}))$. A **Coarse Correlated Equilibrium** (CCE) is a* correlated *policy profile $\boldsymbol{\mu}$ s.t. for all $i \in N$ and $\tilde{\pi}^i \in \Pi^i$, $V^i(\boldsymbol{\mu}) \geqslant V^i((\tilde{\pi}^i, \boldsymbol{\mu}^{-i}))$.*

We write $\mathrm{CCE}(\mathcal{G})$ for the set of all CCEs of a POSG $\mathcal{G}$ and $\mathrm{NE}(\mathcal{G})$ for the set of Nash equilibria of $\mathcal{G}$.

4. Free-Energy Equilibria

We introduce a generative and preferential model for POSGs and derive three free-energy-based solution concepts: Coarse Correlated Free-Energy Equilibrium (CCFEE) and (Logit) Independent Free-Energy Equilibrium (IFEE) based on a novel free-energy functional of a given policy profile. We demonstrate their connection to classic game-theoretic concepts, making assumptions that place the analysis in the realm of perfect rationality. However, the general framework does not require agents to have perfect predictive models or perform exact Bayesian inference.

Free-energy based models of agency generally consist of two models: an unbiased generative model representing the agent's beliefs about the environment's dynamics, and a biased preference model encoding the agent's preferences as a desired distribution over states or observations. Agents aim to minimize the divergence between these distributions through their choice of actions.

Generative Models. Active inference agents, and predictive agents in general, embody a generative model, which furnishes them with beliefs about the trajectories of a game and beliefs about other agents with whom they interact. In general, an agent's generative model consists of probabilistic beliefs $P^i(\mathbf{h}_{0:t}\,;\,\boldsymbol{\mu})$ about possible histories of the system:

$$P^i(\mathbf{h}_{0:t}\,;\,\boldsymbol{\pi}) := P^i(s_0)\Big(\prod_{\tau=0}^{t-1} P^i(\mathbf{o}_\tau|s_\tau)P^i(s_{\tau+1}|s_\tau,\mathbf{a}_\tau) \cdot\,\pi^i(a^i_\tau|o^i_{0:\tau})P^i(\mathbf{a}^{-i}_\tau|\mathbf{o}_{0:\tau})\Big)P^i(\mathbf{o}_t|s_t). \quad (3)$$

(and analogously for correlated $\boldsymbol{\mu}$ and $P^i(\mathbf{h}_{0:t}\,;\,\boldsymbol{\mu})$). Given a history of observations $o^i_{0:t}$ and a policy profile $\boldsymbol{\mu}$, an agent may infer the latent causes of its sensory data by com-

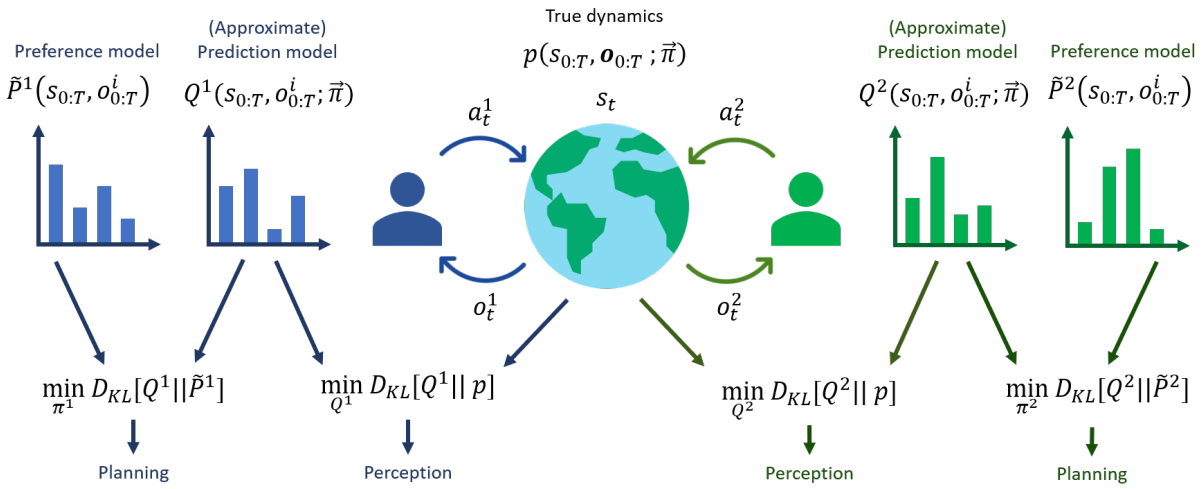

Figure 1. High-level summary of our approach. Each agent embodies a predictive generative model of its environment and other agents, which is approximated by a variational distribution Q^i. Agents also possess a preference model, which assigns higher probabilities to more preferable trajectories. Policies are selected to minimize the Kullback-Leibler divergence between the two distributions, formalising the intuition that agents aim to act in order to minimize the difference between their expectations and their desires.

puting the posterior probability $P^i(\mathbf{h}_t|o^i_{0:t};\boldsymbol{\mu})$. However, even with knowledge of the true generative process, computing such a posterior exactly is computationally intractable in general due to the need to evaluate the model evidence $P^i(o^i_{0:t};\boldsymbol{\mu})$, which involves computing a sum over all possible trajectories that could have given rise to the sequence of observations. Thus, in practice, this generative model is usually approximated by optimizing a variational approximation $Q^i(\mathbf{h}_t;\boldsymbol{\mu})$ called a *recognition model* (Ramstead et al., 2020) to the true posterior $P^i(\mathbf{h}_t|o^i_{0:t};\boldsymbol{\mu})$, which is usually chosen from some tractable family of distributions that admits a factorisation over states, policies, and observations as defined above (Da Costa et al., 2020). A common form of this process is known as variational Bayesian inference, which aims to approximate the posterior P^i by minimising a functional known as the *variational free-energy* (VFE) or the negative *evidence lower bound* (ELBO) (Beal, 2003; Blei et al., 2017), which is an upper bound on the surprise, i.e., the negative log probability of the observed data.

For the purposes of situating our proposed solution concept in relation to standard game-theoretic equilibria, we will make the following assumptions[3] on the predictive models Q^i of agents, which are comprised of a retrospective recognition model and a prospective predictive model (Aumann & Brandenburger, 1995):

1. Agents have perfect knowledge of the generative process and other agents' policies: $Q^i(\mathbf{h};\boldsymbol{\mu}) = p(\mathbf{h};\boldsymbol{\mu})$;
2. The agents' variational approximations Q^i to the true posterior do indeed minimize their VFE and they can thus perform exact Bayesian inference: $Q^i = P^i = p$.

[3]While these assumptions are unrealistic for boundedly-rational agents, we emphasize that they are only utilised to identify the connection between our proposed solution concepts and standard equilibrium concepts in game theory.

Preference Models. In information-theoretic bounded rationality and active inference, preferences are represented as probability distributions over future trajectories, with the interpretation that more preferable trajectories are assigned higher probability (Ortega et al., 2015; Parr et al., 2022). This formulation licenses a description of attaining one's preferences as acting in order to minimize the discrepancy between an agent's predictions and preferences about the unfolding of the system over time, given particular policies. This allows one to express a wider range of preference structures, including risk-aversion, social preferences, and non-Markovian utility functions over trajectories of the game (Skalse & Abate, 2023).

We thus adopt the following general functional form for an agent's preference model, which is defined as a joint probability distribution over states, joint observations, and joint actions:

$$\tilde{P}^i(s_{0:T}, o^i_{0:T}) = \prod_{\tau=0}^{T} \tilde{P}^i(s_\tau) \cdot \tilde{P}^i(o^i_\tau|s_\tau) \tag{4}$$

Notice that in this preference model, the agent adopts an independent prior over states at each time step, which does not take into account transition probabilities between individual states. This reflects a kind of 'wishful thinking', in which an agent's preference model solely reflects the desirability of different states at each point in time, without

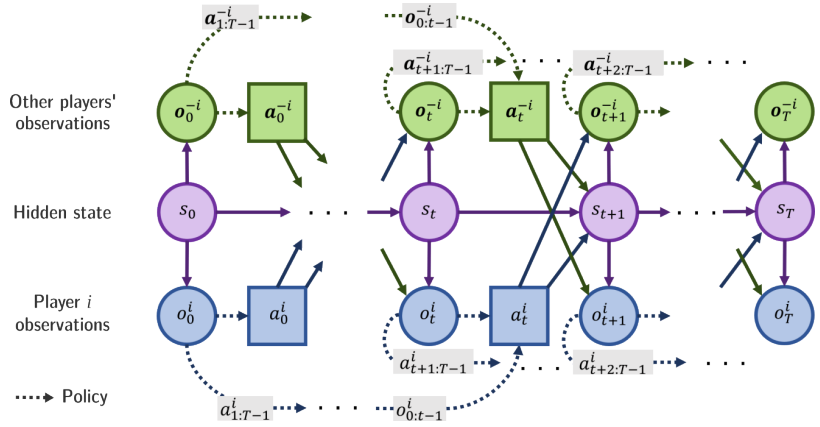

Figure 2. Multi-Agent Influence Diagram representing an agent i's predictive generative model over several time-steps. Square nodes represent action distributions determined by each player's policy $\boldsymbol{\pi}^i$, represented by dashed lines.

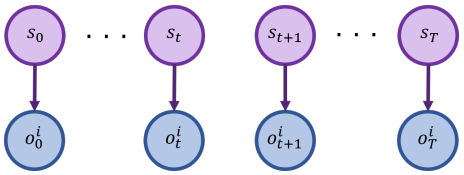

Figure 3. The Probabilistic Graphical Model that we adopt for representing an agent i's preference model.

regard for the particular path that the agent takes over time. Given a POSG with a particular reward function R^i, we can convert the agents' implied preference into a preference model over states, as in Equation 4. To do this, we define a biased prior over states, where for each state $s \in S$, we have $\tilde{P}^i(s) = \exp(\beta^i R^i(s))/Z^i(\beta^i)$, where $\beta^i \in \mathbb{R}_+$ is an agent-specific parameter that determines how motivated agent i is to occupy reward-maximising states and $Z^i(\beta^i) = \sum_{s' \in S} \exp(\beta^i R^i(s'))$ is a normalising constant for each agent's preference model. Thus, under the interpretation that an agent assigns a high (biased) prior probability to states that it prefers, this distribution encodes the intuition that an agent prefers states with higher rewards (for $\beta^i > 0$). In addition, we make no particular assumptions on the form of the preferred observation likelihood function $\tilde{P}^i(o^i_{0:T}|s_{0:T})$. Incorporating these assumptions into the agent's predictive generative model yields:

$$\tilde{P}^i(s_{0:T}, o^i_{0:T}) = \prod_{\tau=0}^{T} \frac{\exp(\beta^i R^i(s_\tau))}{Z^i(\beta^i)} \cdot \tilde{P}^i(o^i_\tau|s_\tau). \quad (5)$$

There is a large design space for preference models to explore. For example, including a biased prior over actions may be used to encode a basic notion of "habits", which bias action towards those that are *a priori* preferred (Friston et al., 2016b; Han et al., 2024).

4.1. Sophisticated Divergence Objective

We define the divergence objective for agents associated with a policy profile $\boldsymbol{\mu}$ recursively by simply taking the expectation of the KL divergence from the preference model to the prediction model with respect to the joint policy. It can be shown that this is equivalent to minimizing the cross entropy of the prediction model with respect to the preference model, subject to information processing costs (Ortega et al., 2015).

Definition 4. *The **Sophisticated Divergence Objective** (SDO) for an agent i in a POSG $\mathcal{G}$ given a correlated policy profile $\boldsymbol{\mu}$ is defined recursively as:*

$$G^i(o^i_{0:T-1}; \boldsymbol{\mu}) = \mathbb{E}_{Q^i(\mathbf{o}^{-i}_{0:T-1}, s_{T-1}|o^i_{0:T-1}; \boldsymbol{\mu}), \boldsymbol{\mu}(\mathbf{a}_{T-1}|\mathbf{o}_{0:T-1})} D_{KL}\left[Q^i(s_T, o^i_T|s_{T-1}, \mathbf{a}_{T-1}) \,||\, \tilde{P}^i(s_T, o^i_T)\right] \quad (6)$$

$$G^i(o^i_{0:t}; \boldsymbol{\mu}) = \mathbb{E}_{Q^i(\mathbf{o}^{-i}_{0:t}, s_t|o^i_t; \boldsymbol{\mu}), \boldsymbol{\mu}(\mathbf{a}_t|\mathbf{o}_{0:t})} \Big[D_{KL}\left[Q^i(s_{t+1}, o^i_{t+1}|s_t, \mathbf{a}_t) \,||\, \tilde{P}^i(s_{t+1}, o^i_{t+1})\right] + \mathbb{E}_{Q^i(o^i_{t+1}|s_t, \mathbf{a}_t)}\left[G^i((o^i_{0:t}, o^i_{t+1}); \boldsymbol{\mu})\right]\Big] \quad (7)$$

$$G^i(\boldsymbol{\mu}) = D_{KL}\left[Q^i(s_0, o^i_0) \,||\, \tilde{P}^i(s_0, o^i_0)\right] + \mathbb{E}_{Q^i(o^i_0)}\left[G^i(o^i_0; \boldsymbol{\mu})\right]. \quad (8)$$

We sometimes later refer to this objective as an agent's 'free-energy'. This objective is closely related to the Free-Energy of the Expected Future (Millidge et al., 2021b) and the Action Perception Divergence objectives (Hafner et al., 2020), which naturally model the tradeoffs between taking actions to resolve uncertainty about the world and maximizing rewards.[4] Also closely related is the notion of expected

[4]For a detailed study of such tradeoffs, we refer the reader to the expositions in (Millidge et al., 2021a;b).

free-energy (Barp et al., 2022; Da Costa et al., 2024), which can also be represented in our framework using a suitably defined preference model.

4.2. Free-Energy Equilibrium

Given this model of objectives, we are now in a position to propose a free-energy-based solution concept for boundedly-rational agents.

Definition 5. *A correlated policy profile $\boldsymbol{\mu}$ is a **Coarse Correlated Free-Energy Equilibrium** (CCFEE) in a POSG $\mathcal{G}$ if for all $\hat{\pi}^i \in \Pi^i$, it holds that*

$$G^i(\boldsymbol{\mu}) \leqslant G^i((\hat{\pi}^i, \boldsymbol{\mu}^{-i})).$$

*An independent policy profile $\boldsymbol{\pi}$ is called an **Independent Free-Energy Equilibrium** (IFEE) if it satisfies the analogous condition (with $\boldsymbol{\pi}$ in place of $\boldsymbol{\mu}$). We let* $\mathrm{CCFEE}(\mathcal{G})$*, resp.* $\mathrm{IFEE}(\mathcal{G})$*, denote the set of all CCFEEs, resp IFEEs, of $\mathcal{G}$. The equilibria are said to be **strict** if the inequality above is strict.*

Since every independent policy profile $\boldsymbol{\pi}$ can be represented as a correlated policy profile, we immediately obtain $\mathrm{IFEE}(\mathcal{G}) \subseteq \mathrm{CCFEE}(\mathcal{G})$.

For any $\beta = (\beta^1, \dots, \beta^n) \in \mathbb{R}^n_+$,, every POSG $\mathcal{G}$ has a Free-Energy Equilibrium. On the other hand, the uniqueness of CCFEE and IFEE is not guaranteed in general, as with many solution concepts. An example demonstrating this non-uniqueness is provided in Appendix B.

4.3. Connections to Existing Solution Concepts

Here, we establish the relationship between the instantiations of the family of free-energy equilibria proposed above to existing solution concepts in the game theory literature. We begin by formally showing how CCFEE and IFEE relate to Coarse Correlated and Nash equilibria in POSGs, and then briefly discuss Quantal Response Equilibrium.

Theorem 6. *For a commonly known POSG $\mathcal{G}$, we have* $\mathrm{CCFEE}(\mathcal{G}) \subseteq \mathrm{CCE}(\mathcal{G})$ *for sufficiently large β^i, $i \in N$.*

Note that since the sets of pure/mixed equilibria are contained within the set of CCEs, it is straightforward to restrict the class of strategies permitted and obtain similar correspondences between modified versions of the CCFEE and the other well-known solution concepts mentioned above.

Corollary 7. *For a commonly known POSG $\mathcal{G}$, we have* $\mathrm{IFEE}(\mathcal{G}) \subseteq \mathrm{NE}(\mathcal{G})$ *for sufficiently large β^i, $i \in N$.*

Moving to models of bounded rationality in games, the Quantal Response Equilibrium (QRE) is a well-known solution concept (McKelvey & Palfrey, 1995; 1998). We focus here on the case of independent policies for simplicity, because a proper treatment of quantal correlation would require the introduction of a mediator sampling signals to coordinate the agents (Černỳ et al., 2022).

Definition 8. *A **Logit Independent Free-Energy Equilibrium** (LIFEE) is an independent joint policy $\boldsymbol{\pi}$ such that for all $i \in N, t \in \mathbb{T}, o^i_{0:t} \in \tilde{\mathbb{O}}_t, a^i_t \in A^i$, we have*

$$\pi^i(a^i_t | o^i_{0:t}) = \frac{\exp(-G^i((o^i_{0:t}, a^i_t); \boldsymbol{\pi}))}{\sum_{a^{i\prime} \in A^i} \exp(-G^i((o^i_{0:t}, a^{i\prime}); \boldsymbol{\pi}))}, \quad (9)$$

where $G^i((o^i_{0:t}, a^t_i); \boldsymbol{\pi})$ is obtained by conditioning the expectations in $G^i((o^i_{0:t}; \boldsymbol{\pi})$ on a^i_t. We show that in the limit as $\beta^i \to \infty$ for each agent, this solution concept tends towards Nash equilibrium as in Theorem 6[5]. However, in the case of finite β^i's, our model is not equivalent to the standard quantal response framework in general. It would thus be useful to compare the empirical predictions of human choice behaviour under our model with existing approaches in the literature.

5. Applications to Cooperation and AI Alignment

Studying cooperation and conflict among participants is crucial for modelling multi-agent systems, both for analysis and for promoting desired outcomes. We propose that free-energy-based models provide a promising framework for studying cooperation and conflict among boundedly-rational players with incomplete information. Unlike game theory and mechanism design, which primarily focus on outcomes by analyzing games in terms of expected utility, the free-energy framework can capture how a more rational player's ability to make precise predictions and inferences about their environment affects their utility, even when the additional utility gained is minimal.

Joint Free-Energy. We introduce the notion of the joint free-energy to study cooperation and conflict among boundedly-rational agents. The *joint free-energy* of a policy profile $G^\Sigma(\boldsymbol{\mu})$ is defined as the sum of individual free-energies under that policy: $G^\Sigma(\boldsymbol{\mu}) = \sum_i G^i(\boldsymbol{\mu})$. A policy profile that minimizes G^Σ is called a *joint free-energy minimizing policy*, denoted as $\mu^\Sigma \in \arg\min G^\Sigma(\boldsymbol{\mu})$.

$G^\Sigma(\boldsymbol{\mu})$ is one way of quantifying the total amount of prediction error or surprisal (Schwartenbeck et al., 2013) (in the sense of a mismatch between beliefs and desires) experienced by all agents collectively under the policy profile $\boldsymbol{\mu}$. Minimizing the joint free-energy can thus be seen as a way for agents to coordinate their behaviour to achieve the best joint outcome, taking into account their beliefs, preferences, and cognitive limitations. In this sense, joint free-energy minimization provides a bounded rationality analogue of utilitarian social welfare maximization.

[5]The full result is presented in Appendix A.

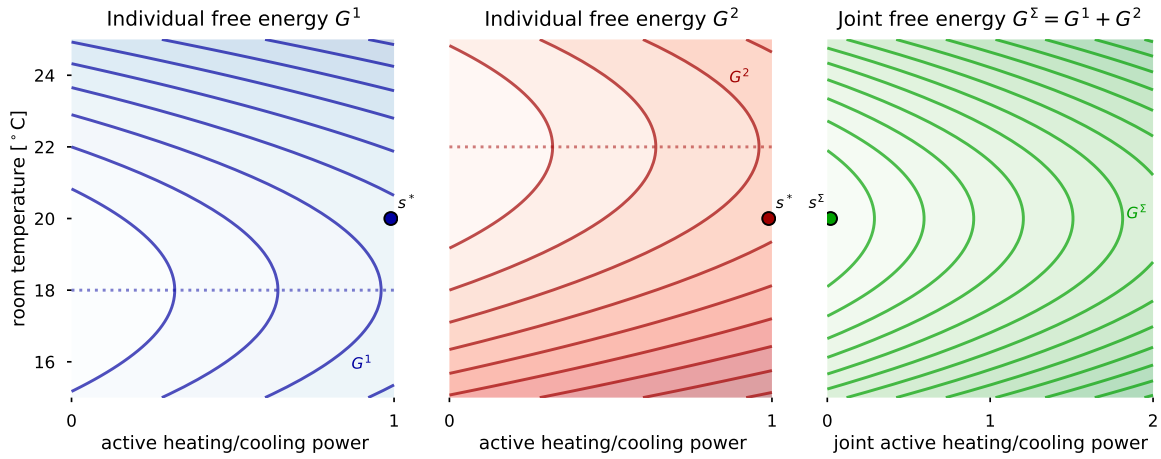

Figure 4. Illustration of individual vs joint free-energy minimization: A full-information scenario with two thermostats in one room, both with equal maximal heating and cooling power, each preferring to use less power themselves. However, they have different preferences for the room temperature (18°C and 22°C). The contour plots show the free-energy corresponding to each state (temperature, power used) of the two thermostats individually (1st and 2nd plots), and their joint free energy (3rd plot). The plotted free energy is the free energy of a policy that would maintain that state. (Note that plotting the space of policies is not feasible due to their complexity.) Darker colours indicate higher free-energy. Assuming a starting temperature of 20°C, selfish play stabilizes at s^* with both players actively opposing each other with maximal power, while the joint free-energy minimizing policy stabilizes at s^Σ, using no energy at the same temperature.

By definition, the joint free-energy of any Free-Energy Equilibrium (FEE) $G^\Sigma(\boldsymbol{\mu})$ is at least $G^\Sigma(\boldsymbol{\mu}^\Sigma)$. This means that if agents could coordinate effectively and share their free energy, they may prefer to play $\boldsymbol{\mu}^\Sigma$ over $\boldsymbol{\mu}$. Indeed, these conditions have been identified as potentially crucial mechanisms for implementing collective agency in biological organisms (McMillen & Levin, 2024; Shreesha & Levin, 2024). See Figure 4 for an example. Studying the gap between individual free-energy minimization (as in FEE) and joint free-energy minimization thus provides a new way to quantify and understand possible tensions between individual and collective rationality in multi-agent settings. This gap, which is analogous to the price of anarchy (Koutsoupias & Papadimitriou, 1999) may serve as one proxy for the degree to which a group of agents may be thought of as a collective agent. Additionally, the distribution of free-energies in equilibrium may be used to understand the fairness implications of diverse 'rationality parameters' β^i among agents.

Applications to AI Alignment. The question of how to align advanced AI systems with humans is a critical open problem in AI. Whilst not resolving the ethical dimensions of this problem, the Free-Energy Equilibrium framework provides a structure which can be used to articulate alignment objectives more precisely. As the relative cognitive capabilities of AI systems to humans increases over time, accurately modelling bounded rationality will become increasingly vital to understanding and shaping the evolving dynamics of human-AI interactions. Current approaches like Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017) assume that human feedback meaningfully reflects long-term preferences, but this assumption may break down as the cognitive gap narrows. As AI advances, this assumption may lead to unintended consequences, such as the promotion of highly-rated policies by humans lacking a view of their broader impacts. Models that faithfully account for these cognitive limitations may help us to identify novel undesirable scenarios that arise specifically from this feature of the agents.

Assistance Games in the Context of Free-Energy. To conclude this discussion, we translate the 'Cooperative Inverse Reinforcement Learning game' (CIRL) (Hadfield-Menell et al., 2016) to the free-energy framework as an example of extending AI alignment research with our model of bounded rationality. CIRL studies a setting where a powerful AI assists a human while initially being uncertain about the human's goals.

Formally, an *assistance game* $\mathcal{G}$ is any fully-cooperative partially observable stochastic (POSG) game with two players: an Assistant A and a Human H, where the reward function $R^H = R^A = R(\cdot \mid \theta)$ depends on a hidden parameter $\theta \in \Theta$, drawn from a prior P_0 at the start of the game, with the value of θ revealed to H but not to A. Players do not observe a reward signal to avoid leaking information about θ, and the players' strategies are generally restricted to independent (uncorrelated) policies.

While this is not an explicit part of $R(\cdot \mid \theta)$, it is often crucial for the Assistant to infer as much as possible about θ. It is thus beneficial for both players to steer the game in a way

that provides the Assistant with useful information about θ.

In the context of free-energy-minimizing agents, the following additional features can be modelled: (i) each of the agents is not just boundedly-rational with parameters β^H and β^A, but also aware of the bounded nature of the other player; (ii) the boundedness of each agent (β^i) may be common knowledge, or it may be sampled privately at the beginning of the game, and the players need to infer it; (iii) agents form (implicit or explicit) posterior estimates of the hidden state at any moment, including θ; and (iv) agents are also motivated by information-seeking (Schwartenbeck et al., 2013; Friston et al., 2015b), in particular focusing on information that helps them predict their reward, which is well-aligned with the motivation of the assistance game.

While the assistance game is a good example in part due to its simplicity and generality, it may rather serve as a starting point for further research into various alignment objectives, particularly those including multi-principal/multi-agent settings (Sourbut et al., 2024) and symbiotic or empathetic relationships between agents (Albarracin et al., 2024b).

6. Conclusion

This paper proposes a model of multiple boundedly-rational agents interacting in partially observable environments, inspired by the active inference and action-perception divergence frameworks. We introduce the Sophisticated Divergence Objective and three free-energy-based solution concepts: Coarse Correlated Free-Energy Equilibrium and (Logit) Independent Free-Energy Equilibrium. We establish relationships between these solution concepts and classical game-theoretic notions in the limit of perfectly rational agents, prove the existence and non-uniqueness of Free-Energy Equilibria, and discuss their potential generalizations. In addition we propose the concept of joint free-energy minimization as a framework for studying cooperation, conflict, and alignment in multi-agent systems. We then discuss the relevance of our framework to AI alignment and demonstrate it by translating the CIRL game to the free-energy setting. This highlights the potential of our model to provide insights into the dynamics of human-AI interactions and address key challenges in AI alignment.

Future research includes investigating the relationship between FEE and other solution concepts (McKelvey & Palfrey, 1995; Harsanyi, 1967), further studying cooperation and conflict through the lens of joint free-energy, applying FEE to model stated vs. revealed collective preferences, extending learning theory and algorithms to multi-agent free-energy-minimizing systems (Friston et al., 2016a; 2023c; Sajid et al., 2022), incorporating Predictive Game Theory to model an external observer's predictions about the system (Wolpert, 2005), quantifying and analyzing collective behavior (Levin, 2019; 2021; 2023; McMillen & Levin, 2024), and studying more realistic formal models of AI alignment.

Acknowledgments

The authors would like to thank James Fox, Lewis Hammond, Matt MacDermott, and Jakub Steiner for helpful feedback and discussions on the work.

References

Albarracin, M., Demekas, D., Ramstead, M. J., and Heins, C. Epistemic communities under active inference. *Entropy*, 24(4):476, 2022.

Albarracin, M., Pitliya, R. J., St. Clere Smithe, T., Friedman, D. A., Friston, K., and Ramstead, M. J. Shared protentions in multi-agent active inference. *Entropy*, 26(4):303, 2024a.

Albarracin, M., Ramstead, M., Pitliya, R. J., Hipolito, I., Da Costa, L., Raffa, M., Constant, A., and Manski, S. G. Sustainability under active inference. *Systems*, 12(5), 2024b. ISSN 2079-8954. doi: 10.3390/systems12050163. URL https://www.mdpi.com/2079-8954/12/5/163.

Albrecht, S. V., Christianos, F., and Schäfer, L. *Multi-Agent Reinforcement Learning: Foundations and Modern Approaches*. MIT Press, 2024. URL https://www.marl-book.com.

Alemi, A. A. Information theory for representation learning, 2023. URL https://nips.cc/virtual/2023/73986. NeurIPS.

Åström, K. J. Optimal control of markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10(1):174–205, 1965.

Aumann, R. and Brandenburger, A. Epistemic conditions for nash equilibrium. *Econometrica: Journal of the Econometric Society*, pp. 1161–1180, 1995.

Barp, A., Da Costa, L., França, G., Friston, K., Girolami, M., Jordan, M. I., and Pavliotis, G. A. Geometric methods for sampling, optimization, inference, and adaptive agents. In *Handbook of Statistics*, volume 46, pp. 21–78. Elsevier, 2022.

Beal, M. J. *Variational algorithms for approximate Bayesian inference*. PhD thesis, UCL (University College London), 2003.

Berrueta, T. A., Pinosky, A., and Murphey, T. D. Maximum diffusion reinforcement learning. *Nature Machine Intelligence*, pp. 1–11, 2024.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Braun, D. A. and Ortega, P. A. Information-theoretic bounded rationality and ε-optimality. *Entropy*, 16(8): 4662–4676, 2014.

Černỳ, J., An, B., and Zhang, A. N. Quantal correlated equilibrium in normal form games. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pp. 210–239, 2022.

Chandrasekaran, M., Chen, Y., and Doshi, P. On markov games played by bayesian and boundedly-rational players. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Constant, A., Ramstead, M. J., Veissière, S. P., and Friston, K. Regimes of expectations: an active inference model of social conformity and human decision making. *Frontiers in psychology*, 10:420184, 2019.

Critch, A. and Krueger, D. Ai research considerations for human existential safety (arches), 2020.

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: 102447, 2020.

Da Costa, L., Sajid, N., Parr, T., Friston, K., and Smith, R. Reward maximization through discrete active inference. *Neural Computation*, 35(5):807–852, 2023.

Da Costa, L., Tenka, S., Zhao, D., and Sajid, N. Active inference as a model of agency. *arXiv preprint arXiv:2401.12917*, 2024.

Demekas, D., Heins, C., and Klein, B. An analytical model of active inference in the iterated prisoner's dilemma. In *International Workshop on Active Inference*, pp. 145–172. Springer, 2023.

Evans, B. P. and Prokopenko, M. A maximum entropy model of bounded rational decision-making with prior beliefs and market feedback. *Entropy*, 23(6):669, 2021.

Eysenbach, B., Khazatsky, A., Levine, S., and Salakhutdinov, R. Mismatched no more: Joint model-policy optimization for model-based rl. *ArXiv*, abs/2110.02758, 2021. URL https://api.semanticscholar.org/CorpusID:238408209.

Foerster, J. N., Chen, R. Y., Al-Shedivat, M., Whiteson, S., Abbeel, P., and Mordatch, I. Learning with opponent-learning awareness. *arXiv preprint arXiv:1709.04326*, 2017.

Fountas, Z., Sajid, N., Mediano, P., and Friston, K. Deep active inference agents using monte-carlo methods. *Advances in neural information processing systems*, 33: 11662–11675, 2020.

Friedman, D. A., Tschantz, A., Ramstead, M. J., Friston, K., and Constant, A. Active inferants: an active inference framework for ant colony behavior. *Frontiers in behavioral neuroscience*, 15:647732, 2021.

Friston, K. and Frith, C. A duet for one. *Consciousness and cognition*, 36:390–405, 2015a.

Friston, K., Levin, M., Sengupta, B., and Pezzulo, G. Knowing one's place: a free-energy approach to pattern regulation. *Journal of the Royal Society Interface*, 12(105): 20141383, 2015a.

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., FitzGerald, T., and Pezzulo, G. Active inference and epistemic value. *Cognitive neuroscience*, 02 2015b. doi: 10.1080/17588928.2015.1020053.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., OḊoherty, J., and Pezzulo, G. Active inference and learning. *Neuroscience and Biobehavioral Reviews*, 68:862–879, 2016a. ISSN 0149-7634. doi: https://doi.org/10.1016/j.neubiorev.2016.06.022. URL https://www.sciencedirect.com/science/article/pii/S0149763416301336.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., Pezzulo, G., et al. Active inference and learning. *Neuroscience & Biobehavioral Reviews*, 68:862–879, 2016b.

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: 1–29, 2023a.

Friston, K., Da Costa, L., Sakthivadivel, D. A., Heins, C., Pavliotis, G. A., Ramstead, M., and Parr, T. Path integrals, particular kinds, and strange things. *Physics of Life Reviews*, 2023b.

Friston, K. J. and Frith, C. D. Active inference, communication and hermeneutics. *cortex*, 68:129–143, 2015b.

Friston, K. J., Daunizeau, J., Kilner, J., and Kiebel, S. J. Action and behavior: A free-energy formulation. 102 (3):227–260, 2010. ISSN 1432-0770. doi: 10.1007/s00422-010-0364-z. URL https://doi.org/10.1007/s00422-010-0364-z.

Friston, K. J., Costa, L. D., Tschantz, A., Kiefer, A., Salvatori, T., Neacsu, V., Koudahl, M. T., Heins, C., Sajid, N., Markovic, D., Parr, T., Verbelen, T., and Buckley, C. L. Supervised structure learning. *ArXiv*, abs/2311.10300, 2023c. URL https://api.semanticscholar.org/CorpusID:265281169.

Friston, K. J., Parr, T., Heins, C., Constant, A., Friedman, D., Isomura, T., Fields, C., Verbelen, T., Ramstead, M., Clippinger, J., et al. Federated inference and belief sharing. *Neuroscience and Biobehavioral Reviews*, pp. 105500, 2023d.

Friston, K. J., Ramstead, M. J., Kiefer, A. B., Tschantz, A., Buckley, C. L., Albarracin, M., Pitliya, R. J., Heins, C., Klein, B., Millidge, B., et al. Designing ecosystems of intelligence from first principles. *Collective Intelligence*, 3(1):26339137231222481, 2024.

Gallagher, S. and Allen, M. Active inference, enactivism and the hermeneutics of social cognition. *Synthese*, 195 (6):2627–2648, 2018.

Gottwald, S. and Braun, D. A. Systems of bounded rational agents with information-theoretic constraints. *Neural computation*, 31(2):440–476, 2019.

Hadfield-Menell, D., Russell, S. J., Abbeel, P., and Dragan, A. Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

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.

Han, D., Doya, K., Li, D., and Tani, J. Synergizing habits and goals with variational bayes. *Nature Communications*, 15(1):4461, 2024.

Hansen, E. A., Bernstein, D. S., and Zilberstein, S. Dynamic programming for partially observable stochastic games. In *AAAI*, volume 4, pp. 709–715, 2004a.

Hansen, E. A., Bernstein, D. S., and Zilberstein, S. Dynamic programming for partially observable stochastic games. In *AAAI*, volume 4, pp. 709–715, 2004b.

Harsanyi, J. and Selten, R. *A General Theory of Equilibrium Selection in Games*. MIT Press Classics. MIT Press, 1988. ISBN 9780262582384. URL https://books.google.cz/books?id=zjwkHAAACAAJ.

Harsanyi, J. C. Games with incomplete information played by "bayesian" players, i–iii part i. the basic model. *Management science*, 14(3):159–182, 1967.

He, H., Boyd-Graber, J., Kwok, K., and Daumé III, H. Opponent modeling in deep reinforcement learning. In *International conference on machine learning*, pp. 1804–1813. PMLR, 2016.

Heins, C., Klein, B., Demekas, D., Aguilera, M., and Buckley, C. L. Spin glass systems as collective active inference. In *International Workshop on Active Inference*, pp. 75–98. Springer, 2022.

Heins, C., Millidge, B., Da Costa, L., Mann, R., Friston, K., and Couzin, I. Collective behavior from surprise minimization. *arXiv preprint arXiv:2307.14804*, 2023.

Hesp, C., Ramstead, M., Constant, A., Badcock, P., Kirchhoff, M., and Friston, K. A multi-scale view of the emergent complexity of life: A free-energy proposal. In *Evolution, development and complexity: multiscale evolutionary models of complex adaptive systems*, pp. 195–227. Springer, 2019.

Hipólito, I. and van Es, T. Enactive-dynamic social cognition and active inference. *Frontiers in Psychology*, 13: 855074, 2022.

Hurwicz, L. The design of mechanisms for resource allocation. *The American Economic Review*, 63(2):1–30, 1973. ISSN 00028282. URL http://www.jstor.org/stable/1817047.

Kahneman, D. Maps of bounded rationality: Psychology for behavioral economics. *The American Economic Review*, 93(5):1449–1475, 2003. ISSN 00028282. URL http://www.jstor.org/stable/3132137.

Kaufmann, R., Gupta, P., and Taylor, J. An active inference model of collective intelligence. *Entropy*, 23(7):830, 2021.

Kimura, H., Miyazaki, K., and Kobayashi, S. Reinforcement learning in pomdps with function approximation. In *ICML*, volume 97, pp. 152–160, 1997.

Korbak, T., Perez, E., and Buckley, C. L. Rl with kl penalties is better viewed as bayesian inference. *arXiv preprint arXiv:2205.11275*, 2022.

Koutsoupias, E. and Papadimitriou, C. Worst-case equilibria. In *Annual symposium on theoretical aspects of computer science*, pp. 404–413. Springer, 1999.

Kovařík, V., Seitz, D., Lisỳ, V., Rudolf, J., Sun, S., and Ha, K. Value functions for depth-limited solving in zero-sum imperfect-information games. *Artificial Intelligence*, 314: 103805, 2023.

Kuchling, F., Friston, K., Georgiev, G., and Levin, M. Morphogenesis as bayesian inference: A variational approach to pattern formation and control in complex biological systems. *Physics of life reviews*, 33:88–108, 2020.

Kuhn, H. W. Extensive form games and the problem of information. *Contributions to the Theory of Games II*, pp. 193–216, 1953.

Kurniawati, H. Partially observable markov decision processes and robotics. *Annual Review of Control, Robotics, and Autonomous Systems*, 5:253–277, 2022.

Levchuk, G., Pattipati, K., Serfaty, D., Fouse, A., and McCormack, R. Active inference in multiagent systems: context-driven collaboration and decentralized purpose-driven team adaptation. In *Artificial Intelligence for the Internet of Everything*, pp. 67–85. Elsevier, 2019.

Levin, M. The computational boundary of a "self": Developmental bioelectricity drives multicellularity and scale-free cognition. *Frontiers in Psychology*, 10, 2019. URL https://api.semanticscholar.org/CorpusID:209324686.

Levin, M. Technological approach to mind everywhere: An experimentally-grounded framework for understanding diverse bodies and minds. *Frontiers in Systems Neuroscience*, 16, 2021. URL https://api.semanticscholar.org/CorpusID:245553743.

Levin, M. Bioelectric networks: the cognitive glue enabling evolutionary scaling from physiology to mind. *Animal Cognition*, 26:1865 – 1891, 2023. URL https://api.semanticscholar.org/CorpusID:258785547.

Maćkowiak, B., Matějka, F., and Wiederholt, M. Rational inattention: A review. *Journal of Economic Literature*, 61(1):226–273, 2023.

Maisto, D., Gregoretti, F., Friston, K., and Pezzulo, G. Active inference tree search in large pomdps. *arXiv preprint arXiv:2103.13860*, 2021.

Maisto, D., Donnarumma, F., and Pezzulo, G. Interactive inference: a multi-agent model of cooperative joint actions. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2023.

Matějka, F. and McKay, A. Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, 105(1):272–298, 2015.

McKelvey, R. D. and Palfrey, T. R. Quantal response equilibria for normal form games. *Games and economic behavior*, 10(1):6–38, 1995.

McKelvey, R. D. and Palfrey, T. R. Quantal response equilibria for extensive form games. *Experimental economics*, 1:9–41, 1998.

McMillen, P. and Levin, M. Collective intelligence: A unifying concept for integrating biology across scales and substrates. *Communications Biology*, 7(1):378, 2024.

Millidge, B., Seth, A., and Buckley, C. Understanding the origin of information-seeking exploration in probabilistic objectives for control. *arXiv preprint arXiv:2103.06859*, 2021a.

Millidge, B., Tschantz, A., and Buckley, C. L. Whence the expected free energy? *Neural Computation*, 33(2): 447–482, 2021b.

Moerland, T. M., Broekens, J., and Jonker, C. M. Model-based reinforcement learning: A survey. *Found. Trends Mach. Learn.*, 16:1–118, 2020. URL https://api.semanticscholar.org/CorpusID:220265929.

Myerson, R. B. Optimal auction design. *Mathematics of Operations Research*, 6(1):58–73, 1981. ISSN 0364765X, 15265471. URL http://www.jstor.org/stable/3689266.

Nash Jr, J. F. Equilibrium points in n-person games. *Proceedings of the national academy of sciences*, 36(1):48–49, 1950.

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):20120683, 2013.

Ortega, P. A., Braun, D. A., Dyer, J., Kim, K.-E., and Tishby, N. Information-theoretic bounded rationality. *arXiv preprint arXiv:1512.06789*, 2015.

Palacios, E. R., Razi, A., Parr, T., Kirchhoff, M., and Friston, K. On markov blankets and hierarchical self-organisation. *Journal of theoretical biology*, 486:110089, 2020.

Parr, T., Pezzulo, G., and Friston, K. J. *Active inference: the free energy principle in mind, brain, and behavior*. MIT Press, 2022.

Pöppel, J., Kahl, S., and Kopp, S. Resonating minds—emergent collaboration through hierarchical active inference. *Cognitive Computation*, 14(2):581–601, 2022.

Ramstead, M. J., Constant, A., Badcock, P. B., and Friston, K. J. Variational ecology and the physics of sentient systems. *Physics of life Reviews*, 31:188–205, 2019.

Ramstead, M. J., Kirchhoff, M. D., and Friston, K. J. A tale of two densities: Active inference is enactive inference. *Adaptive behavior*, 28(4):225–239, 2020.

Ramstead, M. J., Hesp, C., Tschantz, A., Smith, R., Constant, A., and Friston, K. Neural and phenotypic representation under the free-energy principle. *Neuroscience and Biobehavioral Reviews*, 120:109–122, 2021a.

Ramstead, M. J., Kirchhoff, M. D., Constant, A., and Friston, K. J. Multiscale integration: beyond internalism and externalism. *Synthese*, 198(Suppl 1):41–70, 2021b.

Ramstead, M. J., Sakthivadivel, D. A., Heins, C., Koudahl, M., Millidge, B., Da Costa, L., Klein, B., and Friston, K. J. On bayesian mechanics: a physics of and by beliefs. *Interface Focus*, 13(3):20220029, 2023.

Ratliff, L. J., Dong, R., Sekar, S., and Fiez, T. A perspective on incentive design: Challenges and opportunities. *Annual Review of Control, Robotics, and Autonomous Systems*, 2(1):305–338, 2019. doi: 10.1146/annurev-control-053018-023634.

Ray, D., King-Casas, B., Montague, P., and Dayan, P. Bayesian model of behaviour in economic games. *Advances in neural information processing systems*, 21, 2008.

Rogers, B. W., Palfrey, T. R., and Camerer, C. F. Heterogeneous quantal response equilibrium and cognitive hierarchies. *Journal of Economic Theory*, 144(4):1440–1467, 2009.

Russell, S. *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin Publishing Group, 2019. ISBN 9780525558620. URL https://books.google.cz/books?id=M1eFDwAAQBAJ.

Sajid, N., Tigas, P., and Friston, K. J. Active inference, preference learning and adaptive behaviour. *IOP Conference Series: Materials Science and Engineering*, 1261, 2022. URL https://api.semanticscholar.org/CorpusID:252814583.

Schwartenbeck, P., Fitzgerald, T., Dolan, R. J., and Friston, K. Exploration, novelty, surprise, and free energy minimization. 4:710, 2013. ISSN 1664-1078. doi: 10.3389/fpsyg.2013.00710.

Shreesha, L. and Levin, M. Stress sharing as cognitive glue for collective intelligences: a computational model of stress as a coordinator for morphogenesis. 2024.

Simon, H. A. Theories of bounded rationality. pp. 161–176, 1964.

Sims, C. A. Implications of rational inattention. *Journal of monetary Economics*, 50(3):665–690, 2003.

Sims, M. How to count biological minds: symbiosis, the free energy principle, and reciprocal multiscale integration. *Synthese*, 199(1-2):2157–2179, 2021.

Skalse, J. and Abate, A. On the limitations of markovian rewards to express multi-objective, risk-sensitive, and modal tasks. In *Uncertainty in Artificial Intelligence*, pp. 1974–1984. PMLR, 2023.

Sourbut, O., Hammond, L., and Wood, H. Cooperation and control in delegation games. *arXiv preprint arXiv:2402.15821*, 2024.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Vasil, J., Badcock, P. B., Constant, A., Friston, K., and Ramstead, M. J. A world unto itself: human communication as active inference. *Frontiers in psychology*, 11:480375, 2020.

Veissière, S. P., Constant, A., Ramstead, M. J., Friston, K. J., and Kirmayer, L. J. Thinking through other minds: A variational approach to cognition and culture. *Behavioral and brain sciences*, 43:e90, 2020.

Wang, X., Zhang, Z., and Zhang, W. Model-based multi-agent reinforcement learning: Recent progress and prospects. *arXiv preprint arXiv:2203.10603*, 2022.

Wolpert, D. H. A predictive theory of games. *arXiv preprint nlin/0512015*, 2005.

Wolpert, D. H. Information theory—the bridge connecting bounded rational game theory and statistical physics. In *Complex Engineered Systems: Science meets technology*, pp. 262–290. Springer, 2006a.

Wolpert, D. H. What information theory says about bounded rational best response. In *The Complex Networks of Economic Interactions: Essays in Agent-Based Economics and Econophysics*, pp. 293–306. Springer, 2006b.

Yoshida, W., Dolan, R. J., and Friston, K. J. Game theory of mind. *PLoS computational biology*, 4(12):e1000254, 2008.

Yu, X., Jiang, J., Zhang, W., Jiang, H., and Lu, Z. Model-based opponent modeling. *Advances in Neural Information Processing Systems*, 35:28208–28221, 2022.

Yudkowsky, E. *Inadequate Equilibria: Where and How Civilizations Get Stuck*. Machine Intelligence Research Institute, 2017. ISBN 9781939311221. URL https://books.google.co.uk/books?id=zkvutAEACAAJ.

Zhang, B. H., Farina, G., Celli, A., and Sandholm, T. Optimal correlated equilibria in general-sum extensive-form games: Fixed-parameter algorithms, hardness, and two-sided column-generation. In *Proceedings of the 23rd ACM conference on economics and computation*, pp. 1119–1120, 2022.

A. Appendix

Theorem 6. *For a commonly known POSG $\mathcal{G}$, we have* $\mathrm{CCFEE}(\mathcal{G}) \subseteq \mathrm{CCE}(\mathcal{G})$ *for sufficiently large* β^i, $i \in N$.

Proof. The proof goes along the same lines as that of Theorem 16 in (Da Costa et al., 2023). Suppose that $\boldsymbol{\mu}$ is a CCFEE. Then, for all $i \in N$ and alternative policies $\hat{\pi}^i \in \Pi^i$, we have

$$
\begin{aligned}
&\lim_{\beta^i \to \infty} \arg\min_{\hat{\pi}^i \in \Pi^i} G^i((\hat{\pi}^i, \boldsymbol{\mu}^{-i})) \\
&= \lim_{\beta^i \to \infty} \arg\min_{\hat{\pi}^i \in \Pi^i} \sum_{\tau=0}^{T-1} \mathbb{E}_{Q^i(\mathbf{o}_{0:\tau}, s_\tau, \mathbf{a}_\tau | s_{\tau-1}, \mathbf{a}_{\tau-1}, o^i_{0:\tau}; (\hat{\pi}^i, \boldsymbol{\mu}^{-i}))} \left[\mathrm{D}_{\mathrm{KL}} \left[Q^i(s_{\tau+1}, o^i_{\tau+1} | s_\tau, \mathbf{a}_\tau) \,||\, \tilde{P}^i(s_{\tau+1}, o^i_{\tau+1}) \right] \right] \\
&= \lim_{\beta^i \to \infty} \arg\min_{\hat{\pi}^i \in \Pi^i} \sum_{\tau=0}^{T-1} \mathbb{E}_{Q^i(\mathbf{o}_{0:\tau}, s_\tau, \mathbf{a}_\tau | s_{\tau-1}, \mathbf{a}_{\tau-1}, o^i_{0:\tau}; (\hat{\pi}^i, \boldsymbol{\mu}^{-i}))} \left[-H \left[Q^i(s_{\tau+1}, o^i_{\tau+1} | s_\tau, \mathbf{a}_\tau) \right] \right. \\
&\qquad \left. - \mathbb{E}_{Q^i(s_{\tau+1}, o^i_{\tau+1} | s_\tau, \mathbf{a}_\tau)} \left[\log \tilde{P}^i(s_{\tau+1}, o^i_{\tau+1}) \right] \right] \\
&= \lim_{\beta^i \to \infty} \arg\min_{\hat{\pi}^i \in \Pi^i} \sum_{\tau=0}^{T-1} \mathbb{E}_{Q^i(\mathbf{o}_{0:\tau}, s_\tau, \mathbf{a}_\tau | s_{\tau-1}, \mathbf{a}_{\tau-1}, o^i_{0:\tau}; (\hat{\pi}^i, \boldsymbol{\mu}^{-i}))} \left[-H \left[Q^i(s_{\tau+1}, o^i_{\tau+1} | s_\tau, \mathbf{a}_\tau) \right] \right. \\
&\qquad \left. - \mathbb{E}_{Q^i(s_{\tau+1}, o^i_{\tau+1} | s_\tau, \mathbf{a}_\tau)} \left[\log \tilde{P}^i(s_{\tau+1}) + \log \tilde{P}^i(o^i_{\tau+1} | s_{\tau+1}) \right] \right] \\
&= \lim_{\beta^i \to \infty} \arg\min_{\hat{\pi}^i \in \Pi^i} \sum_{\tau=0}^{T-1} \mathbb{E}_{Q^i(\mathbf{o}_{0:\tau}, s_\tau, \mathbf{a}_\tau | s_{\tau-1}, \mathbf{a}_{\tau-1}, o^i_{0:\tau}; (\hat{\pi}^i, \boldsymbol{\mu}^{-i}))} \left[-H \left[Q^i(s_{\tau+1}, o^i_{\tau+1} | s_\tau, \mathbf{a}_\tau) \right] \right. \\
&\qquad \left. + \mathbb{E}_{Q^i(s_{\tau+1}, o^i_{\tau+1} | s_\tau, \mathbf{a}_\tau)} \left[-\beta^i R^i(s_{\tau+1}) + \log(Z^i(\beta^i)) - \log \tilde{P}^i(o^i_{\tau+1} | s_{\tau+1}) \right] \right] \\
&= \lim_{\beta^i \to \infty} \arg\max_{\hat{\pi}^i \in \Pi^i} \sum_{\tau=0}^{T-1} \mathbb{E}_{Q^i(\mathbf{o}_{0:\tau}, s_\tau, \mathbf{a}_\tau | s_{\tau-1}, \mathbf{a}_{\tau-1}, o^i_{0:\tau}; (\hat{\pi}^i, \boldsymbol{\mu}^{-i}))} \left[H \left[Q^i(s_{\tau+1}, o^i_{\tau+1} | s_\tau, \mathbf{a}_\tau) \right] \right. \\
&\qquad \left. + \mathbb{E}_{Q^i(s_{\tau+1}, o^i_{\tau+1} | s_\tau, \mathbf{a}_\tau)} \left[\beta^i R^i(s_{\tau+1}) - \log(Z^i(\beta^i)) + \log \tilde{P}^i(o^i_{\tau+1} | s_{\tau+1}) \right] \right] \\
&\subseteq \lim_{\beta^i \to \infty} \arg\max_{\hat{\pi}^i \in \Pi^i} \sum_{\tau=0}^{T-1} \mathbb{E}_{Q^i(\mathbf{o}_{0:\tau}, s_\tau, \mathbf{a}_\tau | s_{\tau-1}, \mathbf{a}_{\tau-1}, o^i_{0:\tau}; (\hat{\pi}^i, \boldsymbol{\mu}^{-i}))} \left[\mathbb{E}_{Q^i(s_{\tau+1}, o^i_{\tau+1} | s_\tau, \mathbf{a}_\tau)} \left[\beta^i R^i(s_{\tau+1}) - \log(Z^i(\beta^i)) \right] \right] \\
&= \arg\max_{\hat{\pi}^i \in \Pi^i} \sum_{\tau=0}^{T-1} \mathbb{E}_{Q^i(\mathbf{o}_{0:\tau}, s_\tau, \mathbf{a}_\tau, s_{\tau+1}, o^i_{\tau+1} | s_{\tau-1}, \mathbf{a}_{\tau-1}, o^i_{0:\tau}; (\hat{\pi}^i, \boldsymbol{\mu}^{-i}))} \left[R^i(s_{\tau+1}) \right] = \arg\max_{\hat{\pi}^i \in \Pi^i} V^i(\hat{\pi}^i, \boldsymbol{\mu}^{-i}).
\end{aligned}
$$

The first equality follows by unrolling the recursion in the definition of the SDO and noticing that the first KL divergence is not affected by the joint policy. The fourth equality follows from substituting the preference prior over states as defined in Equation 4. The set inclusion follows from the observation that the expected reward term dominates the entropy terms as $\beta^i \to \infty$ and the sixth equality follows from the fact that $Z^i(\beta^i)$ is independent of the chosen policy. The final equality follows from the assumption that the agents know the true transition probability function p of the game and are thus accurately able to estimate states, joint observations, joint actions, and rewards, given a joint policy. □

Quantal Response Equilibrium. The Quantal Response Equilibrium (QRE) is a solution concept developed to model boundedly-rational agents, where the assumption is made that the payoffs/reward functions are observed by agents as independent noisy samples from a probability distribution, rather than their true values (McKelvey & Palfrey, 1995; 1998). We focus here on the case of independent policies for simplicity, because a proper treatment of quantal correlation would require the explicit introduction of a mediator sampling signals upon which the agents would condition their actions (Černý et al., 2022).

Theorem 7. *For a commonly known POSG $\mathcal{G}$, we have* $LIFEE(\mathcal{G}) \subseteq \mathrm{NE}(\mathcal{G})$ *as* $\beta^i \to \infty$ *for all* $i \in N$ *and for some* $\epsilon > 0$.

Proof. We begin by recalling the rollout of the negative SDO given an observation trajectory and an action for an agent

$i \in N$:

$$
\begin{aligned}
-G^i((o^i_{0:t}, a^i_t); \boldsymbol{\pi}) = & \sum_{\tau=t}^{T-1} \mathbb{E}_{Q^i(\mathbf{o}_{0:\tau}, s_\tau, \mathbf{a}_\tau | s_{\tau-1}, \mathbf{a}_{\tau-1}, o^i_{0:\tau}, a^i_t; \boldsymbol{\pi})} \Big[H\left[Q^i(s_{\tau+1}, o^i_{\tau+1} | s_\tau, \mathbf{a}_\tau)\right] \\
& + \mathbb{E}_{Q^i(s_{\tau+1}, o^i_{\tau+1}|s_\tau, \mathbf{a}_\tau)} \left[\beta^i R^i(s_{\tau+1}) - \log(Z^i(\beta^i)) + \log \tilde{P}^i(o^i_{\tau+1} | s_{\tau+1}) \right] \Big] \\
= & \ \beta^i V^i((o^i_{0:t}, a^i_t); \boldsymbol{\pi}) + \sum_{\tau=t}^{T-1} \mathbb{E}_{Q^i(\mathbf{o}_{0:\tau}, s_\tau, \mathbf{a}_\tau | s_{\tau-1}, \mathbf{a}_{\tau-1}, o^i_{0:\tau}, a^i_t; \boldsymbol{\pi})} \Big[H\left[Q^i(s_{\tau+1}, o^i_{\tau+1} | s_\tau, \mathbf{a}_\tau)\right] \\
& + \mathbb{E}_{Q^i(s_{\tau+1}, o^i_{\tau+1}|s_\tau, \mathbf{a}_\tau)} \left[- \log(Z^i(\beta^i)) + \log \tilde{P}^i(o^i_{\tau+1} | s_{\tau+1}) \right] \Big] \\
= & \ \beta^i V^i((o^i_{0:t}, a^i_t); \boldsymbol{\pi}) - (T-t) Z^i(\beta^i) \\
& + \sum_{\tau=t}^{T-1} \mathbb{E}_{Q^i(\mathbf{o}_{0:\tau}, s_\tau, \mathbf{a}_\tau, s_{\tau+1}, o^i_{\tau+1} | s_{\tau-1}, \mathbf{a}_{\tau-1}, o^i_{0:\tau}, a^i_t; \boldsymbol{\pi})} \left[H\left[Q^i(s_{\tau+1}, o^i_{\tau+1} | s_\tau, \mathbf{a}_\tau)\right] + \log \tilde{P}^i(o^i_{\tau+1} | s_{\tau+1}) \right]
\end{aligned}
$$

Let $\mathcal{H}_{t:T-1}(a^t_i) := \sum_{\tau=t}^{T-1} \mathbb{E}_{Q^i(\mathbf{o}_{0:\tau}, s_\tau, \mathbf{a}^{-i}_\tau, s_{\tau+1}, o^i_{\tau+1} | s_{\tau-1}, \mathbf{a}_{\tau-1}, o^i_{0:\tau}, a^i_t; \boldsymbol{\pi})} \left[H\left[Q^i(s_{\tau+1}, o^i_{\tau+1} | s_\tau, \mathbf{a}_\tau) + \log \tilde{P}^i(o^i_{\tau+1} | s_{\tau+1}) \right] \right]$. Then, we have

$$
\begin{aligned}
\pi^i(a^i_t | o^i_{0:t}) &= \frac{\exp(-G^i((o^i_{0:t}, a^i_t); \boldsymbol{\pi}))}{\sum_{a^{i\prime} \in A^i} \exp(-G^i((o^i_{0:t}, a^{i\prime}); \boldsymbol{\pi}))} \\
&= \frac{\exp(\beta^i V^i((o^i_{0:t}, a^i_t); \boldsymbol{\pi}) - (T-t) Z^i(\beta^i) + \mathcal{H}_{t:T-1}(a^t_i))}{\sum_{a^{i\prime} \in A^i} \exp(c + \beta^i V^i((o^i_{0:t}, a^{i\prime}); \boldsymbol{\pi}) - (T-t) Z^i(\beta^i) + \mathcal{H}_{t:T-1}(a_{i\prime}))} \\
&= \frac{\exp(\beta^i V^i((o^i_{0:t}, a^i_t); \boldsymbol{\pi}) + \mathcal{H}_{t:T-1}(a^t_i))}{\sum_{a^{i\prime} \in A^i} \exp(\beta^i V^i((o^i_{0:t}, a^{i\prime}); \boldsymbol{\pi}) + \mathcal{H}_{t:T-1}(a_{i\prime}))}.
\end{aligned}
$$

In general, such a policy does not map neatly onto a LQRE unless further assumptions about the environment are made, such as a uniform transition probability function with respect to actions or even a deterministic transition function. However, taking the limit as $\beta^i \to \infty$, the value function term dominates, and the Boltzmann distribution concentrates around value function maximizing actions. Thus, letting $\beta^i \to \infty$ for all $i \in N$, we observe that every player's policy maximises their value function at the beginning of the game in any LIFEE $\boldsymbol{\pi}$, so any such joint policy is by definition a Nash equilibrium of $\mathcal{G}$. □

B. Properties of Free-Energy Equilibrium

Proposition 8. *For all $\beta = (\beta^1, \ldots, \beta^n) \in \mathbb{R}^n_+$, every commonly known POSG $\mathcal{G}$ with perfect recall has a Free-Energy Equilibrium.*

Proof. We will show the existence of an independent FEE by transforming the given game $\mathcal{G}$ into a modified POSG $\mathcal{G}_H$ whose rewards map directly onto reductions in expected free energy. The existence of a Nash Equilibrium (NE) in $\mathcal{G}_H$, which is guaranteed in any finite-horizon POSG with perfect recall (Kuhn, 1953; Nash Jr, 1950) is thus a witness to the existence of an independent FEE in $\mathcal{G}$, and moreover, an independent FEE can be extracted from any such NE in the transformed game. The construction of $\mathcal{G}_H$ proceeds as follows. For each $s, \mathbf{a}, s' \in S \times A \times S$, we add a new state $\zeta^{s,\mathbf{a},s'}$ to the game such that $R'^i(\zeta^{s,\mathbf{a},s'}) = H[Q^i(s'', o'^i | s, \mathbf{a})] + \beta^i R^i(s') - \log(Z^i(\beta^i)) + \mathbb{E}_{Q^i(s'', o'^i | s, \mathbf{a})} \left[\log \tilde{P}^i(o'^i | s'') \right]$ for all $i \in N$. In addition to this, we let $R'^i(s) = 0$ for all $s \in S$, so that all rewards are obtained in the new states. Then, we modify the transition function p so that $p(s, \mathbf{a}, \zeta^{s,\mathbf{a},s'}) = p(s, \mathbf{a}, s')$ for all $s, \mathbf{a}, s' \in S \times A \times S$ and $p(\zeta^{s,\mathbf{a},s'}, \mathbf{a}', s'') = 1$ if $s'' = s'$ and is 0 otherwise. To ensure that no additional information is leaked to the players in the course of introducing these new states, we let the observation function for the new states be given by $O^i(\zeta^{s,\mathbf{a},s'}, \mathbf{a})(o^i) = O^i(s', \mathbf{a})(o^i)$ for all $o^i \in \Omega^i$. Additionally, we add an additional null observation $o^i_\varnothing$ to each observation set Ω^i and let $O^i(s', \mathbf{a})(\omega^i) = 1$ if $\omega^i = o^i_\varnothing$ and is 0 otherwise, for all $i \in N$ and $\mathbf{a} \in A$.

Under this construction, suppose that the state of the game is s at some point in time, and the agents select the joint action $\mathbf{a}$. Then, the new game $\mathcal{G}_H$ simulates a sample from the transition function $s' \sim p(s, \mathbf{a})$ but instead transitions

to the intermediate state $\zeta^{s,\mathbf{a},s'}$, in which each agent $i \in N$ receives a reward equal to the terms in the sophisticated divergence objective $H[Q^i(s'', o'^i|s, \mathbf{a})] + \beta^i R^i(s') - \log(Z^i(\beta^i)) + \mathbb{E}_{Q^i(s'',o'^i|s,\mathbf{a})}\left[\log \tilde{P}^i(o'^i|s'')\right]$. Additionally, the agents' observations o^i are sampled according to the same distribution $O^i(s', \mathbf{a})$ that would have been sampled from, had the game transitioned into the state s'. Crucially, the new reward from the state $\zeta^{s,\mathbf{a},s'}$ is not included in any of the observations o^i except potentially information about the original reward $R_i(s')$ that was already included in the observation function. This ensures that the agents have no additional information which would not have been available to them otherwise. After this, the game transitions to the previously sampled state s' deterministically, regardless of the joint action chosen. It is clear that under this transformation, the expected reward of an agent $i \in N$ after two timesteps under $(s, \mathbf{a})$ is $\mathbb{E}_{p(s'|s,\mathbf{a})}\left[H[Q^i(s'', o'^i|s, \mathbf{a})] + \beta^i R^i(s') - \log(Z^i(\beta^i)) + \mathbb{E}_{Q^i(s'',o'^i|s,\mathbf{a})}\left[\log \tilde{P}^i(o'^i|s'')\right]\right]$, which is precisely equal to the KL divergence objective $\mathrm{D}_{\mathrm{KL}}\left[Q^i(s', o'^i|s, \mathbf{a}) \,||\, \tilde{P}^i(s', o'^i)\right]$

Thus, given a NE $\hat{\boldsymbol{\pi}}$ in $\mathcal{G}_H$, we can obtain an independent FEE $\boldsymbol{\pi}$ in $\mathcal{G}$ by simply taking $\pi^i(s, a^i) = \hat{\pi}^i(s, a^i)$ for all $i \in N$, $s \in S$, and $a^i \in A^i$. □

B.1. Special Cases

Before concluding, we briefly discuss some further connections to special cases of the general framework presented here that are commonly studied in the game theory literature. In all of the special cases we study below, decision-making under the sophisticated divergence objective is equivalent to reward maximisation, which is widely adopted as the standard of rationality in these scenarios. This highlights the importance of combining three essential features of the POMDP/POSG model in studying divergence minimisation objectives: temporally-extended decision-making, stochastic environmental dynamics, and partial-observability.

B.1.1. One-Shot Games

In normal-form or one-shot games, the situation is much simpler. Such games can be embedded in a MG with a deterministic initial state with reward 0. Each joint action $\mathbf{a}$ deterministically leads to a state $s^{\mathbf{a}}$ where the rewards are assigned according to the payoff matrix of the normal-form game. Let $\boldsymbol{\mu}$ be the policy profile and let the beliefs Q^i trivially reflect common knowledge of this simple embedding. Then, we have the following:

$$G^i(\boldsymbol{\mu}) = \mathbb{E}_{\boldsymbol{\mu}(\mathbf{a}|s_0)}\left[\mathrm{D}_{\mathrm{KL}}\left[Q^i(s_1|s_0, \mathbf{a}) \,||\, \tilde{P}^i(s_1)\right]\right] \tag{10}$$

$$= -\mathbb{E}_{\boldsymbol{\mu}(\mathbf{a}|s_0)\cdot Q^i(s_1|\mathbf{a},s_0)}\left[-\beta^i R^i(s_1) + \log(Z^i(\beta^i))\right] \tag{11}$$

Since $Z^i(\beta^i)$ is a constant with respect to the policy, we observe the following:

Proposition 9. *For any fixed $\beta = (\beta^1, \ldots, \beta^n)$, where each $0 < \beta^i < \infty$, there is a one-to-one correspondence between CCFEEs in normal-form games and CCEs in the same game. The same relationship holds true between IFEEs and NEs.*

B.1.2. Deterministic Games

Next, we can examine the role that probabilistic transitions play in the objective studied here. This can be achieved by considering deterministic POSGs, in which the state transition probability function is deterministic, i.e., $p(s'|s, \mathbf{a}) \in \{0, 1\}$ for all $s, \mathbf{a}, s' \in S \times A \times S$. Following the same line of reasoning as in Theorem 6, we again obtain convergence of FEE in deterministic POSGs to NE:

Proposition 10. *For a commonly known deterministic POSG $\mathcal{G}$, we have* $\mathrm{CCFEE}(\mathcal{G}) \subseteq \mathrm{CCE}(\mathcal{G})$ *and* $\mathrm{IFEE}(\mathcal{G}) \subseteq \mathrm{NE}(\mathcal{G})$ *for sufficiently large β^i, $i \in N$.*

B.1.3. Fully Observable Games

Due to the computational difficulty of multi-agent learning in the fully general framework of POSGs, much of the MARL literature focuses on either fully observable games, known as Markov (or Stochastic) Games (MGs), or POSGs where all agents are assumed to have a common reward function, known as Decentralised-POMDPs (Dec-POMDPs) (Albrecht et al., 2024). In the fully observed setting, the actions of all agents are observed by every agent, as well as the true state of the game. This can be encoded in our setting by letting $\Omega^i = S \times A$ for all $i \in N$, and setting $O^i(\mathbf{o}_t|s_t, \mathbf{a}_{t-1}) = \delta((s_t, \mathbf{a}_{t-1}))$, where δ is the Dirac delta distribution which assigns a probability of 1 to the observation $\mathbf{o}_t$ that corresponds to the true

current state and previous joint action profile that was played, and assigns a probability of 0 to all other observations.

Written in this way, it becomes clear that the entropy terms over the observation likelihood function in the SDO vanish, because it is a deterministic function. However, the entropy terms over the transition probability function remain in the objective, so agents with finite inverse temperature parameters still retain the desire to 'keep options open', as a result of being biased towards actions that lead to a higher path entropy.

In this setting, a history may now be written as a sequence $\mathbf{h}_{t_0:t} = s_{t_0}\mathbf{a}_{t_0}\dots s_t$. Moreover, an agent i's policy becomes a mapping $\pi_i : \bigcup_{t=1}^{T} S^t \to \Delta(A^i)$ from state trajectories to distributions over actions. Thus, an agent i's predictive model simplifies to the following form:

$$Q^i(\mathbf{h}_{0:t}\,;\,\boldsymbol{\pi}) := Q^i(s_0)\cdot\left(\prod_{\tau=0}^{t-1}\pi^i(a^i_\tau|s^i_{0:\tau})\cdot Q^i(\mathbf{a}^{-i}_\tau|s_{0:\tau})\cdot Q^i(s_{\tau+1}|s_\tau,\mathbf{a}_\tau)\right). \tag{12}$$

Likewise, an agent i's preference model that captures state rewards can be encoded by the following:

$$\tilde{P}^i(s_{0:T}) = \prod_{\tau=0}^{T}\frac{\exp(\beta^i R^i(s_\tau))}{Z^i(\beta^i)}. \tag{13}$$

Following the same approach as before, the SDO in this case is given in its unravelled form by:

$$\begin{aligned} G^i(\boldsymbol{\mu}) =&\ \mathrm{D}_{\mathrm{KL}}\left[Q^i(s_0)\,||\,\tilde{P}^i(s_0)\right] + \mathbb{E}_{Q^i(s_0)}\left[G^i(s_0;\boldsymbol{\mu})\right] &(14)\\ =&\ d_0 + \sum_{\tau=0}^{T-1}\mathbb{E}_{Q^i(s_\tau,\mathbf{a}_\tau|s_{0:\tau-1},\mathbf{a}_{\tau-1};\boldsymbol{\mu})}\left[\mathrm{D}_{\mathrm{KL}}\left[Q^i(s_{\tau+1}|s_\tau,\mathbf{a}_\tau)\,||\,\tilde{P}^i(s_{\tau+1})\right]\right]. &(15)\end{aligned}$$

Using this, we obtain an analogous result to Theorem 6:

Proposition 11. *For a commonly known Markov Game $\mathcal{G}$, we have* $\mathrm{CCFEE}(\mathcal{G}) \subseteq \mathrm{CCE}(\mathcal{G})$ *and* $\mathrm{IFEE}(\mathcal{G}) \subseteq \mathrm{NE}(\mathcal{G})$ *for sufficiently large β^i, $i \in N$.*

Proof. Following the same line of reasoning from Theorem 6, suppose that $\boldsymbol{\mu}$ is a CCFEE. Then, for all $i \in N$ and alternative policies $\hat{\pi}^i \in \Pi^i$, we have

$$\begin{aligned}
&\lim_{\beta^i\to\infty}\ \arg\min_{\hat{\pi}^i\in\Pi^i} G^i((\hat{\pi}^i,\boldsymbol{\mu}^{-i}))\\
=&\lim_{\beta^i\to\infty}\ \arg\min_{\hat{\pi}^i\in\Pi^i}\sum_{\tau=0}^{T-1}\mathbb{E}_{Q^i(s_\tau,\mathbf{a}_\tau|s_{0:\tau-1},\mathbf{a}_{\tau-1};(\hat{\pi}^i,\boldsymbol{\mu}^{-i}))}\left[\mathrm{D}_{\mathrm{KL}}\left[Q^i(s_{\tau+1}|s_\tau,\mathbf{a}_\tau)\,||\,\tilde{P}^i(s_{\tau+1})\right]\right]\\
=&\lim_{\beta^i\to\infty}\ \arg\min_{\hat{\pi}^i\in\Pi^i}\sum_{\tau=0}^{T-1}\mathbb{E}_{Q^i(s_\tau,\mathbf{a}_\tau|s_{0:\tau-1},\mathbf{a}_{\tau-1};(\hat{\pi}^i,\boldsymbol{\mu}^{-i}))}\left[-H\left[Q^i(s_{\tau+1}|s_\tau,\mathbf{a}_\tau)\right]-\mathbb{E}_{Q^i(s_{\tau+1}|s_\tau,\mathbf{a}_\tau)}\left[\log\tilde{P}^i(s_{\tau+1})\right]\right]\\
=&\lim_{\beta^i\to\infty}\ \arg\min_{\hat{\pi}^i\in\Pi^i}\sum_{\tau=0}^{T-1}\mathbb{E}_{Q^i(s_\tau,\mathbf{a}_\tau|s_{0:\tau-1},\mathbf{a}_{\tau-1};(\hat{\pi}^i,\boldsymbol{\mu}^{-i}))}\left[-H\left[Q^i(s_{\tau+1}|s_\tau,\mathbf{a}_\tau)\right]\right.\\
&\qquad\left.+\ \mathbb{E}_{Q^i(s_{\tau+1}|s_\tau,\mathbf{a}_\tau)}\left[-\beta^i R^i(s_{\tau+1})+\log(Z^i(\beta^i))\right]\right]\\
=&\lim_{\beta^i\to\infty}\ \arg\max_{\hat{\pi}^i\in\Pi^i}\sum_{\tau=0}^{T-1}\mathbb{E}_{Q^i(s_\tau,\mathbf{a}_\tau|s_{0:\tau-1},\mathbf{a}_{\tau-1};(\hat{\pi}^i,\boldsymbol{\mu}^{-i}))}\left[H\left[Q^i(s_{\tau+1}|s_\tau,\mathbf{a}_\tau)\right]\right.\\
&\qquad\left.+\ \mathbb{E}_{Q^i(s_{\tau+1}|s_\tau,\mathbf{a}_\tau)}\left[\beta^i R^i(s_{\tau+1})-\log(Z^i(\beta^i))\right]\right]\\
\subseteq&\lim_{\beta^i\to\infty}\ \arg\max_{\hat{\pi}^i\in\Pi^i}\sum_{\tau=0}^{T-1}\mathbb{E}_{Q^i(s_\tau,\mathbf{a}_\tau|s_{0:\tau-1},\mathbf{a}_{\tau-1};(\hat{\pi}^i,\boldsymbol{\mu}^{-i}))}\left[\mathbb{E}_{Q^i(s_{\tau+1}|s_\tau,\mathbf{a}_\tau)}\left[\beta^i R^i(s_{\tau+1})-\log(Z^i(\beta^i))\right]\right]\\
=&\arg\max_{\hat{\pi}^i\in\Pi^i}\sum_{\tau=0}^{T-1}\mathbb{E}_{Q^i(s_\tau,\mathbf{a}_\tau,s_{\tau+1}|s_{0:\tau-1},\mathbf{a}_{\tau-1};(\hat{\pi}^i,\boldsymbol{\mu}^{-i}))}\left[R^i(s_{\tau+1})\right] = \arg\max_{\hat{\pi}^i\in\Pi^i} V^i(\hat{\pi}^i,\boldsymbol{\mu}^{-i}),
\end{aligned}$$

thus obtaining the result. It is straightforward to see that this reasoning also applies to establish the relationship between the IFEEs and NEs of a game. □

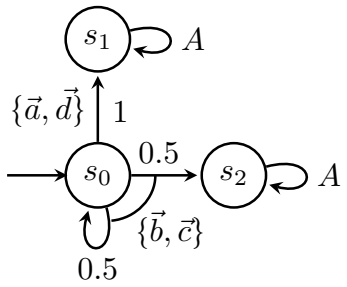

Figure 5. Markov Game illustrating non-uniqueness of the CCFEE and its distinctness from CCEs.

Example 12. *Consider a fully-observable POSG $\mathcal{G}$ (commonly known as a Markov or Stochastic Game) with agents $N = \{1, 2\}$, states $S = \{s_0, s_1, s_2\}$, action sets $Ac^i = \{a, b\}$ for all $i \in N$, transition probability function P and deterministic initial state as illustrated in Figure 5, reward function given by $R^i(s_0) = R^i(s_2) = 1$ and $R^i(s_1) = 2$ for all $i \in N$, and time horizon $T = 1$. Full observability means that at each state of the game, all agents know the state of the game and the joint action profile that was previously played.*

Suppose that in the multi-agent active inference setting, we additionally let $\beta^i = 1$ for all $i \in N$ and $\tilde{P}^i(o'^i|s) = Q^i(o'^i|s)$. Given this simple setting, we can directly compute the EFE associated with each action at the beginning of the game as follows:

$$G^i(\vec{a}|s_0) = G^i(\vec{d}|s_0) = 0 + 1 \cdot 2 - \log(e \cdot (e+2)) + 0 = 2 - \log(e \cdot (e+2))$$
$$G^i(\vec{b}|s_0) = G^i(\vec{c}|s_0) = 1 \cdot 1 + 1 - \log(e \cdot (e+2)) + 0 = 2 - \log(e \cdot (e+2)),$$

Here, we see that the sophisticated divergence objective associated with all joint actions is the same for all agents. Hence, any joint policy in this setting is an IFEE. This illustrates the fact that IFEEs are not necessarily unique.

This example also illustrates the distinctness of the IFEE from the Nash equilibrium solution concept. Observe that from the reward maximisation perspective, only those joint policies that assign zero probability to both $\vec{b}$ and $\vec{c}$ are Nash equilibria. From this, we see that for low enough β^i's, not all IFEEs are Nash equilibria.