UDC 004.896:519.71:623.746-519 **DOI:** 10.31673/2412-9070.2026.053804

M. VOITOVYCH, PhD Student;
ORCID: 0009-0003-0405-0304
O. BYCHKOV, Doctor of Technical Science, Professor,
ORCID: 0000-0002-9378-9535
Taras Shevchenko National University of Kyiv

HYBRID AUTOMATA-GUIDED MULTI-AGENT TD3 FOR UAV SWARM DECISION-MAKING IN ADVERSARIAL ENVIRONMENTS

UAV swarms must coordinate under partial observability, intermittent communication, and adversarial electronic-warfare (EW) interference. End-to-end multi-agent reinforcement learning (MARL) such as MADDPG can learn cooperative behaviors, but monolithic policies are brittle across mission phases and during rare safety-critical events such as sudden jamming. Building on a previously published hybrid framework that couples a supervisory hybrid automaton (HA) with three discrete modes (Search, Track, Safe) and guarded transitions with MARL under centralized training and decentralized execution (CTDE), this work replaces the MADDPG controller with multi-agent Twin Delayed Deep Deterministic Policy Gradient (MATD3). The HA enforces mission-level logic and safety while MATD3 optimizes continuous actions within each mode. In a Gym-PyBullet-Drones testbed with an EW jammer we compare HA+MATD3 against HA+MADDPG under identical HA logic. HA+MATD3 reaches the same final return about 30% sooner and more smoothly, attains success rates of 96% (no jam) and 93% (jam), searches faster (22.3 s vs 24.0 s), and tracks evasive targets more reliably (92.5% vs 89.7% in-view). It also scales more gracefully to larger swarms, retaining 87% mission success under jamming with ten UAVs versus 82% for the baseline controller. The results indicate that pairing formal mode supervision with a twin-critic MARL algorithm yields more robust, sample-efficient swarm control than the baseline.

Keywords: multi-agent reinforcement learning, UAV swarm, hybrid automata, TD3, MADDPG, electronic warfare, CTDE.

Introduction

UAV swarms are increasingly deployed for search, surveillance, and target-tracking missions in environments where communication and sensing are unreliable and an adversary may actively interfere. Deep multi-agent reinforcement learning (MARL) has become a leading approach for learning cooperative swarm behavior [10], [14], yet purely learned policies can behave unpredictably at the boundaries between mission phases and under rare, safety-critical disturbances. Combining learning with explicit high-level decision logic has emerged as a way to obtain both adaptivity and predictability. This paper studies, within such a hybrid framework, how the choice of the underlying continuous-control algorithm affects swarm performance under adversarial electronic-warfare (EW) conditions.

Statement of the problem

We consider a swarm of N cooperative UAVs tasked with searching for and then continuously tracking a mobile ground target in a bounded three-dimensional area, in the presence of an adversarial jammer that intermittently disrupts communication and corrupts sensing. The swarm must explore efficiently to locate the target, maintain coordinated line-of-sight tracking once it is found, and remain safe and regroup when jamming makes coordination unreliable. Each agent acts on local observations under partial observability and under non-stationarity induced both by the other learning agents and by the jammer. A monolithic end-to-end MARL policy (e.g., MADDPG [10]) must entangle exploration, tracking, and disturbance handling within a single network; such policies tend to learn unstably and to react unpredictably to abrupt events such as the onset of jamming, which is unaccep-

© Voitovych M., Bychkov O., 2026

table in safety-critical missions [2]. The problem addressed here is how to obtain robust, sample-efficient, and predictable swarm control under these conditions, and specifically how much the choice of continuous-control algorithm contributes to it.

Analysis of recent studies and publications

Several lines of work inform this study. Actor–critic MARL with centralized training and decentralized execution (CTDE) — notably MADDPG [10] — lets agents learn cooperative policies while acting on local information, and surveys confirm its wide use for multi-UAV control [2], [14]. Decomposing a mission into phases or modes improves learning efficiency and robustness over a single unified policy: two-stage search-then-track designs [9], bio-inspired swarm decision-making schemes [1], and integrated search-and-tracking frameworks [4] all report faster convergence and better tracking, and graph- or attention-based MARL has been applied to collaborative UAV search and tracking [6]. To make mode switching explicit and verifiable, hybrid automata provide a formal model of systems that combine discrete modes with guarded transitions and continuous within-mode dynamics [7]; related logic-based formalisms such as Petri-net models have likewise been used to describe UAV-swarm behavior [13]. On the algorithmic side, value-decomposition methods such as QMIX [11] target the non-stationarity of cooperative MARL, while TD3 reduces the overestimation bias of deterministic actor–critic learning through twin critics, delayed policy updates, and target-policy smoothing [5], [12], improving stability over DDPG and MADDPG in continuous control. In our previous work [3] we combined a supervisory hybrid automaton (Search/Track/Safe) with an MADDPG controller and showed that it markedly outperforms a monolithic MADDPG baseline under jamming and at larger swarm sizes. That study fixed the high-level logic but used MADDPG throughout, leaving open whether a stronger continuous-control algorithm would further improve the hybrid.

Research objectives and tasks

The aim of this work is to isolate and quantify the effect of the continuous-control algorithm within the established hybrid-automaton framework [3]. We retain the three-mode automaton and the existing CTDE training scheme and replace the MADDPG controller with multi-agent TD3 (HA+MATD3), comparing it directly against HA+MADDPG under identical high-level logic; the monolithic baseline, already shown to be substantially inferior [3], is omitted so that the comparison cleanly isolates the algorithmic question (TD3 vs DDPG). The specific tasks are: to design a mode-conditioned MATD3 controller with per-mode reward shaping and per-mode replay buffers, trained under CTDE; to evaluate it against HA+MADDPG on mission success with and without jamming, on search efficiency, and on tracking of evasive targets; and to assess scalability to a larger swarm (N = 10) and compare it with the baseline controller, so that the results are aligned with those reported in [3].

Research results

Problem formulation. We consider N cooperative quadrotor UAVs that search for and then track a mobile ground target in a 20×20×10 m arena simulated with PyBullet [8], in the presence of an adversarial jammer that intermittently degrades sensing and communication. The full formulation — observation model, EW interference model, and the multi-agent Markov decision process — follows [3]; only what is needed below is summarized here. At each step t every agent i receives a local observation that includes its own state, the relative states of neighbors within range, the target bearing and range if detected, and a binary jamming flag, and selects a continuous action. The task is a multi-agent MDP:

$$M = \langle S,\ O_i,\ A_i,\ P,\ R,\ \gamma \rangle\ , \tag{1}$$

where S is the state space, O_i and A_i are the observation and action spaces of agent i, P is the transition kernel, R is the shared reward, and γ is the discount factor. Each agent learns a mode-conditioned policy that maximizes the discounted return:

$$J = \mathbb{E}\left[\sum_t \gamma^t r_t\right]\ , \tag{2}$$

where the discrete mode q at each step is supplied by the hybrid automaton.

Hybrid automaton. The supervisory controller is the same three-mode HA introduced in [3]: Search (coordinated exploration to locate the target), Track (cooperative following once the target is found), and Safe (fail-safe hovering in formation under severe jamming). Guarded transitions are triggered by the events TargetFound, TargetLost, JammingDetected, and JammingCleared (fig. 1); on JammingCleared the automaton returns to the mode active before Safe. The current mode is encoded as a one-hot vector and appended to each agent's policy input, so behavior is conditioned on the mode while the automaton retains authority over the switches. The full formal definition — invariants, resets, and output map — is given in [3].

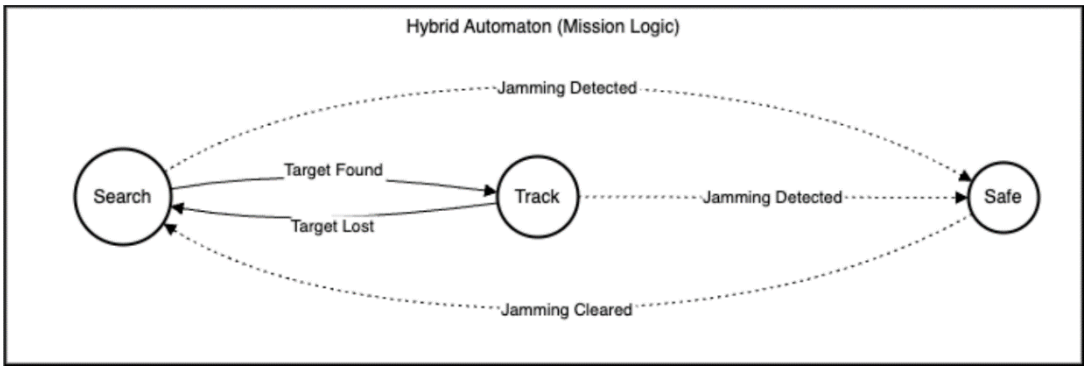


Fig. 1. Hybrid automaton state-transition diagram

MATD3 with CTDE. For continuous control within each mode we use a multi-agent TD3 controller trained under CTDE. Each UAV i has a deterministic actor that maps its observation and the current mode to an action:

$$u_i = \mu_\theta(o_i,\ q) \quad , \tag{3}$$

Actors are shared across the homogeneous agents and run on local observations at execution time. During training, each agent has twin centralized critics that take the global state x, the joint action $u = (u_1, \ldots, u_N)$, and the mode q. Transitions are stored in per-mode replay buffers (b_Search, b_Track, b_Safe), which keeps each critic's updates mode-homogeneous and reduces cross-mode interference; mini-batches are drawn round-robin across modes. Following TD3 [5], the critic target uses the target networks, the minimum of the twin critics, and target-policy smoothing noise added to the next action:

$$y_i = r + \gamma \min_{k=1,2} Q^{k}_{\phi'_i}(x',\ u'_1,\ \ldots,\ u'_N,\ q') \quad , \tag{4}$$

where the smoothing noise ε is clipped Gaussian noise. Each critic minimizes the squared temporal-difference error,

$$L(\phi_i) = \mathbb{E}\left[\left(Q^{k}_{\phi_i}(x,u,q) - y_i\right)^2\right] \quad , \tag{5}$$

for k = 1, 2. The actor is updated less frequently than the critics (policy delay d = 2) by the deterministic policy gradient through the first critic:

$$\nabla_\theta J \approx \mathbb{E}\left[\nabla_\theta\, \mu_\theta(o_i,q)\ \nabla_{a_i} Q^{1}_{\phi_i}(x,\ a_1,\ \ldots,\ a_i = \mu_\theta(o_i,q),\ \ldots,\ a_N)\right] \tag{6}$$

The delayed actor update, twin critics, and target smoothing are the TD3 mechanisms that stabilize learning relative to single-critic MADDPG, particularly around mode boundaries and under jamming-induced observation noise. Reward shaping is applied per mode, as in [3]: Search rewards rapid, well-spread exploration and a one-time detection bonus; Track rewards an in-view, optimal-standoff formation and penalizes collisions or loss of line-of-sight; Safe penalizes drift from a tight holding pattern and rewards prompt resumption once jamming clears. The resulting data flow — HA mode supervision, decentralized actors, and centralized twin critics with per-mode replay — is shown in fig. 2.

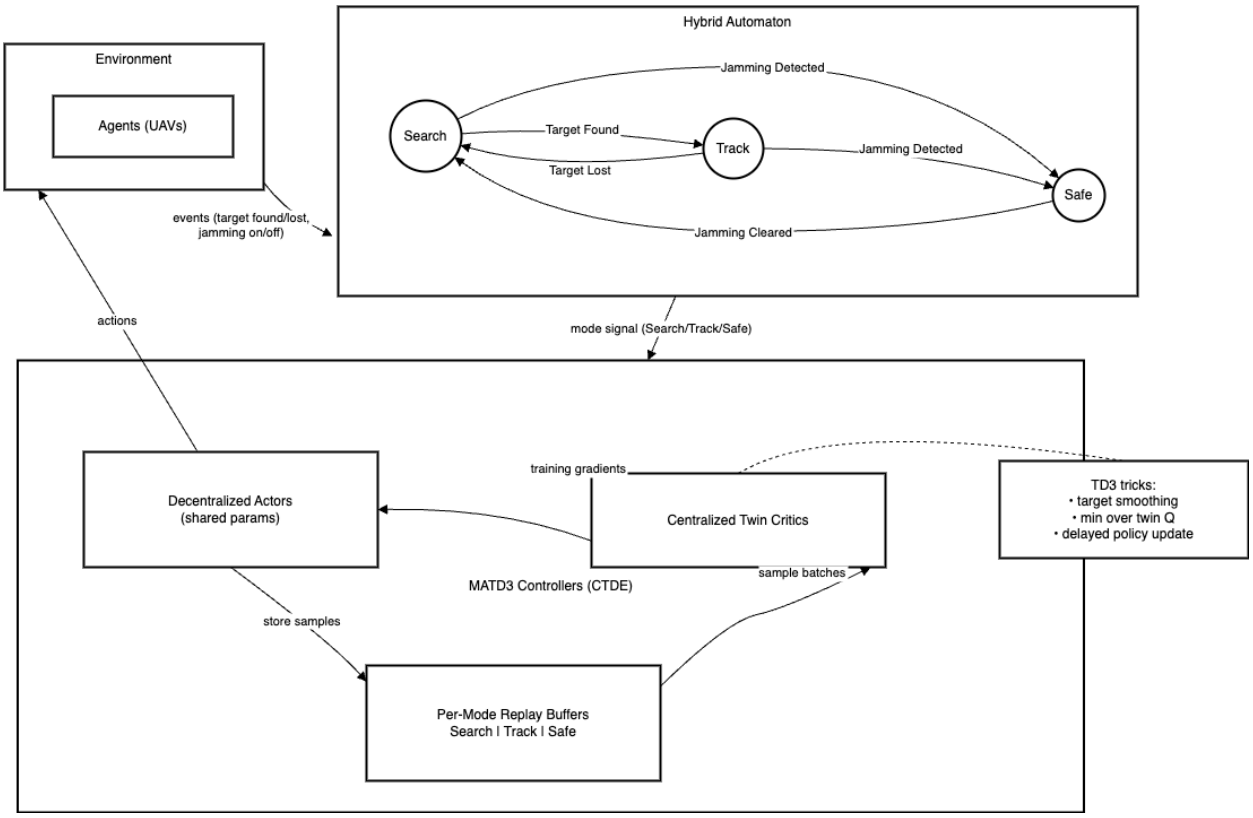


Fig. 2. High-level HA + MATD3 architecture under CTDE

Simulation and training. The environment is built on Gym-PyBullet-Drones [8] with realistic quadrotor dynamics and a PID inner loop; the default team is $N = 4$, and actions are 3D velocity commands at 10 Hz. For the scalability study (below) we additionally train and evaluate with N = 10 UAVs. The jammer activates at a random mid-episode time for about 30 s, during which inter-drone communication is largely lost and target and teammate observations are corrupted — a near-worst-case EW event that the Safe mode is designed to absorb. A run succeeds if the swarm finds the target within 120 s and then maintains continuous line-of-sight for at least 60 s. We compare two controllers that share identical HA logic, dynamics, and sensing and differ only in the RL algorithm: HA+MATD3 (ours) and HA+MADDPG (the configuration of [3], i.e. a single centralized critic per agent without TD3's twin critics, delayed updates, or target smoothing). Actors are MLPs with two hidden layers of 128 units (ReLU); critics use two hidden layers of 256 units. We use Adam with learning rate 10^{-3}, $\gamma = 0.95$, replay capacity 10^6, and batch size 256; for MATD3 the policy delay is 2 and the target-smoothing noise is $\sigma = 0.2$ clipped at ± 0.5. Each controller is trained for up to 5000 episodes over 5 random seeds and evaluated over 100 episodes per condition.

Convergence. HA+MATD3 converges markedly faster than HA+MADDPG (fig. 3), reaching its plateau (average return $\approx$ 5000) by about 2000–2200 episodes versus roughly 3000 for HA+MADDPG — about 30% fewer episodes — and with a visibly smoother curve. Both controllers reach a similar final return, so the advantage of TD3 here is sample efficiency and training stability rather than a higher asymptote: the twin critics and delayed policy updates damp the overestimation and oscillation that make single-critic MADDPG noisier under multi-agent non-stationarity. The automaton's curriculum and per-mode replay benefit both controllers equally, which is why their asymptotic returns coincide.

Mission success. Both controllers perform well without jamming, with a small edge to HA+MATD3 (table 1). Under jamming the gap is preserved: HA+MATD3 retains 93% success and HA+MADDPG 90%, each only 3–4 points below its no-jam value, because both share the Safe-mode logic that holds formation during interference and resumes once JammingCleared fires. The modest edge of HA+MATD3 reflects TD3's reduced overestimation, which yields a slightly more reliable final policy at mode boundaries. For reference, the monolithic MADDPG baseline of [3] fell to 65% under jamming, underscoring the value of the HA structure both controllers share.

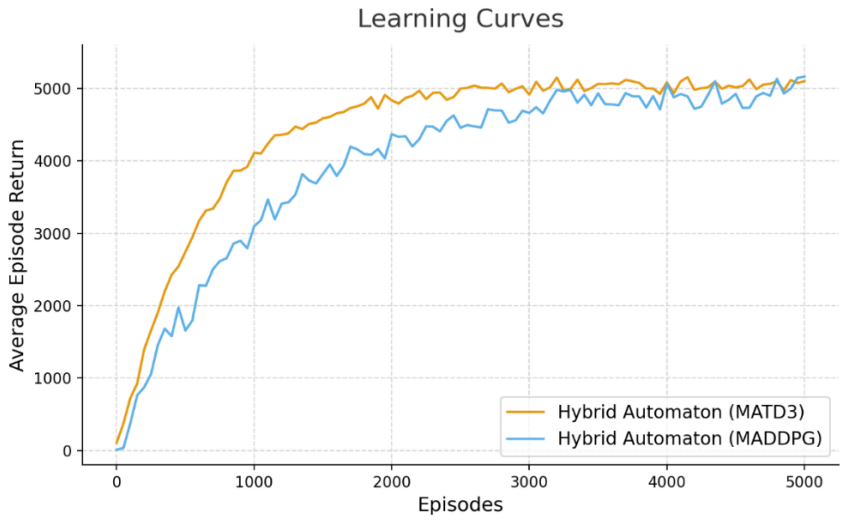


Fig. 3. Learning curves (average episode return vs. training episodes)

Table 1

Mission success rates (100 evaluation runs per condition)

Method	Success – No Jam	Success – With Jam
HA + MATD3	**96%**	**93%**
HA + MADDPG	94%	90%

Search efficiency and tracking continuity. With the jammer disabled and a stationary target, HA+MATD3 finds the target in 22.3 s on average versus 24.0 s for HA+MADDPG, with comparably low variance (table 2); both maintain the full 60 s of continuous tracking in every trial. The faster acquisition reflects more decisive search-to-track transitions under the twin-critic policy.

Table 2

Search and tracking performance (no jamming, stationary target)

Method	Avg. Search Duration (s)	Avg. Continuous Track (s)
HA + MATD3	**22.3 ± 2.2**	**60.0 ± 0.0**
HA + MADDPG	24.0 ± 2.1	60.0 ± 0.0

Dynamic (evasive) target. Against an evasive target, HA+MATD3 keeps the target in view 92.5% ± 4.8% of the time versus 89.7% ± 5.5% for HA+MADDPG (fig. 4).

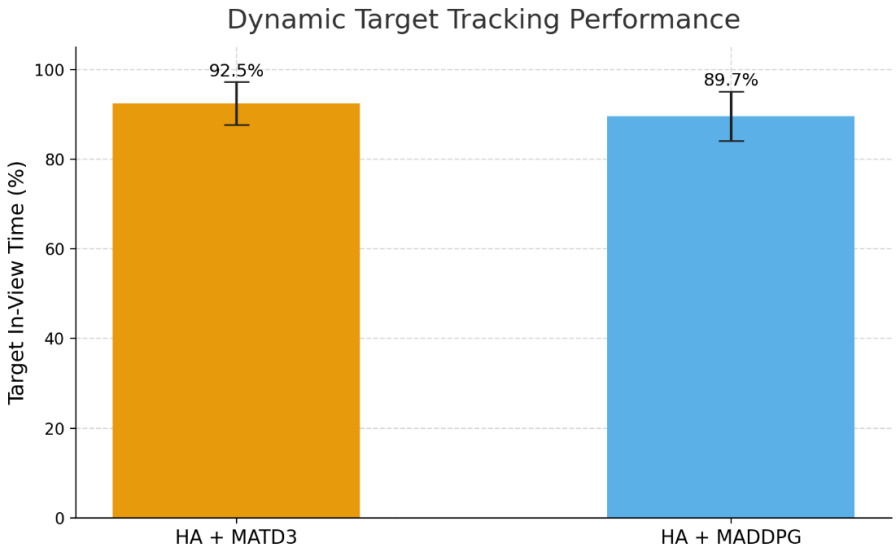


Fig. 4. Evasive-target in-view time (%) by approach

Both hybrids learn encirclement and hand-off tactics and keep the target within sensor range almost continuously, never suffering the prolonged losses characteristic of a monolithic policy; the MATD3 controller holds a marginally tighter and steadier standoff distance, staying closer to the target throughout the episode (fig. 5).

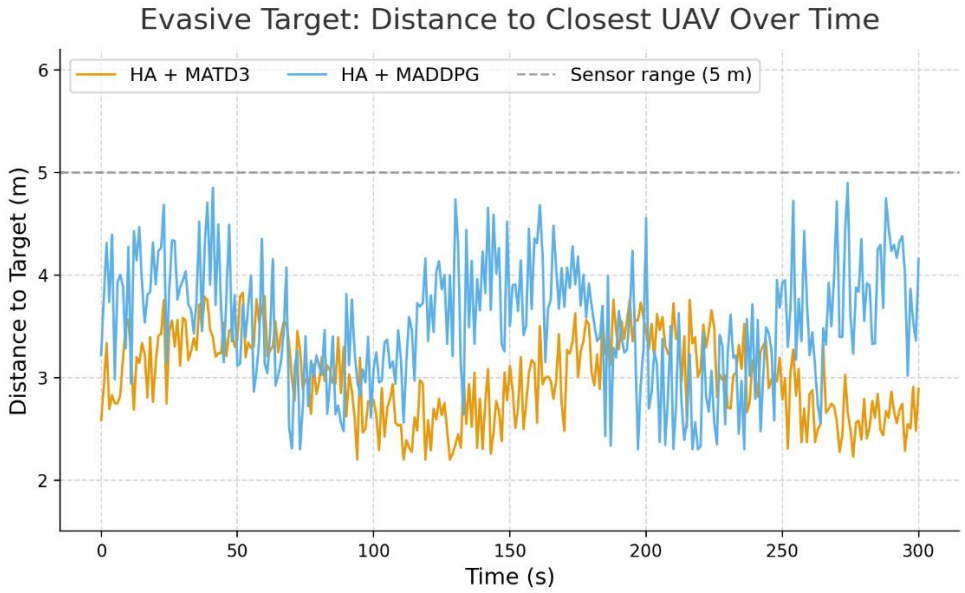


Fig. 5. Distance to the closest UAV over time (evasive target)

Scalability. To check that the algorithmic comparison holds at scale, and to align with [3], we repeated training and evaluation with N = 10 UAVs — with proportionally larger state and action spaces and a larger centralized critic — while keeping the automaton logic unchanged. Both hybrids remained trainable, but TD3's stability advantage grew with swarm size: under jamming, HA+MATD3 retained 87% mission success at N = 10 (down from 93% at N = 4), whereas HA+MADDPG dropped to 82% (from 90%), as shown in table 3. The larger joint-action space amplifies overestimation and credit-assignment difficulty, which the twin critics and delayed updates of TD3 mitigate more effectively than a single critic. Both remain far above the monolithic baseline, which degrades to 38% at N = 10 in [3].

Table 3

Mission success under jamming for different swarm sizes (100 runs per condition)

Method	Success – 4 UAVs	Success – 10 UAVs
HA + MATD3	**93%**	**87%**
HA + MADDPG	90%	82%

Conclusions and prospects for further research

Within an established hybrid-automaton framework for UAV swarm control [3], replacing the MADDPG controller with multi-agent TD3 yields consistent gains while leaving the interpretable, verifiable mode logic untouched. The principal benefit is in training: HA+MATD3 converges to the same final return as HA+MADDPG in roughly 30% fewer episodes and with a smoother curve. The reduced overestimation of TD3 also gives modest but consistent improvements in mission success under jamming (93% vs 90%), search time (22.3 s vs 24.0 s), and tracking of evasive targets (92.5% vs 89.7% in-view), and it scales more gracefully: at ten UAVs HA+MATD3 retains 87% success under jamming versus 82% for HA+MADDPG. Because the high-level logic is identical across both controllers, these improvements are attributable specifically to the twin-critic, delayed-update mechanisms of TD3, which matter most at mode boundaries, under jamming-induced noise, and as the joint-action space grows. The approach inherits the limitations discussed in [3] — reliance on correct guard design and reliable event detection. Future work will explore learned or adaptive guard thresholds,

differentiable mode-switching logic so the automaton structure itself can be optimized, and hardware-in-the-loop validation on physical UAVs.

Authors' contributions

Mykhailo VOITOVYCH – conceptualization of the study, literature review, design and implementation of the HA+MATD3 controller, simulation set-up and evaluation, analysis of results, drafting and final approval of the manuscript; Oleksii BYCHKOV – general supervision, methodology verification, review and editing of the manuscript, and final approval of the manuscript.

Declaration on artificial intelligence

During the preparation of this article, the authors used Anthropic Claude models for content review, formatting, rephrasing, and generating simple code fragments. All AI-generated outputs were carefully reviewed and edited by the authors, who take full responsibility for the final content of the article.

Conflict of interest

The authors declare that there is no conflict of interest and confirm that during the preparation of this work there were no commercial, financial, or other relationships that could be construed as influencing the results of the study or their interpretation. The work was performed in accordance with the principles of academic integrity, ethical standards for conducting scientific research, and editorial policy requirements for preventing conflicts of interest.

References

1. A bio-inspired decision-making method of UAV swarm for attack-defense confrontation via multi-agent reinforcement learning / P. Chi, J. Wei, K. Wu, B. Di, Y. Wang // Biomimetics. – 2023. – Vol. 8, № 2. – Art. 222.

2. A survey on UAV control with multi-agent reinforcement learning / C. C. Ekechi, T. Elfouly, A. Alouani, T. Khattab // Drones. – 2025. – Vol. 9, № 7. – Art. 484.

3. Bychkov O. Robust UAV swarm decision-making under uncertainty via combination of multi-agent reinforcement learning and hybrid automata / O. Bychkov, M. Voitovych // Advanced Information Technology. – 2025. – № 2(5). – DOI: 10.17721/AIT.2025.2.05

4. Deep reinforcement learning for UAV target search and continuous tracking with Gaussian process regression and prior policy embedding / Z. Feng, X. Na, S. Hai, Q. Sun, J. Shi // Electronics. – 2025. – Vol. 14, № 7. – Art. 1330.

5. Fujimoto S. Addressing function approximation error in actor-critic methods / S. Fujimoto, H. van Hoof, D. Meger // Proceedings of the 35th International Conference on Machine Learning (ICML). – 2018. – P. 1587–1596.

6. Graph-based multi-agent reinforcement learning for collaborative search and tracking of multiple UAVs / Y. Cai, F. Jiang, J. Wang, X. Shen // Neurocomputing. – 2022. – Vol. 500. – P. 751–764.

7. Hybrid automata: an algorithmic approach to the specification and verification of hybrid systems / R. Alur, C. Courcoubetis, T. A. Henzinger, P. H. Ho // Theoretical Computer Science. – 1995. – Vol. 138, № 1. – P. 3–34.

8. Learning to fly – a Gym environment with PyBullet physics for reinforcement learning of multi-agent quadcopter control / J. Panerati, H. Zheng, S. Zhou, J. Xu, A. Prorok, A. P. Schoellig // IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). – 2021. – P. 7512–7519.

9. Liu M. A two-stage target search and tracking method for UAV based on deep reinforcement learning / M. Liu, J. Wei, K. Liu // Drones. – 2024. – Vol. 8, № 10. – Art. 544.

10. Multi-agent actor-critic for mixed cooperative-competitive environments / R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, I. Mordatch // Advances in Neural Information Processing Systems (NeurIPS). – 2017. – Vol. 30. – P. 6379–6390.

11. QMIX: monotonic value function factorisation for deep multi-agent reinforcement learning / T. Rashid, M. Samvelyan, C. Schroeder de Witt, G. Farquhar, J. Foerster, S. Whiteson // Proceedings of the 35th International Conference on Machine Learning (ICML). – 2018. – P. 4295–4304.

12. Twin critics and target smoothing for robust multi-robot continuous control / H. Zhang [et al.] // IROS Workshops. – 2022.

13. UAV cluster behavior modeling based on spatial-temporal hybrid Petri net / X. Wang, Y. Guo, N. Lu, P. He // Applied Sciences. – 2023. – Vol. 13, № 2. – Art. 762.

14. Zhang K. Multi-agent reinforcement learning: a selective overview of theories and algorithms / K. Zhang, Z. Yang, T. Başar // Handbook of Reinforcement Learning and Control. – Cham : Springer, 2021. – P. 321–384.

М. В. Войтович, О. С. Бичков

ГІБРИДНІ АВТОМАТИ ТА БАГАТОАГЕНТНИЙ TD3 ДЛЯ ПРИЙНЯТТЯ РІШЕНЬ РОЄМ БПЛА В УМОВАХ ПРОТИДІЇ

Рої безпілотних літальних апаратів (БПЛА) повинні узгоджувати дії в умовах часткової спостережуваності, нестійкого зв'язку та цілеспрямованої радіоелектронної боротьби (РЕБ). Багатоагентні методи навчання з підкріпленням, зокрема MADDPG, здатні формувати ефективну кооперативну поведінку, проте монолітні політики виявляються нестабільними на межах фаз місії та під час рідкісних критичних подій, таких як раптове глушіння. У попередній роботі авторів запропоновано гібридну архітектуру, що поєднує наглядовий гібридний автомат із трьома дискретними режимами (Пошук, Супровід, Безпека) та керованими переходами з багатоагентним навчанням за схемою централізованого навчання і децентралізованого виконання (CTDE), і показано її суттєву перевагу над монолітним базовим рішенням. Дана стаття розвиває цей підхід: за незмінної логіки автомата контролер MADDPG замінюється на багатоагентний алгоритм TD3 (MATD3) з подвійними критиками, згладжуванням цільової політики та відкладеним оновленням актора, що зменшує зміщення переоцінки та підвищує стійкість навчання. Для кожного режиму застосовано окреме формування винагороди та окремі буфери відтворення досвіду, що послаблює взаємну інтерференцію між режимами. Експерименти проведено у середовищі Gym-PyBullet-Drones із моделлю РЕБ-глушіння, де порівнюються два підходи з ідентичною логікою автомата — HA+MATD3 та HA+MADDPG. Запропонований HA+MATD3 досягає такого ж підсумкового виграшу приблизно на 30% швидше та з більш плавною кривою навчання, забезпечує вищі показники успішності місії (96% без глушіння та 93% за наявності глушіння проти 94% і 90%), швидше виявляє ціль (22,3 с проти 24,0 с) та надійніше супроводжує маневрену ціль (частка часу спостереження 92,5% проти 89,7%). Крім того, запропонований підхід краще масштабується: для рою з десяти БПЛА HA+MATD3 зберігає 87% успішних місій за умов глушіння проти 82% для HA+MADDPG. Результати свідчать, що поєднання формального наглядового керування режимами із сучасним алгоритмом TD3 забезпечує більш відмовостійке та ефективне керування роєм за невизначеності, зберігаючи інтерпретованість і гарантії безпеки гібридного автомата.

Ключові слова: багатоагентне навчання з підкріпленням, рій БПЛА, гібридні автомати, TD3, MADDPG, радіоелектронна боротьба, CTDE.

Надійшла до редакції: 31.08.2026
Прийнята до друку: 21.09.2026
Опубліковано: 30.10.2026
© 2026 Voitovych M., Bychkov O.
Цей матеріал ліцензовано за умовами CC BY 4.0.https://creativecommons.org/licenses/by/4.0/