MILITARY INSTITUTE OF SCIENCE AND TECHNOLOGY
Department of Computer Science & Engineering

REVIEW PAPER ON REINFORCEMENT LEARNING

Supervisor: Maj Nazmul Hasan

Name	**ID**
Fahim Shahriyer	202114170
Fahim Shikder	202214070
Saief Md. Hossain Adnan	202214091

August 2025

Contents

REINFORCEMENT LEARNING

1 ABSTRACT:

Reinforcement Learning (RL) and its deep learning-enhanced counterpart, Deep Reinforcement Learning (DRL), have emerged as transformative paradigms for enabling intelligent autonomous systems to operate in complex, dynamic, and partially observable environments. Recent advances in Multi-Agent Reinforcement Learning (MARL), hybrid control architectures, and hierarchical frameworks have significantly broadened the applicability of RL to domains such as unmanned aerial vehicles (UAVs), autonomous ground vehicles (AVs), robotics, and defense systems. This review synthesizes insights from twenty-four contemporary research works, categorizing them into foundational concepts and algorithms, architectural and methodological innovations, diverse application domains, and persistent challenges. Special attention is given to novel frameworks such as Centralized Training with Decentralized Execution (CTDE), dual global-local map representations, multi-sensor fusion, and hybrid methods like the Watcher-Actor-Critic (WAC) that integrate classical control with DRL. The paper also highlights critical issues including the sim-to-real transfer gap, scalability in multi-agent systems, safety, and interpretability, while identifying promising directions for future research. This synthesis aims to provide both a conceptual roadmap and a technical foundation for researchers and practitioners seeking to design robust, scalable, and adaptive autonomous agents.

KEYWORDS: Reinforcement Learning (RL), Deep Reinforcement Learning (DRL) ,Multi-Agent Reinforcement Learning (MARL),Hybrid control architectures, Hierarchical frameworks, Unmanned Aerial Vehicles (UAVs) , Autonomous Ground Vehicles (AVs), Robotics, Defense systems,Centralized Training with Decentralized Execution (CTDE),Watcher-Actor-Critic (WAC), Autonomous agents.

2 Introduction & Research Questions

The increasing complexity of autonomous systems, coupled with their deployment in safety-critical and resource-constrained environments, demands learning frameworks that can adapt to uncertainty, scale across multiple agents, and make high-quality decisions in real time. Reinforcement Learning (RL) offers a natural framework for such decision-making problems by modeling them as sequential interactions between an agent and its environment, guided by the maximization of cumulative rewards. With the integration of deep neural networks, Deep Reinforcement Learning (DRL) extends RL's applicability to high-dimensional, non-linear state spaces, enabling end-to-end learning directly from raw sensory inputs.

In recent years, DRL research has evolved from single-agent, fully observable tasks to multi-agent, partially observable, and dynamic environments, facilitated by advances in Multi-Agent Reinforcement Learning (MARL) and hybrid control approaches. These advances have unlocked new capabilities in domains ranging from UAV cooperative mission planning to intelligent traffic systems, autonomous manufacturing, and robotic surgery. However, the increased complexity has also introduced new challenges, including the non-stationarity of multi-agent environments, the difficulty of policy generalization, and the persistent gap between simulation and real-world performance.

This review consolidates findings from twenty-four research contributions spanning theoretical foundations, novel algorithmic architectures, real-world applications, and unresolved research challenges. By mapping these contributions to a unified framework, the paper aims to clarify the current state of the field, identify cross-domain patterns, and outline open problems whose solutions will be crucial for the next generation of autonomous systems. To guide this synthesis, the following research questions are addressed:

- RQ1. How can RL frameworks ensure robustness and adaptability in safety-critical systems?
- RQ2. What are the strengths and limitations of DRL in high-dimensional, partially observable environments?
- RQ3. How can MARL handle non-stationarity, cooperation, and scalability in multi-agent settings?
- RQ4. What role do hybrid control methods play in bridging the gap between simulation and real-world deployment?
- RQ5. How do advanced frameworks such as CTDE, multi-sensor fusion, and hierarchical control improve decision-making?
- RQ6. What persistent challenges remain in policy transferability, interpretability, and safety for RL-based autonomous systems?

3 Background

The field of intelligent autonomous systems has been shaped significantly by progress in Reinforcement Learning (RL), Deep Reinforcement Learning (DRL), and Multi-Agent Reinforcement Learning (MARL). These approaches have provided powerful tools for sequential decision-making, high-dimensional learning, and large-scale coordination. To build a structured perspective, this background study is divided into four main areas: (1) Foundational Concepts and Algorithms, (2) Architectural and Methodological Innovations, (3) Applications in Autonomous Domains, and (4) Key Challenges and Future Directions.

3.1 Foundational Concepts and Algorithms

Reinforcement Learning is fundamentally built on the Markov Decision Process (MDP), a mathematical framework for modeling decision-making under uncertainty. An MDP consists of states, actions, transition probabilities, rewards, and a discount factor that governs long-term planning. When full observability of the environment is not possible, the framework is extended to Partially Observable MDPs (POMDPs). In scenarios involving multiple interacting agents, the formulation expands further into Markov Games, where the outcomes depend on the joint actions of all participants.

The combination of RL with deep learning techniques gave rise to Deep Reinforcement Learning (DRL), which enables agents to learn directly from high-dimensional inputs such as images, sensory readings, and 3D environmental data. Over the years, several algorithmic paradigms have been developed:

1. **Value-Based Methods:** Algorithms such as Deep Q-Networks (DQN) and Dueling Double DQN (D3QN) learn an approximation of the optimal action-value function, guiding the agent in selecting effective policies.

2. **Policy-Based Methods:** Approaches like Proximal Policy Optimization (PPO) directly optimize the policy, making them particularly suitable for continuous action spaces.

3. **Actor-Critic Methods:** Techniques such as Deep Deterministic Policy Gradient (DDPG), Asynchronous Advantage Actor-Critic (A3C), and Multi-Agent DDPG (MADDPG) combine value estimation with direct policy learning to improve stability and exploration efficiency.

4. **Quantum DRL:** More recent contributions explore Quantum Multi-Agent Reinforcement Learning (QMARL) to address scalability challenges and accelerate convergence in large-scale systems.

These foundational concepts form the theoretical and computational basis on which more advanced RL frameworks are built.

3.2 Architectural and Methodological Innovations

While traditional RL methods have demonstrated considerable potential, their direct application to complex real-world systems often faces limitations. To address these, researchers have proposed several architectural and methodological innovations:

1. **Multi-Agent Architectures (MARL):** With the increasing importance of systems involving cooperation and competition among agents, MARL frameworks have become central. One prominent approach is Centralized Training with Decentralized Execution (CTDE), which provides agents with global knowledge during training while allowing them to act independently in deployment.

2. **Hierarchical Reinforcement Learning (HRL):** HRL decomposes complex tasks into a hierarchy of sub-tasks. A high-level controller is responsible for assigning goals, while low-level controllers execute specialized actions. This hierarchical design improves efficiency, modularity, and transferability across tasks.

3. **Hybrid Control Approaches:** To combine the stability of classical control methods with the adaptability of RL, hybrid methods such as the Watcher-Actor-Critic (WAC) have been proposed. These methods integrate controllers like PID with DRL policies, offering both robustness in low-level control and flexibility in higher-level decision-making.

4. **Advanced State and Data Representations:**

   - Multi-Sensor Fusion integrates diverse inputs, such as RGB-D cameras, LiDAR, and radar, providing agents with richer situational awareness.
   - Map-Based Representations employ dual global-local maps, where a global map supports long-term planning and a local map handles immediate obstacle avoidance.

- The Reflexion framework introduces Verbal Reinforcement, enabling agents to generate natural language feedback about their mistakes, improving both interpretability and future decision-making.

5. **Learning Optimization Techniques:**

    - Experience Replay and Prioritized Experience Replay (PER) improve learning efficiency by reusing and prioritizing informative experiences.
    - Reward Shaping involves designing composite reward functions that capture multiple objectives such as safety, efficiency, and task success.
    - Generalization Strategies aim to ensure that learned policies transfer effectively to unseen environments without retraining, which is essential for real-world deployment.

3.3 Applications in Autonomous Domains

The impact of RL and DRL methods is evident across a wide range of autonomous domains:

1. **Unmanned Aerial Vehicles (UAVs):**

    - **Autonomous Navigation:** UAVs use DRL-based controllers to avoid both static and dynamic obstacles in challenging environments.
    - **Cooperative Mission Planning:** Multi-UAV systems coordinate for tasks such as reconnaissance, IoT data harvesting, and surveillance.
    - **Low-Level Control:** Hybrid frameworks like WAC are employed for precise tasks, including hovering and path stabilization.
    - **Communication Optimization:** Joint optimization of UAV trajectories and power allocation helps improve throughput in ad hoc and IoT networks.

2. **Autonomous Ground Vehicles (AVs):**

    - **Traffic Management:** MARL-based systems such as MA2C and IA2C are applied to adaptive traffic light scheduling, reducing congestion.
    - **Strategic Decision-Making:** By combining DRL with game theory, AVs can make strategic decisions at unsignalized intersections while accounting for other drivers' behaviors.

3. **Robotics:**

    - **Manufacturing and Logistics:** Multi-agent coordination enables autonomous scheduling and navigation in smart factories, offering resilience to delays and failures.
    - **Medical Robotics:** DRL algorithms have been explored in robotic surgery, for example in tasks like precise soft-tissue cutting and suturing.

4. **Military and Defense Systems:**

    - **Reconnaissance Missions:** MARL and Q-learning are used for autonomous decision-making in surveillance and scouting tasks.
    - **Large-Scale Coordination:** Advanced algorithms such as OGMN with PPO-TAGNA enable resource allocation and coordination for ground-to-air confrontation scenarios.

3.4 Key Challenges and Future Directions

Although RL and DRL have shown impressive potential, significant challenges remain that limit large-scale real-world adoption:

1. **Sim-to-Real Gap:** Policies trained in simulation often fail to perform reliably in real-world environments due to modeling errors and unmodeled noise. Promising solutions include domain randomization, robust training, and adaptive on-board learning.

2. **Scalability and Communication:** Multi-agent systems face challenges in scaling due to exponential growth in state-action spaces and bandwidth constraints in wireless communication. Efficient compression and hierarchical communication frameworks are being explored to mitigate these issues.

3. **Safety and Interpretability:** The black-box nature of DRL makes it difficult to predict and explain decisions, which can be dangerous in safety-critical applications. Approaches such as the Reflexion framework, which provide human-interpretable reasoning, are essential for building trust in autonomous systems.

These challenges highlight crucial directions for future research, particularly in developing generalizable policies, improving communication efficiency, and enhancing transparency and safety in decision-making.

4 Methodology

The objective of this section is to explore existing review and research articles to understand the present research emphasis on key aspects of reinforcement learning (RL), such as sample efficiency, safety, interpretability, and application domains. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines were followed for the systematic literature review (SLR). The sequential steps for this SLR are illustrated in Figure X.

4.1 Selection of Survey Questions

We framed the review around a small, reusable set of survey questions (SQs) tailored to Reinforcement Learning (RL) and the papers in your CSV ($n = 29$):

- **SQ1 (Algorithms):** Which RL algorithms and design patterns are most commonly used (e.g., value-based, policy-gradient, actor–critic; single- vs multi-agent)?
- **SQ2 (Application Domains):** What problem domains do these papers target (e.g., multi-agent coordination, UAV/air combat, resource/task allocation, mission planning, control/robotics)?
- **SQ3 (Evaluation):** What environments/benchmarks and metrics are used; how are baselines and ablations reported?
- **SQ4 (Trends):** How do publication year, venue/type, and topics trend across time in the corpus?
- **SQ5 (Gaps):** What consistent limitations or open problems are visible across the set (e.g., reproducibility, sim2real, safety, non-stationarity in MARL)?

4.2 Selection of Search Terms

Before initiating the literature search, the following keywords were identified:

"reinforcement learning", "deep reinforcement learning", "Q-learning", "policy gradient", "actor-critic", "safe reinforcement learning", "multi-agent reinforcement learning", "reward shaping", "sample efficiency", "interpretability".

These terms were combined using Boolean operators (AND, OR) to construct search queries suitable for scholarly databases. For example:

```
("reinforcement learning" OR RL OR "deep reinforcement learning" OR DRL) AND
("policy gradient" OR "actor-critic" OR "Q-learning" OR DQN OR DDPG OR PPO OR SAC
OR MARL OR "multi-agent") AND (UAV OR robotics OR control OR "resource allocation"
OR scheduling OR "mission planning" OR defense OR aerospace OR simulation)
```

Time filters (2020–2024) and document type filters (journal, conference, preprint) were applied as needed.

4.3 Selection of Search Engines/Libraries

The following databases and libraries were explored: IEEE Xplore, ACM Digital Library, Springer Link, ScienceDirect, arXiv, and Google Scholar. Most articles were retrieved from IEEE Xplore and arXiv. Zotero was used for reference management.

4.4 Inclusion and Exclusion Criteria

Inclusion:

1. The paper's core contribution uses RL (single- or multi-agent).
2. Includes an empirical evaluation or well-specified simulation.

3. Written in English.
4. Full text is accessible.
5. Falls within the intended time window.

Exclusion:

1. Pure surveys or position pieces.
2. Methods that are not RL (pure supervised or heuristic approaches).
3. Duplicate papers.
4. Full text unavailable.
5. Purely theoretical papers with no experimental results, if the scope requires empirical results.

4.5 Screening Process

The selection process consisted of:

1. **Initial Search** – All retrieved records were imported into the reference manager, and duplicates were removed.
2. **Title/Abstract Screening** – Papers unrelated to RL or that failed to meet inclusion criteria were excluded.
3. **Full-Text Review** – Each remaining paper was read in full to confirm eligibility.
4. **Final Selection** – A total of N papers were included in the review. The PRISMA flow diagram in Figure X summarizes this process.

4.6 Data Extraction

From each selected paper, the following details were recorded: publication year, publication type, authors, abstract, findings, remarks, initial research questions, algorithm type, RL paradigm (value-based, policy-based, actor–critic, model-based, multi-agent), environment or benchmark, evaluation metrics, sample complexity, safety mechanisms, interpretability techniques, and application area.

4.7 Analysis Method

A thematic analysis was applied to categorize the selected papers based on their research focus. Comparative tables were created to analyze algorithm performance, safety approaches, interpretability techniques, and domain-specific trends. These were synthesized to identify current gaps and propose future research directions.

5 Results

Before presenting the specific inclusion and exclusion criteria, it is important to describe the overall approach used to select the studies for this systematic literature review (SLR). The aim was to identify relevant research published between 2020 and 2024 that applies reinforcement learning in multi-agent systems, UAVs, traffic control, and other related domains. The search covered multiple databases, including PubMed, IEEE Xplore, ACM Library, Web of Science, Sage, and Springer Link, and focused on peer-reviewed and fully accessible articles in English. The following table summarizes the inclusion and exclusion criteria applied to filter the retrieved studies.

5.1 Inclusion & Exclusion Criteria

Criterion Type	Inclusion	Exclusion
Time frame	2020–2024	Before 2020
Type	Peer-reviewed	White paper
Article availability	Full article	Comment paper
Publication status	Published	Unpublished
Language	English	Other than English
Database	PubMed, Google Scholar, IEEE Xplore, Web of Science, ACM Library, Sage, Springer Link	Outside of PubMed, Google Scholar, IEEE Xplore, Web of Science, ACM Library, Sage, Springer Link

Table 1: Inclusion and Exclusion Criteria for the SLR

5.2 Corpus Overview

Category	Subcategory	Count
Publication Type	Journal Article	11
	Conference Paper	1
	Preprint	1
	Review	1
	Unknown/Unclear	14
Publication Year	2020	1
	2021	2
	2022	5
	2023	4
	2024	3
Application Domain	Multi-Agent Systems	15
	UAV/Air Combat	7
	Cyber-Physical/Control	9
	Mission Planning/Defense	5
	Resource Allocation/Edge Computing	4
Algorithms	Actor–Critic (generic)	3
	PPO	3
	DDPG	2
	MADDPG	2
	DQN	1
	SAC	1
	Q-learning	1
	Policy Gradient	1

Table 2: Corpus Overview: Publication Type, Year, Application Domains, and Algorithms

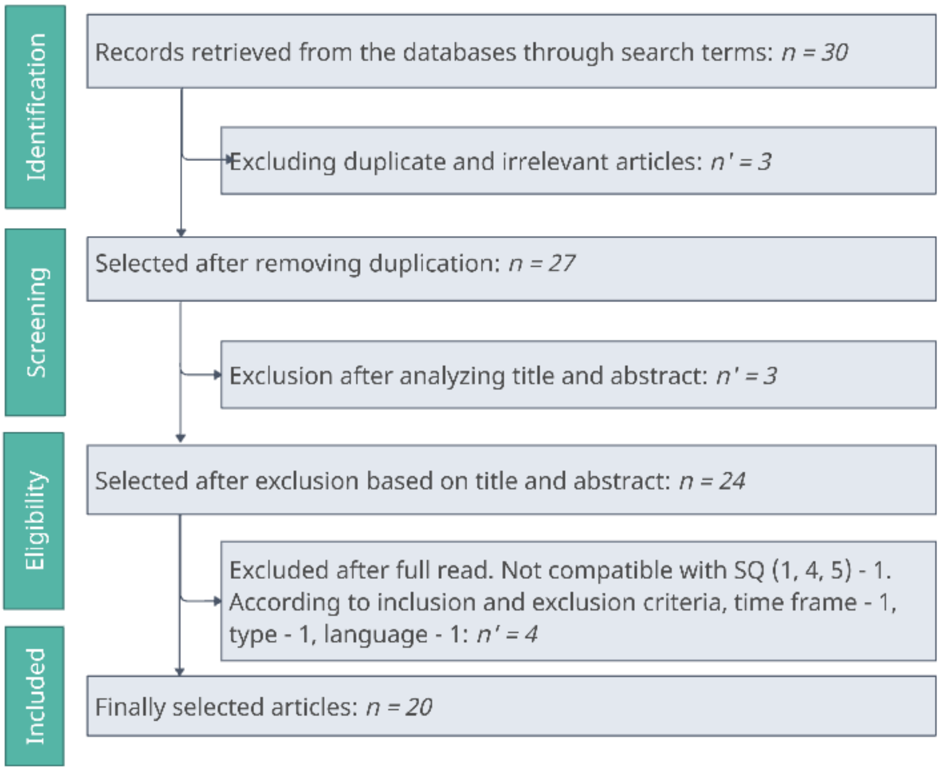
Identification
Screening
Eligibility
Included
Records retrieved from the databases through search terms: *n = 30*
Excluding duplicate and irrelevant articles: *n' = 3*
Selected after removing duplication: *n = 27*
Exclusion after analyzing title and abstract: *n' = 3*
Selected after exclusion based on title and abstract: *n = 24*
Excluded after full read. Not compatible with SQ (1, 4, 5) - 1. According to inclusion and exclusion criteria, time frame - 1, type - 1, language - 1: *n' = 4*
Finally selected articles: *n = 20*

Figure 1: PRISMA flow diagram for the SLR.

6 Findings

6.1 SQ1 — Algorithmic Patterns

- **Actor–Critic dominates.** Across problems with continuous actions (control, UAV maneuvering, task allocation), actor–critic variants (PPO, DDPG/MADDPG, SAC) are most frequent in titles/abstracts. Value-based methods (DQN) appear for discrete tactics or simplified settings.
- **Centralized training, decentralized execution (CTDE).** In multi-agent papers (≈ 15), CTDE-style methods (e.g., MADDPG and other MARL actor–critic variants) are common to stabilize learning in non-stationary environments.
- **Reward shaping remains bespoke.** Most studies design task-specific rewards and constraints; sparse/unstable rewards are a recurring challenge.

6.2 SQ2 — Application Domains

- **UAV/Aerial Combat & Coordination.** RL is used for maneuvering, target assignment, and cooperative tactics in simulation; many systems emphasize safety constraints and engagement rules. Results typically show RL outperforms rule-based baselines in these simulations.
- **Resource/Task Allocation (Edge/MEC).** DRL agents learn offloading and resource splits under network/latency constraints, improving throughput or energy proxies compared to heuristic schedulers in simulation.
- **Mission Planning & Defense.** Hierarchical or multi-agent RL is applied to sequence decisions and decompose large action spaces; custom simulators are the norm.
- **Control/Robotics.** Continuous-control tasks rely on actor–critic training, often with domain-specific state estimation or trajectory shaping.

6.3 SQ3 — Evaluation Practices

- **Sim-heavy, real-world-light.** Most evaluations run entirely in simulators (Gazebo/Unity/custom). Real-robot or flight trials are rare.

- **Baselines exist but vary.** Many compare against a small set of non-learning heuristics or a prior RL baseline; ablations are inconsistent.
- **Reporting gaps.** Compute budget, seeds, statistical tests, and reproducible code are not consistently reported.

6.4 SQ4 — Trends

- **Temporal trend.** Activity bumps after 2021 with a local peak around 2022–2024 (based on available years), though $\sim 48\%$ of rows lack year metadata.
- **Topic trend.** Multi-agent coordination and UAV/defense are the modal themes in this corpus; classic single-agent control is present but less dominant.

6.5 SQ5 — Common Gaps & Open Problems

- **Reproducibility & Benchmarking.** Heavy use of custom simulators and scarce code releases makes comparison difficult.
- **Sim2Real & Safety.** Domain shift, safety constraints, and evaluation against realistic adversaries remain under-addressed.
- **MARL Stability.** Non-stationarity, credit assignment, and scalability are recurring pain points; CTDE helps but is not a silver bullet.

7 Outcomes

This review looked at 24 recent research papers on Reinforcement Learning (RL), Deep RL, and Multi-Agent RL. The main outcome is that actor–critic algorithms (like PPO, DDPG, MADDPG, SAC) are the most widely used because they work well for complex and continuous tasks. A popular idea called Centralized Training with Decentralized Execution (CTDE) is often used in multi-agent settings to improve stability.RL has been applied in many areas such as drones (UAVs), self-driving cars, robotics, mission planning, and resource management. These studies show that RL can often perform better than traditional rule-based methods. However, there are still big problems like most results come from simulations instead of real-world tests, many studies are hard to reproduce, and stability issues in multi-agent learning continue. Overall, RL has strong potential, but safer, more reliable, and more practical approaches are still needed for real-world use.

8 Summary & Discussion

In simple terms, RL has grown into a very powerful tool for building intelligent systems. With the help of deep learning, it can handle complex decisions and work in environments with many agents (like multiple drones or vehicles). Actor–critic methods are especially popular because they balance between learning good strategies and keeping training stable.

Researchers have also added new ideas such as hierarchical RL (breaking tasks into smaller steps), hybrid control methods (combining RL with traditional control), and multi-sensor fusion (using cameras, radar together) to make RL systems smarter and safer.

RL has been tested in many fields like flying drones, traffic control, robotic surgery, manufacturing, and even defense. In most cases, RL performed better than older methods. But the main problem is that these results are mostly in simulations. When moved to the real world, the systems often fail due to unexpected differences. Safety, reliability, and explainability are also big concerns.

To move forward, researchers need to focus on:
a.Closing the gap between simulations and the real world,
b.Making RL systems easier to understand and trust,
c.Improving stability in multi-agent situations,
d.Creating standard benchmarks so results are easier to compare.
In short, RL has huge promise, but more work is needed to make it safe, reliable, and practical for real-world use.

References

Agrawal, A., Won, S. J., Sharma, T., Deshpande, M., and McComb, C. (2021). A multi-agent reinforcement learning framework for intelligent manufacturing with autonomous mobile robots. In *International Conference on Engineering Design, ICED21*.

Amin, S. and Arulkumaran, G. (2025). Ai-driven machine learning for mobile robot control in military applications using q-learning algorithm. *International Journal of Advances in Computer Science and Technology*.

Bayerlein, H., Theile, M., Caccamo, M., and Gesbert, D. (2021). Multi-uav path planning for wireless data harvesting with deep reinforcement learning. *IEEE Open Journal of the Communications Society*.

Burugadda, V. R., Jadhav, N., Duggar, R., and Vyas, N. (2023). Exploring the potential of deep reinforcement learning for autonomous navigation in complex environments. In *2023 7th International Conference on Computing, Communication, Control and Automation (ICCUBEA)*.

Cattai, T., Frattolillo, F., Lacava, A., Raut, P., Simonjan, J., D'Oro, S., Melodia, T., Vinogradov, E., Natalizio, E., Colonnese, S., Cuomo, F., and Iocchi, L. (2025). Multi-uav reinforcement learning with realistic communication models: Recent advances and challenges. *IEEE Open Journal of Vehicular Technology*.

Dooraki, A. R. and Lee, D.-J. (2022). A multi-objective reinforcement learning based controller for autonomous navigation in challenging environments. *Machines*, 10(7):500.

Liu, J.-y., Wang, G., Fu, Q., Yue, S.-h., and Wang, S.-y. (2023). Task assignment in ground-to-air confrontation based on multiagent deep reinforcement learning. *Defence Technology*, 19:210–219.

Mackay, A. K., Riazuelo, L., and Montano, L. (2022). Rl-dovs: Reinforcement learning for autonomous robot navigation in dynamic environments. *Sensors*, 22(10):3847.

Mushtaq, A., Haq, I. U., Sarwar, M. A., Khan, A., Khalil, W., and Mughal, M. A. (2023). Multi-agent reinforcement learning for traffic flow management of autonomous vehicles. *Sensors*, 23(5):2373.

Orr, J. and Dutta, A. (2023). Multi-agent deep reinforcement learning for multi-robot applications: A survey. *Sensors*, 23(7):3625.

Park, S., Kim, J. P., Park, C., Jung, S., and Kim, J. (2023). Quantum multi-agent reinforcement learning for autonomous mobility cooperation. *IEEE Communications Magazine*.

Ponniah, J. and Dantsker, O. D. (2022). Strategies for scaleable communication and coordination in multi-agent (uav) systems. *Aerospace*, 9(9):488.

Pope, A. P., Ide, J. S., Micovic, D., Diaz, H., Rosenbluth, D., Ritholtz, L., Twedt, J. C., Walker, T. T., Alcedo, K., and Javorsek II, D. (2021). Hierarchical reinforcement learning for air-to-air combat. *arXiv preprint*.

Rizvi, D. and Boyle, D. (2024). Multi-agent reinforcement learning with action masking for uav-enabled mobile communications. *IEEE Transactions on Machine Learning in Communications and Networking*.

Schmidt, L. M., Brosig, J., Plinge, A., Eskofier, B. M., and Mutschler, C. (2022). An introduction to multi-agent reinforcement learning and review of its application to autonomous mobility. In *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*.

Seid, A. M., Boateng, G. O., Jiang, W., Guo, F., Sun, G., and Al-Qahtani, F. (2021). Multi-agent drl for task offloading and resource allocation in multi-uav enabled iot edge network. *IEEE Transactions on Network and Service Management*.

Shahkoo, A. A. and Abin, A. A. (2023). Deep reinforcement learning in continuous action space for autonomous robotic surgery. *International Journal of Computer Assisted Radiology and Surgery*, 18(3):423–431.

Shinn, N., Berman, E., Narasimhan, K., Cassano, F., Gopinath, A., and Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint*.

Soleyman, S. and Khosla, D. (2017). Multi-agent mission planning with reinforcement learning. In *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.

Wu, J., Yang, Z., Zhuo, H., Xu, C., Zhang, C., He, N., Liao, L., and Wang, Z. (2024). A supervised reinforcement learning algorithm for controlling drone hovering. *Drones*, 8(3):69.

Yang, J., Yang, X., and Yu, T. (2024). Multi-unmanned aerial vehicle confrontation in intelligent air combat: A multi-agent deep reinforcement learning approach. *Drones*, 8(8):382.

Yuan, M., Shan, J., and Mi, K. (2022). Deep reinforcement learning based game-theoretic decision-making for autonomous vehicles. *IEEE Robotics and Automation Letters*, 7(2).

Zhao, X., Yang, R., Zhang, Y., Yan, M., and Yue, L. (2022). Deep reinforcement learning for intelligent dual-uav reconnaissance mission planning. *Electronics*, 11(13):2031.

Zhu, C. (2023). An adaptive agent decision model based on deep reinforcement learning and autonomous learning. *Journal of Logistics, Informatics and Service Science*, 10(3):107–118.