Reputation-Filtered Reward Reshaping: Encouraging Cooperation in High-Dimensional Semi-Cooperative Multi-agent Settings

Hassan Raissouni
Ai Movement, Mohammed VI Polytechnic University
Rabat, Morocco
hassan.raissouni@um6p.ma

Wissal Bekhti
Ai Movement, Mohammed VI Polytechnic University
Rabat, Morocco
wissal.bekhti@um6p.ma

Btissam El Khamlichi
Ai Movement, Mohammed VI Polytechnic University
Rabat, Morocco
btissam.elkhamlichi@um6p.ma

Amal El Fallah Seghrouchni
Ai Movement, Mohammed VI Polytechnic University
Rabat, Morocco
amal.elfallah-seghrouchni@um6p.ma

ABSTRACT

In semi-cooperative settings, cooperation is induced by appropriate incentives that align individual agents' goals with a common objective. The primary challenge is balancing personal and collective goals, which introduces new complications. A key issue is that cooperating with all agents equally can result in poor decisions, suboptimal cooperation, and inefficiencies in task execution. Furthermore, agents must manage the trade-off between staying connected to share cooperation-related information and pursuing their own objectives. To tackle these issues, we propose a novel framework incorporating a filtered reward-reshaping mechanism with two main components: (1) a reputation system that evaluates trust and competency, allowing agents to assess and filter peers' contributions, collaborate with reliable partners, and improve learning efficiency, and (2) a density-focused Potential-Based Reward Shaping (PBRS) mechanism that promotes connectivity and encourages exploration by adjusting rewards based on the density of agents in the observable space. Our approach was tested against PED-DQN and Independent Q-Learners, demonstrating enhanced performance in high-dimensional semi-cooperative environments. Additionally, theoretical stability analysis confirmed the system's convergence to a desirable equilibrium, ensuring long-term stability.

KEYWORDS

Multi-agent system; Reinforcement Learning; Reputation; Trust; Cooperation; Reward Reshaping.

ACM Reference Format:
Hassan Raissouni, Wissal Bekhti, Btissam El Khamlichi, and Amal El Fallah Seghrouchni. 2025. Reputation-Filtered Reward Reshaping: Encouraging Cooperation in High-Dimensional Semi-Cooperative Multi-agent Settings. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 9 pages.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

1 INTRODUCTION

One of the main challenges in semi-cooperative multi-agent systems arises when agents, despite exhibiting diverse levels of competence and trustworthiness, are treated as equals. This lack of differentiation can lead to suboptimal outcomes, as agents may blindly follow peers who are less competent or even deceptive, resulting in ineffective learning and coordination. Agents often predict their peers' policies to anticipate or influence their behavior [2, 7, 18, 23]. However, when agents base their decisions on inaccurate predictions from peers, they risk adopting suboptimal policies, leading to inefficiencies in task execution. This is especially prominent in high-dimensional environments where agents have to rely on partial observability. Another challenge arises from the conflict between staying connected to other agents and pursuing individual task objectives [10, 11]. In many semi-cooperative MAS environments, agents need to collaborate by sharing information or coordinating actions, yet they also have individual goals that require independent actions [1, 12, 25]. To overcome these challenges, we propose a Filtered Reward Reshaping approach (FRR) based on two key mechanisms.

First, a novel reputation mechanism that bootstraps cooperation based on trust and competency. This mechanism allows agents to assess and filter their peers' contributions by measuring both their trustworthiness and their ability to enhance the overall system's performance. By filtering agents based on trust and competency, the reputation mechanism addresses the issue of suboptimal cooperation by ensuring that agents collaborate with reliable, competent partners, leading to more efficient and productive learning.

Second, to handle the challenge of performing tasks while maintaining the connectivity of the agents, we introduce a density-focused Potential-Based Reward Shaping (PBRS) mechanism. This mechanism encourages agents to remain connected by shaping their rewards based on the density of other agents within their observable area, promoting proximity and preventing isolation. This enhances cooperation and maintains communication channels, which are essential for information exchange and coordinated action. At the same time, the proposed PBRS incentivizes agents to explore new areas of the environment during the early phases of learning, even if it involves some risk. By balancing these incentives, agents are guided to make strategic decisions that promote

exploration without sacrificing the benefits of staying connected to peers.

Together, the reputation mechanism and the density-focused PBRS contribute to addressing the core challenges not only in developing effective cooperation but also in maintaining connectivity within a semi-cooperative MAS. The stability of the learning process is enhanced through the boundedness of the reshaping functions and Lyapunov stability. This stability is essential for preventing the learning dynamics from becoming chaotic or unpredictable, with Lyapunov stability ensuring that once agents' strategies converge, their behavior remains consistent over time [4, 26].

2 RELATED WORK

Semi-cooperative settings in Multi-agent Reinforcement Learning (MARL) frameworks serve as a middle ground between fully cooperative and non-cooperative frameworks [3, 16]. Previous work has examined mechanisms to promote cooperation in the latter. [7] induces cooperation in a non-cooperative setting by incorporating the anticipated learning of neighboring agents into each agent's policy update. [18] proposes a peer-rewarding mechanism, called gifting, which allows agents to guide each other toward prosocial equilibria. While effective, gifting can inadvertently incentivize selfish behavior, as agents act based on separate reward functions. [8] addresses this limitation through distributed reward reshaping (RS), such that the agent's perception of the equilibrium gears towards optimizing social welfare. Given that prior approaches may not fully account for the possibility that some neighbors' assessments may be unreliable or compromised, we propose filtered reward reshaping (FRR), a novel approach that incorporates intelligent filtering mechanisms to induce cooperation in semi-cooperative multi-agent systems (MAS). In what follows, we outline the state of the art relevant to the proposed work.

Reward reshaping is an effective technique for addressing the sample efficiency issue of Reinforcement Learning by incorporating domain knowledge into additional rewards. [13] aims to improve the convergence speed by adjusting rewards based on a potential function through Look-Back Advice and Look-Ahead Advice. The latter is the first approach that guarantees the policy invariance property. Another relevant approach is Difference Rewards [20], designed for fully cooperative multi-agent systems.

Reputation & **trust** Reputation mechanisms are a bridge between Evolutionary Game Theory and Reinforcement Learning. From the perspective of EGT, reputation can be seen as a form of indirect reciprocity, where an agent's fitness (success of strategy) is affected by how they are perceived by others, leading to strategies that promote good reputations. While there are no universal definitions for trust and reputation, [15] defines the latter as "the opinion or view of one about something," formed and updated over time through direct interactions or shared information. Alternatively, [19] describes reputation as "a peer's belief in another peer's capabilities, honesty, and reliability". [9] defines trust as being multifaceted and integrates four distinct sources of trust information to evaluate an agent's performance and trustworthiness.

Lyapunov stability ensures that the reshaping of rewards in a Markov game does not disrupt the system's equilibrium, preventing instability in learned policies and preserving the theoretical convergence of agents. [24] introduced a Lyapunov stability constraint into the MARL framework to guide policy improvement and ensure stability. [21] used candidate Lyapunov functions to perform a detailed stability analysis. [6] focused on accelerating the MARL training process by shaping rewards based on a Lyapunov function, while [5] worked on constructing a Lyapunov function to guarantee policy stability during learning.

3 MATHEMATICAL FRAMEWORK

Notations: We use the following mathematical notations throughout the paper: i refers to an agent, u to the actions taken by agents, r^b is the base reward of agents, and $\hat{r}$ the reshaped reward. The total number of agents in the system is denoted by ν, the set of agents is denoted by ϑ and (x, y) represents the coordinates of an agent's location in the environment. φ denotes inter-agent assessments, δ is the temporal difference error, c represents agent competency, and t_{ij} is the trust score shared between agents. Let $Ne_i(t)$ be the set of neighbors of agent i at time-step t and $\overline{Ne}_i(t)$ the filtered subset of neighbors. $ne_i = \text{Card}(Ne_i)$ and $\overline{ne}_i = \text{Card}(\overline{Ne}_i)$ represent the cardinality of the set of neighbors of agent i and the filtered subset, respectively. $d_i(t)$ represents the agent density in a part of the environment, specifically the density of neighbors $Ne_i(t)$ observable by agent i.

Dec-POMDP We consider a Decentralised Partially Observable Markov Decision Process, where each agent acts based on its local observations and the information shared by its neighbors. We focus on decentralized training and execution, where the training of agent policies and the policies themselves are fully decentralized between the agents. The Dec-POMDP is defined by the tuple $(S, \{U_i\}_{i=1}^n, T, \{R_i\}_{i=1}^n, \{O_i\}_{i=1}^n, O, \gamma)$. The set of states S contains the agents' positions in the environment, as well as local information on nearby obstacles and peers. Each agent i has a set of actions $U_i = \{\text{up}, \text{down}, \text{left}, \text{right}, \text{stay}\}$ and receives observations $o_i \in O_i$ from a dynamic partially observable space centered around its position. The transition function $T : S \times (U_1 \times U_2 \times \cdots \times U_n) \times S \to [0, 1]$, defining the probability $P(s' \mid s, u_1, u_2, \ldots, u_n)$ of moving to a new state s' given the current state s and actions $u_1, u_2, \ldots, u_n$ of all agents. Each agent receives a reward $R_i : S \times U_i \to \mathbb{R}$ based on their contribution to the system, with $R_i(s, u_i) = r_i$, balancing individual goals with the collective objective of all agents. The observation function $O : S \times U_1 \times U_2 \times \cdots \times U_n \times O_1 \times O_2 \times \cdots \times O_n \to [0, 1]$ provides the likelihood $P(o_1, o_2, \ldots, o_n \mid s, u_1, u_2, \ldots, u_n)$ of each agent's observations given the state and actions taken. Finally, the discount factor $\gamma \in [0, 1]$ ensures future rewards are considered, promoting strategic planning.

4 FRR METHOD OVERVIEW

The proposed reward reshaping method is based on two reshaping functions with two filtering mechanisms. The first reshaping function is based on inter-agent assessments, where agents evaluate the contributions of their peers in order to encourage cooperation. The assessment score is computed using the Temporal Difference (TD) error, which reflects how much each agent's action contributes to the system's performance. Agents' reshaped rewards include their base rewards plus the average assessments from their neighbors, promoting teamwork. These assessments are filtered using

trust and competency: trust reflects consistent cooperative behavior, while competency measures the ability to maximize rewards. A competent agent may not always be trustworthy if it acts selfishly, while a less competent agent can still be trustworthy if it consistently demonstrates willingness to cooperate. This ensures that only agents balancing cooperation and performance influence rewards. The second reward-reshaping function introduces a density-focused PBRS method, designed to further promote cooperation by encouraging agents to stay close to one another. Additionally, during the exploration phase, this reshaping function is filtered to encourage agents to explore under-explored states, aligning exploration with cooperative goals.

4.1 Reward Reshaping through Inter-agent Assessments

We propose an inter-agent assessment score that enables agents to evaluate their peers' contributions to the overall system. This approach creates a collaborative environment, where agents work together to improve performance and address the non-stationarity of the environment by sharing these assessments. The goal is to induce cooperation amongst agents in order to achieve faster learning and convergence. We define this assessment as the TD error of each peer i:

$$\begin{aligned}\varphi_i(t) = r_i^b(t) + \ \gamma \ max_u Q_i(o_i(t+1), u|\theta_i(t)) \\ - Q_i(o_i(t), u_i(t)|\theta_i(t))\end{aligned} \tag{1}$$

with r_i^b being the base reward returned by the environment to agent i, $Q_i(o_i, u_i|\theta_i)$ being the Q-value quantifying the quality of the state-action pair $(o_i(t), u_i(t))$ at time-step t, $o_i(t)$, and $u_i(t)$ the observation of agent i and the action it took at time-step t, and $\theta_i(t)$ are the weights of the neural networks. TD errors can be used as an estimation of the Bellman error, which measures the discrepancy between the predicted and actual rewards. By incorporating the TD error as an inter-agent assessment score, agents can better evaluate the impact of their peers' actions on the system's overall performance, allowing for a more collaborative learning environment.

For each agent, we average the assessments received by neighbors, as shown in 1. The reshaped reward becomes:

$$\hat{r}_i(t) \quad = \quad r_i^b(t) + \frac{1}{ne_i(t)} \sum_{j \in Ne_i(t)} \varphi_j(t) \tag{2}$$

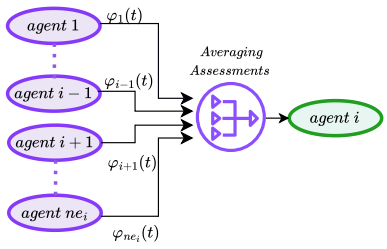


Figure 1: Inter-agent assessments to contribute in peer's reward reshaping

Averaging inter-agent assessments to reshape peers' rewards does not adequately account for the varying reliability of different agents. Even in homogeneous groups, giving equal weight to all assessments—regardless of their quality—can dilute the effectiveness of aggregated feedback, leading to inefficiencies and slowing the learning process. To address this issue, we propose a filtering mechanism for the reshaping function that utilizes agents' reputation scores.

4.2 Introducing a Reputation Mechanism

We propose a reputation mechanism that operates on two key dimensions: **trust** among agents and **competency** of each individual agent. We introduce a novel 2-dimensional trust mechanism that considers not only the level of trust each agent has in its peers but also the potential loss of trust if a peer becomes a Byzantine agent or is subjected to a cyber-attack [22]. We define:

(1) *trust based on mutual winnings:* based on the fact that agents can benefit from cooperation to maximize their own individual rewards.
(2) *trust based on selfishness:* If an agent chooses to conserve its energy and refrain from contributing to the overall system, it is considered untrustworthy by its peers. This decision negatively impacts the trust score that agent $i \in Ne_j$ assigns to agent j.

We represent trust evaluation using individual trust vectors. At each time step t, an agent i monitors and updates the trust scores of up to four neighbors with the highest reputation. To illustrate this concept, we define a square, non-symmetric trust matrix:

$$T_v \quad = \quad \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ t_{21} & t_{22} & \cdots & t_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ t_{n1} & t_{n2} & \cdots & t_{nn} \end{bmatrix} \tag{3}$$

where $T_i = [t_{ij}]_{j \in Ne_i \cup \{i\}}$ represents the trust vector for agent i and $T_v = [T_1, T_2, \cdots, T_n]^T$ serves as a conceptual representation of the system. The trust scores are dichotomous variables: if agents i and j cooperate at time t, $t_{ij}(t)$ is set to 1; otherwise, it is 0. Initially, all trust values are set to 0. The trust scores are updated using Exponential Smoothing:

$$t_{ij}(t) \quad = \quad (1-\lambda)t_{ij}(t-1) + \lambda t_{ij}(t) \tag{4}$$

with λ, the smoothing factor, ensuring that trust values remain within the range [0,1]. This formulation ensures that trust scores retain information from past evaluations while avoiding the need for agents to track the trust values of all peers, making the mechanism scalable to larger systems.

We define competency as the ability of an agent $i \in \vartheta$ to take actions that maximize its individual rewards and help other agents gear towards social success. Quantifying the competency score is based on current Temporal Difference errors: we use the inter-agent assessments of agents from the reputation mechanism.

The competency score is then expressed as:

$$c_i(t) = \frac{1}{\varphi_i(t)} \tag{5}$$

When the temporal difference error increases, it indicates that the agent is not performing competently. This rising error suggests a misalignment between the agent's actions and the optimal strategy. In contrast, as the agent progresses toward convergence, the TD error is expected to approach zero. This reduction in error implies that the agent is effectively learning and optimizing its actions, potentially leading to convergence.

Reputation scores are then calculated using the expression:

$$Rep_i(t) \quad = \quad w_1 \times \frac{1}{ne_i(t)} \sum t_{ij}(t) + w_2 \times c_i(t) \tag{6}$$

with w_1, w_2 being the weights assigned to trust and competency scores. These task-dependent weights can be adjusted to emphasize specific behaviors, such as collaboration or individual efficiency.

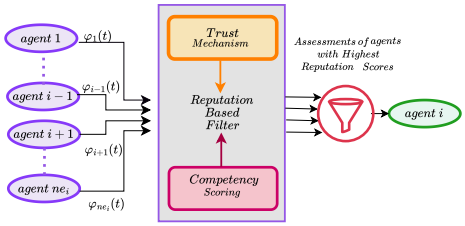


Figure 2: Inter-agent assessments to Contribute in Peer's Reward Reshaping through a Reputation-based Filter.

Reputation scores will serve to filter out agents whose evaluations will not contribute to agent i's learning process, as shown in 2. Let R be the matrix filter such that, if an agent j's reputation ranks among the top four within the neighborhood Ne_i, its assessment $\varphi_j(t)$ will contribute to the calculation of $\hat{r}_i(t)$, and $f = \frac{1}{ne_i(t)} \sum \varphi_j(t)$. The reward $\hat{r}_i(t)$ is then computed as:

$$\hat{r}_i(t) \quad = \quad r_i^b(t) + R \times f \tag{7}$$

4.3 Reward Reshaping with Densities-focused PBRS

To promote agents' execution of individual tasks while maintaining connectivity and enhancing communication, we propose a density-focused Potential-Based Reward Shaping (PBRS) mechanism. This method incentivizes agents to remain connected by adjusting their rewards based on the density of other agents in their observable area. The density of agents $j \in Ne_i(t)$ within the observable space of another agent i serves as a metric to both motivate individual performance and foster cooperation. The PBRS function will be defined as:

$$g \quad = \quad \gamma\phi(s') - \phi(s) = \gamma d(s') - d(s) \tag{8}$$

The density is calculated using:

$$\mathrm{d}(s_i) \quad = \quad \frac{ne_i(t)}{d_l \times d_L} \tag{9}$$

with d_l and d_L: the dimensions of the observable space of agent i (i.e., supposing we are working on a 2-dimensional space). Agent i receives a penalty for moving away from other agents and is rewarded positively for staying near its neighbors. The reshaped reward becomes:

$$\hat{r}_i(t) \quad = \quad r_{b_i}(t) + R \times f + g \tag{10}$$

Algorithm 1 Decentralized Semi-Cooperative MARL with Filtered Reward Reshaping Functions

Inputs: Number of agents v, environment Env, number of episodes N_{ep}, maximum steps per episode T
Initialize replay buffers D_i for each agent $i \in \{1, \ldots, v\}$
Initialize primary Q-networks Q_i and target Q-networks $\hat{Q}_i$ with random weights for each agent $i \in \{1, \ldots, v\}$
for episode = 1 to N_{ep} **do**
  Reset the environment and obtain initial state s_0
  for t = 0 to $T - 1$ **do**
    for each agent $i \in \{1, \ldots, v\}$ **do**
      Select action $u_i(t)$ using ϵ-greedy policy based on $Q_i(s_t, \cdot)$
      Observe next state $s(t+1)$ and reward $r_i^b(t)$
      Store transition $(s(t), u_i(t), r_i(t), s(t+1))$ in replay buffer D_i
      Store encounters op_{ij} with neighbors $j \in Ne_i$
      Calculate inter-agent assessment $\varphi_i(t)$
      Calculate competency $c_i(t)$ and trust scores $t_{ij}(t)$
      Calculate reputation score $Rep_i(t)$
      Filter out the assessment reshaping function $R \times f$
      Calculate densities of neighbors in observable space
      if Exploration Phase **then**
        Filter out the Densities-based PBRS function using Optimism
      end if
      Use Densities-based PBRS with no filter to reshape the reward
      Reshape reward $\hat{r}_i(t) = r_i^b(t) + R \times f + E \times g$ and store it
    end for
    $s(t) \leftarrow s(t+1)$
    if terminal state reached **then**
      break
    end if
    for each agent $i \in \{1, \ldots, v\}$ **do**
      Sample random minibatch of B transitions (s, u, r, s') from D_i
      Set target $y_i = \hat{r}_i + \gamma \max_{u'} \hat{Q}_i(s', u')$
      Periodically update target network $\hat{Q}_i \leftarrow Q_i$
    end for
  end for
end for

4.4 Filtering during Exploration Phase

In MARL, the exploration phase is important for agents to learn effective policies in an environment where multiple agents interact. Exploration allows agents to discover new strategies, states, and rewards that they may not encounter if they only exploit known information.

We add a filter to the second reshaping function g to encourage

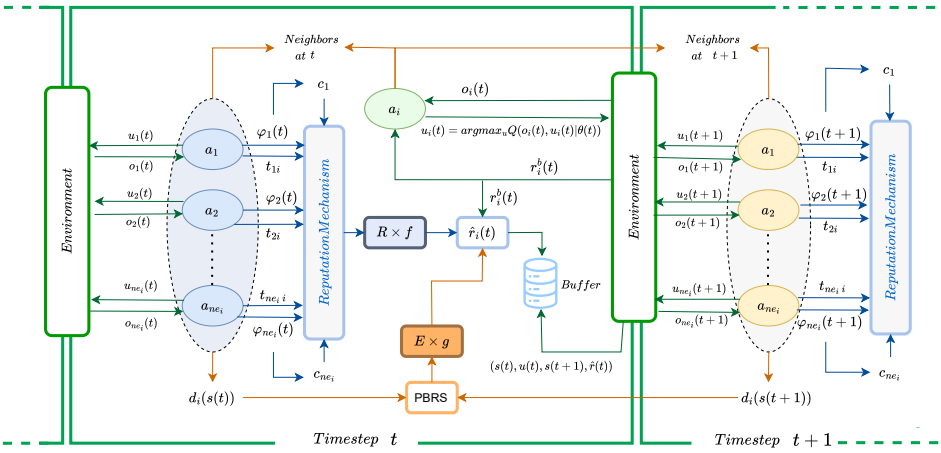


Figure 3: FRR Architecture. At Each time step t, agents interact with the environment by receiving observations $o_{j\neq i}(t)$ and giving back actions $u_j(t)$. Neighbors of agent i share their assessment scores φ_j and their trust scores t_{ji} as inputs to the reputation mechanism. Densities of neighbors in the partially observable space of agent i are calculated in time steps t and $t+1$ to calculate the PBRS function. Reshaped rewards of all agents are then fed to the buffer, along with the current transition $(s(t), u(t), s(t+1), \hat{r}(t))$.

agents to take advantage of the exploration phase:

$$\hat{r}_i(t) \quad = \quad r_{b_i}(t) + R \times f + E \times g \tag{11}$$

Assuming that the reshaping function f that induces cooperation is filtered using the filter matrix R, E is a filter that focuses on bringing positive incentivization to agents during the exploration phase:

$$E \quad = \quad 1_{exploration} \cdot 1_{g\geq 0} + 1_{exploitation} \cdot 1_{g\neq 0} \tag{12}$$

with the function $1_{exploration}$ indicating that the reshaping function g will only be incorporated into the reshaped reward when agents are engaged in the exploration phase. This exploration-guided reshaping mechanism utilizes a function that encourages agents to work towards individual and collective objectives. Initially, agents can explore the state space, and once the exploration phase (i.e., determined by the number of exploration episodes during training) concludes, they are rewarded or penalized based on the PBRS value. If the PBRS value g is negative during the exploitation phase, the agent is guided to reconsider its decisions regarding the exploitation of states that may not yield positive rewards.

Adding optimism to the intrinsic filter: By integrating an optimism factor into the proposed exploration filter, we seek to motivate agents to engage with those they have not sufficiently collaborated with in the past. This approach provides a ARLore robust exploration strategy that balances discovering new opportunities and achieving individual objectives. Optimism is inspired by the Upper Confidence Bound (UCB) approach, which encourages exploration by favoring less-visited state-action pairs. The constant (0.9) in our optimism factor promotes collaboration while avoiding excessive exploration. The exploration-based filter becomes:

$$\begin{aligned} E = {} & 1_{exploration} \cdot 1_{g\geq 0} \cdot \text{optimism} \\ & + 1_{exploitation} \cdot 1_{g\neq 0} \end{aligned} \tag{13}$$

An interaction occurs when agent j is within the observable space of agent i (i.e., the distance between them is below a specified threshold) and results in an encounter if the agents collaborate and achieve mutual winnings. Encounters are incrementally updated based on the number of successful collaborative interactions between agents i and j. We define optimism as follows:

$$\text{optimism} \quad = \quad \frac{0.9}{\sqrt{1 + \sum op_{ij}}} \tag{14}$$

with op_{ij}: the number of all encounters between agent i and agent $j \in Ne_i(t)$. Increased cooperation with new neighbors results in higher rewards for agent i during the exploration phase (see Algorithm 1). Both optimism and the reputation filtering parameters are environment-dependent and task-specific. For instance, higher optimism values are suited to cooperative scenarios to encourage collaboration, while stricter filtering enhances reliability in adversarial settings.

5 CONVERGENCE AND STABILITY ANALYSIS

We employ Lyapunov stability to ensure that the system's overall state, including all agents' states, converges to a desirable equilibrium and remains stable. In this work, instead of considering a stability-constrained MARL, we allow agents to explore/exploit the environment and update their policies in a Dec-POMDP with no pre-defined constraints. We start with a Boundedness analysis, which ensures that the system doesn't experience unbounded growth in Q-values or TD errors, which could lead to instability in learning. Additionally, Bounded TD errors allow the updates to converge gradually to an equilibrium rather than diverging.
We prove that the proposed reshaped reward is bounded:

$$\hat{r} \quad = \quad r + R \times f + E \times g \leq c_1 \tag{15}$$

with c_1 being a constant. Let's note: $\tilde{r}_j^b$ the maximum base reward an agent j can get back from the environment (since agents are homogeneous: $\tilde{r}_j^b = \tilde{r}_i^b, \forall (i,j) \in \vartheta$), d_{max} and d_{min} the maximum and minimum densities of agents $\{j \in \vartheta - \{i\}\}$ in the observable space of an agent i respectively, (s', u', s, u) being, respectively, the next and actual pair of state-action of agent j, and $\lambda_1, \lambda_2, \lambda_3$ positive constants.

Let's note that the filter R does not affect the boundedness proof for the first filtered reshaping function, as the reputation mechanism allows agent i to consider only the four best assessments from its neighbors:

$$\begin{aligned} |f| &= |\sum_{j=1}^{4} \delta_j(r_j^b)| \\ &= |\sum_{j=1}^{4} (r_j^b + \gamma max_{u'} Q_j(s', u') - Q_j(s, u_j)| \\ &\leq |\sum_{j=1}^{4} (\tilde{r}_j^b + max_{u'}(\gamma Q_j(s', u') - Q_j(s, u_j))| \\ &\leq |4 \times \tilde{r}_j^b + \sum_{j=1}^{4} (max_{u'}(\gamma Q_j(s', u') - Q_j(s, u_j))| \end{aligned} \tag{16}$$

Since the Q-values are bounded (as will be proven in the theoretical convergence analysis), we get:

$$\begin{aligned} &\forall (s,u) \in S \times U,\ Q(s,u) \leq \lambda_3 \\ &\rightarrow max_{u'}(\gamma Q_j(s', u') - Q_j(s, u_j)) < 2\lambda_3 \end{aligned} \tag{17}$$

This implies the first reshaping function is bounded such that:

$$|f| \quad \leq \quad 4(\tilde{r}^b + 2\lambda_3) \tag{18}$$

The second reshaping function is a densities-based PBRS. For an agent i:

$$\begin{aligned} |g| &= |\gamma d(s_i') - d(s_i)| \\ &\leq |\gamma d_{min}(s_i') - d_{max}(s_i)| \\ &\leq \lambda_2 \end{aligned} \tag{19}$$

Since E filters g during the exploration phase and we induced a constrained optimism such that $\frac{0.9}{\sqrt{1+\sum op_{ij}}} < 1$, consequently:

$$|E \times g| \quad \leq \quad |g| \leq \lambda_2 \tag{20}$$

We get an upper bound of the reshaped reward:

$$\begin{aligned} \hat{r} &\leq |r_j^b + R \times f + E \times g| \\ &\leq |\tilde{r}_j^b + \lambda_1 + \lambda_2| \end{aligned} \tag{21}$$

5.1 Theoretical Convergence

For the system to converge, we assume the following conditions hold as $t \rightarrow \infty$ with π_i^* being the optimal policy for agent i:

$$\lim_{t \to \infty} \pi_i(t) \quad = \quad \pi_i^* \tag{22}$$

Under the optimal policy π_i^*, as $t \rightarrow \infty$:

$$\begin{aligned} Q_i(s,u) &\longrightarrow Q^*(s,u) \\ \theta &\longrightarrow \theta^* \\ \delta(\hat{r}) &\longrightarrow 0 \end{aligned} \tag{23}$$

The Q-values $Q_i(s,u)$ are expected to converge to their optimal values $Q^*(s,u)$, and the TD error $\delta(\hat{r})$ should asymptotically approach zero. This implies that the agent's policy $\pi_i(t)$ converges to the optimal policy π_i^* over time.

The convergence of the Q-values is guaranteed under certain conditions on the learning rate α. The Q-values will converge to their optimal values if α satisfies the Robbins-Monro conditions [14]:

$$\sum_{t=0}^{\infty} \alpha_t \quad = \quad \infty \quad \text{and} \quad \sum_{t=0}^{\infty} \alpha_t^2 < \infty \tag{24}$$

These conditions ensure that the learning rate decreases sufficiently over-time to stabilize the learning process while still allowing enough updates for convergence.

As established in the boundedness analysis, the reshaped reward $\hat{r}$ is bounded by a constant c_1, ensuring that the TD error $\delta(\hat{r})$ remains bounded as well. Specifically:

$$|\delta(\hat{r})| \quad \leq \quad \tilde{r}_j^b + \lambda_1 + \lambda_2 + 2\lambda_3 \tag{25}$$

Since the TD error $\delta(\hat{r})$ is bounded, the Q-value updates remain stable, preventing divergence and ensuring that the Q-values $Q_i(s,u)$ gradually converge to their optimal values $Q^*(s,u)$. Introducing the reshaped reward accelerates learning by providing additional guidance to the agents through reward shaping, leading to faster learning of agents. However, the bounded nature of the reshaped reward ensures that this acceleration does not destabilize the learning process.

5.2 Lyapunov Stability Analysis

We propose a Lyapunov function $L(s,u) = V(Q(t))$, which takes strictly positive values and is formulated as the sum of squared errors in Q-values. We define $\Phi_i = Q_i(s,u) - Q^*(s,u)$ such that:

$$\begin{aligned} V(Q(t)) &= \frac{1}{2} \sum_{i=1}^{\nu} \Phi_i^2 \\ &= \frac{1}{2} \sum_{i=1}^{\nu} \left(Q_i(s,u) - Q^*(s,u)\right)^2 \end{aligned} \tag{26}$$

This function represents a cost function that must decrease over time for the multi-agent system to achieve Lyapunov stability. This condition is equivalent to the time derivative $\dot{V}(Q(t))$ having negative values:

$$\dot{V}(Q(t)) \quad = \quad \sum_{i=1}^{\nu} (\Phi_i) \cdot \dot{\Phi}_i \tag{27}$$

With respect to the Q-value updates, we have:

$$\begin{aligned} \dot{\Phi} = \alpha([r + R \times f + E \times g] \\ + \gamma \max_{u'} Q_i(s', u') - Q_i(s,u)) \end{aligned} \tag{28}$$

The Lyapunov function's time derivative becomes:

$$\begin{aligned} \dot{V}(Q(t)) = \sum_{i=1}^{\nu} [\ (Q_i(s,u) - Q^*(s,u)) \\ \times \alpha([r + R \times f + E \times g] \\ + \gamma \max_{u'} Q_i(s', u') - Q_i(s,u))\ ] \end{aligned} \tag{29}$$

The TD error with the reshaped reward $\hat{r}$ is proven to be bounded:

$$\begin{aligned} |\delta_i(\hat{r})| &\leq [\ |r_i^b + R \times f + E \times g| \\ &\quad + max_{u'} |\gamma Q_j(s', u') - Q_j(s, u_j)|\ ] \\ &\leq \tilde{r}_j^b + \lambda_1 + \lambda_2 + 2\lambda_3 \end{aligned} \tag{30}$$

We introduce a new positive constant λ that reflects the discrepancy between the optimal Q-value and the agent's current Q-value at each time step:

$$\begin{aligned} \lambda = &\quad Q^*(s(t), u(t)) - Q_i(s(t), u(t)) > 0 \\ \leftrightarrow -\lambda = &\quad Q_i(s(t), u(t)) - Q^*(s(t), u(t)) < 0 \end{aligned} \tag{31}$$

with α the learning rate, λ and $|\delta_i(\hat{r})|$ being all positive, we prove that $V(Q(t))$ does decrease in time, leading to the convergence of the agents:

$$\begin{aligned} \dot{V}(Q(t)) &\leq -\sum \alpha\lambda |\delta_i(\hat{r})| \\ &< -\sum \alpha\lambda |\delta_i(\hat{r})|^2 < 0 \end{aligned} \tag{32}$$

6 EXPERIMENT

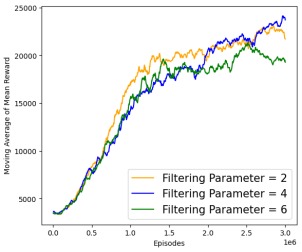
6.1 Environment

The environment is defined as a dynamic grid world, characterized by a variety of elements including empty cells, walls, and entities such as hunters and prey. The focus is centered on training the hunters within this grid. Each hunter is equipped to navigate and interact within the grid, using local information and observations limited to a partially observable space centered around its position. Prey move freely, avoiding predators when possible. For a prey to be captured, it must be fully surrounded by hunters. The grid's dimensions, as well as the number of prey and hunters, are varied to test the adaptability of our algorithm to high-dimensional and semi-cooperative settings. The hunters must balance the objectives of surrounding and capturing prey while conserving their battery levels.

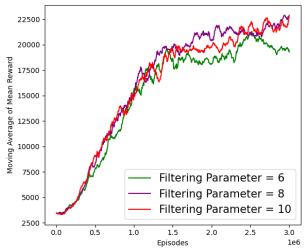
6.2 Numerical Results

We use two baselines to evaluate the performance of agents using our approach, both are adjusted to encourage cooperation in partially observable semi-cooperative environments. The first baseline is the *PED-DQN* framework [8], which is a value-based MARL method designed to enhance cooperation by reshaping rewards through peer evaluation. This peer evaluation metric incorporates feedback from other agents to encourage socially optimal actions. The second baseline is *Independant Q-learners (IQL)* [17]. IQL employs the classic Q-learning algorithm, where agents update their Q-values based solely on their experiences, which consist of their own state-action pairs and the rewards received.

We define the filtering parameter as a reputation-related user-defined value that specifies the number of peers whose assessments will be included in the reward reshaping process. Figures 5a and 5b show that different filtering parameters of the reputation mechanism lead to varying rates of improvement of the learning of agents. Setting the filtering parameter to 4 demonstrates a better performance of hunters compared to the others, showcasing that choosing a high filtering parameter is not necessarily the best option. In what follows, we set the filtering parameter to 4 by default.



(a) Results with reputation filtering parameters $= \{2, 4, 6\}$.



(b) Results with reputation filtering parameters $= \{6, 8, 10\}$.

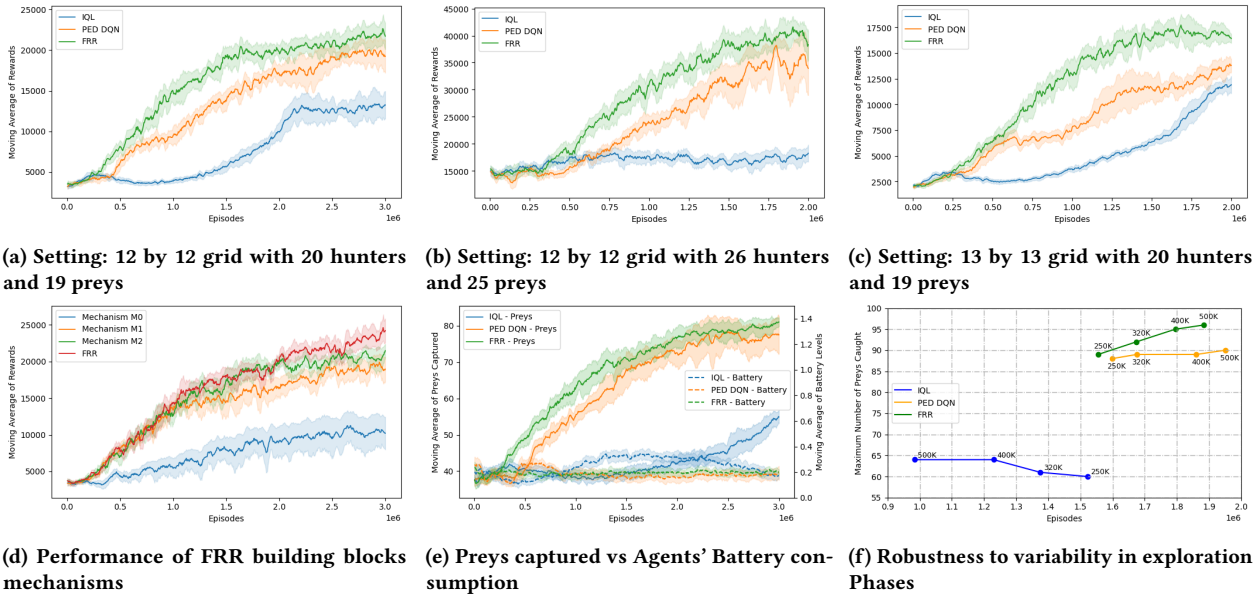
Figure 5: Smoothed results of the FRR mechanism in a 12 by 12 grid with 20 hunters and 19 preys.

We evaluate the performance of agents using the different building blocks of our mechanism in the same setting (as shown in 4d) to demonstrate that agents are more encouraged to cooperate when their rewards are reshaped using both filtered reward reshaping functions:

- **Mechanism M0:** We average the inter-agent assessments $\varphi_j(t)$ received by neighbors in the partially observable space of each agent $\hat{r}_i(t) = r_i^b(t) + \frac{1}{\overline{ne_i}(t)} \sum_{j \in \overline{Ne_i}(t)} \varphi_j(t)$, such that $\overline{Ne_i}$ does not account for all neighbors but only the neighbors with the least assessments (i.e., as an indicator of good learning).
- **Mechanism M1:** We filter the proposed first reshaping function using the reputation mechanism. The reshaped reward becomes: $r_i(t) = r_i^b(t) + R \times f$.
- **Mechanism M2:** We add the second reshaping function with only the exploration-based filter to reshape the reward: $\hat{r}_i(t) = r_i^b(t) + R \times f + (1_{exploration} \cdot 1_{g \geq 0} + 1_{exploitation} \cdot 1_{g \neq 0}) \times g$.

6.3 Interpretations

6.3.1 ***Cooperation and scalability.*** Figures 4a, 4b, and 4c illustrate the moving average of rewards across different settings. IQL experiences slow learning and adopts suboptimal policies due to interference among independent learners. While IQL agents improve gradually in 4a and 4c, their performance in 4b stagnates because the high environment density impedes effective learning and convergence. PED-DQN uses peer evaluations to encourage cooperation among agents, resulting in better performance compared to IQL. FRR outperforms both IQL and PED-DQN by filtering assessments, allowing agents to rely on the most competent collaborators. In the scenario shown in 4c, FRR achieves a 78% increase in rewards over IQL and a 16% increase over PED-DQN. FRR effectively transforms the scalability challenges often encountered in multi-agent systems into a significant advantage, enabling the system to achieve higher rewards. This benefit of cooperation is especially evident in Figure 4b, where FRR's density-focused PBRS mechanism encourages agents to remain close to their collaborators. 4e shows that FRR indeed enhances prey capture rates while efficiently conserving energy. Conversely, IQL initially focuses on conserving energy but eventually shifts its strategy towards prey capture as it yields to higher rewards.



(a) Setting: 12 by 12 grid with 20 hunters and 19 preys

(b) Setting: 12 by 12 grid with 26 hunters and 25 preys

(c) Setting: 13 by 13 grid with 20 hunters and 19 preys

(d) Performance of FRR building blocks mechanisms

(e) Preys captured vs Agents' Battery consumption

(f) Robustness to variability in exploration Phases

Figure 4: Performance evaluation of the proposed FRR approach compared to baseline methods (PED-DQN and IQL) across different grid settings and exploration phases. (a), (b), and (c) show the moving average of rewards over training episodes for environments with varying grid sizes and numbers of agents. (d) illustrates the performance of FRR building block mechanisms (M0, M1, and M2 explained in 6.2). (e) presents a comparison of preys captured versus agents' battery consumption in a setting with 23 hunters and 22 preys in a 12 by 12 grid. (f) demonstrates the robustness of the proposed FRR approach and baselines across different exploration phases.

6.3.2 ***Comparison between mechanisms.*** Figure 4d shows that mechanism M0, which averages inter-agent assessments without filtering, leads to the slowest learning progress. Mechanism M1, with Reputation-based filtering, improves performance by allowing agents to rely on more competent collaborators. Mechanism M2, by adding an exploration-based filter, improves learning by dynamically adjusting agent behavior in the exploration and exploitation phases. FRR, which combines both filters, outperforms all mechanisms by encouraging cooperation using the reputation mechanism and the densities-focused PBRS, demonstrating the most effective learning and achieving the highest rewards. This combination accelerates learning and ensures more consistent cooperation between agents in high-dimensional settings.

6.3.3 ***Exploration phase.*** Figure 4f illustrates the robustness of FRR and the baseline methods in a 12x12 grid with 23 hunters and 22 preys over 2 million steps. FRR's optimistic exploration filter accelerates improvement beyond the exploration phase and achieves a maximum of 96 preys captured during a 500,000-step exploration phase. Similarly, PED-DQN also shows improvement with sufficient exploration, capturing 90 preys at its peak with the same exploration length. IQL reaches a maximum prey capture of 64 more quickly with sufficient exploration but still falls short of the performance achieved by PED-DQN and FRR due to its lack of coordination and independent learning approach.

7 CONCLUSION AND FUTURE WORK

In this work, we introduced a novel framework for promoting cooperation and optimizing task performance in high-dimensional semi-cooperative multi-agent systems. By integrating a reputation-based reward reshaping mechanism that filters agent contributions based on trust and competency, we ensure that only the most reliable agents influence their peers' learning process. Additionally, our density-focused PBRS mechanism encourages agents to maintain connectivity by adjusting rewards according to the density of their neighbors, allowing both individual and collective success. The combination of these mechanisms allows for more effective coordination in multi-agent settings, addressing challenges such as communication, connectivity, and the conflict between individual objectives and group connectivity. By reshaping rewards using competency assessments, reputation filtering, and neighborhood densities, we create an environment where agents can better balance exploration with task achievement and collaboration. In future work, we aim to improve the estimation of the reputation filtering parameter by enabling agents to learn and adaptively choose the filtering parameter best suited to each scenario, ensuring optimal learning dynamics across different settings. Furthermore, we plan to extend this framework to environments with heterogeneous agents that may have different capabilities, objectives, or action spaces, allowing us to explore the challenges of maintaining cooperation and stability in multi-agent systems with more complex tasks.

REFERENCES

[1] Sara Amini and Mohsen Afsharchi. 2014. Finding Better Teammates in a Semi-cooperative Multi-agent System. In *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, Vol. 3. 143–150. https://doi.org/10.1109/WI-IAT.2014.161

[2] Ariyan Bighashdel, Daan de Geus, Pavol Jancura, and Gijs Dubbelman. 2023. Off-Policy Action Anticipation in Multi-Agent Reinforcement Learning. arXiv:2304.01447 [cs.MA] https://arxiv.org/abs/2304.01447

[3] Noam Buckman, Sertac Karaman, and Daniela Rus. 2023. Studying the Impact of Semi-Cooperative Drivers on Overall Highway Flow. In *2023 IEEE Intelligent Vehicles Symposium (IV)*. 1–8. https://doi.org/10.1109/IV55152.2023.10186563

[4] Yinlam Chow, Ofir Nachum, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. 2018. A Lyapunov-based Approach to Safe Reinforcement Learning. arXiv:1805.07708 [cs.LG] https://arxiv.org/abs/1805.07708

[5] Julio B. Clempner. 2022. A Lyapunov approach for stable reinforcement learning. *Computational and Applied Mathematics* 41, 279 (2022). https://doi.org/10.1007/s40314-022-01988-y

[6] Yunlong Dong, Xiuchuan Tang, and Ye Yuan. 2020. Principled reward shaping for reinforcement learning via lyapunov stability theory. *Neurocomputing* 393 (2020), 83–90. https://doi.org/10.1016/j.neucom.2020.02.008

[7] Jakob Foerster, Richard Y. Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. 2018. Learning with Opponent-Learning Awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems* (Stockholm, Sweden) *(AAMAS '18)*. International Foundation for Autonomous Agents and Multiagent Systems, 122–130.

[8] David Earl Hostallero, Daewoo Kim, Sangwoo Moon, Kyunghwan Son, Wan Ju Kang, and Yung Yi. 2020. Inducing Cooperation through Reward Reshaping based on Peer Evaluations in Deep Multi-Agent Reinforcement Learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems* (Auckland, New Zealand) *(AAMAS '20)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 520–528.

[9] Trung Dong Huynh. 2006. Trust and reputation in open multi-agent systems. https://api.semanticscholar.org/CorpusID:38269536

[10] Jiechuan Jiang and Zongqing Lu. 2018. Learning Attentional Communication for Multi-Agent Cooperation. arXiv:1805.07733 [cs.AI] https://arxiv.org/abs/1805.07733

[11] Daewoo Kim, Sangwoo Moon, David Hostallero, Wan Ju Kang, Taeyoung Lee, Kyunghwan Son, and Yung Yi. 2019. Learning to Schedule Communication in Multi-agent Reinforcement Learning. arXiv:1902.01554 [cs.AI] https://arxiv.org/abs/1902.01554

[12] Adam Lerer and Alexander Peysakhovich. 2018. Maintaining cooperation in complex social dilemmas using deep reinforcement learning. arXiv:1707.01068 [cs.AI] https://arxiv.org/abs/1707.01068

[13] Andrew Y Ng, Daishi Harada, and Stuart Russell. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the 16th International Conference on Machine Learning (ICML'99)*. 278–287.

[14] Herbert E. Robbins. 1951. A Stochastic Approximation Method. *Annals of Mathematical Statistics* 22 (1951), 400–407. https://api.semanticscholar.org/CorpusID:16945044

[15] Jordi Sabater and Carles Sierra. 2001. REGRET: A reputation model for gregarious societies. https://api.semanticscholar.org/CorpusID:749615

[16] Reid Sawtell, Sarah Kitchen, Timothy Aris, and Chris McGroarty. 2024. Learning Cohesive Behaviors Across Scales for Semi-Cooperative Agents. *The International FLAIRS Conference Proceedings* (2024). https://api.semanticscholar.org/CorpusID:269900826

[17] Ming Tan. 1993. Multi-agent reinforcement learning: independent versus cooperative agents. In *Proceedings of the Tenth International Conference on International Conference on Machine Learning* (Amherst, MA, USA) *(ICML'93)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 330–337.

[18] Woodrow Z. Wang, Mark Beliaev, Erdem Bıyık, Daniel A. Lazar, Ramtin Pedarsani, and Dorsa Sadigh. 2021. Emergent Prosociality in Multi-Agent Games Through Gifting. arXiv:2105.06593 [cs.MA] https://arxiv.org/abs/2105.06593

[19] Y. Wang and J. Vassileva. 2003. Trust and reputation model in peer-to-peer networks. In *Proceedings Third International Conference on Peer-to-Peer Computing (P2P2003)*. 150–157. https://doi.org/10.1109/PTP.2003.1231515

[20] David H. Wolpert, Kevin R. Wheeler, and Kagan Tumer. 2000. Collective Intelligence for Control of Distributed Dynamical Systems. *EPL (Europhysics Letters)* 49, 6 (2000), 708.

[21] Federico M. Zegers, Matthew T. Hale, John M. Shea, and Warren E. Dixon. 2021. Event-Triggered Formation Control and Leader Tracking With Resilience to Byzantine Adversaries: A Reputation-Based Approach. *IEEE Transactions on Control of Network Systems* 8, 3 (2021), 1417–1429. https://doi.org/10.1109/TCNS.2021.3068348

[22] Federico M. Zegers, Matthew T. Hale, John M. Shea, and Warren E. Dixon. 2021. Event-Triggered Formation Control and Leader Tracking With Resilience to Byzantine Adversaries: A Reputation-Based Approach. *IEEE Transactions on Control of Network Systems* 8, 3 (2021), 1417–1429. https://doi.org/10.1109/TCNS.2021.3068348

[23] Chongjie Zhang and Victor Lesser. 2010. Multi-Agent Learning with Policy Prediction. *Proceedings of the National Conference on Artificial Intelligence* 2.

[24] Qingrui Zhang, Hao Dong, and Wei Pan. 2020. Lyapunov-Based Reinforcement Learning for Decentralized Multi-agent Control. In *Distributed Artificial Intelligence*, Matthew E. Taylor, Yang Yu, Edith Elkind, and Yang Gao (Eds.). Springer International Publishing, Cham, 55–68.

[25] Xiaoqin Zhang and Victor Lesser. 2011. Solving Negotiation Chains in Semi Cooperative Multi-Agent Systems. (05 2011).

[26] Liqun Zhao, Konstantinos Gatsis, and Antonis Papachristodoulou. 2023. Stable and Safe Reinforcement Learning via a Barrier-Lyapunov Actor-Critic Approach. arXiv:2304.04066 [eess.SY] https://arxiv.org/abs/2304.04066