Adversarial Attacks and Defenses in Multi-Agent Deep Reinforcement Learning: A Comprehensive Survey

Hamza Ba-mohammed[1], Latifa El Bouga[1], Mostapha Essoulami[1], Douae Benhlima[1], Amine Andam[1], Jamal Bentahar[2,3], Mustapha Hedabou[1]

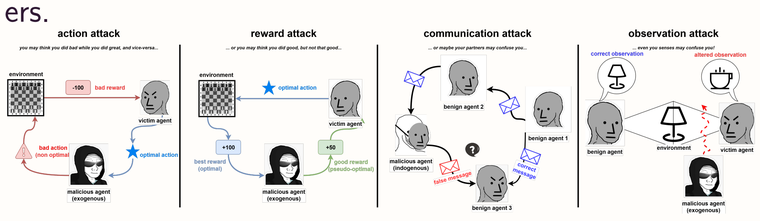
[1]College of Computing, Mohammed VI Polytechnic University, Benguerir, Morocco

[2]Concordia Institute for Information Systems Engineering, Concordia University, Montreal, Canada

[3]Department of Electrical Engineering and Computer Science, Khalifa University, Abu Dhabi, UAE

Introduction

Background **Deep Reinforcement Learning (DRL)** provides a foundation where agents learn optimal behaviors by interacting with an environment and maximizing rewards. When extended to **Multi-Agent Systems (MAS)**, this naturally evolves into **Multi-Agent Deep Reinforcement Learning (MADRL)**, where multiple agents learn simultaneously in a shared setting. In the context of **cyber-security**, such systems often involve adversarial interactions, giving rise to **adversarial MADRL**, where some agents actively seek to disrupt or exploit others.



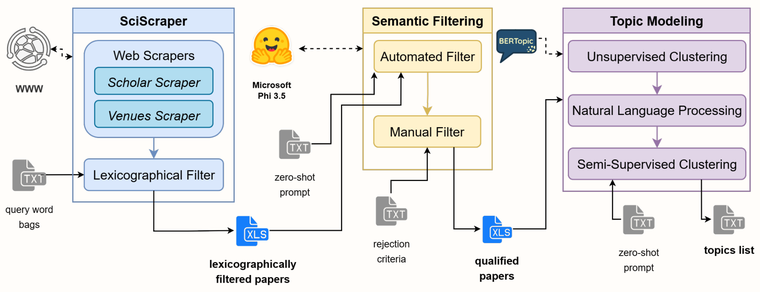
Motivation Existing surveys on adversarial MARL are insufficient. One lacks dedicated focus on adversarial aspects[1], while another is limited to test-time attacks and ignores train-time attacks[2].

Research Questions Our study is guided by the following three fundamental questions:

1. What are the known vulnerabilities of MADRL?
2. How can multi-agent systems be made more robust?
3. How does current research translate into real-world?

Methodology

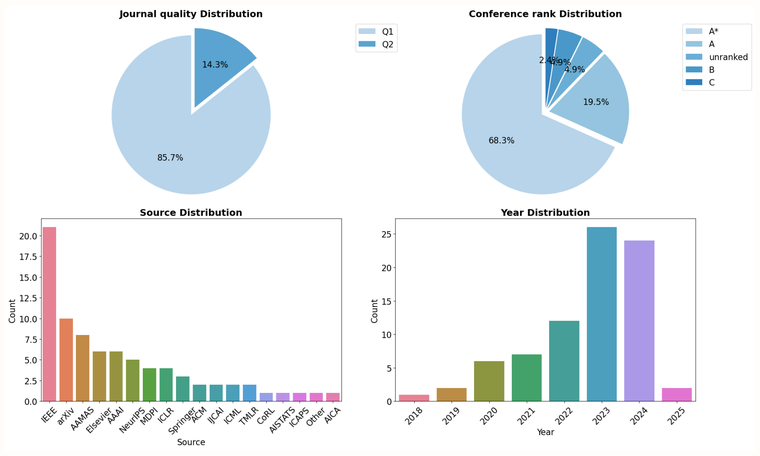
We conduct this survey through a reproducible, four-phase methodology (Collection, Filtering, EDA, Topic Modeling) designed for comprehensive analysis and easy adaptation to other research domains. **80 papers** were finally selected.



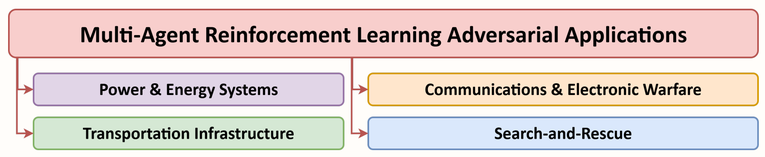
The final list of topics T, comprising **Attacks, Defenses, Communication, and Applications**, was then used to classify the papers into distinct themes.

Findings

Statistical Overview

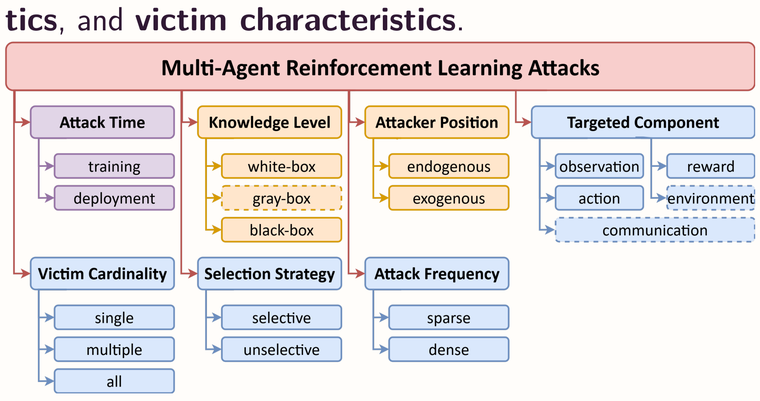


Applications

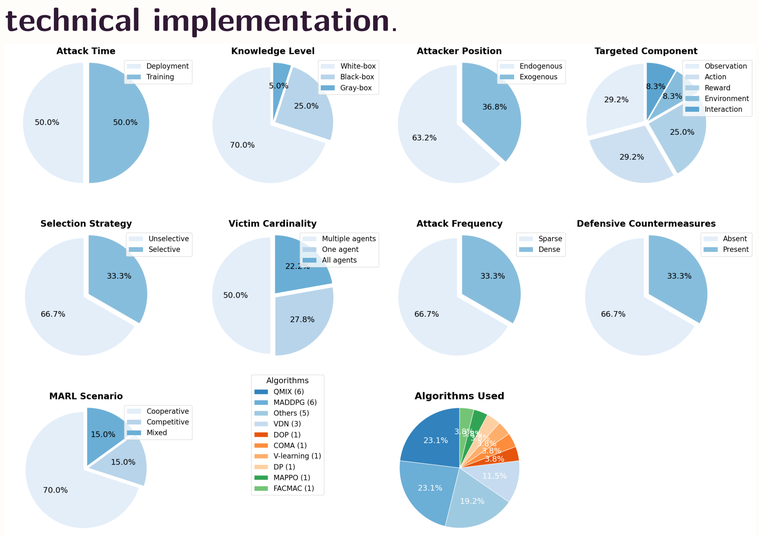


Attacks

Our taxonomy categorizes adversarial interactions along three core axes: **attack timing**, **attacker characteristics**, and **victim characteristics**.

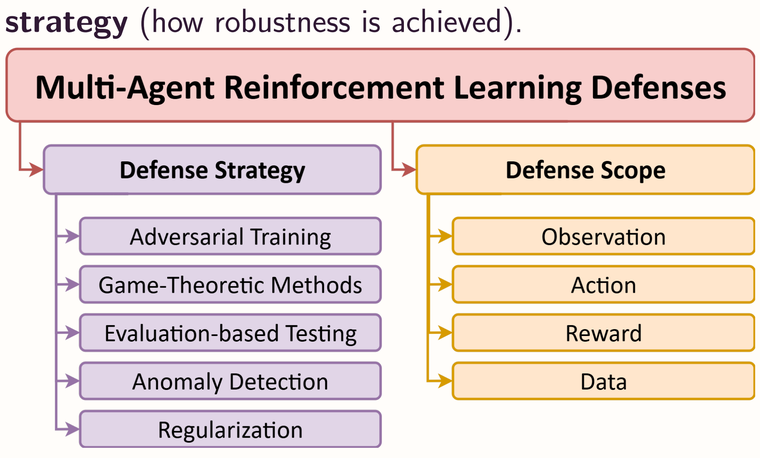


It's further enriched by three complementary dimensions: **mathematical modeling**, **defense perspective**, and **technical implementation**.

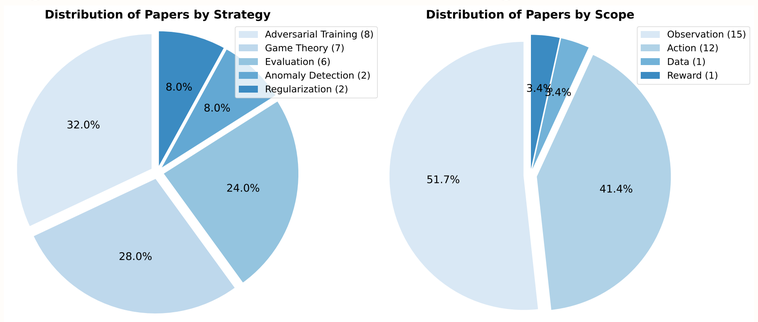


Defenses

We categorize MARL defenses along two key dimensions: the **defense scope** (what is protected) and **the defense strategy** (how robustness is achieved).

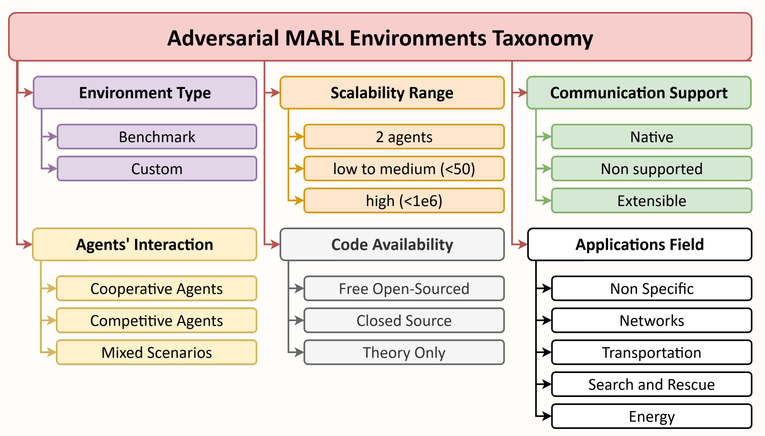


Current defenses predominantly target observation/action spaces, with adversarial training and game theory emerging as the dominant robustness strategies.



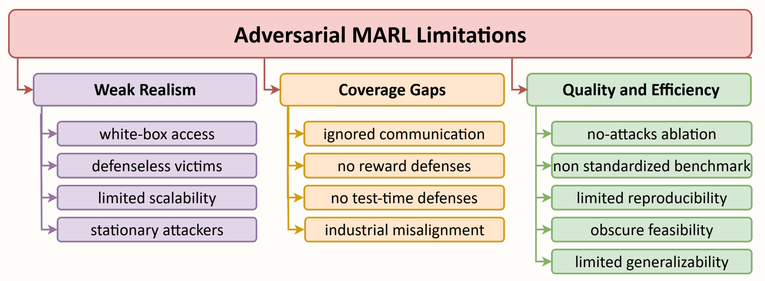
Environments

We evaluate MARL environments using six key criteria:



Limitations

We categorize current challenges in adversarial MARL into three core gaps:



Future Directions

- **Enhance Realism:** Develop scalable, black-box environments with adaptive defenses and multi-vector adversaries
- **Expand Coverage:** Address communication attacks and create runtime mitigation strategies for emerging domains (Agentic AI, 6G Networks, Smart Cities...)
- **Improve Rigor:** Establish standardized evaluations and report computational costs for deployable systems

*

References

[1] Standen, Maxwell, Junae Kim, and Claudia Szabo. "Adversarial machine learning attacks and defences in multi-agent reinforcement learning." ACM Computing Surveys 57.5 (2025): 1-35.

[2] Zhou, Ziyuan, Guanjun Liu, and Ying Tang. "Multi-agent reinforcement learning: Methods, applications, visionary prospects, and challenges." arXiv preprint arXiv:2305.10091 (2023).