Asymmetric Communication Policies in Multi-Agent Reinforcement Learning for Tethered Robots

Chintan Shah

February 2025

Abstract

Multi-Agent Reinforcement Learning (MARL) has traditionally focused on environments with bidirectional communication, yet real-world applications often face one-way or unreliable transmissions. This study explores asymmetric communication strategies in MARL, specifically for tethered robots operating in a partially observable 2D grid world. The system consists of a Leader (R1) with full map awareness and a Follower (R2) that relies solely on local sensing and R1's directional signals. A key challenge is ensuring effective coordination despite limited feedback and differing perception capabilities.

To address this, we propose a Multi-Agent PPO (MAPPO) framework combined with Graph Neural Networks (GNNs) for message encoding and a Sparse Attention Mechanism to optimize communication. Our approach is tested against two baseline models: a fully communicative MARL model and a no-communication model, using metrics such as completion rate, navigation efficiency, tether constraint violations, and collision rate.

1 Project Description

1.1 Environment Description

This project involves a 2D grid-based environment filled with obstacles, where two robots, R1 (Leader) and R2 (Follower), must navigate from Point A to Point B while staying within a fixed tethered distance. Since the environment is only partially observable, each robot can only see a limited area around itself.

1.2 Agent Interaction

R1 (Leader) Has access to a full map and a pre-planned navigation path but does not receive any feedback from R2. This agent can only send simple directional commands (go, stop, left, right) to R2.

R2 (Follower) Has local sensing capabilities (ray-based detection, grid-based occupancy awareness) but cannot see the entire map. It depends entirely on R1's instructions and its own local observations to decide its movement.

1.3 Action Space

Both robots have discrete movement options: {Up, Down, Left, Right, Stay}.
R1's movements are based on its independent decision-making and R2's movements are influenced by both its own perception and the signals received from R1.

1.4 Tether Constraint

The robots must remain within a predefined maximum distance. If they exceed this limit, they incur a penalty.

2 Objectives & Challenges

2.1 Agent's Goal

The main goal of a robot is to reach the goal position (B) from an arbitrary starting position (A) successfully.

2.2 Agent's Constraints

Robots should avoid obstacles and maintain the tether constraint. Moreover, it is important to minimize the number of steps to improve efficiency.

2.3 Key Challenges

Asymmetric communication R2 cannot send feedback, making it difficult for R1 to adjust its strategy dynamically.

Different levels of perception R1 has a global view, while R2 can only see its immediate surroundings.

Hence, coordinating movement effectively within these constraints will make an agent achieving its goal challenging.

3 Why This Problem Matters?

This project could potentially be extended to several real-world applications.

3.1 Real-World Applications

Warehouse & Supply Chain Robotics [BA23] Refers to Leader-follower robot systems where a primary unit sends commands, and secondary units must interpret and execute them autonomously.

Military & Search-and-Rescue Operations Potentially extends to the idea of drones guiding ground robots in disaster zones where the communication is unreliable [Cal+24].

Autonomous Convoys [Mas+24] Implies groups of vehicles following a leader, where only the lead vehicle has access to GPS or the satellite data. Those groups of vehicles could employ possibly a multitude of algorithms, namely the sparse communication graph model [SSH20], which scales up a multi-agent reinforcement learning based robotic system by reducing the communication overhead and at the same time ensuring coordinated convoy movements. Apart from that, those vehicles are likely employed effective collision avoidance mechanisms for safety [Na+22].

4 Research Significance

Most multi-agent reinforcement learning (MARL) research assumes bidirectional communication, but real-world systems often rely on one-way or unreliable transmissions. This study challenges existing MARL approaches by investigating optimal learning strategies under such constraints.

5 Baseline Environment & Algorithms

5.1 Baseline Environment

Fully Communicative MARL Model Both R1 and R2 can exchange full-bidirectional messages. It can be used as a benchmark to compare against the one-way communication constraint.

No Communication Model R2 relies only on local detection which completely ignores messages from R1.

5.2 Baseline Algorithms

Multi-Agent Deep Q-Network (MADQN) This algorithm uses independent Q-learning for each agent. D-MARL [Cal+24] explores independent Q-learning variants through **Heterogeneous-Agent Proximal Policy Optimization (HAPPO)** and shared critics. In our context, R1 learns what signals to send, and R2 learns how to interpret them. As suggested by D-MARL [Cal+24], it is a type of decentralized Q-learning strategy where agents decide their own actions independently.

Independent PPO (IPPO) Inspired by the idea of independent agents [Cal+24] learning decentralized policies within minimal communication, we intend to design a mechanism where both robots are trained independently, and at the same time, R2 make decisions based on the signals received from R1.

6 Proposed Solution Approach

6.1 Learning Algorithms

Firstly, **multi-Agent Proximal Policy Optimization (MAPPO)** [KCS21] allows agents to share experiences while maintaining decentralized decision-making. Then, **Graph Neural Networks (GNNs)** [Low+20] are used for message encoding which helps R2 learn patterns in R1's signals over time. In addition, a **Sparse Attention Mechanism** [DMS19] optimizes when R1 should communicate, preventing unnecessary messages.

6.2 Project Backup Plan

If the MARL approach struggles, alternative solutions include:

Hierarchical Reinforcement Learning (HRL) R1 acts as a high-level planner, while R2 executes specific actions. This is a joint control-and-communication optimization mechanism where decisions are divided between high-level planners and low-level executors. [Luo+24] For example, in autonomous systems (like connected vehicles), hierarchical structures are often implied to differentiate route planning (leader) from control execution (follower).

Self-Supervised Learning for Signal Interpretation If MARL performs poorly, a supervised loss function can be introduced to refine R2's decision-making.

6.3 Testing & Performance Comparisons

6.4 Evaluation Metrics

- **Completion Rate**: Percentage of successful goal-reaching attempts.
- **Navigation Efficiency**: Steps taken relative to the optimal path.
- **Tether Constraint Violations**: Number of times the robots exceed the maximum allowed distance.
- **Collision Rate** Number of times R2 collides with obstacles.

6.5 Comparative Baselines

No Communication Model Tests on whether one-way signals improve performance over local sensing alone.

Fully Communicative Model Measures whether bidirectional communication leads to better results.

Asymmetric Model (Ours) Evaluates learned policies against both baselines.

6.6 Validation

If our approach improves task completion, reduces collisions, and enhances efficiency under one-way communication constraints, it proves its practical viability.

7 Research Roadmap

The leader is responsible for dividing the subtasks, managing time and preparing the report and analysis for the assigned section among the other team members.

Research Roadmap Table

Phase	Task	Lead	Time
1	**Develop a 2D MARL simulation (using VMAS or PettingZoo)**: Train basic PPO agents for R2's navigation.	Kelvin	Week 1-2
2	**Train MADQN and MAPPO models**: Integrate one-way communication constraints.	Kimia	Week 3-4
3	**Evaluate performance against baseline models**: Optimize message encoding and sparse attention mechanisms.	Chintan	Week 5-6

Future Work (Beyond 6 Weeks)

- **Scaling to Multi-Agent Navigation [Qiu+23]**: Extending the model to handle convoys of robots following R1.
- **Simulating Real-World Signal Loss [Cal+24]**: Testing how disruptions in communication affect performance.
- **Hardware Deployment on Physical Robots**: Transitioning from simulation to real-world implementation.

8 Implementation consideration

- Computationally feasible using NVIDIA 2080/3060 GPUs.
- No real-world sensor data required, fully simulated environment.
- Aligns with Project Stubs B & C by modifying communication structures.
- Extends MARL research into realistic asymmetric communication scenarios.

References

[BA23] Marc-André Blais and Moulay A. Akhloufi. "Reinforcement Learning for Swarm Robotics: An Overview of Applications, Algorithms, and Simulators". In: *Cognitive Robotics* 3 (2023), pages 226–256. https://doi.org/10.1016/j.cogr.2023.07.004.

[Cal+24] Gabriele Calzolari, Vidya Sumathy, Christoforos Kanellakis, and George Nikolakopoulos. "D-MARL: A Dynamic Communication-Based Action Space Enhancement for Multi Agent Reinforcement Learning Exploration of Large Scale Unknown Environments". In: *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Abu Dhabi, UAE, Oct. 2024, pages 3470–3475. https://doi.org/10.1109/IROS58592.2024.10801319.

[DMS19] Abhijit Das, Sarthak Mittal, and Gaurav Sukhatme. "Tarmac: Targeted Multi-Agent Communication". In: *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*. 2019. https://proceedings.mlr.press/v97/das19a.html.

[KCS21] Jungsoo Kim, Kyunghyun Cho, and David Sontag. "Communication-Efficient Multi-Agent Reinforcement Learning via Signaling". In: *Advances in Neural Information Processing Systems (NeurIPS 2021)*. 2021. https://proceedings.neurips.cc/paper/2021/hash/486c0401c56bf7ec2daa9eba58907da9-Abstract.html.

[Low+20] Ryan Lowe, Jakub Sygnowski, Alexander I. Cowen-Rivers, Wendelin Böhmer, Jost Tobias Springenberg, Nicolas Heess, and Yuhuai Wu. "Multi-Agent Policy Optimization with Distributional Reinforcement Learning". In: *Advances in Neural Information Processing Systems (NeurIPS 2020)*. 2020. https://proceedings.neurips.cc/paper_files/paper/2020/hash/8b5c8441a8ff8e151b191c53c1842a38-Abstract.html.

[Luo+24] Ruyu Luo, Hui Tian, Wanli Ni, Julian Cheng, and Kwang-Cheng Chen. "Deep Reinforcement Learning Enables Joint Trajectory and Communication in Internet of Robotic Things". In: *IEEE Transactions on Wireless Communications* 23.12 (Dec. 2024), pages 18154–18165. https://doi.org/10.1109/TWC.2024.3462450.

[Mas+24] Federico Mason, Federico Chiariotti, Andrea Zanella, and Petar Popovski. "Multi-Agent Reinforcement Learning for Coordinating Communication and Control". In: *IEEE Transactions on Cognitive Communications and Networking* 10.4 (Aug. 2024), pages 1566–1578. https://doi.org/10.1109/TCCN.2024.3384492.

[Na+22] Seongin Na, Hanlin Niu, Barry Lennox, and Farshad Arvin. "Bio-Inspired Collision Avoidance in Swarm Systems via Deep Reinforcement Learning". In: *IEEE Transactions on Vehicular Technology* 71.3 (Mar. 2022), pages 2511–2525. https://doi.org/10.1109/TVT.2022.3145346.

[Qiu+23] Wei Qiu et al. "Off-Beat Multi-Agent Reinforcement Learning". In: *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023)*. IFAAMAS. London, United Kingdom, May 2023, pages 2424–2426.

[SSH20] Chuangchuang Sun, Macheng Shen, and Jonathan P. How. "Scaling Up Multiagent Reinforcement Learning for Robotic Systems: Learn an Adaptive Sparse Communication Graph". In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Las Vegas, NV, USA, Oct. 2020, pages 11755–11762. https://doi.org/10.1109/IROS45743.2020.9341303.

Appendix: Project Stubs

Project A

Multiple speakers are aware of their relative position in space. However, the follower/listener does not know this relative positioning and has to learn how to fuse the information from multiple guidance sources in a coherent fashion. Furthermore, the order in which the guidance arrives is not indicative of spatial relationship between the speakers. In fact, the order of these messages is scrambled.
Goal: Create an algorithm that will allow the speakers and the listener to learn how facilitate the listener's task through communication. Demonstrate that either speaker's identity or spacial position is coded in their learned message.

Project B

Multiple speakers are aware of their relative position in space. The follower/listener does recognise their relative position in space as well, and has to fuse the information from multiple guidance sources in a coherent fashion. Furthermore, the order in which the guidance arrives is fixed and is clearly perceived as coming from a specific speaker. However, random communication failures sometimes block the original message from a speaker (usually one, but not always the same one). The non-blocked speakers can identify this occurence and contribute to the signal-channel of the blocked speaker. The non-blocked speakers make the decision to contribute independently of each other, i.e., there's not central decision maker to decide who and what will contribute. All contributes are superimposed on each other.
Goal: Create an algorithm that will allow the speakers and the listener to learn how facilitate the listener's task through communication. Demonstrate that the speakers' contribution strategies to the blocked-spaker signal-channel are correlated.

Project C

Multiple speakers are aware of their relative position in space. The follower/listener does recognise their relative position in space as well, and has to fuse the information from multiple guidance sources in a coherent fashion. Furthermore, the –order– in which the guidance arrives is fixed, but does not contain speaker id a priori. Random communication failures sometimes block the original message from one or more speakers. Thus, the number of messages that arrive can vary.
Goal: Create an algorithm that will allow the speakers and the listener to learn how facilitate the listener's task through communication. Demonstrate how the communication strategy and performance depend (if at all) on the probability of communication failure.

Project D

Consider a MARL that uses full parameter sharing (swarm-like behaviour), but no communication, and is capable of solving at least one of the following scenarios of VMAS:

- balance

- wheel

Now, modify the scenario so that at random intervals one random agent (not necessarily the same) loses control for 'k' steps – a freeze event. More specifically, while the "frozen" agent continues to receive its usual perceptions and makes an action decision, the chosen action is not implemented. Instead, a default "zero-accelleration" action takes affect.
Goal: What will be the performance of the solution algorithm under these conditions as a function of the "freeze" probability? Can you modify the algorithm to recover the performance? You can expand the sensory input of the agent to include information about other agents, but this information should be anonymous and non-ordered.

Project E

Consider a MARL that uses full parameter sharing (swarm-like behaviour), but no communication, and is capable of solving at least one of the following scenarios of VMAS:

- balance
- wheel

Now, modify the scenario so that at random intervals one random agent (not necessarily the same) loses perception for 'k' steps – a "blind" event. More specifically, while the "blinded" agent's chosen actions continue to be implemented as usual, the observations it receives are supplanted by a neutral "all-zeros" signal or "white noise".
Goal: What will be the performance of the solution algorithm under these conditions as a function of the "blind" probability? Can you modify the algorithm to recover the performance? You can expand the (non-blinded) sensory input of the agent to include information about other agents, but this information should be anonymous and non-ordered.