UDC 629.735.2:681.324(045)
DOI:10.18372/1990-5548.87.20909

Igor Yudenko

MULTI-AGENT DEEP REINFORCEMENT LEARNING IN THE COLLISION AVOIDANCE PROBLEM

Department of Avionics and Control Systems, Faculty of Air Navigation Electronics and Telecommunications, State University "Kyiv Aviation Institute", Kyiv, Ukraine
E-mail: ioyudenko@gmail.com ORCID 0000-0002-7022-366X

Abstract—*Obstacle avoidance is crucial for the succesfull completion of unmanned aerial vehicles missions. This article is devoted to the research of the multi-agent deep reinforcement learning in the collision avoidance problem. It is considered unmanned aerial vehicle swarms encounter diverse obstacles categorized into: static large-scale and small-scale obstacles, dynamic large-scale and small-scale obstacles, complex terrain, thin/low-visibility obstacles, partially-occluded/transparent obstacles. To address the abvove problem, a multi-agent deep reinforcement learning based trajectory control algorithm is proposed for managing the trajectory of each unmanned aerial vehicle independently. It was researched different approaches and multiple 3D simulation environments with help of reinforcement learning for the swarm of unmanned aerial vehicles.*

Keywords—Multi-agent reinforcement learning; swarm of unmanned aerial vehicles; obstacle avoidance; non-flying zones; deep reinforcement learning.

I. INTRODUCTION

Unmanned aerial vehicles (UAVs) have become crucial in solving a variety of complex civil and military tasks. These aircraft operate without human pilots, offering varying degrees of autonomy, from remote control to fully autonomous operation. Achieving high levels of autonomy is crucial for UAVs to perform missions in unpredictable situations without human intervention.

A unmanned aerial vehicle swarm is a group of many UAVs, often small and inexpensive, that operate cooperatively and autonomously (or semi-autonomously) to achieve a common goal.

Rather than simply having many drones flying independently or individually controlled by multiple operators, a true UAV swarm possesses several key characteristics:

1) Collective Intelligence: Individual UAVs interact and coordinate with each other, often using distributed algorithms and local interactions, to make collective decisions and exhibit complex group behavior.

2) Autonomy / Decentralization: While high-level control from a single operator or central system may exist, individual UAVs in a swarm typically possess significant autonomy. They make local decisions based on their sensor data and communications with nearby swarm members, rather than relying on constant individual commands from a central controller. This decentralization makes them resilient.

3) Cooperation and Coordination: UAVs work together to achieve a common goal. This may include maintaining specific formations, sharing data, collectively surveying an area, or suppressing a target from multiple directions.

4) Emergent Behavior: Complex, intelligent group behavior patterns can emerge from simple rules and interactions between individual UAVs, similar to natural swarms of insects or birds.

5) Robustness and Resilience: If one or more UAVs in a swarm fail or are lost, the remaining members can often adapt and continue the mission because information and tasks are shared.

6) Scalability: The system is often designed to function effectively regardless of whether it consists of dozens or hundreds of UAVs, without a proportional increase in human oversight.

II. RATIONALE FOR USING REINFORCEMENT LEARNING IN UAV SWARM CONTROL PROBLEMS

A multi-agent system (MAS) is a system consisting of many interacting intelligent agents that cooperate to achieve a common goal or solve complex problems that are difficult or impossible for a single agent to solve alone.

Key characteristics of a multi-agent system often include.

1) Multiple agents: The system consists of multiple individual entities, each capable of perceiving its environment, making decisions, and performing actions.

2) Intelligent agents: Each individual agent in the system is typically autonomous, meaning it can make its own decisions. It perceives its environment (via sensors or inputs), processes that information,

and then acts on the environment (via actuators or actions) to achieve its goals.

3) Interaction: This is the defining characteristic of a multi-agent system. Agents don't simply operate in parallel; they interact with each other and with the environment. These interactions can include:

- *Collaboration:* Agents work together to achieve a common goal (e.g., a swarm of drones cooperatively mapping an area).
- *Competition:* Agents have conflicting goals and may attempt to outperform or hinder each other (e.g., AI opponents in a game).
- *Coordination:* Agents adjust their behavior to avoid collisions, share resources, or synchronize actions, even when they have independent goals.

4) Autonomy: Individual agents typically have some degree of independence in decision making and actions, even within a cooperative structure.

5) Decentralization: Responsibility for control and decision making is often distributed among agents, rather than being managed by a single central authority. This contrasts with centralized systems, where one entity controls all others.

6) Emergent Behavior: Complex global behavior patterns can emerge from the local interactions of individual agents, even if these global behavior patterns were not explicitly programmed into any individual agent.

A swarm of UAVs is a prime example of a multi-agent system. Each UAV acts as an agent, and they interact (for example, to form a specific formation or intercept a target) without relying on a single operator to control each one individually. Their collective actions result in the behavior of a coordinated defense system.

Reinforcement learning (RL) is a paradigm in which an agent learns to achieve a goal by performing actions and receiving feedback (rewards or penalties) from the environment, striving to maximize cumulative reward over time. Deep reinforcement learning (DRL) has played a significant role in enabling drones to autonomously perform certain tasks, including navigation, obstacle avoidance, and mission planning, with minimal human intervention.

When multiple agents are involved, the problem becomes significantly more complex, leading to problems addressed by Multi-Agent Reinforcement Learning (MARL). Multi-Agent Reinforcement Learning extends the RL framework to scenarios with multiple interacting agents, where complexities such as non-stationarity (due to the changing strategies of other agents) and scaling issues become relevant. In MARL, the environment is typically modeled as a Markov (or stochastic) game. Multi-Agent Reinforcement Learning algorithms generally fall into three main types:

- Value-based methods: Focus on learning and updating optimal value functions (e.g., Q-functions) for state-action pairs.
- Policy-based methods: Directly optimize policy functions that map states to actions.
- Actor-critic methods: Combine elements of both approaches, using an actor to update the policy and a critic to evaluate the policy actions by evaluating value functions.

Value-based MARL methods primarily focus on learning optimal state and action value functions (Q-functions) to infer optimal strategies that are particularly efficient in discrete action spaces.

III. Swarm Obstacle Avoidance

Drone swarms encounter a variety of obstacles, which can be divided into several categories.

- *Static large-scale objects:* buildings, cliffs, bridges, towers, power lines, trees.
- *Static small-scale objects:* poles, signs, fences, antennas, guy wires.
- *Dynamic large-scale objects:* vehicles, boats, other aircraft, crowds of people.
- *Dynamic small-scale objects:* pedestrians, animals, migratory birds, drones.
- *Complex terrain:* urban canyons, forests, mountain ridges, ravines.
- *Thin / inconspicuous obstacles:* wires, guy wires, tree branches, cables.
- *Partially obscured/transparent objects:* glass facades, mesh / grille, gaps in vegetation.

Environmental factors such as severe weather conditions and non-physical constraints such as no-fly zones (NFZ) or communication jamming also act as significant navigational, physical, operational and electronic obstacles, requiring decentralized collision avoidance.

Merey et al. [1] provided a comprehensive review of UAV obstacle detection and avoidance methods, with a particular focus on the consideration of obstacle-free zones (OFZs), a gap in previous reviews. Their findings can be summarized as follows:

1) Comprehensive classification of obstacle detection and avoidance methods: We classify and review state-of-the-art approaches to obstacle detection and avoidance, including those addressing static and dynamic obstacles, as well as OFZ avoidance strategies.

2) Obstacle detection and avoidance process: We present an overview of the components of a UAV obstacle detection and avoidance mission, as

shown in Fig. 1, emphasizing the integration of trajectory planning, environment mapping, and obstacle detection.

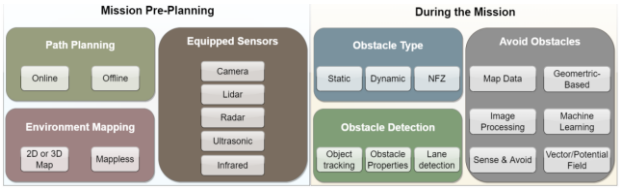

Fig. 1. Overview of ODA mission components [1]

3) Comparison of obstacle detection and avoidance methods: We compare different obstacle detection and avoidance methods based on performance metrics such as efficiency, power consumption, and adaptability to dynamic conditions.

4) Analysis of OFZ processing: We analyze how OFZs are modeled and incorporated into planning algorithms, highlighting effective approaches and identifying research gaps. • Evolution of ODA Methods: We trace the development of ODA methodologies over the past decade, describing trends and emerging issues in UAV navigation.

In article [2], a Markov decision process (MDP) is used to represent the decision-making process of UAVs, embedding MCPPs as high-cost areas from which UAVs are diverted. MCPPs lead to large penalty terms in the cost function, inducing UAVs to choose alternative routes that maintain formation while avoiding MCPPs. UAVs predict future state vectors to estimate the position ahead and choose paths with minimal penalties.

In article [3], we find that robust trajectory and resource allocation schemes for UAV-enabled wireless networks are developed to improve spectral efficiency while addressing real-world challenges such as UAV positioning errors and signal dead zones. The deep learning-based framework efficiently models UAV trajectories and resource allocation, ensuring reliable signal dead zone avoidance.

A multi-agent reinforcement learning method that trains UAVs to consider obstacle-free zones (OFZs) as areas to avoid during trajectory optimization is described in [4]. Obstacle-free zones are modeled in the simulation as obstacles or areas with negative rewards, so UAVs learn to avoid these areas when planning trajectories. Through continuous interaction with the environment, UAVs dynamically adjust their routes to find efficient trajectories that avoid OFZs.

IV. Literature Review

Ekechi, et.al. [5] provide a focused review of multi-agent reinforcement learning (MARL) applied to UAV control, synthesizing algorithms, architectures, benchmarks, and open problems related to UAV use cases (search, coverage, tracking, resource allocation). The paper categorizes the methods into three broad groups – value-based (QMIX, MADQN), policy-based (MADDPG, MAPPO, TRPO/PPO variants), and federated MARL – and explains their suitability for UAV tasks (continuous control, partial observability, privacy/communication constraints). Key explanatory figures illustrate UAV control modes and reinforcement learning flows for single and multiple agents; Tables systematically compare the algorithms based on scalability, fault tolerance, convergence, and typical UAV application areas.

This review highlights practical solutions in the field of MARL UAV design: Dec-POMDP/POMDP modeling, CTDE as a common learning paradigm, continuous action space algorithms (MADDPG, MAPPO) for flight control, and reward generation / role decomposition for handling credit allocation. The advantages of value factorization (QMIX/VDN) for discrete cooperative problems and the strengths of policy gradient methods in continuous dynamic environments are discussed. The authors devote significant attention to federated MARL, arguing that they are suitable for distributed UAV systems with privacy and capacity constraints, and summarize recent work on federated trajectories, resource allocation, and MEC.

Limitations include reliance on results obtained from diverse, non-standardized benchmarks (many comparisons are qualitative or environment-specific) and less emphasis on incorporating theoretical guarantees (convergence, sampling efficiency) or standardized evaluation protocols for continuous multi-UAV control.

Ku et al. [6] on the other hand, propose PMI-MADDPG, an extension of the CTDE (centralized learning, decentralized execution) MADDPG framework adapted for collaborative UAV swarm trajectory planning under partial observability.

The experiments use an improved particle multi-agent environment with 2–3 UAVs and continuous control; PMI-MADDPG is compared to conventional MADDPG and MAPPO. The results show faster convergence and improved task rewards for PMI-MADDPG (the paper reports an average reward improvement of approximately 7.5% compared to MADDPG), while the critic/actor loss curves indicate more stable learning for small swarms. The main advantage of this work lies in a new approach to evaluating dependencies based on mutual information in reward generation, which provides a

principled way to encourage beneficial cooperation without explicit communication protocols.

Limitations include evaluation only on small swarm sizes and limited discussion of scalability, sampling efficiency, or theoretical guarantees of convergence. Overall, the paper represents a practical, well-founded contribution to the development of MARL reward systems for continuous multi-robot control and provides a useful reference for approaches that use learned dependency measures to encourage cooperation.

Dey and Xu [7] consider the control of very large swarms of UAVs from a different perspective, reformulating the system as a leader-follower multi-group system (LS-MAS) and introducing a "mixed game" hierarchy that combines cooperative play between group leaders, leader-follower Stackelberg communication, and mean-field games (MFGs) between multiple followers.

The authors validate the approach in large simulations: four groups, each with one leader and 500 followers, tracking a time-varying anchor point.

Advantages: principled combination of cooperative, Stackelberg, and mean-field games for scaling to very large swarms; specific hierarchical reinforcement learning architecture (actor-critic-mass) with stability guarantees; realistic emphasis on the cost of communication and coordination at the group level. Limitations: cumbersome mathematical formalism and numerous hyperparameters; dependence on an accurate probability density function / mass estimate and constant excitation conditions for convergence; experiments are only simulations and are tuned to the presented scenario (four groups, specific dynamics), so generalization, sampling efficiency, and robustness to real-world conditions (communication loss, adversarial agents, turbulence) remain open. Utility: A valuable benchmark for scalable multi-layer reinforcement learning / control systems that combine game-theoretic decompositions with online actor-critic learning for continuous control of large-swarm UAVs.

V. Reinforcement Learning for Obstacle Avoidance by a Swarm of UAVs

Applying deep reinforcement learning to UAVs involves training an agent to navigate an environment by performing actions and receiving rewards. The agent interacts with its environment through a Markov decision process: from the current state s_t, it performs action a_t, moves to state s_{t+1}, and receives reward r_t. States are often represented as depth prediction images, and the reward magnitude determines how strongly the UAV updates its policy. For obstacle avoidance, [8] showed that deep Q-learning (as in [9]) works well for 2D robotic navigation. Other methods include the work of Han et al. [10], who proposed a model-based scheme that estimates the probability of collision in unknown environments and selects actions based on the confidence in the prediction, although its generality is limited as tests focused primarily on static obstacles. Long et al. [11] considered collisions with moving objects using a decentralized multi-robot navigation system, but did not conduct explicit evaluation for UAVs. The value of deep reinforcement learning, and DQN in particular for UAVs, was highlighted in [12], where the benefits of combining DQN with an actor-critic model using two networks to handle short-term and long-term decisions were noted. The D3QN algorithm in [8] applied DQN with two networks to obstacle avoidance, estimating $Q(s_t, a_t)$ to reflect the accumulated, immediate, and expected future rewards.

$$Q'(s_t, a_t) = Q(s_t, a_t) + \alpha\left(r_t + \gamma \cdot \max_a Q\left(s_{t+1}, a\right)\right). \quad (1)$$

Equation (1) gives the Q-network update, where α is the learning rate and γ is the discount factor [13].

The UAV's goal is to maximize its expected total reward by quickly moving through the environment and exploring as many state spaces as possible without collisions.

$$r_t = v \cdot \cos(\psi) \cdot \delta t + \left(\lambda \cdot \mathrm{BB}_{\mathrm{distance}}\right) - \left(\rho \cdot \mathrm{BB}_{\mathrm{penalty}}\right). \quad (2)$$

Equation (2) defines the reward for each non-terminal state. An episode ends when the UAV collides with an object (as specified in the interaction module) or when the agent reaches the 500-step limit.

The unmanned aerial vehicle is incentivized to move quickly by a reward proportional to the distance traveled $\left(v \cdot \cos(\psi)\right)$. It receives a positive reward for detecting a person outside the center of its field of view ($\mathrm{BB}_{\mathrm{distance}}$), which induces the UAV to deviate from them. Conversely, a penalty based on the aspect ratio of the bounding box ($\mathrm{BB}_{\mathrm{penalty}}$) increases as the person approaches, leading to an increase in the negative reward. A collision results in a reward of –10. This heuristic reward function combines object detection features with obstacle avoidance and depends on tunable hyperparameters

(δt, λ, ρ), which must be adjusted during training to obtain the appropriate result.

VI. Conclusions

This paper analyzes existing approaches to solving the problem of obstacle avoidance by a swarm of UAVs.

Reinforcement learning algorithms struggle in three-dimensional settings due to the significantly increased state space. Improved exploration strategies are needed to improve the performance of reinforcement learning in large state spaces in various three-dimensional modeling environments.

It is shown that deep machine reinforcement learning offers the greatest advantage.

Various exploration strategies accelerate training, but performance still needs improvement and greater stability when exploring more complex, dynamic environments.

References

[1] A. Merei, H. Mcheick, A. Ghaddar, and D. A. Rebaine, "Survey on Obstacle Detection and Avoidance Methods for UAVs," *Drones. 2025*, 9(3):203. https://doi.org/10.3390/drones9030203

[2] F. Trotti, A. Farinelli, and R. Muradore, "A Markov Decision Process Approach for Decentralized UAV Formation Path Planning," *In Proceedings of the 2024 European Control Conference (ECC)*, Stockholm, Sweden, 25–28 June 2024; IEEE: Piscataway, NJ, USA, 2024, pp. 436–441. https://doi.org/10.23919/ECC64448.2024.10591307

[3] W. Lee, and K. Lee, "Robust Trajectory and Resource Allocation for UAV Communications in Uncertain Environments With No-Fly Zone: A Deep Learning Approach," *IEEE Trans. Intell. Transp. Syst.*, 2024, 25, 14233–14244. https://doi.org/10.1109/TITS.2024.3399913

[4] Y. Gao, S. Wang, M. Liu, and Y. Hu, "Multi-agent reinforcement learning for UAVs 3D trajectory designing and mobile ground users scheduling with no-fly zones," *In Proceedings of the 2023 IEEE/CIC International Conference on Communications in China (ICCC)*, Dalian, China, 10–12 August 2023, IEEE: Piscataway, NJ, USA, 2023; pp. 1–6. https://doi.org/10.1109/ICCC57788.2023.10233375

[5] C. C. Ekechi, T. Elfouly, A. Alouani, & T. Khattab, "A Survey on UAV Control with Multi-Agent Reinforcement Learning," *Drones*, 9(7), 484, 2025. https://doi.org/10.3390/drones9070484

[6] Pingping Qu, Huan Liu, Song Xu et al., "Multi-Agent Deep Reinforcement Learning for Cooperative Path Planning of UAV Swarms," 15 October 2025, PREPRINT (Version 1) available at Research Square. https://doi.org/10.21203/rs.3.rs-6508231/v1

[7] S. Dey and H. Xu, "Intelligent Distributed Swarm Control for Large-Scale Multi-UAV Systems: A Hierarchical Learning Approach," *Electronics*, 12(1):89, 2023. https://doi.org/10.3390/electronics12010089

[8] L. Xie, S. Wang, A. Markham, and N. Trigoni, "Towards monocular vision based obstacle avoidance through deep reinforcement learning," arXiv preprint arXiv:1706.09829 (2017)

[9] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, and N. Freitas, "Dueling network architectures for deep reinforcement learning," *Machine Learning*, arXiv:1511.06581, pp. 1995–2003, 2016. https://doi.org/10.48550/arXiv.1511.06581

[10] G. Kahn, A. Villaflor, B. Ding, P. and S. Levine, "Self-supervised deep reinforcement learning with generalized computation graphs for robot navigation," *In 2018 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE 2018, pp. 1–8. https://doi.org/10.1109/ICRA.2018.8460655

[11] P. Long, T. Fanl, X. Liao, W. Liu, H. Zhang, and J. Pan, "Towards optimally decentralized multi-robot collision avoidance via deep reinforcement learning," *In 2018 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE 2018, pp. 6252–6259. https://doi.org/10.1109/ICRA.2018.8461113

[12] C. Wang, J. Wang, X. Zhang, and X. Zhang, "Autonomous navigation of UAV in large-scale unknown complex environment with deep reinforcement learning," *In 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, IEEE 2017, pp. 858–862. https://doi.org/10.1109/GlobalSIP.2017.8309082

[13] R. S. Sutton and A. G. Barto, Reinforcement learning: An introduction. MIT press (2018).

Received: February 06, 2026
Accepted: February 23, 2026
Published: March 06, 2026

Yudenko Igor. ORCID 0000-0002-7022-366X. Postgraduate Student.
Department of Avionics and Control Systems, Faculty of Air Navigation, Electronics and Telecommunications, State University "Kyiv Aviation Institute", Kyiv, Ukraine.
Education: National Aviation University, Kyiv, Ukraine, (2021).
Research interests: artificial neural networks, artificial intelligence, programming.
Publications: 4.
E-mail: ioyudenko@gmail.com

І. О. Юденко. Багатоагентне глибоке навчання з підкріпленням в задачі уникнення зіткнень

Уникнення перешкод має вирішальне значення для успішного виконання місій безпілотних літальних апаратів. Ця стаття присвячена дослідженню багатоагентного глибокого навчання з підкріпленням в задачі уникнення зіткнень. У роботі розглянуто, що рої безпілотних літальних апаратів стикаються з різноманітними перешкодами, які класифікуються на: статичні великомасштабні та дрібномасштабні перешкоди, динамічні великомасштабні та дрібномасштабні перешкоди, складний рельєф місцевості, тонкі/маловидні перешкоди, частково закриті/прозорі перешкоди. Для вирішення цієї проблеми пропонується багатоагентний алгоритм керування траєкторією на основі глибокого навчання з підкріпленням для управління траєкторією кожного безпілотного літального апарата незалежно. Були досліджені різні підходи та кілька середовищ 3D-моделювання за допомогою навчання з підкріпленням для рою безпілотних літальних апаратів.

Ключові слова: багатоагентне навчання з підкріпленням; рій безпілотних літальних апаратів; уникнення перешкод; нельотні зони; глибоке навчання з підкріпленням.

Юденко Ігор Олександрович. ORCID 0000-0002-7022-366X. Аспірант.
Кафедра авіоніки та систем управління, Факультет аеронавігації, електроніки та телекомунікацій, Державний університет «Київський авіаційний інститут», Київ, Україна.
Освіта: Національний авіаційний університет, Київ, Україна, (2021).
Напрям наукової діяльності: штучні нейронні мережі, штучний інтелект, програмування.
Публікації: 4.
E-mail: ioyudenko@gmail.com