Empathy in Multi-Agent Reinforcement Learning

Fixed Empathy and Endogenous Adaptation

Christina Eirini Christodoulou Simone De Giorgi Lavinia Maria Alexandra Skandali

Bocconi Students for Machine Learning

Università Bocconi, Milan, Italy

December 21, 2025

Abstract

In multi-agent reinforcement learning, individually rational behavior can lead to inefficient collective outcomes. Empathy, in this context defined as valuing other agents' rewards in addition to one's own, has been proposed as a mechanism to promote cooperation, but its interaction with learning dynamics remains insufficiently researched. We study empathy-weighted reward shaping in the Prisoner's Dilemma, the Stag Hunt, and a Renewable Resource Sharing environment, comparing fixed empathy with learned empathy. High empathy consistently improves cooperation and collective welfare across environments. However, agents that learn empathy endogenously fail to converge to this optimal regime and instead stabilize at suboptimal intermediate values. Our results show that while high empathy is globally optimal under fixed regimes, standard learning dynamics fail to reliably discover and sustain this level.

1 Introduction

Multi-agent reinforcement learning (MARL) studies how autonomous decision-makers adapt within shared settings. When each agent optimizes only its own reward, the system often converges to equilibria that are socially inefficient. These dynamics reproduce classic social dilemmas, where cooperation is fragile, coordination breaks down, and shared resources are overused.

One approach to addressing these failures is to modify agents' objectives instead of their learning rules. Specifically, assigning agent preferences that positively weight others' outcomes may promote cooperation. Although such preferences are intuitively appealing and socially desirable, their effectiveness within learning systems remains insufficiently understood.

A central challenge is that socially optimal preferences need not be dynamically attainable. In multi-agent systems, agents learn in non-stationary environments shaped by other learners, where short-term incentives and local optima can dominate long-run social benefits. As a result, even globally optimal preferences may be difficult for standard learning processes to discover and sustain.

This work investigates the tension between the social optimality of empathy and its learnability under standard reinforcement learning dynamics. By examining this tension across canonical social dilemmas and shared-resource settings, the study clarifies how prosocial preferences improve collective outcomes and why standard learning dynamics may nevertheless fail to discover and sustain them.

1.1 Contributions

- We show that high empathy can be globally optimal in social dilemmas, maximizing cooperation, welfare, and equality.

- We show that in commons dilemmas, empathy improves sustainability and welfare, while inequality need not decrease monotonically.
- We demonstrate that learning empathy fails to recover the global optimal, converging instead to suboptimal intermediate regimes.
- We show that limitations arise from learning dynamics rather than empathy itself.

1.2 Related Work

- Social preferences and fairness. Models of inequity aversion formalize how agents trade off personal and social outcomes, motivating empathy-based objectives in MARL (Bolton & Ockenfels, 2000; Fehr & Schmidt, 1999).
- Cooperation in social dilemmas. The emergence of cooperation has been extensively studied in evolutionary systems (Axelrod & Hamilton, 1981; Nowak, 2006). Even without empathy or reward shaping, cooperation can emerge under repeated strategic interaction (Axelrod & Hamilton, 1981; Nowak, 2006). Sequential social dilemmas extend these ideas to temporally extended learning environments, where agents must balance short-term incentives against long-term cooperation (Leibo et al., 2017).
- Prosocial MARL. Prosocial reward shaping and inequity-averse preferences have been shown to improve coordination and cooperation in certain multi-agent reinforcement learning settings, though results are sensitive to scale and learning dynamics (Hughes et al., 2018; Peysakhovich & Lerer, 2018).
- Learning instability. Deep MARL systems are known to suffer from instability and local optima, motivating careful multi-seed evaluation and robustness analysis (Henderson et al., 2018).

2 Methodology

2.1 Learning Setup

All agents are trained using standard tabular Q-learning. Each agent learns how good each possible action is in a given state by updating an action-value table over time.

At each environment step (episode), agents choose actions using an ε-greedy exploration strategy. The exploration parameter ε is initialized at 1.0 (fully random action selection) and decayed multiplicatively by a factor of 0.995 at each step until reaching a minimum value of 0.05, allowing agents to explore early and behave more consistently as learning progresses. Training is run for a total of 5,000 environment steps in all experiments.

After each step, agents update their action values using standard Q-learning based on the received reward and the estimated value of the next state. We use the same learning rate and discount factor across all experiments. Training is episodic in all environments.

The learning setup is kept identical in the Prisoner's Dilemma, Stag Hunt, and Renewable Resource Sharing environments. This ensures that differences in behavior across experiments are due to the structure of the environments and the empathy-based reward shaping, rather than differences in learning or optimization settings.

2.2 Empathy-Weighted Reward

Let $r_i(t)$ denote the environment (selfish) reward received by agent i at time t, and let N be the number of agents. We define an empathy-weighted shaped reward:

$$\tilde{r}_i(t) = a_i r_i(t) + (1 - a_i)\frac{1}{N-1}\sum_{j \neq i} r_j(t)$$

where $a_i \in [0,1]$ controls selfishness.

Convention.

Throughout the paper:

α = 0 corresponds to maximal empathy,

α = 1 corresponds to pure selfishness.

When empathy is active, agents update their policies using the empathy-weighted reward $\tilde{r}_i(t)$. However, all reported performance metrics (cooperation, welfare, inequality) are computed using the original environment rewards $r_i(t)$, ensuring that evaluation reflects true game outcomes rather than the reshaped learning signal.

2.3 Fixed vs Learned Empathy

Fixed empathy.

In fixed-empathy experiments, all agents use the same empathy level throughout training. This value is set at the beginning and remains constant for the entire duration of learning. We evaluate four fixed settings, ranging from fully selfish behavior to highly empathic behavior: $\alpha \in \{1.0, 0.8, 0.5, 0.2\}$. This allows us to directly compare how different degrees of empathy affect learning dynamics and long-run outcomes, holding all other aspects of the learning process constant.

Learned empathy.

In learned-empathy experiments, agents adapt their empathy level over time rather than keeping it fixed. Empathy learning and action learning operate on separate time scales.

We model empathy adaptation as a discrete multi-armed bandit over the set $\alpha \in \{1.0, 0.8, 0.5, 0.2\}$, where lower values correspond to greater empathy.

Every 50 environment steps, each agent selects an empathy level and holds it fixed for the duration of that block. Selection is governed by a separate ε_α-greedy strategy with constant exploration rate $\varepsilon_\alpha = 0.1$, meaning that at the beginning of each block the agent explores with probability 0.1 and exploits with probability 0.9.

During the block, agents learn actions using standard Q-learning under the reward signal defined by the selected empathy level. Action-level exploration follows the decaying ε schedule described in Section 2.1.

At the end of each block, the chosen empathy level is evaluated using the average team reward obtained during that block. Empathy preferences are updated via the incremental mean rule:

$$Q_a(a_k) \leftarrow Q_a(a_k) + \frac{1}{n_k}(R_{team} - Q_a(a_k))$$

where $Q_a(a_k)$ is the estimated value of empathy level a_k, n_k is the number of times a_k has been selected so far, and R_{team} is the average team reward over the block.

This procedure creates a clear separation of time scales: action policies update at every step, while empathy values update once per block using aggregate feedback. No gradients or parameter sharing occur between the two processes.

Throughout training, we track the population mean of α to visualize how learned empathy evolves.

Algorithm 1 provides pseudocode for the bandit-based empathy adaptation procedure used in the learned-empathy experiments. The algorithm presents the bandit-based procedure used to select and update empathy levels from block-averaged team rewards.

Algorithm 1 Endogenous Empathy Learning (Bandit-based)

Initialize:
  Candidate empathy coefficients $\mathcal{A} = \{0.0, 0.2, 0.5, 0.8, 1.0\}$
  Bandit values $V(\alpha) \leftarrow 0$, counts $N(\alpha) \leftarrow 0 \ \ \forall \alpha \in \mathcal{A}$
for each block of $B = 50$ steps **do**
  Select empathy coefficient α**:**
    With probability $\epsilon_\alpha = 0.1$, sample α uniformly from $\mathcal{A}$
    Otherwise, $\alpha \leftarrow \arg\max_{\alpha' \in \mathcal{A}} V(\alpha')$
  Interact: Execute B steps using shaped reward
    $\tilde{r}_i(t) = \alpha r_i(t) + (1 - \alpha)\bar{r}_{-i}(t)$
  Observe signal (block-average team reward):
    $G \leftarrow \frac{1}{B \cdot N} \sum_{t=1}^{B} \sum_{i=1}^{N} r_i(t)$
  Update bandit:
    $N(\alpha) \leftarrow N(\alpha) + 1$
    $V(\alpha) \leftarrow V(\alpha) + \frac{1}{N(\alpha)}(G - V(\alpha))$
end for

2.4 Environments

We evaluate three environments: Prisoner's Dilemma (PD), Stag Hunt (SH), and Renewable Resource Sharing (RS). For PD and SH we use a multi-agent matrix-game construction: at each episode, every unordered pair of agents (i,j) plays a simultaneous two-player game. Each agent therefore participates in N-1 pairwise interactions per episode. Pairwise rewards are summed and normalized by N-1, yielding the episode reward:

$$r_i(t) = \frac{1}{N-1} \sum_{j \neq i} r_i^{(i,j)}(t)$$

This normalization keeps reward magnitudes comparable across population sizes.

2.4.1 Prisoner's Dilemma (PD)

Actions: $A = \{C, D\}$. Pairwise payoffs are:

- (C, C) → (3, 3)
- (C, D) → (0, 5)

- (D, C) → (5, 0)
- (D, D) → (1, 1)

2.4.2 Stag Hunt (SH)

Actions: $A = \{S, H\}$. Pairwise payoffs are:

- (S, S) → (4, 4)
- (S, H) → (0, 3)
- (H, S) → (3, 0)
- (H, H) → (2, 2)

2.4.3 Renewable Resource Sharing (RS)

RS is a commons dilemma with a shared renewable stock $x(t)$. The stock starts at $x(0) = 10$, regenerates by 2 units per episode, and is capped at 10. Each agent chooses an extraction level $a_i(t) \in \{0, 1, 2\}$. Let total extraction be $E(t) = \sum_i a_i(t)$. Rewards are:

- If $E(t) \leq x(t)$, each agent receives $r_i(t) = a_i(t)$.
- If $E(t) > x(t)$, the resource collapses that episode and all agents receive $r_i(t) = -0.5$.

The stock then updates according to the extraction and regeneration dynamics, with upper bound 10. Cooperation in RS is defined as maintaining extraction at or below the sustainable threshold $E(t) \leq 4$ under the given regeneration rate.

2.5 State Representations

Prisoner's Dilemma and Stag Hunt are one-shot games and therefore do not have a natural evolving state: the stage game itself is independent of past interactions. However, tabular Q-learning requires a state representation in order to update action values over time. To address this, we introduce a simple synthetic state.

In these environments, the state observed by each agent corresponds to the joint action profile from the previous episode, i.e., the actions chosen by all agents in the previous round. This provides agents with minimal information about recent behavior, allowing them to condition their decisions on others' prior behavior. Importantly, this state representation does not alter the game's strategic structure or introduce additional dynamics; it merely provides the minimal memory required for stable learning and enables the emergence of reactive strategies.

In the Renewable Resource Sharing environment, the situation is different. Here, outcomes depend on the evolution of a shared resource over time. For this reason, agents observe the previous resource stock level as the state. This gives agents direct information about scarcity and about how past collective extraction decisions affect future availability of the resource, allowing them to learn sustainable extraction policies.

2.6 Metrics and Reporting

We report the following metrics (computed using the environment rewards $r_i(t)$):

Cooperation fraction: moving average of cooperative episodes over a window of 100 episodes.

Strict cooperation: an indicator for full coordination, defined as:

- PD: (C, C,…, C)
- SH: (S, S,…, S)
- RS: sustainable extraction $E(t) \leq 4$

Total reward (welfare): $\sum_{i=1}^{N} r_i(t)$, reported as a moving average.

Inequality: variance of individual rewards across agents at each episode.

RS stock level: moving average of $x(t)$ over training.

RS collapse frequency: moving average of $\{E(t) > x(t)\}$.

Learned empathy: population mean of the empathy parameter α, averaged across agents, in experiments where empathy is learned.

All experiments are repeated across 10 independent random seeds (seeds 0-9). Reported learning curves show the mean across seeds, smoothed using a 100-episode moving average to reduce per-step noise. Shaded regions in all plots represent 95% confidence intervals, computed as 1.96 times the standard error of the mean across seeds.

Note that in the reported figures, we also include two baseline conditions for comparison. The random baseline corresponds to agents selecting actions

uniformly at random. The hetero_fixed baseline assigns agents fixed but heterogeneous empathy levels sampled from the same candidate set. These baselines provide reference points for interpreting performance differences but are not the primary focus of the analysis.

2.7 Hyperparameters and Training Schedule

All experiments use the same learning rates, exploration schedules, and training conditions unless otherwise stated. Table 1 summarizes the hyperparameters used for action learning and empathy adaptation across all environments.

Parameter	Value
Q-learning learning rate	0.2
Discount factor	0.95
Action exploration ε (initial)	1.0
Action exploration ε (minimum)	0.05
Action exploration decay	0.995 per environment step
Empathy exploration ε_α (constant)	0.1
Empathy update block size	50 steps
Training horizon	5,000 environment steps

Table 1: *Hyperparameters and training schedule used across all experiments.*

3 Prisoner's Dilemma

3.1 Prisoner's Dilemma (N = 2)

Environment and Game Structure

The Prisoner's Dilemma is a social dilemma in which individually rational behavior leads to a collectively inefficient outcome. Each agent chooses between cooperation and defection, where defection strictly dominates cooperation at the individual level. However, mutual cooperation yields higher collective welfare than mutual defection. Specifically in our experiment, mutual cooperation (C,C) yields 3 to each agent, for a total welfare of 6. If one agent defects while the other cooperates (D,C), the defector receives 5 and the cooperator receives 0, resulting in total welfare of 5. Mutual defection (D,D) yields 1 to each agent, for a total welfare of 2.

We analyze the Prisoner's Dilemma at two scales. We first consider the standard two-agent case, which serves as a baseline without coordination scaling issues. We then extend the environment to four agents, where cooperation becomes more fragile due to the increased sensitivity to unilateral deviations. Both settings use identical learning dynamics and reward shaping, allowing for a direct comparison.

Cooperation Dynamics

Figure 1: *PD (N = 2): strict cooperation. Maximal empathy yields near-perfect cooperation, while selfish agents collapse to zero.*

Figure 1 reports strict cooperation, defined as episodes in which both agents simultaneously choose to cooperate. Selfish agents (α = 1.0) rapidly converge to zero strict cooperation, reproducing the classical mutual defection equilibrium of the PD. Introducing empathy leads to a sharp and monotonic increase in strict cooperation. Under high empathy (α = 0.2), agents achieve near-perfect cooperation after a short transient period. The learned-empathy curve converges to a lower-intermediate empathy level, achieving cooperation rates close to the welfare-maximizing

state, but remaining slightly below the performance attained under fixed high empathy.

Figure 2: *PD (N = 2): cooperation fraction. Cooperation decreases monotonically with increasing selfishness.*

Figure 2 reports the cooperation fraction, defined as the moving average of episodes in which both agents choose to cooperate. The curve exhibits the same monotonic relationship. As selfishness increases, cooperation decreases smoothly and consistently across training. The learned-empathy curve exhibits similar performance to that observed in graph 1. Unlike in larger populations, cooperation in the two-agent case is relatively stable once achieved, as deviations quickly lead to the loss of mutual cooperation.

Welfare and Inequality

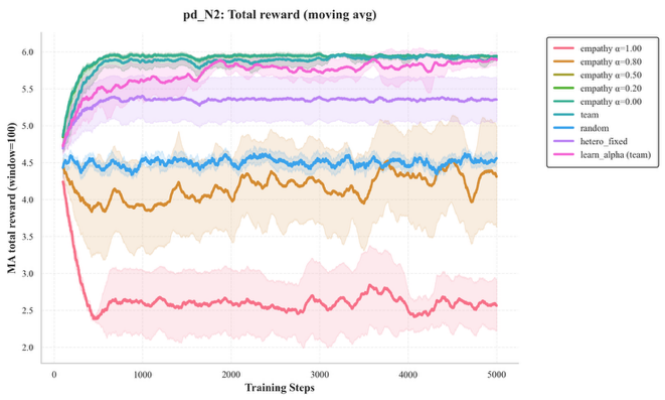

Figure 3: *PD (N = 2): total reward. Empathy maximizes social welfare and cooperation.*

Figure 3 reports total reward, defined as the sum of environment rewards $r_i(t)$ across both agents. Welfare is maximized under high empathy and decreases monotonically with increasing selfishness. Importantly, there is no trade-off between cooperation and welfare in this setting: the same empathy regimes that maximize cooperation also maximize total reward. Learned-empathy curve continues to closely track intermediate to high empathy levels, while remaining slightly below them.

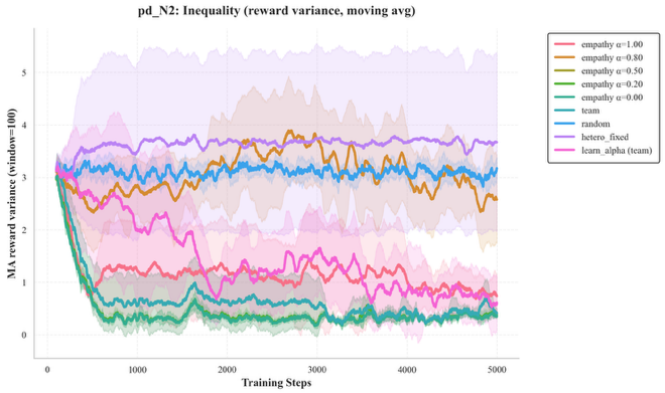

Figure 4: *PD (N = 2): inequality. Increasing empathy sharply reduces reward variance.*

Figure 4 shows reward inequality, measured as the variance of individual rewards. Inequality increases sharply as agents move away from strong empathy, but it does not peak under full selfishness. Instead, the highest variance arises at intermediate levels of selfishness, where asymmetric exploitation and occasional cooperation generate the largest payoff dispersion. Under full selfishness, agents more consistently defect, producing uniformly low but relatively similar payoffs. The learned-empathy curve appears close to the high-empathy case, but exhibits greater variability throughout training and shows slower convergence. In contrast, high empathy compresses reward variance, leading to both cooperative and equitable outcomes. This confirms that in the two-agent setting, strong empathy simultaneously improves efficiency and fairness.

Learned Empathy

Figure 5: *PD (N = 2): learned empathy. Despite $\alpha = 0$ being globally optimal, learning converges to intermediate values.*

Figure 5 reports the evolution of learned empathy when α is adapted endogenously during training using the bandit-based procedure described in Section 2.3. Despite maximal empathy being globally optimal across cooperation, welfare, and inequality, learning dynamics consistently converge to lower-intermediate values of α. This indicates that even in the simplest Prisoner's Dilemma setting, the bandit-based empathy adaptation dynamics converge to suboptimal intermediate regimes rather than to maximal empathy.

This two-agent case therefore establishes that while empathy is unambiguously beneficial, endogenously learning the optimal empathy level remains difficult even in the absence of coordination complexity.

3.2 Prisoner's Dilemma (N = 4)

Environment and Game Structure

We now turn to the four-agent Prisoner's Dilemma, where coordination becomes substantially more challenging. In each episode, every agent plays the standard pairwise Prisoner's Dilemma against each of the other three agents, using the same payoff matrix as in the two-agent case. Rewards are summed and normalized to maintain comparability across settings.

Agents again choose between cooperation (C) and defection (D) and are trained using the same Q-learning setup described in Section 2. However, in this larger population, strict cooperation requires all four agents to choose C simultaneously. A single defection lowers payoffs in multiple pairwise interactions, making the cooperative outcome significantly more fragile than in the two-agent case.

Cooperation Dynamics

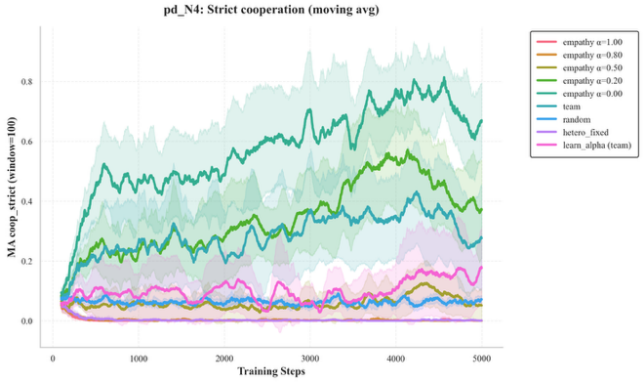

Figure 6: *PD (N = 4): strict cooperation. Empathy dominates across all levels of cooperation.*

Figure 6 shows strict cooperation, highlighting that selfish agents fail entirely to sustain joint cooperation as the population size increases. This failure directly translates into welfare outcomes. Learned-empathy tends to mimic the behavior of the moderate-empathy curve but lies a little higher and with larger variations.

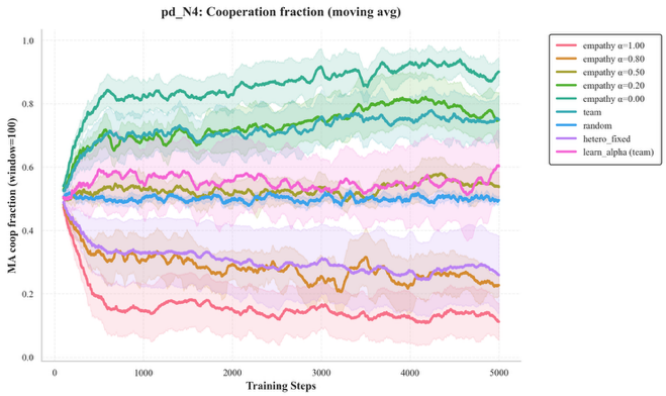

Figure 7: *PD (N = 4): cooperation fraction. Empathy dominates across all levels of cooperation.*

Figure 7 reports the cooperation fraction, defined as the moving average of episodes in which all four agents choose to cooperate. Selfish agents rapidly converge to near-zero cooperation, reproducing the mutual defection Nash equilibrium. Weak empathy does not qualitatively change this behavior, as cooperation remains unstable and quickly converges after exploration. Moderate empathy (α = 0.5) increases the frequency of cooperative episodes but fails to produce stable cooperation, as occasional defections repeatedly disrupt coordination. In contrast, strong empathy (α = 0.2) leads to a sharp increase in cooperation, reaching approximately 75-85% in late training, although outcomes remain more volatile than in the two-agent case. Overall, these results indicate that empathy must be sufficiently strong to overcome the increased sensitivity of cooperation to unilateral deviations.

Welfare and Inequality

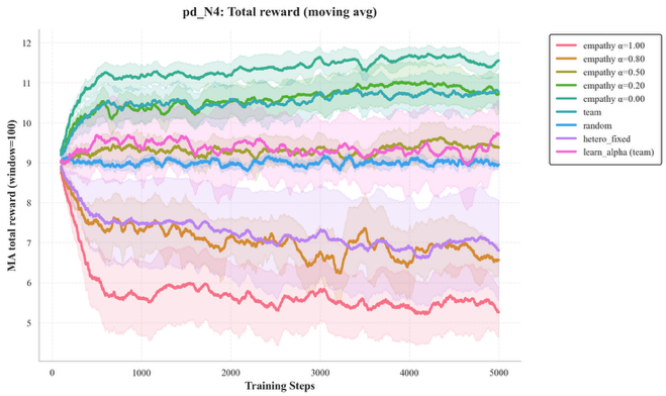

Figure 8: *PD (N = 4): total reward. Empathy maximizes welfare, while selfishness yields the worst outcomes.*

Figure 8 reports total reward, defined as the sum of selfish rewards across all agents. Under full defection, each agent receives a reward of 1, yielding a total reward of 4. Under full cooperation, each agent receives 3, yielding a total reward of 12.

Consistent with the cooperation results, selfish agents stabilize near the low-welfare defection regime, while weak empathy yields only marginal improvements. Moderate empathy, matched by learned empathy, achieves intermediate welfare levels, whereas strong empathy consistently approaches the social optimum. These findings confirm that empathy alters the game's effective incentives. By incorporating others' rewards into the learning signal, empathic agents have different incentives that favor cooperative and high-welfare outcomes.

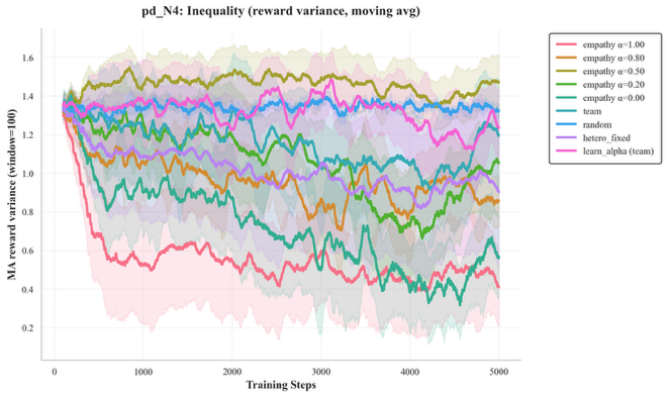

Figure 9: *PD (N = 4): inequality. Reward variance is highest at intermediate empathy levels.*

Figure 9 shows reward inequality, measured as the variance of individual rewards. Inequality does not increase monotonically with selfishness. Instead, it follows a non-monotonic pattern, as in the two-agents model. Variance is relatively low under extreme selfishness (α = 1.0), where agents converge to mutual defection and receive similar payoffs, and also low under strong empathy (α = 0.2), where coordination on cooperation produces equal outcomes. Thus, strong empathy compresses payoff disparities, but extreme selfishness can also yield low inequality due to uniformly poor outcomes. Inequality peaks at intermediate empathy levels, where asymmetric cooperation and defection generate greater payoff dispersion. Again here, learned empathy behaves similarly to moderate empathy.

Learned Empathy

Figure 10: *PD (N = 4): learned empathy. Learning stabilizes at suboptimal intermediate regimes.*

Finally, Figure 10 reports the evolution of learned empathy over training. Despite strong empathy yielding the best overall outcomes, learning dynamics converge to intermediate values of α, specifically around 0.5, rather than to full empathy. This pattern indicates that the limitation lies not in empathy itself, but in the learning dynamics as the bandit-based empathy adaptation stabilizes to suboptimal intermediate states in a changing multi-agent environment. Note that this is a slightly higher level to the one obtained in the two-agents environment, meaning that in the simpler setup learned empathy converged closer to the optimally high empathy.

In conclusion, across both the two-agent and four-agent Prisoner's Dilemma, higher empathy consistently increases cooperation and total welfare, with strong empathy (α = 0.2) approaching the social optimum even under increased coordination complexity. Inequality exhibits a non-monotonic pattern, peaking at intermediate empathy levels but remaining low under

both extreme selfishness and strong empathy. However, when empathy is learned endogenously, agents converge to intermediate α values rather than the welfare-maximizing regime, indicating that the primary limitation arises from learning dynamics rather than from empathy itself.

4 Stag Hunt (N = 4)

Environment and Game Structure

The Stag Hunt is a coordination game in which agents choose between a safe action that yields a guaranteed payoff and a risky action that produces higher rewards only under successful coordination.

Unlike in the Prisoner's Dilemma, where defection yields a direct strategic advantage by exploiting a cooperating opponent, cooperation in the Stag Hunt does not create an incentive to exploit others. Instead, success depends on mutual coordination, and failure arises from misalignment rather than opportunistic defection.

We consider a four-agent Stag Hunt implemented through pairwise interactions. In each episode, every agent plays the standard Stag Hunt against each of the other three agents, using the same payoff matrix as defined above. Mutual Stag (S,S) yields a payoff of 4 to both agents, mutual Hare (H,H) yields 2 to both, and mismatched actions yield 3 to the Hare player and 0 to the Stag player. Resulting rewards are summed and normalized to ensure comparability across environments.

This game admits two pure-strategy equilibria. The risk-dominant equilibrium corresponds to all agents choosing Hare, which guarantees a positive payoff regardless of others' actions but yields low collective welfare. The payoff-dominant equilibrium corresponds to all agents choosing Stag, which maximizes total welfare but requires tight coordination since a single agent choosing Hare collapses the high-payoff outcome for the entire group.

As in the Prisoner's Dilemma experiments, agents are trained using the same learning dynamics, state representation, and empathy-weighted reward shaping described in Section 2.

Cooperation Dynamics

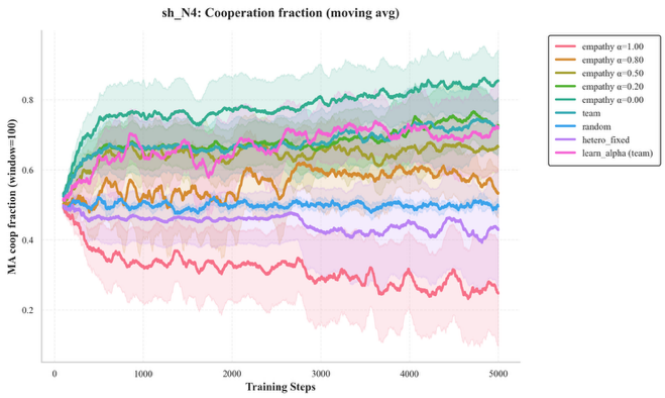

Figure 11: *SH (N = 4): cooperation fraction. Higher empathy reaches the highest cooperation levels.*

Figure 11 reports the cooperation fraction, defined as the moving average of episodes in which all four agents choose Stag. Selfish agents fail to coordinate on the payoff-dominant equilibrium and remain at low levels of cooperation, thus converging the safer all-Hare outcome. As empathy increases, cooperation improves steadily with high empathy achieving the highest and most stable cooperation levels among the fixed-empathy regimes. The learned-empathy curve tends to closely follow the high-empathy one but with larger variations, demonstrating relatively high levels of cooperation. Overall, cooperation increases monotonically as agents become more empathic.

Strict Cooperation

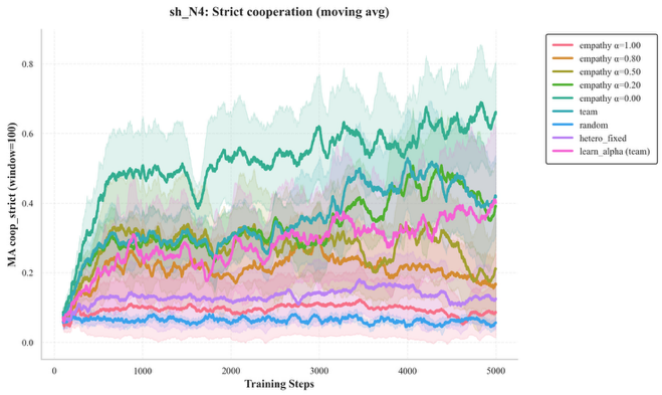

Figure 12: *SH (N = 4): strict cooperation. High empathy promotes stable coordination.*

Figure 12 shows strict cooperation, defined as the fraction of episodes in which all four agents choose Stag simultaneously, capturing full coordination on the payoff-dominant equilibrium. Selfish agents almost never achieve strict cooperation, instead stabilizing at

persistently low coordination levels. As with strict cooperation, as empathy increases, strict cooperation improves steadily. In contrast to strict cooperation, the learned-empathy curve more closely resembles moderate empathy, though it remains relatively close to the optimal level. Again, coordination improves monotonically as agents become more empathic.

Welfare and Stability

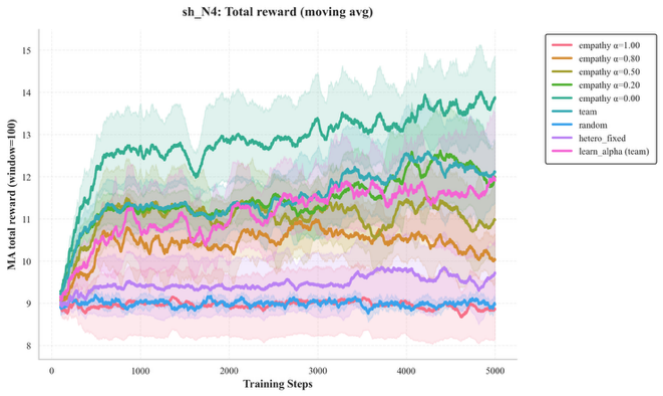

Figure 13: *SH (N = 4): total reward. Empathy yields the highest collective payoff.*

Figure 13 models the total reward across training. The welfare dynamics mirror the cooperation results. Selfish agents stabilize near the low-welfare all-Hare equilibrium, while total reward increases monotonically with empathy, with strong empathy achieving the highest long-run welfare. The learned-empathy curve exhibits substantial variability, at times crossing the weak, moderate, and high empathy curves, reflecting unstable convergence toward an intermediate level.

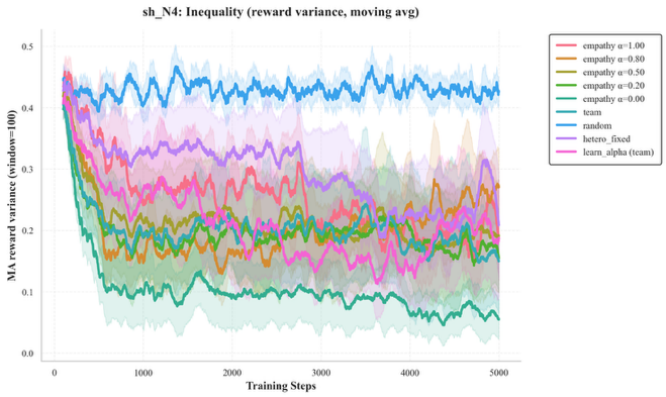

Figure 14: *SH (N = 4): inequality. Empathy compresses payoff disparities without sacrificing welfare.*

Figure 14 shows reward inequality. In contrast to the Prisoner's Dilemma, inequality in the Stag Hunt increases monotonically as agents become more selfish. The highest discrepancy arises under strong selfishness, where persistent miscoordination generates uneven payoffs, while higher empathy compresses disparities by stabilizing symmetric coordination on Stag. The learned-empathy curve exhibits substantial variability overall but gradually converges toward the behavior of the more empathic settings as training progresses.

Learned Empathy

Figure 15: *SH (N = 4): learned empathy. Learning stabilizes at a suboptimal empathy level.*

The learned-empathy results on figure 15 show that, as in PD, although strong empathy consistently yields the highest coordination and welfare under fixed regimes, agents do not reliably converge to this optimal level when empathy is adapted endogenously. Instead, learning stabilizes at intermediate α values (around 0.5), slightly below the level reached in the Prisoner's Dilemma, and falls short of the payoff-dominant coordination achieved under fixed high empathy. This indicates that even in a pure coordination game, where cooperation is not strategically exploited, the learning dynamics struggle to discover and maintain the welfare-maximizing level of empathy.

In conclusion, the four-agent Stag Hunt, higher empathy consistently improves coordination and total welfare, with strong empathy achieving the payoff-dominant equilibrium under fixed regimes. Inequality declines monotonically as empathy increases, reflecting stable and symmetric coordination outcomes. However, when empathy is learned endogenously, agents again converge to intermediate α values rather

than the welfare-maximizing regime. As in the Prisoner's Dilemma, the primary limitation lies not in empathy's effectiveness, but in the learning dynamics that govern its adaptation.

5 Renewable Resource Sharing (N = 4)

Environment and Dynamics

The Renewable Resource Sharing environment models a common-pool resource dilemma in which agents interact indirectly through a shared stock rather than through pairwise strategic games. Four agents repeatedly extract from a renewable resource that evolves over time, creating a tension between short-term individual gain and long-term collective sustainability.

At the beginning of each episode, the resource stock is initialized at 10 units. Each agent independently selects an extraction level from the discrete set {0, 1, 2}, corresponding to low, medium, and high extraction. If the total extraction does not exceed the available stock, each agent receives a selfish reward equal to the amount extracted. If total extraction exceeds the stock, the resource collapses for that episode and all agents receive a penalty of −0.5, independent of their individual extraction. After rewards are assigned, the stock regenerates by 2 units, up to a maximum of 10.

Cooperation in this environment is defined as maintaining total extraction at or below (≤) 4 units per episode, which corresponds to the sustainable threshold under the given regeneration dynamics. Agents observe the previous stock level as the state, providing minimal temporal information about the long-term consequences of their actions.

Learning dynamics and empathy-weighted reward shaping follow the same setup as in the previous environments.

Cooperation Dynamics

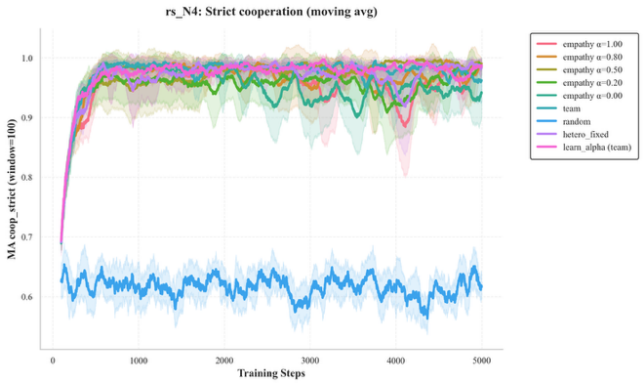

Figure 16: *RS (N = 4): strict cooperation. High empathy exhibits very strong variations, while other levels tend to cluster at higher cooperation levels.*

Figure 16 demonstrates strict cooperation, defined as episodes in which total extraction remains within the sustainable threshold. Although high empathy yields the highest long-run welfare and lowest collapse frequency (figures 17 & 18), it exhibits greater variability in strict cooperation compared to moderate empathy. In the dynamic commons setting, strong empathy tightly aligns agents' incentives, leading them to adjust extraction in a highly synchronized manner in response to stock fluctuations. This collective responsiveness can occasionally produce small, coordinated overshoots around the sustainability threshold. Under lower empathy levels, agents respond more independently, and these less synchronized adjustments dampen collective oscillations, resulting in slightly smoother strict-cooperation dynamics. The learned-empathy condition does not converge to this highly aligned state, instead stabilizing near intermediate empathy levels and therefore failing to replicate the high-empathy performance.

Collapse Frequency and Stock Dynamics

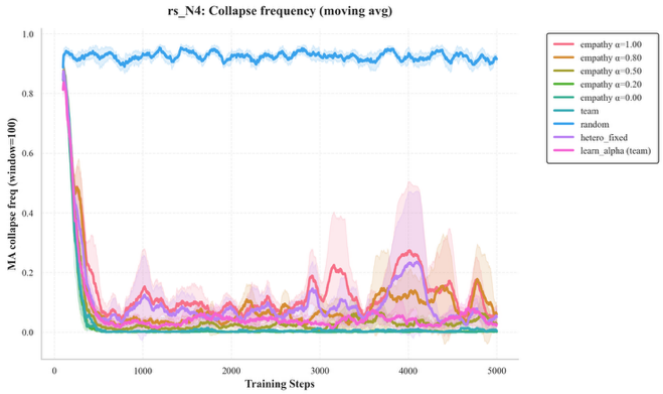

Figure 17: *RS (N = 4): collapse frequency. Empathy nearly eliminates collapses, while selfishness induces systemic failure.*

Figure 17 reports the frequency of resource failure, defined as episodes in which total extraction exceeds the available stock. Selfish agents display frequent failures early in training due to aggressive over-extraction and continue to experience occasional breakdowns even after convergence. The learned-empathy curve mimics this behavior. As empathy increases, collapse frequency declines steadily with high empathy nearly eliminating collapses after a brief transient.

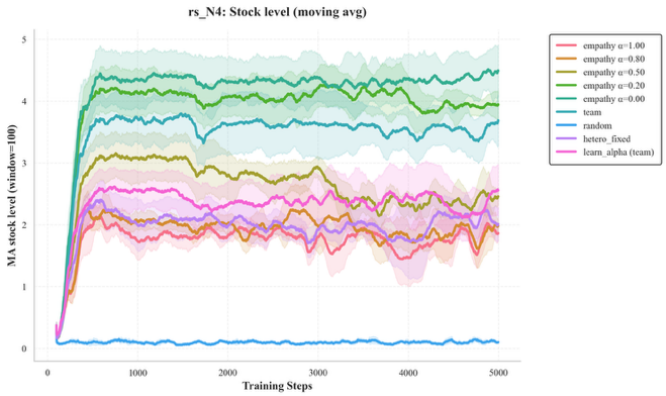

Figure 18: *RS (N = 4): stock level. Empathy sustains higher resource levels over time.*

Figure 18 shows the evolution of the resource stock over training. Selfish agents maintain the lowest average stock levels, reflecting repeated episodes of over-extraction and incomplete recovery. As empathy increases, the long-run stock level rises monotonically. High empathy sustains the resource near its maximum capacity with some but low variance, indicating consistent restraint across agents. Learned empathy converges to behavior consistent with moderate empathy.

Welfare and Inequality

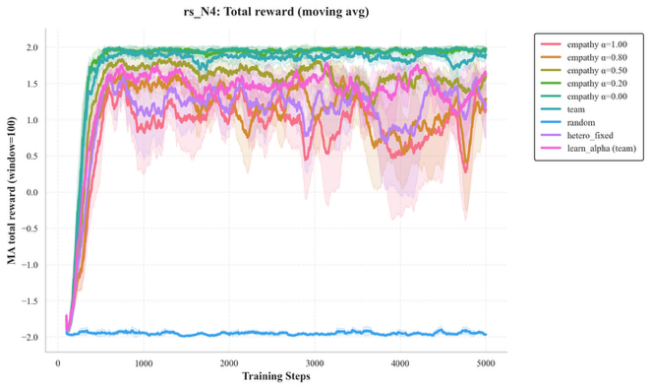

Figure 19: *RS (N = 4): total reward. Total welfare increases monotonically with empathy, with high empathy achieving the highest and most stable long-run performance.*

Figure 19 shows that total reward increases with empathy in the Renewable Resource environment. Strong empathy achieves the highest long-run welfare and the most stable performance, as reduced collapse frequency and higher stock levels translate directly into improved total reward. In contrast, more selfish settings stabilize at lower welfare and experience deeper downturns due to recurrent over-extraction. The learned-empathy condition settles near intermediate empathy levels, attaining moderate performance but falling short of the high-empathy outcome.

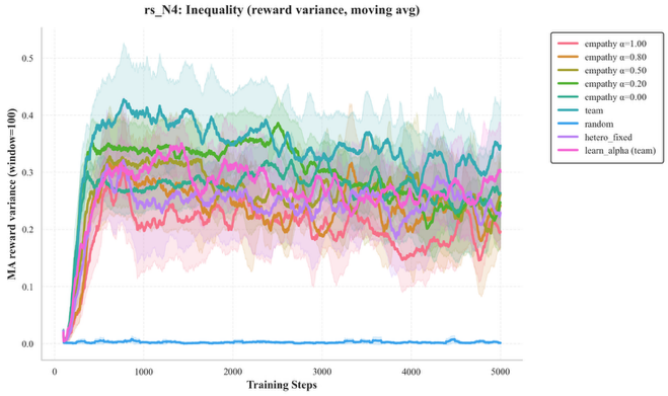

Figure 20: *RS (N = 4): inequality. Empathy tends to occasionally favor inequality.*

Figure 20 reports reward inequality, measured as the variance of selfish rewards across agents. Similarly to the Stag Hunt, inequality in the Renewable Resource environment changes monotonically with empathy, but in the opposite direction as higher empathy is associated with greater reward variance. Strong empathy leads to the highest dispersion, while selfish agents exhibit lower variance. This pattern arises because under strong empathy the resource is sustained

and total extraction remains high, allowing differences in individual extraction choices to translate into meaningful payoff differences. Under selfishness, frequent collapses compress payoffs toward uniformly low values (−0.5 or low extraction rewards), mechanically reducing variance despite poor collective performance. Thus, lower inequality under selfishness reflects uniformly bad outcomes rather than equitable coordination. The learned-empathy condition stabilizes near intermediate empathy levels and therefore exhibits inequality levels comparable to moderate empathy, rather than matching the higher dispersion observed under strong empathy.

Learned Empathy

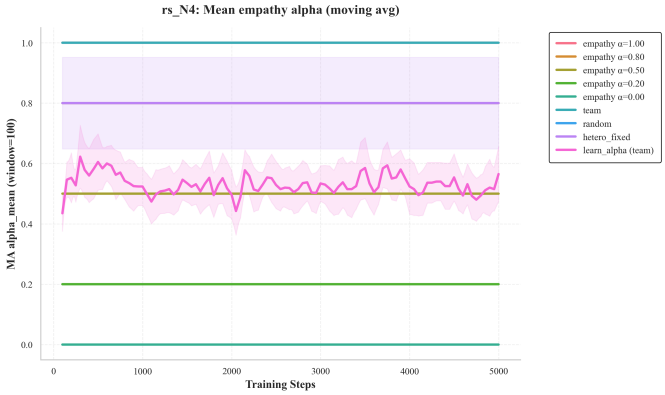

Figure 21: *RS (N = 4): learned empathy. Learning stabilizes below the welfare-maximizing level.*

Figure 21 reports the evolution of learned empathy. Once again, although strong empathy yields the highest sustainability and welfare under fixed regimes, agents fail to converge to this level when empathy is adapted endogenously. Instead, learning stabilizes around intermediate α values, fluctuating near moderate empathy (α = 0.5) rather than approaching the welfare-maximizing level. The learned-empathy dynamics closely resemble those observed in the Stag Hunt, converging to similar intermediate values across both environments.

To conclude, as in the Prisoner's Dilemma and the Stag Hunt, higher empathy improves collective welfare under fixed regimes, with α = 0.2 achieving the strongest performance. However, when empathy is learned endogenously, agents stabilize at intermediate levels rather than converging to the welfare-maximizing regime. In the Renewable Resource setting, this gap is particularly consequential, as empathy operates through long-run sustainability rather than strategic coordination, yet standard learning dynamics still fail to discover and maintain the globally optimal level of prosociality.

6 Comparative Analysis Across Environments

Across all three environments, fixed high empathy (α = 0.2) consistently yields the strongest collective outcomes, but the mechanism through which it improves performance differs by environment. In the Prisoner's Dilemma, empathy offsets direct incentives to defect; in the Stag Hunt, it resolves coordination failure; and in the Renewable Resource setting, it stabilizes long-run sustainability dynamics. Despite these structural differences, a common pattern emerges: when empathy is allowed to adapt endogenously, agents systematically converge to intermediate levels rather than to the welfare-maximizing value.

We interpret this gap as arising from the bandit-based adaptation process, which evaluates empathy locally over short performance windows. In the Prisoner's Dilemma, intermediate empathy can generate sufficient cooperation to produce reasonable rewards without fully overcoming defection incentives, potentially creating a locally stable but suboptimal compromise. In the Stag Hunt, partial coordination may yield moderate payoffs that reduce the pressure to explore stronger empathy levels. In the Renewable Resource environment, intermediate empathy can already sustain the resource at acceptable levels, limiting the apparent marginal benefit of converging fully to high empathy. Across environments, the learning process appears to settle at levels that are locally satisfactory in the short run rather than globally optimal in the long run.

Table 2 provides a comparative analysis across all three environments:

Environment	Fixed High Empathy ($\alpha = 0.2$)	Learned Empathy	Structural Role of Empathy
Prisoner's Dilemma	Near-optimal cooperation and welfare; inequality low (non-monotonic overall)	Converges to intermediate α (~0.5); furthest from optimal; moderate welfare and cooperation	Overcomes direct defection incentives
Stag Hunt	Achieves payoff-dominant equilibrium; welfare maximized; inequality decreases monotonically	Converges to intermediate α (~0.5), slightly closer to high empathy than in PD; unstable coordination	Resolves miscoordination
Renewable Resource	Maximizes long-run sustainability and welfare; lowest collapse frequency; inequality increases with empathy	Converges near intermediate α (~0.5), again slightly closer to high empathy than in PD; moderate sustainability	Stabilizes dynamic commons

Table 2: Comparative analysis across environments

7 Ablation Considerations

While this study keeps the empathy-learning architecture fixed to compare environments, the consistent gap between fixed optimal empathy and learned empathy suggests that the adaptation mechanism itself may influence convergence. We consider several ablation directions that could clarify whether suboptimal stabilization arises from structural properties of bandit-based adaptation or from structural properties of the environments themselves.

Representation of Empathy.

The current design restricts empathy to a discrete set of candidate values. A finer discretization or a continuous empathy parameter would allow incremental adjustments rather than coarse switches between fixed levels. Such an ablation would test whether convergence to intermediate empathy reflects genuine learning dynamics or simply the limited resolution of the empathy grid.

Update interval.

Empathy values are evaluated using average team reward over short (50 steps) fixed blocks. This relatively short interval may favor levels that provide stable but locally sufficient performance, even if stronger empathy would yield higher long-run welfare. Extending the evaluation period or increasing the length of empathy blocks would allow the adaptation mechanism to account for longer-term outcomes, helping determine whether the convergence to intermediate empathy levels is driven by short-term performance signals rather than true long-run optimality.

Exploration of Empathy (ε_α Schedule).

Empathy exploration is currently governed by a constant ε_α-greedy strategy. Varying this schedule, for example by increasing exploration or allowing it to decay over time, would help determine whether convergence to intermediate empathy levels is driven by insufficient exploration rather than inherent limitations of the adaptation mechanism.

Time-Scale Separation.

The architecture enforces separation between fast action learning and slower empathy adaptation. Allowing empathy and actions to update on comparable time scales could reveal whether delayed preference updates reinforce local optima. If synchronized updates reduce persistent intermediate stabilization, this would indicate that time-scale separation plays a central role in the observed gap.

Alternative Adaptation Mechanisms.

Finally, replacing the discrete bandit-based adaptation with a gradient-based update, in which empathy is adjusted incrementally in the direction that improves long-run reward, would test whether the discrete grid structure itself induces stabilization at intermediate levels rather than allowing convergence toward the welfare-maximizing empathy value.

Taken together, these ablations would help determine whether the observed gap between fixed optimal empathy and learned empathy arises from architectural design choices, training duration, or deeper limitations of endogenous preference learning in multi-agent systems.

8 Conclusion

Across all environments, fixing empathy at a high level consistently maximizes cooperation, welfare, and stability. However, when empathy is learned endogenously, agents systematically converge to intermediate values rather than to the welfare-maximizing level, revealing a persistent gap between what is socially optimal and what learning dynamics can sustain. Under the present adaptation mechanism, the central limitation therefore lies not in empathy itself, but in the mechanism used to learn it. These findings highlight the challenge of explaining how stable prosocial norms and coordination mechanisms emerge in multi-agent systems driven by decentralized learning.

References

Fehr, Ernst, and Klaus M. Schmidt. *A Theory of Fairness, Competition, and Cooperation*. *The Quarterly Journal of Economics*, vol. 114, no. 3, 1999, pp. 817-868. https://web.stanford.edu/~niederle/Fehr.Schmidt.1999.QJE.pdf

Hardin, Garrett. The Tragedy of the Commons. Science, vol. 162, no. 3859, 1968, pp. 1243–1248. https://math.uchicago.edu/~shmuel/Modeling/Hardin,%20Tragedy%20of%20the%20Commons.pdf

Henderson, Peter, et al. *Deep Reinforcement Learning That Matters*. *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018. https://ojs.aaai.org/index.php/AAAI/article/view/11694

Hughes, Edward, et al. *Inequity Aversion Improves Cooperation in Intertemporal Social Dilemmas*. arXiv, 2018. https://arxiv.org/pdf/1803.08884

Leibo, Joel Z., et al. *Multi-Agent Reinforcement Learning in Sequential Social Dilemmas*. arXiv, 2017. https://arxiv.org/pdf/1702.03037

Lerer, Adam, and Alexander Peysakhovich. *Maintaining Cooperation in Complex Social Dilemmas Using Deep Reinforcement Learning*. arXiv, 2017. https://arxiv.org/pdf/1707.01068

Lowe, Ryan, et al. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. arXiv, 2017. https://arxiv.org/abs/1706.02275

Matignon, Laëtitia, et al. *Independent Reinforcement Learners in Cooperative Markov Games*. *Journal of Autonomous Agents and Multi-Agent Systems*, vol. 24, no. 1, 2012, pp. 1-51. https://hal.science/hal-00720669/file/Matignon2012independent.pdf

Ng, Andrew Y., Daishi Harada, and Stuart J. Russell. Policy Invariance under Reward Transformations: Theory and Application to Reward Shaping. Proceedings of the Sixteenth International Conference on Machine Learning, 1999, pp. 278–287. https://people.eecs.berkeley.edu/~pabbeel/cs287-fa09/readings/NgHaradaRussell-shaping-ICML1999.pdf

Nowak, Martin A. *Five Rules for the Evolution of Cooperation*. *Science*, vol. 314, no. 5805, 2006, pp. 1560-1563. https://pmc.ncbi.nlm.nih.gov/articles/PMC3279745/

Peysakhovich, Alexander, and Adam Lerer. *Prosocial Learning Agents Solve Coordination Problems*. arXiv, 2018. https://arxiv.org/pdf/1709.02865

Rabin, Matthew. Incorporating Fairness into Game Theory and Economics. American Economic Review, vol. 83, no. 5, 1993, pp. 1281–1302. https://econweb.ucsd.edu/~jandreon/Econ264/papers/Rabin%20AER%201993.pdf

Sutton, Richard S., and Andrew G. Barto. Reinforcement Learning: An Introduction. 2nd ed., MIT Press, 2018. https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf