Modeling Cooperation in Heterogeneous Multi-Agent Systems

Abstract

Modeling the complexity of global economics and politics requires capturing the diverse interactions of real-world actors with varying capabilities. To address this, we introduce a framework for running Multi-Agent Reinforcement Learning scenarios with heterogeneous agents. These agents vary in learning rates, risk tolerances, and observational abilities, allowing them to more accurately represent the diverse behaviors of real-world actors. Preliminary results suggest that sustainable cooperation can emerge even in the presence of significant asymmetries in agent capabilities. This highlights the potential of advanced computational models to help inform policies that improve global economic and political collaboration.

1 Introduction

Cooperation has been the key factor allowing humans to flourish Hare and Woods [2020], Bar-Yam [2000]. The complexity of modern societies often surpasses our evolutionary capacity for effective coordination, leading to social dilemmas where individual incentives and collective welfare are misaligned. These dilemmas manifest themselves in various domains, from the struggle to implement effective climate policies Kersting [2017], to the difficulties in agreeing to resolutions in order to deal with pandemics in a timely manner Glennerster et al. [2024]. Coordination problems are also prevalent in politics, where divergent interests and asymmetric information cause a misalignment of priorities among stakeholders Frieden and Lake [2015], leading to policy paralysis or, in extreme cases war Dafoe et al. [2020]. Similarly, in economics cooperation is often undermined by institutional failures Acemoglu and Robinson [2012]. Institutions may fail either because some members exploit others or because a lack of trust prevents cooperation, despite mutual cooperation maximizing collective welfare.

Recent advancements in economic modeling have increasingly focused on utilizing Markov Decision Processes (MDPs) to address these coordination problems Glennerster et al. [2024], Charpentier et al. [2021]. MDPs, coupled with Reinforcement Learning (RL) techniques like Temporal Difference (TD) learning, offer a promising framework for capturing the dynamic interactions among agents. In particular, these approaches allow for modeling strategic behavior where agents act based on current states without full knowledge of future outcomes. Agents in these models can adjust their strategies in response to the evolving economic landscape and other agents, much like governments, firms, and consumers in the global economy.

In this work, we introduce the novel capability of running TD simulations with heterogeneous agents. This method allows for the exploration of scenarios where heterogeneity—whether in terms of access to information, risk tolerance, or market reputation—plays a pivotal role in shaping outcomes. By simulating environments that reflect these disparities, we can better understand how different agent characteristics influence overall market dynamics and the potential for cooperation.

To model these different interaction scenarios, we utilize general-sum game environments, where the total payoff for all participants does not necessarily sum to zero. In contrast to zero-sum games, where one agent's gain is exactly another agent's loss, general-sum games allow for a range of outcomes

Preprint. Under review.

in which cooperation or competition can emerge. In these settings, cooperation may arise even in situations where agents have incentives to act in their own self-interest Barfuss [2022]. Over time, self-interested strategies can converge toward a common equilibrium that benefits both sides, even in the absence of a clear immediate incentive to cooperate, as is characteristic of general-sum games.

This paper is still a work-in-progress, and our current focus has primarily been on adjusting the agents' varying capacities to observe and react to their environment. Our early findings suggest that cooperation between heterogeneous agents that have a knowledge imbalance is feasible. This preliminary result indicates that even in games with heterogeneous agents, there is potential for cooperation, and it points to the broader applicability of these techniques in shaping more equitable economic outcomes.

2 Methods

2.1 Learning Algorithm

We employ the deterministic Expected SARSA Sutton and Barto [2018] algorithm to simulate the interactions of heterogeneous agents within a general-sum game scenario. Here we build upon previous work by Barfuss and Meylahn [2023].

Our framework considers two agents, $i \in \{1, 2\}$, where each agent selects from a set of possible actions, $a^i \in \{c, d\}$, representing a cooperating or defecting action, respectively. The joint action vector is denoted as $\mathbf{a} = \{a^1, a^2\}$. In the traditional SARSA algorithm the state $s \in S$ is shared by both agents and together with the agent's actions, determines the reward $r^i(s, \mathbf{a})$ received by agent i and the transition to the next state $s' \in S$.

However in our project, we then model the differences in state observability among agents by introducing an observed state $\tilde{s} \in S$. This clarifies the fact that one agent might not be observing the same state as another agent, even if they co-exist in the same environment.

At each time step t, agent i selects action a with a probability $x_i^t(a|\tilde{s})$, a frequency distribution over actions dependent on the observed state $\tilde{s}$. The action selection is guided by an ϵ-greedy exploration strategy, where the agent chooses the action with the highest state-action value with probability $1 - \epsilon_i$. Next, the agent selects an action at random with probability ϵ_i to encourage exploration. Specifically, for the two-action case, the action probabilities are defined as:

$$x_t^i(c|\tilde{s}) = \begin{cases} 1 - \frac{\epsilon_i}{2} & \text{if } q_t^i(c, a) > q_t^i(d, a), \\ \frac{\epsilon_i}{2} & \text{otherwise,} \end{cases} \tag{1}$$

where $q_i^t(a, \tilde{s})$ represents the state-action value function for agent i, updated according to the temporal difference learning rule.

The state-action values are updated after each time step as follows:

$$q_{t+1}^i(\tilde{s}_t, a_t) = (1 - \alpha_i) q_t^i(\tilde{s}_t, a_t) + \alpha \left[r_t^i + \delta \sum_a x_t^i(a|\tilde{s}_{t+1}) q_t^i(\tilde{s}_{t+1}, a) \right], \tag{2}$$

where $\alpha_i \in [0, 1]$ is the learning rate and $\delta_i \in [0, 1]$ is the discount factor for agent i, which controls how much weight the agent places on future rewards. These parameters can be set differently for each agent to reflect individual differences in learning rates and future reward valuations.

Furthermore, the varying ϵ_i and δ_i parameters allow us to simulate agents with differing boldness in their decision-making processes, creating a parallel with the risk tolerance concept observed in real-world markets.

2.2 Environment

The environment we utilize to test our method is based on the Iterated Prisoner's Dilemma, a well-known framework for examining cooperation. This model serves as a representative example of how individual incentives can conflict with collective welfare, as it captures the tension between

cooperating for mutual benefit and defecting for individual gain. The reward matrix used for the two-agent case is given specified in the Appendix in Table 1.

For the model simulations, we utilize the framework of Multi-Agent Reinforcement Learning (MARL). While powerful, agent based models and Deep Reinforcement Learning (DRL) systems have significant disadvantages for improving the understanding of the learning agent's emergent collective behavior: they are noisy, sometimes hard to explain, sample-inefficient, and computationally intense. We, therefore, employ an approach to MARL inspired by complex system science and evolutionary game theory: collective reinforcement learning dynamics (CRLD) Barfuss [2023].

CRLD is characterized by two types of idealization. First, CRLD uses dynamic learning equations as a computationally efficient alternative to traditional RL updates, allowing for simulations to be run on personal computers without the need for extensive computational resources. Second, CRLD can be used to study the principles behind emergent collective behavior in idealized, low-dimensional environments.

3 Results

For our initial experiments, we chose to use the IPD with a two-memory period as a test bed. Our approach leverages the concept of direct reciprocity in cooperation, as theorized by Schmid et al. [2021], in which agents base their cooperative decisions on personal historical interactions.

In a first experiment with 5,000 iterations we reliably found that agents that are more knowledgeable receives more rewards than less knowledgeable agents. See appendix for more details.

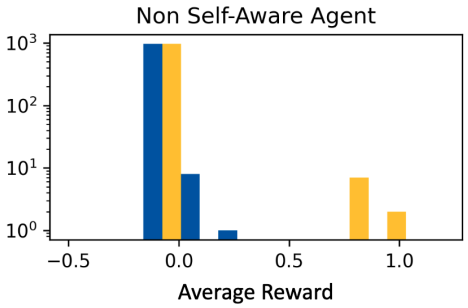

Figure 1: Distribution of average rewards for non self-aware agents. Agent 1 (blue) has complete observability while Agent 2 (yellow) has partial observability.

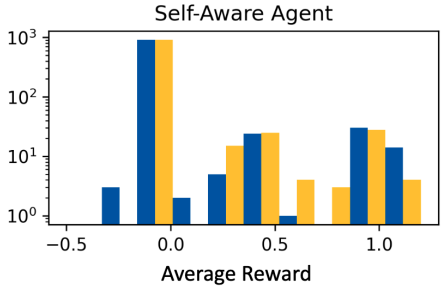

Figure 2: Distribution of average rewards for self-aware agents. Agent 1 (blue) has complete observability while Agent 2 (yellow) has partial observability.

In our second experimental setup involving 8,000 iterations with two agents. Agent 1 is always able to completely observe its environment, it has perfect memory and perfect understanding. Agent 2's observational capabilities varied across two conditions: in the first, Agent 2 could observe the full market in the first game but in the second it loses the ability to observe its own actions (Table 2). In the second condition, Agent 2 retained knowledge of its own actions but could not observe Agent 1's states (Table 3).

Our results, reveal striking differences in cooperative outcomes based on these observational asymmetries. Figure 1 shows the distribution of average rewards for the "Non Self-Aware Agent" scenario. We observe a strong tendency toward non-cooperative outcomes, with most interactions yielding low or zero rewards, showing that when an agent lack awareness of its own actions the results for cooperation are catastrophic. Figure 2 illustrates the outcomes for the case in which Agent 2 is a self-aware agent. This graph displays a broader distribution of rewards, with significant occurrences at higher values, suggesting that self-awareness facilitates better strategic decision-making and increases cooperation.

In a third experiment, an agent's perception of the environment influences their roles as either exploiters or the exploited. When Agent 2 is only aware of instances when Agent 1 cooperates (Table 4), it tends to overestimate its own appropriate level of cooperation. This miscalculation leads to

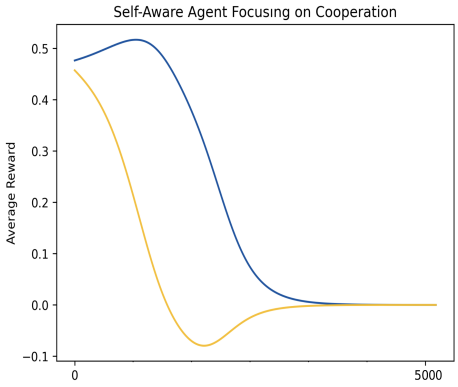

Figure 3: Average reward trends with Agent 2 (yellow) with partial observability focusing on cooperation and being exploited by Agent 1 (blue) with full observability.

Figure 4: Average reward trends showing Agent 2 (yellow) with partial observability focusing on defection and exploiting Agent 1's (blue) despite Agent 1's greater awareness.

Agent 2 being exploited by Agent 1, as illustrated in Figure 3. Conversely, when Agent 2 is only aware of Agent 1 defections (Table 5), it becomes overly skeptical, choosing to exploit Agent 1.

4 Future work

Moving forward, we aim to further test our framework in a variety of strategic interaction settings, such as public goods games, cooperative bargaining and common pool resource games. We also hope to further analyze the strategies developed by these agents to gain deeper insights in the development of cooperation through the simple, low-dimensional scenarios typical of RL. Building on this understanding, we plan to integrate DRL algorithms into our stack to enhance our model's capability to handle more complex environments and possibly incorporate Large Language Model agents. We hope that this could make our models more realistic in order to aid policymakers in crafting more informed and effective policies. Our ultimate goal is to develop a comprehensive suite of tools that combine MDP and game-theoretical approaches to better support decision-making processes in economics and public policy.

5 Limitations

The IPD framework, by design, encapsulates a limited set of possible interactions and may not capture the full spectrum of strategic behaviors that agents exhibit in more nuanced scenarios. Moreover, the computational model assumes that all agents rationalize and make decisions based solely on predefined rules and their perception of the environment, which might not accurately reflect the often irrational and unpredictable nature of human decision-making. The model's outcomes are also highly dependent on the parameters chosen for the simulations, such as learning rates and discount factors, which might limit the generalizability of the findings to real-world applications, where such parameters are not as easily quantifiable or consistent.

6 Acknowledgments

This project was supported by Impact Academy and the University of Bonn. The authors don't have any conflict of interest to disclose.

References

Daron Acemoglu and James A. Robinson. *Why nations fail: The origins of power, prosperity, and poverty*. Currency, New York, 2012.

Yaneer Bar-Yam. Complexity rising: From human beings to human civilization, a complexity profile. 2000.

Wolfram Barfuss. Dynamical systems as a level of cognitive analysis of multi-agent learning. *Neural Computing and Applications*, 34(3):1653–1671, 2022. ISSN 1433-3058. doi: 10.1007/s00521-021-06117-0. URL `https://doi.org/10.1007/s00521-021-06117-0`.

Wolfram Barfuss. pyCRLD: Collective reinforcement learning dynamics in python, 2023. URL `https://wbarfuss.github.io/pyCRLD/`.

Wolfram Barfuss and Janusz M. Meylahn. Intrinsic fluctuations of reinforcement learning promote cooperation. *Scientific Reports*, 13(1):1309, 2023.

Arthur Charpentier, Romuald Elie, and Carl Remlinger. Reinforcement learning in economics and finance. *Computational Economics*, pages 1–38, 2021.

Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R McKee, Joel Z Leibo, Kate Larson, and Thore Graepel. Open problems in cooperative ai. *arXiv preprint arXiv:2012.08630*, 2020.

Jeffry A. Frieden and David A. Lake. *World Politics: Interests, Interactions, Institutions*. W. W. Norton & Company, New York, 2015.

Rachel Glennerster et al. Quantifying the social value of a universal covid-19 vaccine and incentivizing its development. Working Paper 32059, National Bureau of Economic Research, 2024.

Brian Hare and Vanessa Woods. Survival of the friendliest. *Scientific American*, Aug 1 2020.

Jan Kersting. *Stability of cooperation in the international climate negotiations: An analysis using cooperative game theory*. KIT Scientific Publishing, 2017.

Laura Schmid, Krishnendu Chatterjee, Christian Hilbe, and Martin A. Nowak. A unified framework of direct and indirect reciprocity. *Nature Human Behaviour*, 5(10):1292–1302, 2021.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: An introduction*. MIT Press, 2018.

A Appendix / supplemental material

A.1 Software

Code will be made available upon publication. A link to the repository will be provided in the camera-ready version.

A.2 Parameters

For all experiments a constant learning rate of 0.1 and discount factor of 0.9 were used. Though it's possible to change these values as we have noted, in the interest of space we only presented findings maintaining these constant.

A.3 Figures

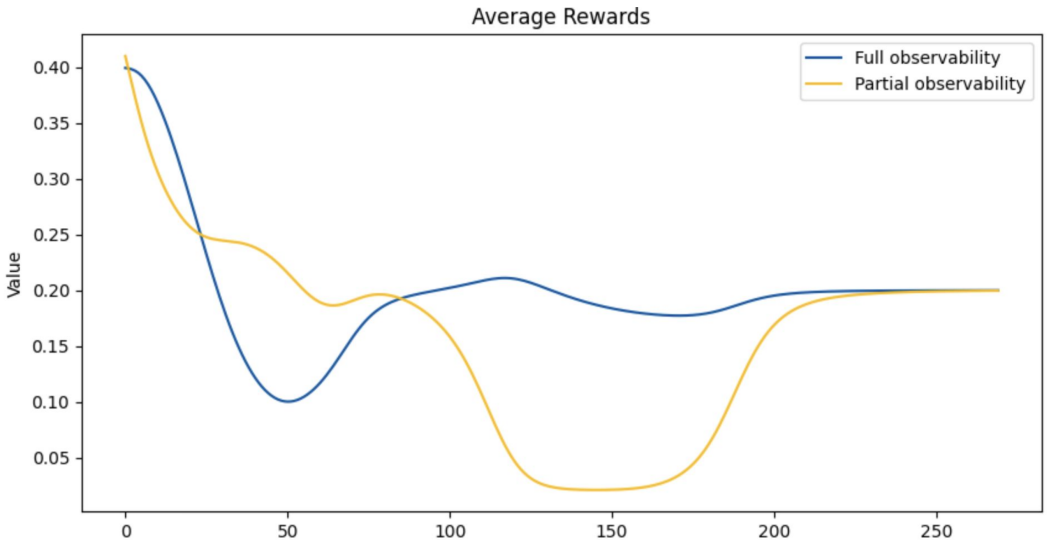

Figure 5: Average reward for agents within mixed observability environments, ran over 5000 iterations. We observe that more knowledgeable agents, receive better rewards on the long run than less knowledgeable agents. It is also clear from our analysis that exploitation is not a dominating behavior as agents tend towards mostly defecting or oscillating between cooperating and defecting.

A.4 Tables

Table 1: Payoff Matrix for IPD

	c	**d**
c	(1, 1)	(1.2, -0.5)
d	(-0.5, 1.2)	(0, 0)

Table 2: 1st Experiment: States observed by agents in two iterations of the Prisoner's Dilemma

Agent 1, Game 1	Agent 2, Game 1	Agent 1, Game 2	Agent 2, Game 2
(C, C)	(C, C)	(C, C)	(C, *)
(C, C)	(C, D)	(C, C)	(C, *)
(C, C)	(D, C)	(C, C)	(C, *)
(C, C)	(D, D)	(C, C)	(C, *)
(C, D)	(C, C)	(C, D)	(C, *)
(C, D)	(C, D)	(C, D)	(C, *)
(C, D)	(D, C)	(C, D)	(C, *)
(C, D)	(D, D)	(C, D)	(C, *)
(D, C)	(C, C)	(D, C)	(D, *)
(D, C)	(C, D)	(D, C)	(D, *)
(D, C)	(D, C)	(D, C)	(D, *)
(D, C)	(D, D)	(D, C)	(D, *)
(D, D)	(C, C)	(D, D)	(D, *)
(D, D)	(C, D)	(D, D)	(D, *)
(D, D)	(D, C)	(D, D)	(D, *)
(D, D)	(D, D)	(D, D)	(D, *)

Table 3: 2nd Experiment: States observed by agents in two iterations of the Prisoner's Dilemma

Agent 1, Game 1	Agent 2, Game 1	Agent 1, Game 2	Agent 2, Game 2
(C, C)	(C, C)	(C, C)	(*, C)
(C, C)	(C, D)	(C, C)	(*, D)
(C, C)	(D, C)	(C, C)	(*, C)
(C, C)	(D, D)	(C, C)	(*, D)
(C, D)	(C, C)	(C, D)	(*, C)
(C, D)	(C, D)	(C, D)	(*, D)
(C, D)	(D, C)	(C, D)	(*, C)
(C, D)	(D, D)	(C, D)	(*, D)
(D, C)	(C, C)	(D, C)	(*, C)
(D, C)	(C, D)	(D, C)	(*, D)
(D, C)	(D, C)	(D, C)	(*, C)
(D, C)	(D, D)	(D, C)	(*, D)
(D, D)	(C, C)	(D, D)	(*, C)
(D, D)	(C, D)	(D, D)	(*, D)
(D, D)	(D, C)	(D, D)	(*, C)
(D, D)	(D, D)	(D, D)	(*, D)

Table 4: 3rd Experiment: States observed by agents in two iterations of the Prisoner's Dilemma

Agent 1, Game 1	Agent 2, Game 1	Agent 1, Game 2	Agent 2, Game 2
(C, C)	(C, C)	(C, C)	(C, C)
(C, C)	(C, D)	(C, C)	(C, D)
(C, C)	(D, C)	(C, C)	(*, C)
(C, C)	(D, D)	(C, C)	(*, D)
(C, D)	(C, C)	(C, D)	(C, C)
(C, D)	(C, D)	(C, D)	(C, D)
(C, D)	(D, C)	(C, D)	(*, C)
(C, D)	(D, D)	(C, D)	(*, D)
(D, C)	(C, C)	(D, C)	(C, C)
(D, C)	(C, D)	(D, C)	(C, D)
(D, C)	(D, C)	(D, C)	(*, C)
(D, C)	(D, D)	(D, C)	(*, D)
(D, D)	(C, C)	(D, D)	(C, C)
(D, D)	(C, D)	(D, D)	(C, D)
(D, D)	(D, C)	(D, D)	(*, C)
(D, D)	(D, D)	(D, D)	(*, D)

Table 5: 4th Experiment: States observed by agents in two iterations of the Prisoner's Dilemma

Agent 1, Game 1	Agent 2, Game 1	Agent 1, Game 2	Agent 2, Game 2
(C, C)	(C, C)	(C, C)	(*, C)
(C, C)	(C, D)	(C, C)	(*, D)
(C, C)	(D, C)	(C, C)	(D, C)
(C, C)	(D, D)	(C, C)	(D, D)
(C, D)	(C, C)	(C, D)	(*, C)
(C, D)	(C, D)	(C, D)	(*, D)
(C, D)	(D, C)	(C, D)	(D, C)
(C, D)	(D, D)	(C, D)	(D, D)
(D, C)	(C, C)	(D, C)	(*, C)
(D, C)	(C, D)	(D, C)	(*, D)
(D, C)	(D, C)	(D, C)	(D, C)
(D, C)	(D, D)	(D, C)	(D, D)
(D, D)	(C, C)	(D, D)	(*, C)
(D, D)	(C, D)	(D, D)	(*, D)
(D, D)	(D, C)	(D, D)	(D, C)
(D, D)	(D, D)	(D, D)	(D, D)