Inference of Altruism and Intrinsic Rewards in Multi-Agent Systems

Victor Villin
University of Neuchâtel
Neuchâtel, Switzerland
victor.villin@unine.ch

Christos Dimitrakakis
University of Neuchâtel
Neuchâtel, Switzerland
christos.dimitrakakis@unine.ch

ABSTRACT

Human interactions are influenced by emotions, temperament, and affection, often conflicting with individuals' underlying preferences. Without explicit knowledge of those preferences, judging whether behaviour is appropriate becomes guesswork, leaving us highly prone to misinterpretation. Yet, such understanding is critical if autonomous agents are to collaborate effectively with humans. We frame this as a multi-agent inverse reinforcement learning problem and show that even a simple model, where agents weigh their own welfare against that of others, can cover complex social behaviours. Using novel Bayesian techniques, we find that intrinsic rewards and altruistic tendencies can be reliably identified by placing agents in varied groups. Crucially, this disentanglement of intrinsic motivation from altruism enables the synthesis of new behaviours aligned with any desired level of altruism, even when demonstrations are drawn from restricted behaviour profiles.

KEYWORDS

Multi-Agent Inverse Reinforcement Learning; Altruism; Reward Identification; Bayesian Inference

ACM Reference Format:
Victor Villin and Christos Dimitrakakis. 2026. Inference of Altruism and Intrinsic Rewards in Multi-Agent Systems. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 20 pages. https://doi.org/10.65109/DWDD1205

1 INTRODUCTION

Multi-Agent Inverse Reinforcement Learning (MAIRL) seeks to uncover the hidden reward structures that govern interacting agents. By inferring these latent motivations, we can either interpret observed behaviours [12] or train policies that align with them [19, 20, 50]. Yet, human interactions are rarely straightforward. Assuming that an agent's actions directly reflect their personal rewards is dangerously simplistic: the same behaviour could stem from a desire to help, harm, or manipulate others. Misreading these intentions can lead to fundamentally flawed models of social behaviour.

As Artificial Intelligence (AI) systems become increasingly pervasive, aligning them with human values is no longer optional [15]. AI must not only infer agents' preferences but also evaluate whether these preferences are socially beneficial. Detecting harmful or counterproductive behaviours is essential for designing policies that reliably elevate human welfare.

Reward inference is notoriously challenging in multi-agent systems [16]. Classical approaches assume zero-sum or fully cooperative structures [28]. Real-world interactions, however, are far more nuanced: humans are neither perfectly competitive nor purely cooperative. A chess player may deliberately hold back while coworkers may occasionally act selfishly, deviating from collective benefit. Interactions in society are mostly general-sum.

Among the factors that shape observed preferences, *altruism* plays a central role. Psychologically, altruism is defined as 'any behavior that increases another person's welfare' [3, 35]. The *altruism scale* [44] frames this as a continuum: negative altruism corresponds to antagonistic behaviour, zero to pure self-interest, and positive altruism to prosociality. More precisely, Sawyer [44] interprets altruism as the weight an individual places on the welfare of others relative to their own.

Motivated by this perspective, we study the case where rewards of an agent are a linear combination of its intrinsic rewards and those of others, weighted by a latent altruism level. This gives rise to a two-fold inference problem: (1) recovering the intrinsic rewards that drive each agent's behaviour, and (2) estimating how each agent balances self-interest with concern for others.

Understanding altruism is crucial for interpreting and guiding social behaviour. It explains deviations from cooperative goals, supports team management to encourage prosociality, and helps diagnose misaligned AI agents. We show that disentangling altruism from rewards enables the design of agents that are consistently altruistic, by acting according to the inferred intrinsic rewards of others. Ultimately, this leads to AI systems that are not only effective, but also interpretable, trustworthy, and socially aware.

Contributions. In summary, our main contributions are:

(1) We formally introduce MAIRL with altruism-structured rewards (Section 4), forming a general-sum problem constrained by the fact that agents' rewards are modeled as linear combinations of their intrinsic rewards and those of others.
(2) We analyse identifiability under the altruism scale model (Section 5) and reveal that rewards are generally hard to recover. We prove that leveraging demonstrations from multiple agent groups closes this identifiability gap.
(3) We propose two novel Bayesian methods for learning from multi-group demonstrations (Section 6). The first adapts Bayesian inverse reinforcement learning to multiple agents by constructing a reward posterior assuming Boltzmann rationality. The second, our main contribution, does *not* require such a behavioural

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/DWDD1205

assumption: it first infers a policy posterior from demonstrations, then derives a reward posterior conditioned on the policy posterior.

(4) We conduct extensive experiments on challenging sets of random Markov Games (Section 7). We further consider a practical Overcooked scenario [9], where anti-social chefs must collaborate, and test whether we can synthesise behaviours at new altruism levels. We compare with other state-of-the-art MAIRL methods, namely Multi-Agent Marginal Q-Learning (MAMQL [20]) and Multi-Agent Adversarial IRL (MAAIRL [50]).

(5) We show that our approach can accurately disentangle intrinsic rewards from altruism levels. Interestingly, we demonstrate that, without disentaglement, trying to optimise an agent for co-operation can result in *adversarial* behaviour. In contrast, by leveraging demonstrations from multiple groups, our approach identifies rewards and robustly generalises to behaviours aligned with previously unseen altruism levels.

2 RELATED WORK

IRL and MAIRL aim to infer reward functions that explain observed behaviours, assuming these are near-optimal [32, 34]. Reward identification in IRL is inherently ill-posed: even under entropy regularisation, rewards are only identifiable up to potential-based shaping transformations [33]. Recovering them up to additive constants further requires demonstrations under varying dynamics [6, 8, 42, 45]. In multi-agent settings, the notion of optimality is considerably more intricate. Rather than maximising individual rewards, agents interact strategically, leading to different formulations of MAIRL. Some works simplify the problem by decomposing it into independent single-agent IRL tasks [17, 28], while others explicitly model equilibrium concepts such as Nash equilibria [4, 30, 31, 40]. Despite these advances, theoretical understanding of reward identifiability in MAIRL remains limited. Freihaut and Ramponi [16] characterised the feasible set of reward functions consistent with demonstrations but did not establish identification results. In this work, we prove that we can achieve identification under widely adopted reward assumptions [16, 40], by placing agents in different groups. This is similar to how changing a partner's policy reveals information about the rewards of others [7].

MAIRL has seen successful applications in fully cooperative and zero-sum domains [17, 24, 30], where reward structures are tightly constrained. Now, extending it to general-sum games poses new challenges. While maximum-entropy MAIRL and inverse Q-learning approaches can reproduce expert-like behaviours [20, 50], their learned rewards are not guaranteed to be interpretable or socially consistent. Nevertheless, it has been successfully applied to analyse human driving habits [31], and related efforts have explored theory of mind formulations for inferring others' intentions [12, 49]. We demonstrate however that such methods fail to learn a disentangled understanding of social preferences, and misunderstand agent intrinsic rewards.

Bayesian IRL is a framework for inferring rewards through probabilistic reasoning [38]. Despite its success in single-agent scenarios [13, 14, 43], its application to MAIRL remains underexplored [29, 30]. Common drawbacks to existing MAIRL methods are strong assumption about behaviours, such as strict rationality or a specific amount of entropy regularisation [28, 50]. *We propose* a Bayesian modelling approach that only places a prior on the optimality of the policies [c.f. 14], which we show performs significantly better than approaches making behavioural assumptions.

Altruism plays a fundamental role in decision-making and social interaction. The problem of inferring altruistic behaviour has been explored across both behavioural economics and psychology. In these fields, altruism is typically modeled as a linear concern for others' welfare. Charness and Rabin [11] showed that, under such models, individuals tend to increase collective payoffs even at a personal cost. From a psychological standpoint, Sawyer [44] proposed representing altruism as a weight on others' welfare along a continuum ranging from negative (adversarial) to neutral (egoistic) to positive (altruistic). However, these studies are limited to controlled surveys or stylised economic games. In multi-agent reinforcement learning, altruism has been incorporated in a similar linear form, combining an agent's own reward with those of its teammates to promote cooperation and prosocial behaviour [1, 5, 21, 22, 36]. Related approaches based on reputation dynamics also encourage cooperative behaviour, and can be viewed as a more complex generalisation of altruism: agents prefer to act altruistically toward others with good reputations [2, 41]. In MAIRL, Fukumoto et al. [18] showed that cooperative policies can be induced from selfish demonstrations by augmenting expert data with generated samples, though this approach does not explicitly model altruism.

We study the case where each agent has its own intrinsic level of altruism, independent of the individuals it interacts with. This aligns with prior approaches while extending the continuous altruism scale of Sawyer [44]. To the best of our knowledge, we are the first to employ MAIRL to infer altruism in dynamic multi-agent games and to generate behaviours across the full altruism spectrum.

3 PRELIMINARIES

Rewardless Markov Games. An n-player Rewardless Markov Game (RMG) can be formalised as a tuple $\mathcal{G} = \mathcal{G}(\cdot) = (\mathcal{S}, \boldsymbol{\mathcal{A}}, \mathcal{T}, \gamma, \omega_0, \cdot)$, where $\mathcal{S}$ is a set of states, $\boldsymbol{\mathcal{A}} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$ is a set of discrete actions for each player, $\mathcal{T} : \mathcal{S} \times \boldsymbol{\mathcal{A}} \to \Delta^{\mathcal{S}}$ is a transition function, $\gamma \in [0, 1)$ is a discount factor, and ω_0 an initial state distribution. Without loss of generality, we focus on games where all players share the same action space, i.e. $\boldsymbol{\mathcal{A}} = \mathcal{A}^n$. We label the discrete actions as $\{a^1, \ldots, a^{|\mathcal{A}|}\} = \mathcal{A}$. For clarity, we study player-permutation invariant games, meaning dynamics remain unchanged upon player reordering (e.g. agent A playing with agent B is the same as B playing with A).[1]

Policies. A policy $\pi : \mathcal{S} \to \Delta^{\mathcal{A}}$ is a probability distribution over a single agent's actions. A joint policy is given by $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n) = (\pi_i, \boldsymbol{\pi}_{-i})$, where $\boldsymbol{\pi}_{-i}$ refers to the joint policy of all policies except policy i. We denote the set of policies by Π. We assume actions are conditionally independent, so that the probability of the joint action $\mathbf{a} \in \mathcal{A}^m$ at state s under the joint policy $\boldsymbol{\pi}$ is

[1] Player-permutation invariant games are not restrictive. Many real-world social interactions (e.g., trading in markets, sports such as football) are inherently symmetric in dynamics. Here, no structural advantage or disadvantage arises purely from the dynamics, but rather from the agents' preferences.

$\boldsymbol{\pi}(\mathbf{a} \mid s) = \prod_i \pi_i(a_i \mid s)$. We denote by $\mathcal{T}^{\boldsymbol{\pi}} = \mathbb{E}_{\boldsymbol{\pi}}[\mathcal{T}]$ transitions induced by the joint policy, and $\mathcal{T}_a^{\boldsymbol{\pi}_{-i}}(s) = \mathbb{E}_{\mathbf{a}_{-i} \sim \boldsymbol{\pi}_{-i}}[\mathcal{T}(\cdot \mid s, a, \mathbf{a}_{-i})]$ transitions induced when agent i picks action a.

Rewards and regularised values. The reward function of an agent i, $R_i : \mathcal{S} \times A \to \mathbb{R}$, outputs a bounded real value given a state and a joint action. We write the joint reward function as $\mathbf{R} = (R_i)_i$. Given a reward function $\mathbf{R}$, a general-sum Markov game (MG) is defined as $\mathcal{G}(\mathbf{R})$. For a joint policy $\boldsymbol{\pi}$ on $\mathcal{G}(\mathbf{R})$, the entropy-regularised value function of agent i is

$$V_i^{\boldsymbol{\pi}}(s) \coloneqq \mathbb{E}_{\mathcal{G}}^{\boldsymbol{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t \left(R_i(s_t, \mathbf{a}_t) + \frac{1}{\beta}\mathcal{H}\left(\pi_i(\cdot \mid s_t)\right)\right) \,\middle|\, s_0 = s\right],$$

where the expectation is over game transitions and the joint policy, β is the entropy coefficient, $\mathcal{H}(\pi) = -\sum_a \pi(a) \log \pi(a)$ the entropy of π. The corresponding Q-function is

$$Q_i^{\boldsymbol{\pi}}(s, \mathbf{a}) \coloneqq R_i(s, a_i) + \gamma \sum_{s'} \mathcal{T}(s' \mid s, \mathbf{a}) V_i^{\boldsymbol{\pi}}(s').$$

We further denote $\bar{Q}_i^{\boldsymbol{\pi}}(s, a) = \sum_{\mathbf{a}_{-i}} \boldsymbol{\pi}_{-i}(\mathbf{a}_{-i} \mid s) Q_i^{\boldsymbol{\pi}}(s, \mathbf{a})$ the expected Q-value over the other agents' policies.

Agents and groups. An agent i is an individual decision-maker with its own reward function R_i, and an entropy parameter β_i controlling its stochasticity. For simplicity, we assume all agents share the same entropy parameter $\beta_i = \beta$. A group $\mathbf{g}$ is a subset of agents of fixed size n, i.e. $\mathbf{g} \subseteq \{1, \dots, m\}$, where m is the number of available agents. When a group $\mathbf{g}$ plays a game, we define its joint policy $\boldsymbol{\pi}_{\mathbf{g}} = (\pi_{\mathbf{g},i})_{i \in \mathbf{g}}$ and joint reward $\mathbf{R}_{\mathbf{g}} = (R_{\mathbf{g},i})_{i \in \mathbf{g}}$, where the group subscript on $\pi_{\mathbf{g},i}$ and $R_{\mathbf{g},i}$ indicates both policies adopted and effective rewards are group dependent.

Optimality. In single-agent settings, a policy is commonly defined optimal if it maximises expected cumulative rewards. In multi-agent games, however, this notion becomes insufficient, since each agent's objective depends on the strategies of others. A more meaningful solution concept is the Quantal Response Equilibrium (QRE), an entropy-regularised equivalent of the Nash Equilibrium. Formally, agents play a QRE $\boldsymbol{\pi}^*$ if no agent can unilaterally improve their regularised value by deviating, assuming the other agents' policies remain fixed. That is, for every state s:

$$V_i^{\boldsymbol{\pi}^*}(s) \geq V_i^{\{\pi_i\} \cup \boldsymbol{\pi}^*_{-i}}(s), \quad \forall \pi_i \in \Pi. \tag{1}$$

Suboptimality can manifest in two ways: (1) as the *distance from a QRE*, reflecting that some agents could improve their regularised value by adjusting their policies, and (2) through *higher stochasticity* [10, 25]. QRE policies are optimal under entropy-regularisation, while in the limit $\beta \to \infty$, they approach standard Nash equilibrium, corresponding to 'raw' optimality.

4 PROBLEM FORMULATION

We study a group $\mathbf{g}$ of n agents interacting in a Markov game $\mathcal{G}$. We are given a set of demonstrations

$$\mathcal{D}_{\mathbf{g}} = \{\tau_k\}_k, \quad \tau_k = (s_0^k, \mathbf{a}_0^k, s_1^k, \mathbf{a}_1^k, \dots),$$

where each trajectory τ_k is generated by the agents following the joint policy $\boldsymbol{\pi}_{\mathbf{g}}$. Each agent $i \in \mathbf{g}$ is assumed to act near-optimally with respect to their unknown, individual reward function R_i.

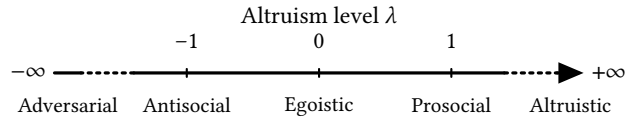


Figure 1: The altruism scale model. We highlight three key values of λ. -1 : agent values its own welfare as much as it harms others. 0 : agent ignores others' welfare. 1 : agent values its own and others' welfare equally.

4.1 Altruism-Structured Rewards

To capture the social nature of agent interactions, we assume that each agent's reward depends not only on its own intrinsic preferences but also on the outcomes experienced by other agents [44].

Assumption 4.1 (Altruism-structured rewards). *Each agent i has an* intrinsic reward function $r_i : \mathcal{S} \times \mathcal{A}^n \to [r_{min}, r_{max}]$, *and an* altruism level $\lambda_i \in [\lambda_{\min}, \lambda_{\max}]$. *The effective reward of agent i when acting within group $\mathbf{g}$ is*

$$R_{\mathbf{g},i}(s, \mathbf{a}) \;=\; r_i(s, a_i) \;+\; \frac{\lambda_i}{n-1} \sum_{\substack{k \in \mathbf{g} \\ k \neq i}} r_k(s, a_k).$$

This simple yet expressive formulation captures the trade-off between self-interest and concern for others. Positive values of λ_i correspond to altruistic behaviour, $\lambda_i = 0$ to purely selfish agents, and negative values reflect adversarial tendencies. Figure 1 illustrates the continuum of behaviours induced by different levels of altruism. The model makes the following simplifying assumptions:

(1) Agents have access to each other's intrinsic rewards, which is reasonable whenever familiar individuals interact. This assumption is important, because otherwise agent behaviour would change as they learned more about each other.
(2) An agent's altruism level is independent of who it interacts with. This allows us to capture social tendencies. For example, individuals can be globally pro-social. While this assumption can be relaxed, it allows us to keep the setting simple.
(3) Each agent's intrinsic reward only depends on that agent's action and the state. This is not a restrictive assumption. While rewards could depend on joint actions, this would not result in a richer setting, as any interacting terms can be modelled as part of the joint state.

While in general, altruism-structured rewards induce general-sum games, we observe that in the specific cases where altruism levels are either all equal to -1 or $n-1$, we obtain zero-sum and fully cooperative game dynamics, respectively. Denoting $\mathbf{z}$ and $\mathbf{c}$ groups of antisocial and cooperative agents respectively, we have :

$$\sum_{i \in \mathbf{z}} R_{\mathbf{z},i}(s, \mathbf{a}) = 0, \quad \text{and} \quad R_{\mathbf{c},1}(s, \mathbf{a}) = \cdots = R_{\mathbf{c},n}(s, \mathbf{a}).$$

This insight indicates that altruism generalises from structured games to a wide variety of general sum games. This implies that our results cover a wide range of type of games.

4.2 Objective

Our goal is to interpret observed social behaviour. We therefore wish to recover each agent's ground-truth reward function R_i from the demonstrations $\mathcal{D}_{\mathbf{g}}$, under the assumption that these rewards are altruism-structured. Formally, we aim to disentangle each agent's reward into its intrinsic component and altruism level $R_i = (r_i, \lambda_i)$. By recovering these components, we obtain a concise and interpretable description of each agent's preferences and social tendencies, which generalises across different interaction contexts.

5 IDENTIFYING ALTRUISM AND INTRINSIC REWARDS

Before diving into reward inference, we must ask: when agents are altruistic, can we disentangle intrinsic rewards from social incentives? Altruism entangles agents' motivations, possibly complicating inference. We show that even when altruism levels are known, reward ambiguity persists unless agents are observed in multiple interaction groups. Interestingly, hiding altruism adds no further ambiguity under the same conditions. Proofs are given in Appendix A.

5.1 QRE Policies

To reason formally about reward identifiability, we first examine the structure of QRE policies under entropy-regularised reinforcement learning. As noted in Section 3, entropy-regularisation admits QREs as equilibria, which, unlike general-sum Nash equilibria, are unique for a given reward function [26]. This uniqueness simplifies the problem of tracking feasible rewards, as it removes ambiguity over which equilibrium is observed [16]. Formally, QRE policies have the following characterisation:

$$\pi_i^*(s,a) \propto \exp\left(\beta \bar{Q}_i^*(s,a)\right). \tag{2}$$

Note that the setting introduces entropy parameter β into the inference process. For the purpose of the analysis presented in this section, we will assume β is known. We later show in Section 6 that this poses no fundamental obstacle: we can infer over the entropy parameter using a suitable prior. Later, we will remove the QRE assumption completely.

5.2 Known Altruism

We first consider the case where altruism levels are known.

Proposition 1. *Assume we observe a QRE $\boldsymbol{\pi}^*$ for the game $\mathcal{G}(\mathbf{R})$, and that we know the altruism levels of agents. Then, intrinsic rewards are identifiable up to potential shaping transformations $\tilde{r}_i(s,a) = r_i(s,a) + \delta r_i(s,a)$, with $\phi : \mathcal{S} \to \mathbb{R}$ and*

$$\delta r_i(s, a_i) = \gamma \sum_{s'} \mathcal{T}_{a_i}^{\boldsymbol{\pi}^*_{-i}}(s')\phi(s') - \phi(s).$$

Even with known altruism, intrinsic rewards remain difficult to identify, mirroring the same ambiguities encountered in single-agent IRL [33]. To reduce the remaining ambiguity, one approach in single-agent IRL is to observe multiple environments [8]. In our multi-agent setting, we recycle the idea by constructing different transition dynamics through *agent groups*. For example, in a 2-player game with 3 agents, we can observe up to three distinct equilibria instead of just one. These multiple equilibria can then be exploited for identification: if a candidate reward explains one group's behaviour but not another's, it can be ruled out. This idea, observing agents across diverse groups to constrain feasible rewards, is novel in MAIRL. Specifically, by observing an agent in two groups that induce sufficiently different transition dynamics, we can reduce its reward ambiguity up to state specific shifts.

Corollary 1. *Let $\mathbf{g}$ and $\mathbf{g}'$ be two distinct groups containing agent i, and assume we observe QREs $\boldsymbol{\pi}^*_{\mathbf{g}}$ and $\boldsymbol{\pi}^*_{\mathbf{g}'}$ with known altruism. Then, the intrinsic rewards of agent i, can be recovered up to some non-trivial state-dependent shifts $\tilde{r}_i(s,a) = r_i(s,a) + \delta r_i(s)$, if and only if the rank condition*

$$\operatorname{rank}\begin{pmatrix} I - \gamma \mathcal{T}_{a^1}^{\boldsymbol{\pi}^*_{\mathbf{g}-i}} & I - \gamma \mathcal{T}_{a^1}^{\boldsymbol{\pi}^*_{\mathbf{g}'-i}} \\ \vdots & \vdots \\ I - \gamma \mathcal{T}_{a^{|\mathcal{A}|}}^{\boldsymbol{\pi}^*_{\mathbf{g}-i}} & I - \mathcal{T}_{a^{|\mathcal{A}|}}^{\boldsymbol{\pi}^*_{\mathbf{g}'-i}} \end{pmatrix} = 2|\mathcal{S}| - 1$$

holds. If $\lambda_i = 0$, the intrinsic rewards can be identified up to a constant.

Intuitively, the condition in Corollary 1 requires the two groups to generate sufficiently distinct dynamics, that is, agents must behave meaningfully differently across the two groups. Even with two induced dynamics, some ambiguity may persist when an agent is not purely egoistic ($\lambda_i \neq 0$), because its consideration for others' intrinsic rewards introduces additional degrees of freedom. Intuitively, the intrinsic rewards of other agents must also be pinned down to fully resolve the ambiguity.

Corollary 2. *Under the conditions of Corollary 1, if each agent is observed in two groups and the rank condition of Corollary 1 holds for every pair, the intrinsic rewards can be recovered up to a constant.*

Importantly, forming multiple groups to meet Corollary 2's requirement is surprisingly not expensive. By introducing one extra agent on top of the original agents, we can form three groups (e.g., $\{1,\ldots,n\}$, $\{2,\ldots,n+1\}$, and $\{1,3,4,\ldots,n+1\}$) such that every agent is observed at least twice. If, for each i agent, the rank condition of Corollary 1 is verified (using two groups in which i appears), then Corollary 2 holds for all $n+1$ agents. With this, intrinsic rewards can be reliably recovered, allowing us to proceed to our main theoretical result on disentangling *latent* altruism.

5.3 Latent Altruism

We now tackle the case where altruism is latent. Intuition suggests that unknown altruism makes inference much harder, but remarkably, simply observing agents across sufficiently diverse groups is enough to fully disentangle altruism from intrinsic rewards.

Theorem 1. *Let us have a set of n+1 agents, and assume we observe enough QREs such that we observed each agent act in two separate groups $(\mathbf{g}_i, \mathbf{g}'_i)$. Then, if and only if the rank condition from Corollary 1 is verified on every pair $(\mathbf{g}_i, \mathbf{g}'_i)$, the altruism levels can be perfectly disentangled from intrinsic rewards for all agents. Specifically, the altruism levels can be exactly identified, and the intrinsic rewards can be recovered up to a constant.*

Intuition. To see why Theorem 1 holds, consider perturbing agent i's altruism: $\tilde{\lambda}_i = \lambda_i + \delta\lambda_i$. In principle, this change could be compensated by adjusting i's intrinsic rewards, the intrinsic

rewards of other agents, or the value function. However, the rank condition ensures that the value function can only absorb *constant shifts* across states, while the effect of $\delta\lambda_i$ depends on the other agents' rewards in a state-dependent way, so it cannot be canceled by the value function.

Next, observing agent i in multiple groups with different compositions rules out compensating via the intrinsic rewards of others: the perturbation interacts differently in each group. Likewise, agent i's intrinsic reward cannot simultaneously offset $\delta\lambda_i$ across groups. Together, these constraints uniquely pin down λ_i, allowing us to disentangle altruism from intrinsic rewards before recovering the latter with Corollary 2.

While richer characterisations of transition dynamics exist [45], verifying the rank condition in practice is challenging, especially with sub-optimal demonstrations or unknown β, and we cannot freely choose agents to satisfy it. As multiple groups generally reduce ambiguity, we suggest adopting a more practical approach using Bayesian inference on groups of the agents available.

6 TWO BAYESIAN METHODS FOR INFERRING ALTRUISM AND INTRINSIC REWARDS.

We now turn to the practical question of how to infer intrinsic rewards and altruism from observed behaviour. Crucially, we introduce methods that can ingest demonstrations collected from *multiple groups*, leveraging results from Section 5.

We adopt a *Bayesian framework*, which offers a principled way to represent uncertainty over rewards [38]. Given a prior $\mathbb{P}(\cdot)$ over reward functions, Bayesian MAIRL/IRL seeks the posterior $\mathbb{P}(\cdot \mid \mathcal{D})$ conditioned on demonstrations $\mathcal{D}$. We perform posterior sampling using Stochastic Gradient Langevin Dynamics (SGLD) [48], a gradient-informed sampler well-suited for efficiently exploring complex posteriors.[2] We propose two complementary Bayesian approaches: (1) extending standard Bayesian IRL [6, 7, 23] to the multi-agent setting with altruistic rewards, by evaluating the likelihood of observed demonstrations under QRE policies. (2) Our main contribution, which (similarly to [14]) first infers a posterior over policies given demonstrations, and then calculates reward posterior by marginalising over the inferred policies.

In the following, we consider demonstrations from multiple groups. We denote $\mathcal{X} = \bigcup_{\mathrm{g}} \mathcal{D}_{\mathrm{g}}$ the full demonstration set, and $\mathcal{R} = (R_i)_{i=1}^m = ((r_i, \lambda_i))_{i=1}^m$ the unknown reward vector where m is the total number of agents observed. Altruism is assumed independent of intrinsic rewards, and demonstrations are independent across groups. Figure 2 illustrates the overall inference process.

6.1 Direct Reward Posterior (DRP)

One way to write the reward posterior over rewards is to directly model the likelihood of the demonstrations under those rewards, i.e., $\mathbb{P}(\mathcal{X} \mid \mathcal{R}) = \prod_{\tau \in \mathcal{X}} \mathbb{P}(\tau \mid \mathcal{R})$. Since demonstrations are independent across groups, we can rearrange the product accordingly:

$$\mathbb{P}(\mathcal{X} \mid \mathcal{R}) = \prod_{\mathrm{g}} \prod_{\tau \in \mathcal{D}_{\mathrm{g}}} \mathbb{P}(\tau \mid \mathbf{R}_{\mathrm{g}}).$$

[2] Additional details on this choice and a description of the algorithmic implementation are provided in Appendix B.

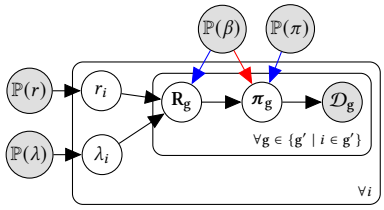


Figure 2: Graphical model of policy and altruism-structured rewards across groups. Latent nodes are white. Priors $\mathbb{P}(r)$ and $\mathbb{P}(\lambda)$ generate agent rewards r_i and altruism λ_i, which determine group rewards $\mathbf{R}_{\mathrm{g}}$ and policies π_{g}, producing demonstrations $\mathcal{D}_{\mathrm{g}}$. Priors $\mathbb{P}(\pi)$ and $\mathbb{P}(\beta)$ also participate in generating rewards and policies. Coloured edges indicate how and which prior is used (red is for DRP, blue for PORP).

This formulation is conceptually similar to that of Buening et al. [6], though their focus is on multiple environments, whereas we consider different agent groups. Somehow, we need to specify the likelihood $\mathbb{P}(\tau \mid \mathbf{R})$, modelling how agents generate trajectories based on rewards. We propose to model demonstrated policies as QRE. In single-agent settings, it reduces to the standard Boltzmann form used in IRL. We therefore assume all agents adopt QRE policies, and denote by $\boldsymbol{\pi}^*_{\mathbf{R}}(\,\cdot\,;\beta)$ the QRE policy under joint reward $\mathbf{R}$ with entropy parameter/optimality β. Given a prior belief $\mathbb{P}(\beta)$ over β, we can marginalise it out, and obtain the trajectory likelihood:

$$\mathbb{P}(\tau \mid \mathbf{R}) = \prod_{(s,\mathbf{a}) \in \tau} \int_{\beta_{\min}}^{\infty} \boldsymbol{\pi}^*_{\mathbf{R}}(\mathbf{a} \mid s;\beta) \mathrm{d}\mathbb{P}(\beta),$$

with $\beta_{\min} > 0$. Substituting this expression back into the demonstration likelihood, Bayes'rule yields the full DRP:

$$\mathbb{P}(\mathcal{R} \mid \mathcal{X}) \propto \mathbb{P}(\mathcal{R}) \prod_{\mathrm{g}} \prod_{\tau \in \mathcal{D}_{\mathrm{g}}} \prod_{(s,\mathbf{a}) \in \tau} \int_{\beta_{\min}}^{\infty} \boldsymbol{\pi}^*_{\mathbf{R}_{\mathrm{g}}}(\mathbf{a} \mid s;\beta) \mathrm{d}\mathbb{P}(\beta). \quad (3)$$

where the prior over rewards can be expressed as a product of priors over altruism and intrinsic rewards $\mathbb{P}(\mathcal{R}) = \prod_i^m \mathbb{P}(r_i)\mathbb{P}(\lambda_i)$. We next propose a novel reward posterior that bypasses the need to model policies as QREs.

6.2 Policy-Oriented Reward Posterior (PORP)

DRP evaluates trajectory likelihoods assuming agents adopt QRE policies. It is both a strong assumption, and intractable as computing QREs grows expensive. In contrast, in this section we place a prior over policies, and infer a posterior over policies from the demonstrations. This allows us to simply use a prior on the *suboptimality* of the policies demonstrated, rather than assume a behavioural model such as QRE. Within any group, we can express the posterior over rewards by marginalising over joint policies:

$$\mathbb{P}(\mathbf{R} \mid \mathcal{D}) = \int_{\Pi^n} \mathbb{P}(\mathbf{R} \mid \boldsymbol{\pi}) \mathrm{d}\mathbb{P}(\boldsymbol{\pi} \mid \mathcal{D}), \quad (4)$$

assuming that rewards are conditionally independent of demonstrations given policies $\mathbb{P}(\mathbf{R} \mid \boldsymbol{\pi}, \mathcal{D}) = \mathbb{P}(\mathbf{R} \mid \boldsymbol{\pi})$. This requires us to specify how likely a reward is, given a policy. Intuitively, a reward is more plausible if the policy is near-optimal under it. We formalise this via a *gap function* $\Delta_{\mathbf{R},\beta} : \Pi \to \mathbb{R}^+$, which measures

the suboptimality of policy $\boldsymbol{\pi}$ under joint reward $\mathbf{R}$ for a specific entropy parameter β. The likelihood of $\mathbf{R}$ given $\boldsymbol{\pi}$ and β is then:

$$\mathbb{P}(\mathbf{R} \mid \boldsymbol{\pi}, \beta) = \frac{1}{Z_{\boldsymbol{\pi},\beta}} \mathbb{P}(\mathbf{R}) \cdot e^{-c\Delta_{\mathbf{R},\beta}(\boldsymbol{\pi})}, \tag{5}$$

where $Z_{\boldsymbol{\pi},\beta}$ is a partition function, and where c controls how strictly we believe agents are near-optimal. Plugging this into (4) and marginalising over β, we obtain:

$$\mathbb{P}(\mathbf{R} \mid \mathcal{D}) = \int_{\Pi^n} \frac{1}{Z_{\boldsymbol{\pi}}} \int_{\beta_{\min}}^{\infty} \mathbb{P}(\mathbf{R}) \cdot e^{-c\Delta_{\mathbf{R},\beta}(\boldsymbol{\pi})} \mathrm{d}\mathbb{P}(\beta) \mathrm{d}\mathbb{P}(\boldsymbol{\pi} \mid \mathcal{D}),$$

Essentially, this posterior says a reward is more plausible if the observed demonstrations could have come from policies that are nearly optimal under it, averaged over all plausible policies and levels of stochasticity across groups. A practical difficulty here is that $Z_{\boldsymbol{\pi}}$ varies with each policy, and thus cannot be pulled outside of the integral over policies. However, we show that using an appropriate gap function, we get $Z_{\boldsymbol{\pi}} \approx Z$ across policies, partly because the posterior over policies $\mathbb{P}(\boldsymbol{\pi} \mid \mathcal{D})$ tends to be sharply peaked, especially when enough demonstration data is available.[3] This makes the following approximation reasonable:

$$\mathbb{P}(\mathcal{R} \mid \mathcal{X}) \propto\!\!\!\sim \prod_{\mathbf{g}} \int_{\Pi^n} \int_{\beta_{\min}}^{\infty} \mathbb{P}(\mathbf{R}_{\mathbf{g}}) \cdot e^{-c\Delta_{\mathbf{R}_{\mathbf{g}},\beta}(\boldsymbol{\pi})} \mathrm{d}\mathbb{P}(\beta) \mathrm{d}\mathbb{P}(\boldsymbol{\pi} \mid \mathcal{D}_{\mathbf{g}}). \tag{6}$$

This posterior is easy to compute, as we show in the next paragraph. Thereafter, we only need to choose an appropriate gap function, which we address in the rest of this section.

Two-step posterior sampling. We now outline a practical procedure to sample from the PORP (6). This proceeds in two steps:

(1) Given that the likelihood of a trajectory under a policy is simply the probability of the policy generating it, we have: $\mathbb{P}(\boldsymbol{\pi} \mid \tau) \propto \mathbb{P}(\boldsymbol{\pi}) \prod_{(s,\mathbf{a})\in\tau} \boldsymbol{\pi}(\mathbf{a} \mid s)$. Factoring over a group's dataset:

$$\mathbb{P}(\boldsymbol{\pi} \mid \mathcal{D}) \propto \prod_{\tau\in\mathcal{D}} \mathbb{P}(\boldsymbol{\pi}) \prod_{(s,\mathbf{a})\in\tau} \boldsymbol{\pi}(\mathbf{a} \mid s). \tag{7}$$

In the first step, we draw N samples $\hat{\boldsymbol{\pi}}_{\mathbf{g},1}, \ldots, \hat{\boldsymbol{\pi}}_{\mathbf{g},N}$ from this posterior for each group $\mathbf{g}$.

(2) We obtain the reward posterior (6) via Monte Carlo integration:

$$\mathbb{P}(\mathcal{R} \mid \mathcal{X}) \propto\!\!\!\sim \prod_{\mathbf{g}} \sum_{k=1}^{N} \int_{\beta_{\min}}^{\infty} \mathbb{P}(\mathbf{R}_{\mathbf{g}}) \cdot e^{-c\Delta_{\mathbf{R}_{\mathbf{g}},\beta}(\hat{\boldsymbol{\pi}}_{\mathbf{g},k})} \mathrm{d}\mathbb{P}(\beta). \tag{8}$$

We are then left with sampling from that approximate posterior to generate reward candidates.

This two-step approach enables efficient approximation of the posterior over rewards: first by inferring plausible policies from data, and then by weighting rewards according to how well these policies align with optimal behaviour under each reward hypothesis.

Gap functions. Gap functions are constructed such that a value of zero indicates an equilibrium, while larger values reflect deviations from optimality. The Nash Imitation Gap [39], recently standardised in the MAIRL setting [16], satisfies this property. It measures the maximum incentive any agent has to unilaterally deviate from the current policy. A gap of zero implies that no agent can benefit from deviating, thus identifying a Nash equilibrium. We extend this to the entropy-regularised setting.

Definition 1 (QRE Imitation Gap (QIG)). *The QIG is given by:*

$$\Delta^{QIG}_{\mathbf{R},\beta}(\boldsymbol{\pi}) \coloneqq \max_{i\in[n]} \max_{\pi_i\in\Pi} \sum_{s\in\mathcal{S}} V_i^{\{\pi_i\}\cup\boldsymbol{\pi}_{-i}}(s) - V_i^{\boldsymbol{\pi}}(s), \tag{9}$$

where the values are computed on $\mathbf{R}$ *and* β.

Compared to DRP, the QIG requires computing best responses instead of full equilibria. However, as the number of players grows, even the QIG can become computationally prohibitive. To address this, we introduce the Policy Stability Gap (PSG), which exploits the known structure of optimal entropy-regularised policies.

Definition 2 (Policy Stability Gap (PSG)). *Let* D_{KL} *be the Kullback-Leibler (KL) divergence. Furthermore, let the 'soft response' of the* i*-th agent to the joint policy* $\boldsymbol{\pi}_{-i}$*, under rewards* $\mathbf{R}$ *and regularisation parameter* β*, be:*

$$\sigma^{\boldsymbol{\pi}_{-i}}_{\mathbf{R}}(a \mid s; \beta) \coloneqq \frac{\exp\left(\beta \tilde{Q}_i^{\boldsymbol{\pi}}(s,a)\right)}{\sum_{a'} \beta \exp\left(\beta \tilde{Q}_i^{\boldsymbol{\pi}}(s,a')\right)}$$

where the Q-values are computed on $\mathbf{R}$ *and* β*. We define the PSG as:*

$$\Delta^{PSG}_{\mathbf{R},\beta}(\boldsymbol{\pi}) \coloneqq \max_{i\in[n]} \sum_{s\in\mathcal{S}} D_{KL}\Big(\pi_i(\cdot \mid s) \,\Big\|\, \sigma^{\boldsymbol{\pi}_{-i}}_{\mathbf{R}}(\cdot \mid s; \beta)\Big). \tag{10}$$

The PSG measures how far a joint policy deviates from its soft response. A QRE policy inherently has a PSG of 0 by definition (Eq.(2)). Crucially, PSG is computationally efficient, requiring only the evaluation of Q for the current joint policy.

It is unclear whether these gap functions truly capture human irrationality. Nonetheless, they naturally capture suboptimality: small gaps indicate that agents act largely rationally but occasionally make mistakes. By measuring deviations from best (QIG) or soft (PSG) responses, we can infer rewards from demonstrations *without* assuming perfect QRE behaviour.

6.3 Priors

To enable proper and stable Bayesian inference, we need to specify the priors. We place independent Gaussian priors on intrinsic rewards and policies. The reward prior is centered at zero, reflecting the assumption that states with little information carry no inherent reward, an assumption that is generally safer than positing there are incentives in unreachable states. Similarly, the policy prior is centered around uniform action probabilities, capturing the idea that, in the absence of evidence, agents behave without clear preference. For altruism, we use a uniform prior to enable exploration across the entire altruism spectrum, though more informative choices (e.g., a Gaussian centered around egoism) could be employed when empirical data are available. Finally, we assign an exponential prior to the entropy parameter β, assuming demonstrations are mostly low-entropy, reflecting confident yet occasionally inconsistent behaviour. Further details are provided in Appendix D.

7 EXPERIMENTS

Our experiments aim to answer two key questions: (1) Can we *disentangle altruism from intrinsic rewards* by observing agents in

[3] Both a theoretical and empirical study on the near constant nature of $Z_{\boldsymbol{\pi}}$ is provided in Appendix C.

different groups? (2) Can we synthesise behaviours corresponding to *any level of altruism* by doing so ?

To address the first question, Section 7.1 studies challenging randomised MGs, providing a robust benchmark and enabling ablation studies on using groups of agents for demonstrations. To answer the second question, Section 7.2 considers a scenario where conflicting kitchen employees must work together, testing whether we can generate altruistic behaviours from inherently anti-social agents. We compare our approach against state-of-the-art maximum entropy and inverse Q-learning methods, MAAIRL [46] and MAMQL [20], which we adapt to use an altruism-structured reward model. These baselines, however, are not extended to multiple-group inference. To ensure a fair comparison, all methods are provided with the same total number of demonstrations. For methods using multiple groups, the demonstration budget is evenly split across all available groups (algorithms do not choose which group they use). Trajectories have a fixed length of 1000 timesteps, and all metrics are computed with respect to the first group, which is shared across methods. Demonstrations are generated from QRE policies using a hidden entropy parameter β. For MAMQL and MAAIRL, which do not explicitly account for the uncertainty in β, we provide the true value. For all of our experiments, we set $\lambda_{\max} = -\lambda_{\min} = 5$, $r_{\max} = 1$ and $r_{\min} = 0$. All errors are reported as mean squared errors and standard errors, rescaled such that 1 corresponds to the expected error of random guessing. Note that our experiments do not attempt to address MAIRL scalability in general. Instead, we focus on the largest environments that can be handled without approximations such as neural networks, leaving broader scalability questions to future work. Additional results and experimental details, including priors, are provided in Appendices C and D.

7.1 Disentangling Altruism from Intrinsic Rewards

In this experiment, we validate DRP and PORP on randomised MGs.

Experimental setup. We consider two sets of randomised MGs:

(1) A 3-player, 512-state, 5-action set, which serves as our main benchmark domain. Here, we provide 4 agents, allowing methods to infer from 4 different combinations of agents. We provide algorithms with a budget of 200 trajectories.
(2) A 4-player, 16-state, 5-action set, with 6 agents, to study the benefits of using a subset or all possible groups. We also evaluate methods on an increasingly large budget for demonstrations.

Transition functions are sampled from a Dirichlet distribution. Each agent's altruism parameter is sampled uniformly, while intrinsic rewards are defined by randomly selecting a subset of state-action pairs with reward 1, with all other pairs set to 0. Those are naturally hidden from algorithms. Metrics are averaged over 10 seeds.

Results. Table 1 and Figure 3 both demonstrate that PORP-PSG is the most reliable method for disentangling altruism from intrinsic rewards, consistently achieving much lower errors than all other approaches. Figure 3 further shows that inferring altruism and intrinsic rewards is extremely difficult without multiple groups. Introducing as few as 5 groups substantially reduces ambiguity. When increasing the number of groups, there appears to be a trade-off between the information gained from additional groups and the information lost due to fewer demonstrations per group. DRP performs surprisingly well in 3-player games using a single group (Table 1) but exhibits limited robustness in 4-player environments (Figure 3). Its performance declines as trajectory budget increases, likely because fitting QREs becomes harder with larger datasets. Finally, using the QIG gap function improve over basic baselines but remain far behind PSG. Overall, PORP with PSG emerges as the most promising method. It bears low computational cost and has the highest sample efficiency with groups, which is particularly valuable when demonstrations are costly. Additional results on randomised MGs are provided in Appendix C.

Table 1: Altruism and intrinsic reward recovery errors on Random MGs. Lower is better.

Method	Altruism Error	Intrinsic Rewards Error ($\times 10^3$)
PORP-PSG	**0.024 ± 0.003**	**0.065 ± 0.004**
PORP-QIG	0.392 ± 0.156	0.338 ± 0.081
DRP	0.040 ± 0.007	0.091 ± 0.005
MAMQL	0.855 ± 0.156	1.526 ± 0.041
MAAIRL	3.313 ± 0.435	1.555 ± 0.042
Ablations (without groups)		
PORP-PSG	0.352 ± 0.133	0.474 ± 0.134
PORP-QIG	1.431 ± 0.118	1.003 ± 0.062
DRP	0.059 ± 0.008	0.120 ± 0.007

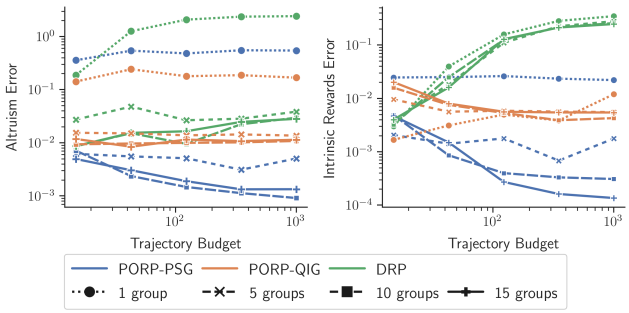
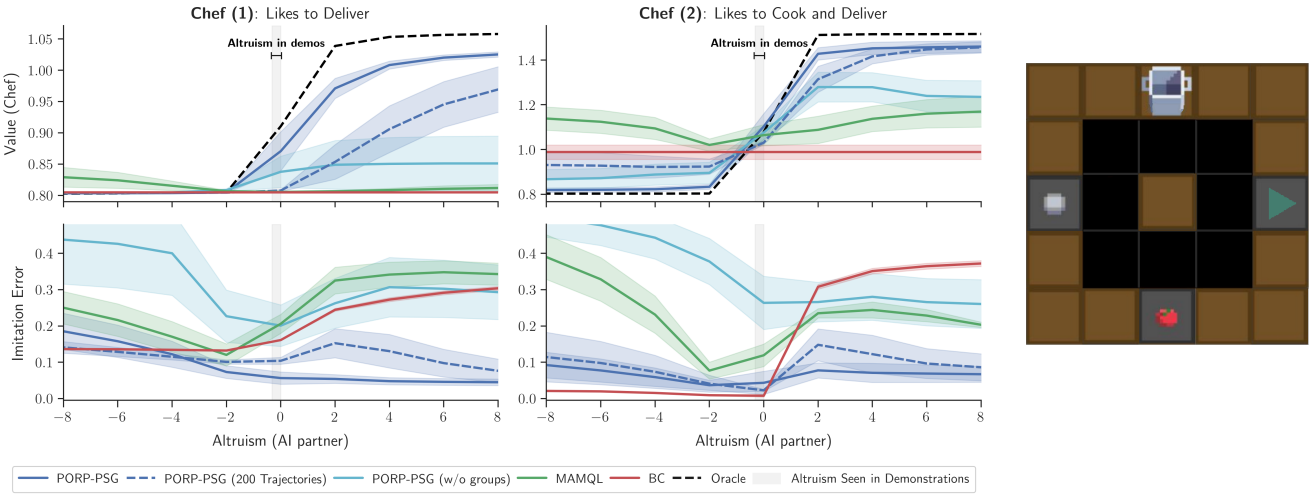


Figure 3: Sample efficiency of the proposed Bayesian inference methods on 4-player randomised games, using different numbers of groups. Error bars are provided in Appendix C.

7.2 Achieving Generalised Altruism Imitation in a Collaborative Cooking Task

We consider a simplified kitchen scenario inspired by Overcooked [9], with three chefs, of which only two work per shift. Interpersonal tensions cause some chefs to act unreasonably, reducing productivity. Our goal is to improve welfare by identifying each chef's intrinsic rewards and introducing an AI partner that can exhibit any desired level of altruism toward its teammate.

Experimental setup. We design a compact but challenging kitchen (Figure 4b) with 3420 states and 5 actions (4 for moving, and 1 for interacting). Agents can cooperate (e.g., passing ingredients) or act adversarially (e.g., blocking the cooking pot). Each chef have



(a) Imitation error of synthesised policies for various altruism levels and the value of their partner. **(b) Kitchen layout.**

Figure 4: Synthesis of policies from anti-social demonstrations in Overcooked, ranging from adversarial to altruistic.

their own intrinsic rewards: chef (1) receives +1 for *delivering*, chef (2) for both *cooking and delivering*, and chef (3) for *cooking*. To assess how well the inferred rewards generalise to novel altruism levels, chefs latent altruism parameters are exclusively sampled uniformly from $[-0.25, 0]$ (slightly antisocial profiles). Each algorithm receives 1000 trajectories, allocated either to a single group (1,2) or across all three groups (1,2), (2,3), and (1,3), from which intrinsic rewards $\hat{r}_i$ are inferred. We reconstruct joint reward functions by plugging an altruism level λ_{AI} of our choice. We then replace either chef 1 or 2 in the (1,2) group with an AI partner optimising $\hat{R}_{(1,2),1}(s, \mathbf{a}) = \hat{r}_1(s, a_1) + \lambda_{\text{AI}}\hat{r}_2(s, a_2)$ or $\hat{R}_{(1,2),2}$ analogously, and let the remaining chef optimise their own true intrinsic reward egoistically. We measure the imitation quality through the chef's policy *value* when partnered with the AI, compared to an oracle agent using ground-truth rewards, and the *policy imitation error*, measured as the mean KL divergence from the oracle policy. The protocol is conducted with values of λ_{AI} interpolated between -8 and 8, to assess the imitation capabilities of methods across the whole altruism spectrum. Due to the prohibitive computation requirements of DRP and QIG, we do not include them in this study. Instead, we include a Behaviour Cloning (BC) baseline, which replicates demonstration policies while allowing the partner to best respond. Experiments are repeated across five random seeds.

Results. Table 2 shows that PORP-PSG can reliably imitate *any desired level of altruism*, although it learned exclusively from anti-social demonstrations. Remarkably, as illustrated in Figure 4a, the method hugs the oracle's curve, indicating near perfect imitation on the whole altruism continuum. By contrast, MAMQL exhibits a striking misalignment: as the AI's intended altruism increases, the partner's utility does not improve, and in some cases even declines. This reveals a critical pitfall of incomplete reward disentanglement, where models tuned to 'be kind' may in fact behave adversarially. Even under reduced data, PORP-PSG consistently produces agents whose altruism reliably translates into higher welfare for chefs, demonstrating both robustness and genuine social alignment.

Table 2: Reward and policy recovery errors on Overcooked. The chef value error measures the difference in the chef's policy value when paired with the AI versus the oracle, averaged over the full altruism spectrum. Lower is better.

Method	Altruism Err.	Intrinsic Rewards Err.	Policy Imitation Err.	Chef Value Err.
PORP-PSG	**0.008 ± 0.001**	**0.016 ± 0.001**	**0.082 ± 0.007**	**0.025 ± 0.005**
PSG (w/o groups)	0.076 ± 0.004	0.225 ± 0.014	0.340 ± 0.018	0.330 ± 0.050
MAMQL	0.080 ± 0.002	0.330 ± 0.005	0.254 ± 0.011	0.450 ± 0.050
BC	–	–	0.188 ± 0.013	0.548 ± 0.055

8 CONCLUSION

We have demonstrated that observing agents across multiple groups is crucial for disentangling altruism from intrinsic rewards in MAIRL. Importantly, such multi-group demonstrations often arise naturally, incurring little extra cost. We propose a practical Bayesian method for constructing a posterior over policies from these demonstrations, we can infer accurate, disentangled rewards without restrictive assumptions on agent rationality. This enables *generalised altruism imitation*, yielding socially aligned AI agents that are both more interpretable and trustworthy. Our results pave the way for AI systems that can *reliably understand and replicate social behaviours*, providing a principled foundation for aligned, cooperative multi-agent interactions.

While our current model assumes context-independent altruism and perfect knowledge of others' preferences, these simplifications point to exciting future directions. Extensions include learning from learning agents, modelling altruism as a function of state and action histories, actively selecting informative groups, scaling to larger environments without dynamics model access [10], and leveraging human proxies such as LLMs to model decision-making.

REFERENCES

[1] John P. Agapiou, Alexander Sasha Vezhnevets, Edgar A. Duéñez-Guzmán, Jayd Matyas, Yiran Mao, Peter Sunehag, Raphael Köster, Udari Madhushani, Kavya Kopparapu, Ramona Comanescu, DJ Strouse, Michael B. Johanson, Sukhdeep Singh, Julia Haas, Igor Mordatch, Dean Mobbs, and Joel Z. Leibo. 2023. Melting Pot 2.0. arXiv:2211.13746 [cs.MA]

[2] Nicolas Anastassacos, Julian García, Stephen Hailes, and Mirco Musolesi. 2021. Cooperation and reputation dynamics with reinforcement learning. *arXiv preprint arXiv:2102.07523* (2021).

[3] C Daniel Batson. 1991. *The altruism question: Toward a social-psychological answer*. Lawrence Erlbaum Associates.

[4] Sage Bergerson. 2021. Multi-agent inverse reinforcement learning: Suboptimal demonstrations and alternative solution concepts. *arXiv preprint arXiv:2109.01178* (2021).

[5] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. 2019. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680* (2019).

[6] Thomas Kleine Buening, Victor Villin, and Christos Dimitrakakis. 2024. Environment design for inverse reinforcement learning. In *ICML* (Vienna, Austria). JMLR.org, Article 994, 21 pages.

[7] Thomas Kleine Büning, Anne-Marie George, and Christos Dimitrakakis. 2022. Interactive inverse reinforcement learning for cooperative games. In *ICML*. PMLR, 2393–2413.

[8] Haoyang Cao, Samuel Cohen, and Lukasz Szpruch. 2021. Identifiability in inverse reinforcement learning. *NeurIPS* 34 (2021), 12362–12373.

[9] Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. 2019. On the utility of learning about humans for human-ai coordination. *NeurIPS* 32 (2019).

[10] Alex J Chan and Mihaela van der Schaar. 2021. Scalable bayesian inverse reinforcement learning. *arXiv preprint arXiv:2102.06483* (2021).

[11] Gary Charness and Matthew Rabin. 2002. Understanding social preferences with simple tests. *Quarterly journal of Economics* (2002), 817–869.

[12] Yusi Chen, Angela Radulescu, and Herbert Zheng Wu. 2024. Unveiling the latent dynamics in social cognition with multi-agent inverse reinforcement learning. *bioRxiv* (2024), 2024–10.

[13] Jaedeug Choi and Kee-Eung Kim. 2011. Map inference for bayesian inverse reinforcement learning. *NeurIPS* 24 (2011).

[14] Christos Dimitrakakis and Constantin A Rothkopf. 2011. Bayesian multitask inverse reinforcement learning. In *European workshop on reinforcement learning*. Springer, 273–284.

[15] Leonard Dung. 2023. Current cases of AI misalignment and their implications for future risks. *Synthese* 202, 5 (2023), 138.

[16] Till Freihaut and Giorgia Ramponi. 2025. On Feasible Rewards in Multi-Agent Inverse Reinforcement Learning. arXiv:2411.15046 [cs.LG] https://arxiv.org/abs/2411.15046

[17] Justin Fu, Andrea Tacchetti, Julien Perolat, and Yoram Bachrach. 2021. Evaluating strategic structures in multi-agent inverse reinforcement learning. *Journal of Artificial Intelligence Research* 71 (2021), 925–951.

[18] Yukiko Fukumoto, Masakazu Tadokoro, and Keiki Takadama. 2020. Cooperative multi-agent inverse reinforcement learning based on selfish expert and its behavior archives. In *IEEE Symposium Series on Computational Intelligence*. IEEE, 2202–2209.

[19] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. 2016. Cooperative inverse reinforcement learning. *NeurIPS* 29 (2016).

[20] Nathaniel Haynam, Adam Khoja, Dhruv Kumar, Vivek Myers, and Erdem Bıyık. 2025. Multi-Agent Inverse Q-Learning from Demonstrations. *arXiv preprint arXiv:2503.04679* (2025).

[21] David Earl Hostallero, Daewoo Kim, Sangwoo Moon, Kyunghwan Son, Wan Ju Kang, and Yung Yi. 2020. Inducing cooperation through reward reshaping based on peer evaluations in deep multi-agent reinforcement learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*. 520–528.

[22] Edward Hughes, Joel Z Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez-Guzman, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin McKee, Raphael Koster, et al. 2018. Inequity aversion improves cooperation in intertemporal social dilemmas. *NeurIPS* 31 (2018).

[23] Hong Jun Jeon, Smitha Milli, and Anca Dragan. 2020. Reward-rational (implicit) choice: A unifying formalism for reward learning. *NeurIPS* 33 (2020), 4415–4426.

[24] Wonseok Jeon, Paul Barde, Derek Nowrouzezahrai, and Joelle Pineau. 2020. Scalable multi-agent inverse reinforcement learning via actor-attention-critic. *arXiv preprint arXiv:2002.10525* (2020).

[25] Cassidy Laidlaw and Anca Dragan. 2022. The boltzmann policy distribution: Accounting for systematic suboptimality in human models. *arXiv preprint arXiv:2204.10759* (2022).

[26] Stefanos Leonardos, Georgios Piliouras, and Kelly Spendlove. 2021. Exploration-exploitation in multi-agent competition: convergence with bounded rationality. *NeurIPS* 34 (2021), 26318–26331.

[27] Chunyuan Li, Changyou Chen, David Carlson, and Lawrence Carin. 2016. Preconditioned stochastic gradient Langevin dynamics for deep neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 30.

[28] Xiaomin Lin, Stephen C Adams, and Peter A Beling. 2019. Multi-agent inverse reinforcement learning for certain general-sum stochastic games. *Journal of Artificial Intelligence Research* 66 (2019), 473–502.

[29] Xiaomin Lin, Peter A Beling, and Randy Cogill. 2017. Multiagent inverse reinforcement learning for two-person zero-sum games. *IEEE Transactions on Games* 10, 1 (2017), 56–68.

[30] Carlos Martin and A Sandholm. 2021. Bayesian Multiagent Inverse Reinforcement Learning for Policy Recommendation. In *AAAI Workshop on Reinforcement Learning in Games*.

[31] Negar Mehr, Mingyu Wang, Maulik Bhatt, and Mac Schwager. 2023. Maximum-entropy multi-agent dynamic games: Forward and inverse solutions. *IEEE transactions on robotics* 39, 3 (2023), 1801–1815.

[32] Sriraam Natarajan, Gautam Kunapuli, Kshitij Judah, Prasad Tadepalli, Kristian Kersting, and Jude Shavlik. 2010. Multi-agent inverse reinforcement learning. In *2010 ninth international conference on machine learning and applications*. IEEE, 395–400.

[33] Andrew Y Ng, Daishi Harada, and Stuart Russell. 1999. Theory and application to reward shaping. In *ICML*.

[34] Andrew Y Ng, Stuart Russell, et al. 2000. Algorithms for inverse reinforcement learning.. In *ICML*, Vol. 1. 2.

[35] Ellen Frankel. Paul, Fred Dycus Miller, and Jeffrey. Paul. 1993. *Altruism*. Cambridge University Press.

[36] Alexander Peysakhovich and Adam Lerer. 2018. Prosocial Learning Agents Solve Generalized Stag Hunts Better than Selfish Ones. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems* (Stockholm, Sweden). International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2043–2044.

[37] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. 2017. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*. PMLR, 1674–1703.

[38] Deepak Ramachandran and Eyal Amir. 2007. Bayesian Inverse Reinforcement Learning.. In *IJCAI*, Vol. 7. 2586–2591.

[39] Giorgia Ramponi, Pavel Kolev, Olivier Pietquin, Niao He, Mathieu Laurière, and Matthieu Geist. 2023. On imitation in mean-field games. *NeurIPS* 36 (2023), 40426–40437.

[40] Tummalapalli Sudhamsh Reddy, Vamsikrishna Gopikrishna, Gergely Zaruba, and Manfred Huber. 2012. Inverse reinforcement learning for decentralized non-cooperative multiagent systems. In *2012 ieee international conference on systems, man, and cybernetics*. IEEE, 1930–1935.

[41] Tianyu Ren and Xiao-Jun Zeng. 2023. Reputation-based interaction promotes cooperation with reinforcement learning. *IEEE Transactions on Evolutionary Computation* 28, 4 (2023), 1177–1188.

[42] Paul Rolland, Luca Viano, Norman Schürhoff, Boris Nikolov, and Volkan Cevher. 2022. Identifiability and generalizability from multiple experts in inverse reinforcement learning. *NeurIPS* 35 (2022), 550–564.

[43] Constantin A Rothkopf and Christos Dimitrakakis. 2011. Preference elicitation and inverse reinforcement learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part III 22*. Springer, 34–48.

[44] Jack Sawyer. 1966. The altruism scale: A measure of co-operative, individualistic, and competitive interpersonal orientation. *Amer. J. Sociology* 71, 4 (1966), 407–416.

[45] Andreas Schlaginhaufen and Maryam Kamgarpour. 2024. Towards the transferability of rewards recovered via regularized inverse reinforcement learning. *NeurIPS* 37 (2024), 21461–21501.

[46] Jiaming Song, Hongyu Ren, Dorsa Sadigh, and Stefano Ermon. 2018. Multi-agent generative adversarial imitation learning. *NeurIPS* 31 (2018).

[47] Santosh Vempala and Andre Wibisono. 2019. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. *NeurIPS* 32 (2019).

[48] Max Welling and Yee W Teh. 2011. Bayesian learning via stochastic gradient Langevin dynamics. In *ICML*. 681–688.

[49] W. Yoshida, R. J. Dolan RJ, and K.J. Friston. 2008. Game Theory of Mind. *PLoS Computational Biology* 4, 12 (2008).

[50] Lantao Yu, Jiaming Song, and Stefano Ermon. 2019. Multi-agent adversarial inverse reinforcement learning. In *ICML*. PMLR, 7194–7201.

[51] Difan Zou, Pan Xu, and Quanquan Gu. 2021. Faster convergence of stochastic gradient langevin dynamics for non-log-concave sampling. In *Uncertainty in Artificial Intelligence*. PMLR, 1152–1162.