Empathy-Like Behaviors in Multi-Agent Environments

Can reward sharing induce prosocial emergence?

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 levels with agents that learn empathy endogenously. High empathy consistently improves cooperation, welfare, and reward. However, agents that learn empathy fail to converge to this optimal standard and instead stabilize at suboptimal intermediate values. In shared-resource environments, empathy reduces over-extraction, preventing the resource from being depleted over time. Our results show that while high empathy is globally optimal, standard learning dynamics converge to and remain in suboptimal empathy regimes.

1 Introduction

Multi-agent reinforcement learning (MARL) examines how autonomous agents adapt and interact in shared environments. In such settings, learning alone does not resolve conflicts between individual incentives and collective welfare: agents reproduce classic social dilemmas, in which cooperation is fragile and coordination fails, leading to overuse of shared resources.

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 when prosocial preferences

meaningfully improve collective outcomes and when they are undermined by learning dynamics.

Contributions.

- We show that high empathy can be globally optimal in social dilemmas, maximizing cooperation, welfare, and equality.
- We demonstrate that learning empathy fails to recover this optimum, converging instead to suboptimal intermediate regimes.
- We identify a stability - efficiency trade-off in commons dilemmas.
- We show that limitations arise from learning dynamics rather than empathy itself.

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.

Cooperation in social dilemmas. The emergence of cooperation has been extensively studied in evolutionary systems. Sequential social dilemmas extend these ideas to learning agents.

Prosocial MARL. Prosocial reward shaping can improve outcomes in some cooperative set- tings, but results are sensitive to scale and learning dynamics.

Learning instability. Deep MARL is known to suffer from sensitivity and local optima, motivating multi-seed evaluation.

3 Methodology

3.1 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 every episode, agents choose actions using an ε-greedy exploration strategy. With probability ε, an agent selects a random action to explore. Otherwise, it chooses the action with the highest current value estimate. Exploration starts high (ε=1.0), so agents initially behave almost randomly, and then gradually decreases during training until it reaches a minimum value of 0.05. This allows agents to explore early on and become more consistent as learning progresses.

After each episode, agents update their action values based on the reward they receive 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.

3.2 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:

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

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 learn using $\tilde{r}_i(t)$ as the learning signal, while we still report

evaluation metrics based on the underlying environment rewards $r_i(t)$.

3.3 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 does not change. We evaluate four fixed settings, ranging from fully selfish behavior to highly empathic behavior: α = 1.0, 0.8, 0.5, and 0.2. This allows us to directly compare how different degrees of empathy affect learning and long-run outcomes.

Learned empathy.

In learned-empathy experiments, agents are not assigned a fixed empathy level. Instead, each agent learns how much weight to place on others' rewards during training, alongside learning its action policy. The empathy parameter is optimized by sharing gradients with the action-value updates in Q-learning, allowing these components to be learned simultaneously. To ensure that empathy values remain between 0 and 1, we parameterize it using a bounded transformation. Empathy is updated with the learning signal in each iteration, and we track the average empathy level across agents over time.

3.4 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.

3.4.1 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)

3.4.2 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)

3.4.3 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.

3.5 State Representations

Prisoner's Dilemma and Stag Hunt are one-shot games and therefore do not have a natural evolving state: each episode 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.

3.6 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 when empathy is optimized.

All learning curves are reported as moving averages over 100 episodes, consistent with the plots presented in the Results section.

4 Prisoner's Dilemma

4.1 Prisoner's Dilemma (N = 2)

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. For example, if two agents both choose to cooperate, they each receive a payoff of 3, resulting in a total welfare of 6. In contrast, if one agent defects while the other cooperates, the defector receives a higher payoff of 5, but total welfare falls to 5, disadvantaging the group as a whole. This tension between individual incentives and collective outcomes makes cooperation fragile and highly sensitive to incentives.

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 and Strict Cooperation

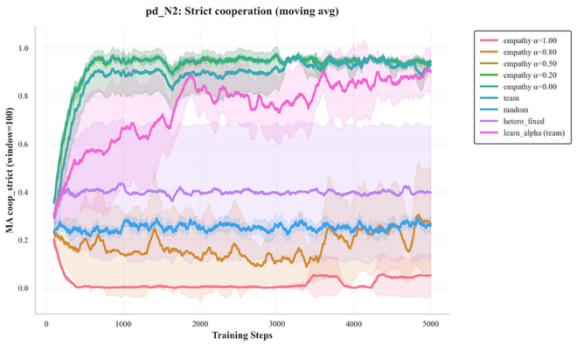

Figure 1: *PD (N = 2): strict cooperation. Maximal empathy (α= 0) yields near-perfect cooperation, while selfish agents (α= 1) collapse to zero.*

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 maximal empathy (α = 0), agents achieve near-perfect cooperation after a short transient period.

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

Figure 2 reports the cooperation fraction, which exhibits the same monotonic relationship. As selfishness increases, cooperation decreases smoothly and consistently across training. Unlike in larger populations, cooperation in the two-agent case is relatively stable once achieved, as deviations are immediately punished by the loss of mutual cooperation.

Welfare and Inequality

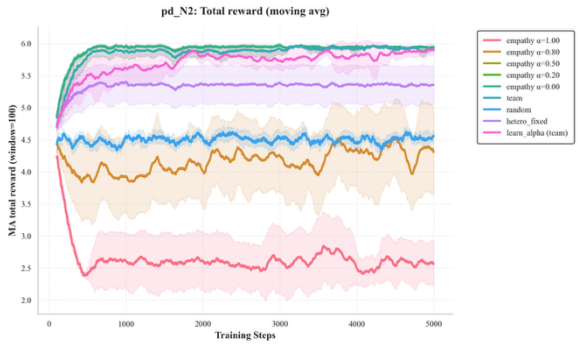

Figure 3: *PD (N = 2): total reward. Empathy simultaneously maximizes social welfare and cooperation.*

Figure 3 reports total reward, defined as the sum of selfish rewards 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.

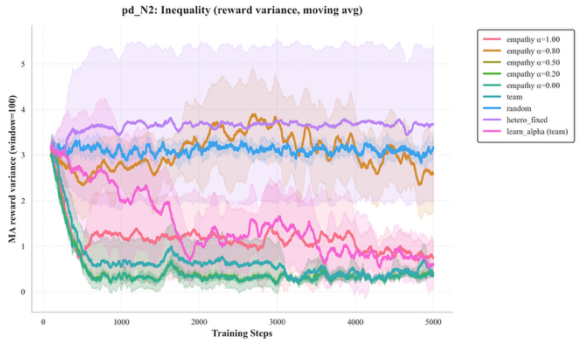

Figure 4: *PD (N = 2): inequality. Increasing empathy sharply reduces reward variance.*

Figure 4 shows reward inequality, measured as the variance of individual rewards. Inequality increases sharply with selfishness, reflecting asymmetric outcomes during exploitation and defection. High empathy compresses reward variance, leading to both cooperative and equitable outcomes. This confirms that in the two-agent setting, empathy simultaneously improves efficiency and fairness.

Learned Empathy

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

Figure 5 reports the evolution of learned empathy when α is optimized jointly with policy learning. Despite maximal empathy (α = 0) being globally optimal across cooperation, welfare, and inequality, learning dynamics consistently converge to intermediate values of α. This indicates that even in the simplest Prisoner's Dilemma setting, standard learning dynamics struggle to discover and maintain the globally optimal empathic regime.

This two-agent case therefore establishes a clear baseline: empathy is unambiguously beneficial, but learning it endogenously remains difficult even in the absence of coordination complexity.

4.2 Prisoner's Dilemma (N = 4)

We now turn to the four-agent Prisoner's Dilemma, where coordination becomes substantially more challenging. In the four-agent setting, each agent plays a Prisoner's Dilemma game against each of the other three agents in every episode, using the standard payoff matrix. Mutual cooperation yields a reward of 3 to both agents, unilateral defection yields 5 to the defector and 0 to the cooperator, and mutual defection yields 1 to both. Each agent therefore participates in three simultaneous pairwise interactions per episode. The resulting rewards are summed and normalized by the number of interactions to maintain comparability with the two-agent case.

The action space is binary, with each agent choosing either to cooperate (C) or defect (D) at every episode. Although the Prisoner's Dilemma is formally stateless, Q-learning requires a state representation. We therefore provide agents with a synthetic state consisting of the joint action vector from the previous episode. This does not alter the strategic structure of the game, but allows agents to condition their behavior on recent collective outcomes, which helps stabilize learning in the multi-agent setting.

Each agent is trained using standard Q-learning with an ε-greedy exploration policy. Exploration starts high (ε = 1.0) and decays gradually to a minimum of 0.05. We evaluate fixed empathy regimes with α ∈ {1.0, 0.8, 0.5, 0.2}, where lower values correspond to greater empathy in the shaped reward.

A key feature of the four-agent Prisoner's Dilemma is that the fully cooperative outcome is particularly sensitive to unilateral deviations. An episode is classified as strictly cooperative only if all four agents choose C simultaneously. A single defection by any agent collapses the cooperative outcome for the entire group, making coordination substantially more difficult than in the two-agent case.

Cooperation Dynamics

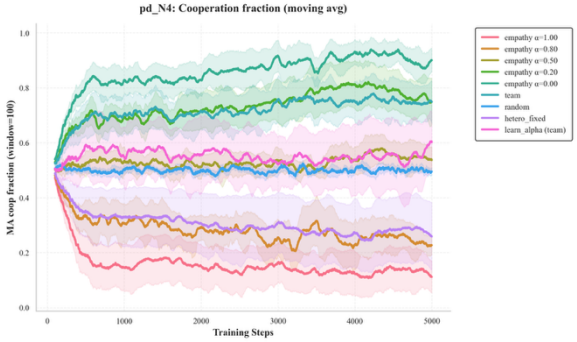

Figure 6: *PD (N = 4): cooperation fraction. Empathy dominates across all levels of cooperation.*

Figure 6 reports the cooperation fraction, defined as the moving average of episodes in which all

four agents choose to cooperate. Selfish agents (α = 1.0) rapidly converge to near-zero cooperation, reproducing the mutual defection Nash equilibrium. Weak empathy (α = 0.8) does not qualitatively change this behavior, as cooperation remains unstable and quickly collapses 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 80–90% 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 in the multi-agent setting.

Strict Cooperation and Welfare

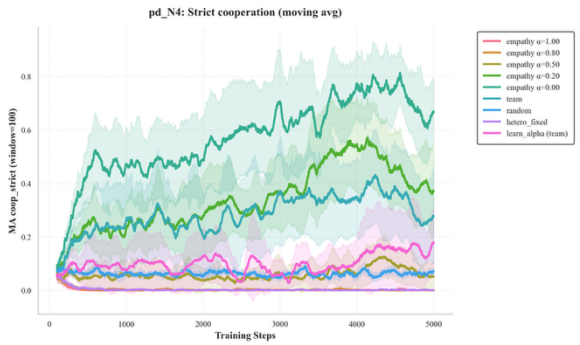

Figure 7: *PD (N = 4): strict cooperation. Selfish agents fail to sustain cooperation at scale.*

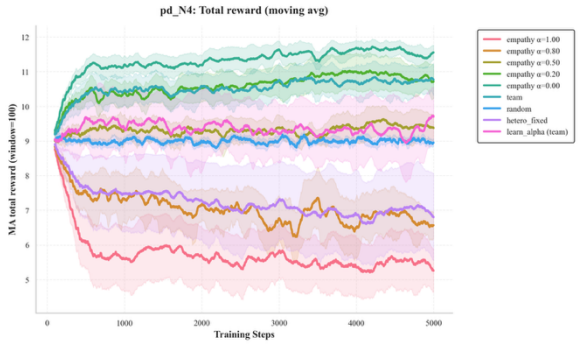

Figure 8: *PD (N = 4): total reward. Empathy maximizes welfare, while selfishness yields the worst outcomes.*

Figure 7 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. 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 (α = 1.0) stabilize near the low-welfare defection regime, while weak empathy (α = 0.8) yields only marginal improvements. Moderate empathy (α = 0.5) achieves intermediate welfare levels, whereas strong empathy (α = 0.2) consistently approaches the social optimum. These findings confirm that empathy reshapes the effective incentives of the game. By incorporating others' rewards into the learning signal, empathic agents shift away from individually rational but collectively inefficient strategies toward high-welfare cooperative behavior.

Inequality and Learned Empathy

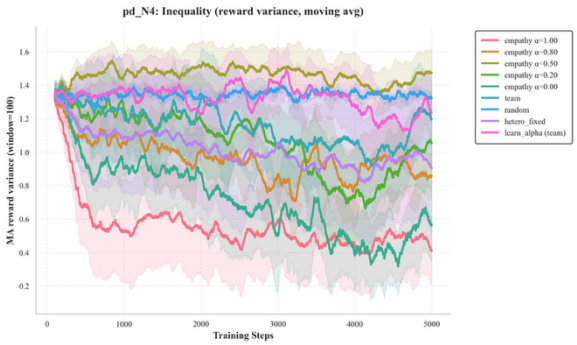

Figure 9: *PD (N = 4): inequality. Reward variance increases monotonically with selfishness.*

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

Figure 9 shows that reward inequality, measured as the variance of individual rewards, increases monotonically with selfishness. Strong empathy substantially compresses payoff disparities, reflecting more equitable and coordinated outcomes. Finally, Figure 10 reports the evolution of learned empathy when α is optimized jointly with the policy. Despite strong empathy yielding the best overall outcomes, learning dynamics converge to intermediate values of α rather than to full empathy.

This pattern indicates that the limitation lies not in empathy itself, but in the learning dynamics: gradient-based optimization struggles to discover and sustain the globally optimal empathic regime in a non-stationary multi-agent environment.

5 Stag Hunt (N = 4)

Environment and Game Structure

The Stag Hunt is a well-known coordination game in which agents face a trade-off between a safe, individually reliable action and a risky action that yields higher payoffs only under successful coordination. Unlike the Prisoner's Dilemma, cooperation in the Stag Hunt is not exploited by defection. Instead, failure arises from miscoordination.

We consider a four-agent Stag Hunt implemented through pairwise interactions. In each episode, every agent plays a Stag Hunt game with each of the other three agents. The pairwise payoff matrix is given by: 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. Similar to the Prisoner's Dilemma environment, pairwise rewards are summed and normalized so that episode-level rewards remain comparable 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: a single agent choosing Hare collapses the high-payoff outcome for the entire group.

In principle the environment is stateless, but agents are provided with a synthetic state equal to the joint action profile from the previous episode. This allows agents to condition their behavior on recent coordination outcomes without altering the structure of the game. The learning setup and empathy-weighted reward shaping are identical to the Prisoner's Dilemma experiments.

Cooperation and Coordination Dynamics

Figure 11: *SH (N = 4): cooperation fraction. Empathy improves coordination on the payoff-dominant equilibrium.*

Figure 11 reports the cooperation fraction, defined as the moving average of episodes in which all four agents choose Stag. Selfish agents ($\alpha = 1.0$) fail to coordinate on the payoff-dominant equilibrium and remain near zero cooperation, converging instead to the risk-dominant all-Hare outcome. Moderate empathy ($\alpha = 0.8$) produces a sharp transition during training: once coordination emerges, agents rapidly converge to the cooperative equilibrium and stabilize at cooperation levels around 80–90%. On the contrary, stronger empathy does not improve outcomes. For $\alpha \leq 0.5$, cooperation remains unstable or highly volatile, with coordination repeatedly disrupted by small deviations.

Strict Cooperation

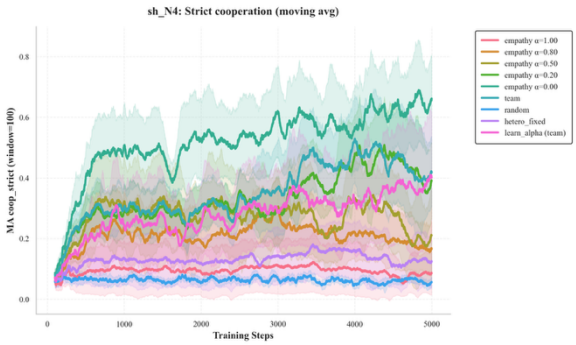

Figure 12: *SH (N = 4): strict cooperation. High empathy promotes stable coordination.*

Figure 12 shows strict cooperation, defined as the fraction of episodes in which all four agents choose Stag, capturing the stability of coordination on the payoff-dominant equilibrium. Selfish agents ($\alpha = 1.0$) almost never achieve strict cooperation, remaining trapped in the risk-dominant all-Hare outcome. Moderate empathy ($\alpha = 0.8$) yields the highest and most stable strict cooperation, with agents coordinating on (S,S,S,S) in a majority of episodes after convergence. In contrast, stronger empathy ($\alpha \leq 0.5$) produces unstable coordination; although high cooperation is occasionally reached, outcomes exhibit large fluctuations and repeated collapses. This indicates that in the Stag Hunt, coordination stability is the primary challenge. Empathy must be strong enough to overcome risk dominance but not so strong that it destabilizes learning, with moderate empathy achieving this balance.

Welfare and Stability

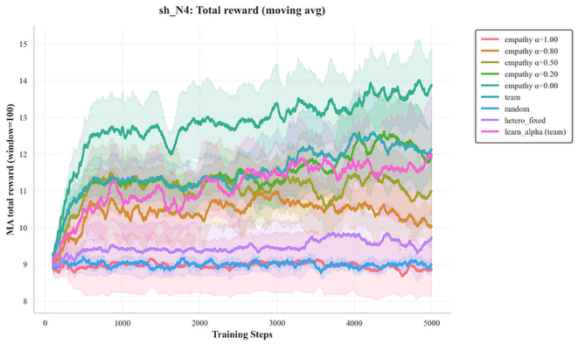

Figure 13: *SH (N = 4): total reward. Empathy yields the highest collective payoff.*

Figure 13 models the total reward across training. The welfare dynamics mirror the cooperation results. Selfish agents stabilize near a total reward of approximately 8, corresponding to the all-Hare equilibrium. Moderate empathy ($\alpha = 0.8$) achieves the highest long-run welfare, frequently approaching the theoretical maximum of 16, indicating extended periods of full coordination on Stag. For $\alpha = 0.5$, total reward stabilizes around intermediate values, reflecting partial and inconsistent coordination. For $\alpha = 0.2$, welfare reaches high levels intermittently but exhibits large downward spikes, resulting in lower average performance despite high peak values. This instability reflects the fragility of the cooperative equilibrium under excessive sensitivity to others' rewards.

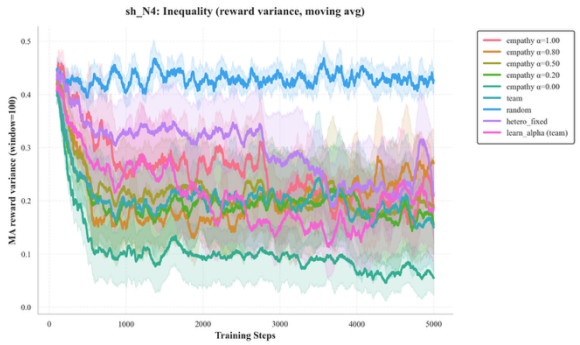

Figure 14: *SH (N = 4): inequality. Empathy compresses payoff disparities without sacrificing welfare.*

Figure 14 shows reward inequality. As in the Prisoner's Dilemma, increased empathy compresses payoff disparities. However, in the Stag Hunt, reduced inequality alone is insufficient to guarantee high welfare. Stable coordination on the payoff-dominant equilibrium requires both aligned incentives and a sufficiently clean learning signal.

The Stag Hunt reveals a non-monotonic relationship between empathy and performance that contrasts sharply with the Prisoner's Dilemma. In PD, increasing empathy monotonically improves cooperation and welfare. In the Stag Hunt, moderate empathy is optimal.

6 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. 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.

Resource Collapse and Stability

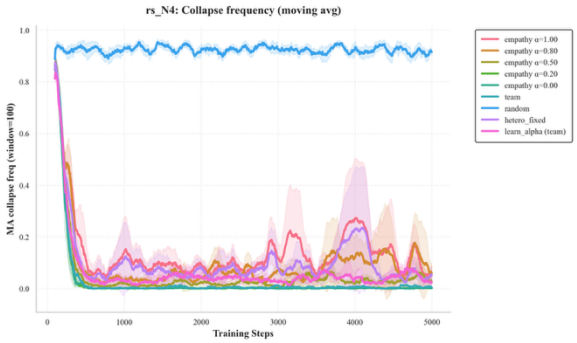

Figure 15: *RS (N = 4): collapse frequency. Empathy nearly eliminates collapses, while selfishness induces systemic failure.*

Figure 15 reports the collapse frequency, defined as the fraction of episodes in which total extraction exceeds the available stock. Selfish agents (α = 1.0) exhibit frequent collapses early in training, reflecting aggressive over-extraction during exploration. Although collapse frequency decreases over time, selfish agents continue to experience occasional systemic failures throughout training. Introducing empathy dramatically alters this behavior. Even moderate

empathy levels (α = 0.8 and α = 0.5) almost entirely eliminate collapses after a short transient period. Highly empathic agents (α = 0.2) also prevent collapse, but exhibit small bursts of instability associated with fluctuations in extraction behavior.

Stock Dynamics and Coordinated Restraint

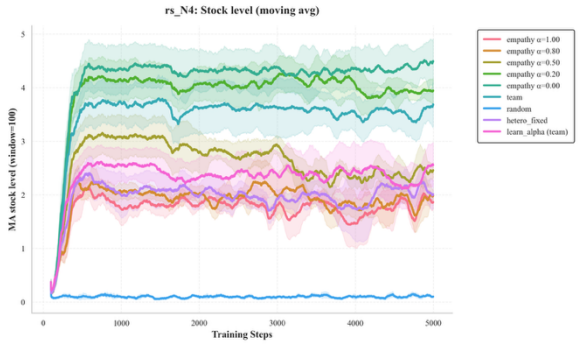

Figure 16: *RS (N = 4): stock level. Empathy sustains higher resource levels over time.*

Figure 16 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. Moderate empathy sustains the resource near its maximum capacity with low variance, indicating consistent restraint across agents. High empathy achieves similarly high stock levels but with greater volatility.

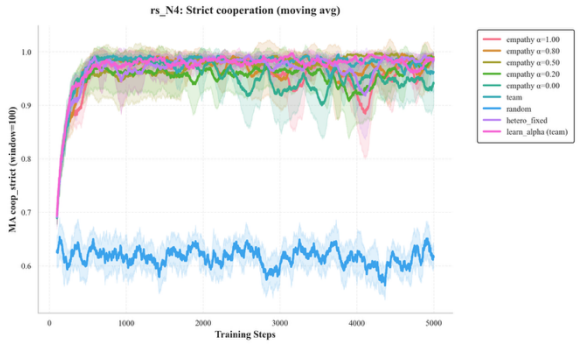

Figure 17: *RS (N = 4): strict cooperation. Empathy promotes coordinated restraint in resource extraction.*

Figure 17 demonstrates strict cooperation, defined as episodes in which total extraction remains within the sustainable threshold. All empathy regimes eventually reach high levels of strict cooperation, but convergence speed and stability differ substantially. Moderate empathy produces the fastest and smoothest convergence to near-perfect coordination. In contrast, very high empathy exhibits oscillations, with temporary drops in strict cooperation despite maintaining high long-run averages.

Welfare and Inequality

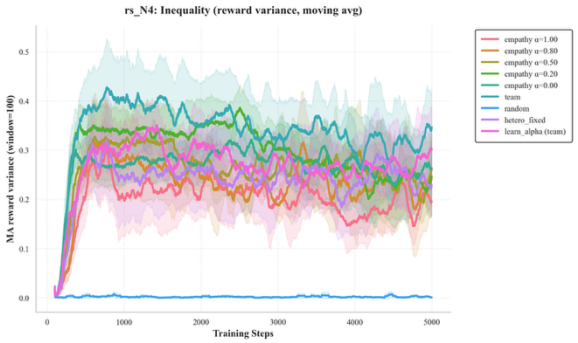

Figure 18: *RS (N = 4): inequality. Empathy reduces payoff dispersion while stabilizing the commons.*

Figure 18 reports reward inequality, measured as the variance of selfish rewards across agents. As in the other environments, increased empathy compresses payoff disparities. Highly empathic agents exhibit the lowest inequality, reflecting synchronized extraction behavior and shared access to the resource.

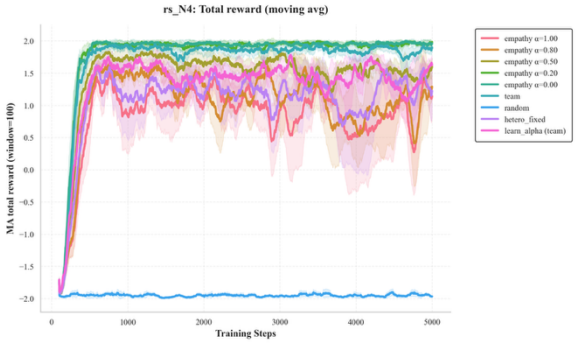

Figure 19: *RS (N = 4): total reward. Intermediate selfishness can maximize reward, revealing a stability - efficiency trade-off.*

Figure 19 shows total reward. Unlike in the Prisoner's Dilemma and Stag Hunt, maximal empathy does not maximize collective welfare. Instead, intermediate selfishness achieves the highest average reward. Highly empathic agents extract conservatively to maintain high stock levels, prioritizing long-term stability over immediate gain. While this prevents collapse, it limits short-term payoff. Moderate empathy strikes a balance, sustaining the resource while allowing agents to extract efficiently.

Unlike the Prisoner's Dilemma, empathy is not required to escape a low-welfare equilibrium. Instead, it governs a trade-off between efficiency and stability. Moderate empathy provides the best balance, while extreme empathy sacrifices short-term reward for robustness.

7 Comparative Analysis Across Environments

The three environments illustrate distinct mechanisms by which empathy interacts with multi-agent learning, as summarized in the following table:

Environment	Primary Failure Mode	Effect of Empathy	Optimal Regime
Prisoner's Dilemma	Exploitation	Monotonic improvement in cooperation, welfare, equality	High empathy
Stag Hunt	Miscoordination	Non-monotonic; stabilizes coordination only at intermediate levels	Moderate empathy
Resource Sharing	Over-extraction	Improves stability but trades-off efficiency	Intermediate empathy

8 Conclusion

Empathy influences multi-agent learning in ways that depend on the environment's structure. While increasing empathy monotonically improves outcomes in exploitation-based dilemmas, it induces non-monotonic trade-offs in coordination and resource-sharing settings. Furthermore, learning dynamics consistently fail to converge to the globally optimal empathic regime.

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

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

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

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