Ben-Gurion University of the Negev

Faculty of Computer and Information Science

Department of Software and Information Systems Engineering

Final Project Report

Tag, You're Nash!
Analyzing the Impact of Reward Structures on Strategic Stability and Nash Convergence in MARL

Submitted by:

Omer Toledano, Eliya Naomi Aharon

Department of Software and Information Systems Engineering

Ben-Gurion University of the Negev

Emails: omertole@post.bgu.ac.il, eliyaah@post.bgu.ac.il

February 15, 2026

1 Introduction

In the field of Multi-Agent Reinforcement Learning (MARL), a gap exists between theoretical stability and empirical outcomes. While the objective is to achieve stable equilibria, results are sensitive to reward configurations, often leading to learned strategies that favor strategic stagnation, manifested as a "stalling" state in the Simple Tag scenario, over intended task objectives. These challenges are further amplified in mixed cooperative-competitive environments, where agents must cooperate with teammates while simultaneously competing against adversaries. This project explores the line between coordination and strategic paralysis, analyzing how shifts in reward distribution influence the stability and convergence of agents.

To explore these dynamics, this project addresses two research questions: (1) How does shifting the reward mechanism influence cooperation in a competitive environment?, and (2) To what extent does the observed 'stalling' behavior constitute a true game-theoretic equilibrium?

To address these, we utilize a Cooperation Factor (α) that balances an agent's individual reward against the collective performance of its team, evaluating its impact on emergent behaviors and strategic stability.

Our empirical findings reveal a trade-off between stability and task performance. While a competitive reward structure leads to consistent but suboptimal behavior, we demonstrate that introducing reward sharing disrupts this equilibrium and fosters coordinated pursuit strategies. Specifically, a cooperation factor of $\alpha = 0.25$ emerges as the optimal balance, achieving peak capture frequency while preserving the individual drive necessary for effective performance.

This project offers two primary contributions. First, we characterize the impact of reward redistribution on emergent multi-agent dynamics within the Simple Tag domain, illustrating how the shift from individual to collective incentives, modulated by the cooperation factor, influences agent rewards, spatial occupancy, and the emergence of specialized roles such as blocking. Second, we establish a methodological framework for Empirical Nash Equilibrium (ENE) analysis, utilizing trajectory-based counterfactual deviations to evaluate the strategic stability of learned policies and differentiate between true theoretical equilibria and suboptimal convergence points.

Implementation and source code for this project are available on GitHub [1].

[1] `https://github.com/omertol/Tag-Youre-Nash`

2 Related Work

2.1 Mixed Cooperative-Competitive MARL and Benchmarks

The Multi-Agent Particle Environment (MPE) was introduced as a benchmark for studying emergent coordination and communication in multi-agent systems [5, 4]. The `simple_tag` scenario involves a team of predator agents coordinating to capture an actively evading prey, combining intra-team cooperation with inter-team competition. This setup has been shown to foster collaborative strategies such as encirclement, making it ideal for analyzing incentive alignment and strategic interaction [4].

2.2 Independent Learning and Non-Stationarity

To address the non-stationarity inherent in multi-agent settings, Independent Learning (IL) treats other agents as part of the environment. While IL lacks theoretical convergence guarantees in mixed multi-agent settings, it remains widely used due to its simplicity and scalability. Empirical results show that methods such as Independent PPO can achieve competitive task performance on mixed MARL benchmarks [3]. However, prior evaluations typically focus on metrics such as cumulative reward or capture frequency, leaving the stability of the resulting policies under unilateral deviation unexplored.

2.3 Reward Design and Emergent Behavior

Since independent learners rely only on their own signals, the reward structure is the main factor guiding their behavior. Studies have shown that systematically adjusting the balance between individual and shared rewards can induce transitions along the cooperation-competition spectrum [7, 1].

In cooperative or partially cooperative settings, agents often share global rewards, which introduces the multi-agent credit assignment problem: an individual agent cannot easily determine whether a positive or negative outcome is attributable to its own action or to the actions of teammates [3, 4]. In mixed environments such as `simple_tag`, this issue is compounded by the presence of both cooperative and competitive incentives [3, 6]. Shared rewards can diffuse responsibility for outcomes, potentially encouraging free-riding or risk-averse behavior [2, 7].

2.4 Strategic Stability and Robustness

From a game-theoretic perspective, stability is characterized by the absence of profitable unilateral deviations, i.e., Nash equilibrium. Empirical studies rarely evaluate MARL policies using deviation-based analysis, focusing instead on average task performance [2, 3], likely due to the

computational cost of verifying such stability. As a result, the relationship between reward design, emergent cooperation, and equilibrium stability remains underexplored, particularly in mixed environments such as `simple_tag`.

2.5 Research Gap and Contribution

The literature establishes three key foundations: (1) `simple_tag` as a benchmark for mixed MARL coordination [4, 6, 3]; (2) the sensitivity of emergent behavior to reward design [7, 1]; and (3) that independent learning can work in non-stationary multi-agent settings [3].

Prior work evaluates coordination primarily through performance metrics [6], often overlooking strategic robustness [2]. Furthermore, no modular computational library currently exists to identify Empirical Nash Equilibria (ENE) in simultaneous-move MARL environments.

This project addresses these gaps by linking reward structures to cooperative performance and stability in `simple_tag`. Additionally, we employ a three-phased approach, detailed in chapter 3, to assess whether the observed 'stalling' behavior represents a stable Nash Equilibrium.

3 Approach

3.1 Environment Setup

Standardized libraries such as PettingZoo [8] and JaxMARL [6] provide accessible implementations of MPE scenarios, enabling reproducible empirical evaluation. Accordingly, the experiments in this project were conducted using the MPE `simple_tag` scenario, accessed via the PettingZoo library. `Simple_tag` simulates a competitive interaction in a continuous 2D space between two groups of agents with opposing objectives:

- **Predators**: A team of three slower agents aiming to capture the prey through physical contact. A +10 reward is awarded to the predator that successfully captures the prey.
- **Prey**: A single, faster agent whose objective is to evade capture. The prey receives a -10 penalty whenever it is captured by a predator.
- **Obstacles**: Objects within the environment that influence movement.

The agent's action space is discrete and consists of five possible actions: move up, down, left, right, or stationary. The continuous observation space consists of the agent's own velocity and position, along with the relative coordinates of other agents and obstacles.

3.2 Reward Structure Transition: From Competition to Cooperation

To address RQ1, we explore the impact of shifting the reward structure. This approach is grounded in established literature demonstrating that the strategic weighting of individual and group-level incentives can effectively modulate multi-agent behaviors along the competitive-cooperative spectrum [1, 7]. By implementing a reward sharing mechanism, we hypothesize that redistributing incentives will alter the learning dynamics, potentially preventing the convergence toward a stalling equilibrium and instead facilitating the emergence of coordinated predatory strategies.

To formally implement this mechanism, we introduce a Cooperation Factor ($\alpha \in [0, 1]$) as a weighting parameter to define the training objective function. The redistributed reward for each predator, R_i, which serves as the primary signal for policy optimization during the learning phase, is defined as:

$$R_i = \alpha \cdot r_i + (1 - \alpha) \cdot \sum_{j=1}^{N} r_j \tag{1}$$

In this context, R_i acts as a synthetic incentive designed to modulate learning dynamics and facilitate the emergence of coordinated strategies , where r_i represents the individual reward received by predator i and $\sum_{j=1}^{N} r_j$ denotes the total collective reward of the entire predator team.

This framework facilitates the evaluation of whether reducing α from 1.0 shifts agent objectives toward collaboration. We hypothesize that this reward redistribution will destabilize the stalling equilibrium and facilitate the emergence of coordinated strategies within the predator team. To analyze the resulting dynamics, we established five experimental profiles with $\alpha \in \{0.0, 0.25, 0.5, 0.75, 1.0\}$, representing the spectrum from pure competition to full reward sharing.

3.3 Empirical Stability Analysis: Identifying Stalling and Nash Convergence

To address RQ2 regarding the alignment between empirical strategies and theoretical stability, we developed a method to extract and analyze game-theoretic properties of learned behaviors via three key phases: (1) trajectory extraction, (2) event detection, and (3) game abstraction.

Trajectory Extraction We constructed a trajectory matrix for each cooperation factor α to record the spatial state of the environment at every simulation cycle. This matrix tracked the position coordinates of all four agents. The magnitude of the velocity vector, $||v_t||$, was derived from the Euclidean distance between consecutive spatial coordinates.

Event Detection Using this data, we sought to identify specific "Stalling" states in which the agents exhibit strategic paralysis, a phenomenon linked to penalty-driven risk aversion [7]. A state

is classified as a stalling event if the velocity vector magnitude, $||v_t||$, of at least one agent remains below a fixed threshold across a sliding window of 15 simulation cycles:

$$||v_t|| < 0.1, \quad \forall t \in [T, T + 15] \tag{2}$$

where t represents the simulation cycle. The threshold of 15 cycles was chosen to filter out momentary pauses and isolate genuine equilibrium states where agents refuse to engage or move. For each α configuration, the sampling process was conducted until a total of 100 unique stalling events were identified and recorded for further analysis.

Game Abstraction Once a stalling event was identified at time T, we evaluated its game-theoretic stability by performing a counterfactual deviation analysis. The goal was to determine if the observed stalling state constitutes an Empirical Nash Equilibrium (ENE), which is a state where no agent can improve its long-term outcome by unilaterally changing its action. This evaluation framework addresses the non-stationarity challenges in independent learning [3, 1] by adapting action deviation methodologies used for robustness testing [2].

For each agent i involved in a stalling event, we re-executed the simulation up to cycle T. At this point, the original recorded action was replaced with each of the four alternative discrete actions $a' \in \mathcal{A}$ (excluding the action taken in the original trajectory). Following the deviation, the simulation continued until the end of the episode ($T_{end} = 50$). We calculated the cumulative reward for each alternative path and compared it against the baseline reward obtained from the original trajectory.

Formally, we identified the best possible alternative action as:

$$a^*_{best} = \arg\max_{a' \in \mathcal{A}} \left(\sum_{t=T}^{T_{end}} r_t(a') \right) \tag{3}$$

A stalling state was classified as an Empirical Nash Equilibrium if the regret, denoted as Δ, and representing the potential gain from a unilateral deviation, satisfied the following condition:

$$\Delta = \text{Reward}(a^*_{best}) - \text{Reward}_{baseline} \leq 0.1 \tag{4}$$

This threshold ensures that minor fluctuations are filtered out, identifying only those states where there is no significant incentive for the agent to move. In cases where multiple actions yielded the same maximum cumulative reward, the first action encountered in the search space was selected as the representative deviation.

3.4 Training Process

The training framework utilizes Proximal Policy Optimization (PPO) as the core learning algorithm, implemented within an adversarial self-play loop. This methodology involves alternating training cycles between the predator team and the prey, ensuring that as one group improves, the other is compelled to adapt its strategy. By leveraging this iterative process, we establish a baseline for analyzing the emergence of multi-agent dynamics across our various reward redistribution profiles.

Each experimental profile was trained for a total of 1,000,000 timesteps, divided into ten sequential rounds of 100,000 steps. This training budget represents a trade-off between computational constraints and the empirical convergence requirements for stable policies in the MPE environment [6]. This iterative schedule facilitates the analysis of policy adaptation and strategy stability under varying α configurations.

4 Evaluation

To address the first research question (RQ1) regarding the impact of reward redistribution on agent cooperation, we conducted a systematic evaluation across all five experimental α profiles. The evaluation protocol consists of 50 independent episodes, with each episode capped at 50 steps per agent. Episodes terminate upon reaching this step limit, independent of capture occurrences.

To assess the impact of the reward transition on the predators' cooperation, we use **Cumulative Raw Team Reward** as our performance metric. This metric represents the aggregate success of the predator team in accomplishing the task of capturing the prey. The metric is calculated as follows:

$$\text{Team Performance} = \sum_{j=1}^{N} r_j \tag{5}$$

In this formula, N denotes the number of predator agents ($N = 3$) and r_j represents the raw individual reward received from the environment by agent j. Crucially, while the training objective R_i utilizes the cooperation factor α to modulate learning dynamics, this evaluation metric omits such weighting to provide a standardized baseline for comparing absolute performance across all experimental profiles.

Figure 1 presents the total accumulated reward collected by the agents during the evaluation episodes across different α values. A performance trade-off is observed: for the model trained with $\alpha = 1.0$, performance remains stable and consistent across episodes, though the total reward

magnitude is lower compared to other configurations. Conversely, for $\alpha = 0.0$, the agents achieve a significantly higher total reward, yet exhibit sharper fluctuations between episodes. This suggests that $\alpha = 0.0$ introduces sensitivity to initial episode conditions, whereas $\alpha = 1$ prioritizes strategic stability.

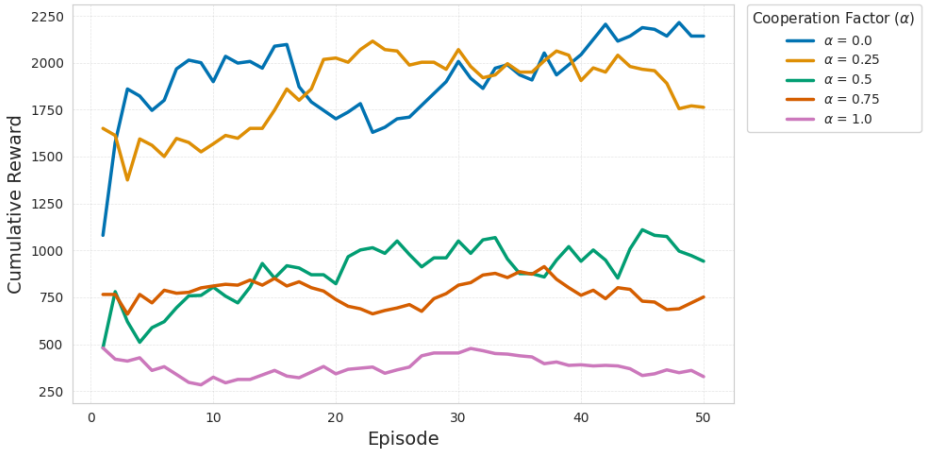

Figure 1: Total reward accumulation across training episodes for different α values.

In addition, we evaluate **Capture Frequency** as a direct indicator of task completion. This metric is intended to assess the agents' ability to break the stalling equilibrium, potentially providing deeper insight into how reward sharing facilitates effective cooperation and active pursuit of the target.

The empirical results, as illustrated in the capture frequency distribution (Figure 2), reveal a significant correlation between the reward redistribution factor α and predatory captures. Under the pure competition configuration of $\alpha = 1.0$, the team exhibited the lowest performance, recording an average capture frequency of 6.83 ± 4.92. A substantial performance leap is observed when reward sharing is introduced at $\alpha = 0.75$, with the average capture frequency more than doubling to 14.84 ± 4.0. This upward trend continued, with the predators reaching their peak effectiveness at $\alpha = 0.25$ by achieving a mean capture frequency of 20.26 ± 4.51. Finally, the transition to full reward redistribution at $\alpha = 0.0$ resulted in a slight performance decline to 18.42 ± 5.51.

To address the second research question (RQ2), which examines the alignment between empirically learned strategies and theoretical stability, we applied the empirical stability analysis framework described in Section 3.3. Specifically, we evaluated the robustness of the learned stalling behaviors by measuring the **Regret** (Δ) associated with forced action deviations.

A state was defined as violating the Nash equilibrium condition if the calculated Regret exceeded a numerical threshold ($\epsilon = 0.1$), indicating the existence of a strictly better alternative action for the agent. Based on this criterion, we computed the **Violation Rate**, defined as the proportion

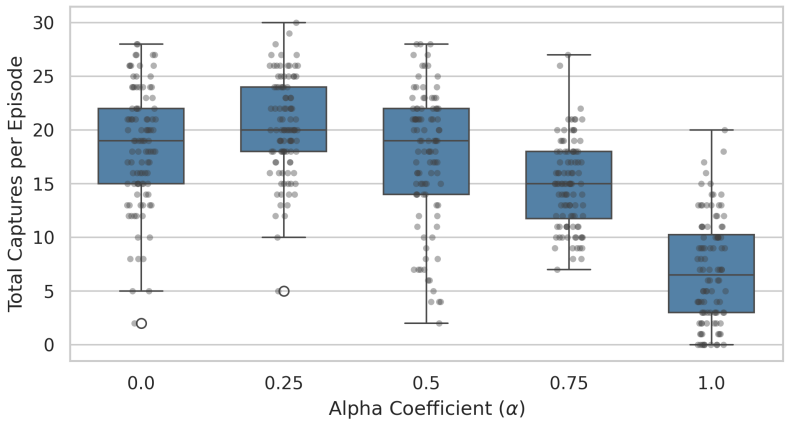

Figure 2: Average Capture Frequency per episode across experimental α profiles.

Table 1: Nash Equilibrium Stability Analysis. Regret is shown as Mean ± Std.

Alpha	Baseline Reward	Deviated Reward	Regret (Mean ± Std)	Violation Rate (%)
0.00	488.01	677.31	189.29 ± 191.28	78.00
0.25	429.29	580.04	150.75 ± 145.94	81.00
0.50	210.18	348.98	138.80 ± 133.99	73.00
0.75	159.15	238.99	79.84 ± 81.19	74.00
1.00	72.52	124.25	51.73 ± 49.71	74.00

of states exhibiting strategic incentives to deviate from the stalling policy. This metric quantifies the empirical frequency of deviations from equilibrium behavior.

We sampled $N = 100$ states where agents exhibited stalling behavior and calculated the mean regret for deviating from the static policy. Table 1 summarizes these findings across different α profiles.

The results show a clear downward trend in Mean Regret as α increases. In the fully cooperative profile ($\alpha = 0.0$), the Mean Regret reached its peak at 189.29, representing a potential reward increase of $\approx 38\%$ through deviation. Conversely, in the fully competitive profile ($\alpha = 1.0$), the Mean Regret reached its minimum at 51.73. Notably, while the magnitude of the regret varied significantly, the Violation Rate remained relatively stable across all configurations, ranging between 73% and 81%.

5 Discussion

The average capture frequency trends indicate that reward redistribution facilitates coordination by aligning agent incentives with collective success. While the $\alpha = 1.0$ baseline results in risk-averse passivity, the transition to shared rewards at $\alpha = 0.75$ triggers coordination. Peak performance at

$\alpha = 0.25$ highlights an optimal individual incentive. By preserving a degree of individual incentive, this configuration ensures that agents maintain a consistent pursuit drive. This approach effectively prevents the 'lazy agent' effect observed at $\alpha = 0.0$, where the total decoupling of rewards from individual actions results in a minor decrease in individual effort.

5.1 Spatial Analysis

The influence of α on agent policy is qualitatively evidenced by the spatial dynamics shown in Figure 3. At $\alpha = 1$, the predators exhibit extensive spatial dispersion with overlapping trajectories. This overlap suggests a lack of coordination between the predators; instead, each agent appears to prioritize individual reward maximization. Consequently, the prey maintains a relatively large operational area, indicating that the predators' uncoordinated pursuit is less effective at constraining the prey's movement.

In contrast, $\alpha = 0.5$ reveals cooperative dynamics. While two predators (1 and 2) engage in strategic flanking maneuvers from opposing sides, the third predator adopts a 'blocking' role, oscillating within a confined region to intercept potential escape routes. This collective behavior effectively restricts the prey's spatial occupancy to a minimal area, demonstrating a clear transition from individual pursuit to a coordinated encircling strategy.

Finally, at $\alpha = 0$, the predators operate in adjacent yet distinct spatial partitions, working in close proximity without redundant overlapping. This coverage indicates a highly coordinated approach where the agents effectively partition the environment to encircle the prey, validating that lower α values encourage complex inter-agent coordination.

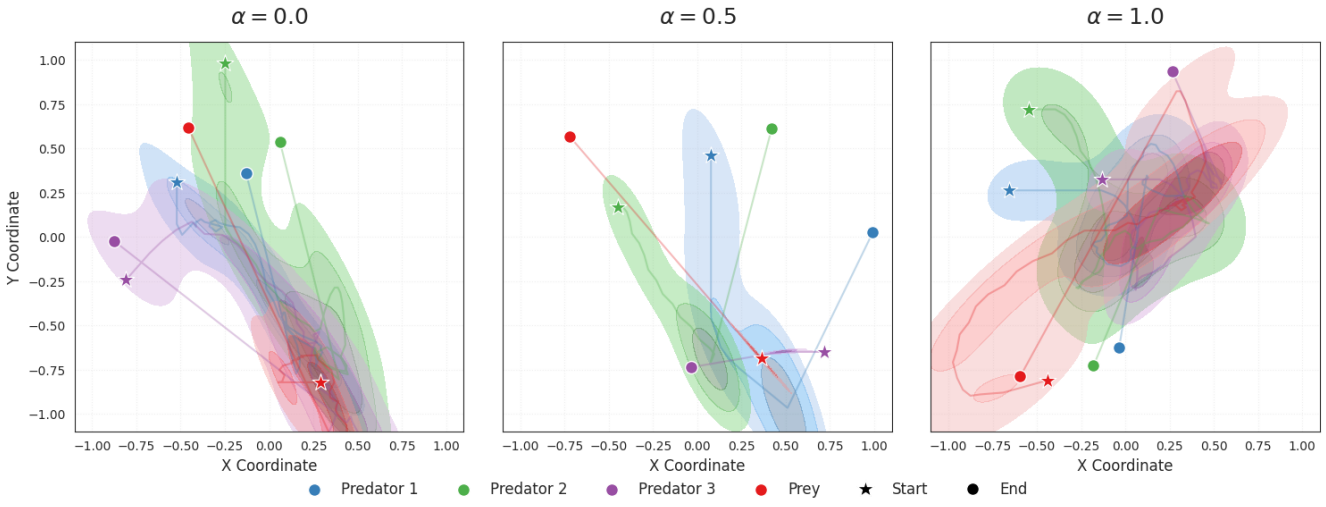

Figure 3: Spatial occupancy heatmaps for Episode 1 across $\alpha \in \{0, 0.5, 1\}$. The markers denote starting (*) and ending (o) positions of each agent.

5.2 Stability Analysis

Stability analysis reveals a trade-off: shifting toward cooperation ($\alpha \to 0$) improves capture frequency but destabilizes equilibrium. Mean Regret rose from 51.73 ($\alpha = 1.0$) to 189.29 ($\alpha = 0.0$), suggesting that cooperative rewards incentivize individual deviations at the expense of team coordination, which is a classic credit assignment challenge [3].

The consistently high standard deviation, often near the mean (e.g., 189.29 ± 191.28 for $\alpha = 0$), shows large outcome variability. This is due to stalling states being sensitive to small differences in agent positions, and because a single agent's action change can strongly affect the whole team [2].

Furthermore, the persistent Violation Rate (73%–81%) suggest that stalling is not a true Nash equilibrium. Instead, it appears to be a suboptimal convergence point caused by the PPO learning process itself. As independent PPO is known to be highly sensitive to hyperparameters in mixed environments [3], the agents likely settled on this strategy because the algorithm failed to explore more stable alternatives, highlighting the fragility of policies under shared rewards.

5.3 Limitations & Future Work

A technical limitation of this project is the unavailability of collision data. Although intended as a metric for spatial coordination, the PettingZoo library consistently reports zero collisions regardless of agent interactions, precluding the use of collisions to further validate coordination patterns.

Future work should focus on developing a modular computational library based on the methodology proposed in this study to provide a general-purpose framework for identifying Empirical Nash Equilibria (ENE) in simultaneous-move games. Standardizing this approach would facilitate broader stability analysis across diverse MARL environments beyond the scope of the Simple Tag scenario.

5.4 Use of Generative AI

This project utilized AI-assisted tools:

- Gemini: Employed for refining phrasing and assisting in code implementation.
- NotebookLM: Utilized as a research assistant to evaluate the relevance and contextual alignment of literature sourced from Google Scholar.

6 Appendices

For appendices please see the attached files.

References

[1] Ishan Durugkar, Elad Liebman, and Peter Stone. Balancing individual preferences and shared objectives in multiagent reinforcement learning. International Joint Conference on Artificial Intelligence, 2020.

[2] Ishan Honhaga and Claudia Szabo. A simulation and experimentation architecture for resilient cooperative multiagent reinforcement learning models operating in contested and dynamic environments. *Simulation*, 100(6):563–579, 2024.

[3] Ken Ming Lee, Sriram Ganapathi Subramanian, and Mark Crowley. Investigation of independent reinforcement learning algorithms in multi-agent environments. *Frontiers in Artificial Intelligence*, 5:805823, 2022.

[4] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in Neural Information Processing Systems*, 30, 2017.

[5] Igor Mordatch and Pieter Abbeel. Emergence of grounded compositional language in multi-agent populations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[6] Alexander Rutherford, Benjamin Ellis, Matteo Gallici, Jonathan Cook, Andrei Lupu, Garar Ingvarsson Juto, Timon Willi, Ravi Hammond, Akbir Khan, Christian Schroeder de Witt, et al. Jaxmarl: Multi-agent rl environments and algorithms in jax. *Advances in Neural Information Processing Systems*, 37:50925–50951, 2024.

[7] Ardi Tampuu, Tambet Matiisen, Dorian Kodelja, Ilya Kuzovkin, Kristjan Korjus, Juhan Aru, Jaan Aru, and Raul Vicente. Multiagent cooperation and competition with deep reinforcement learning. *PLOS ONE*, 12(4):e0172395, 2017.

[8] Jordan Terry, Benjamin Black, Nathaniel Grammel, Mario Jayakumar, Ananth Hari, Ryan Sullivan, Luis S Santos, Clemens Dieffendahl, Caroline Horsch, Rodrigo Perez-Vicente, et al. Pettingzoo: Gym for multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34:15032–15043, 2021.