КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 4 (32), 2026
ISSN 2663 – 4023
DOI 10.28925/2663-4023.2026.32.1099
УДК 629.7:004.8:681.5

Трембовецький Роман Сергійович
аспірант
асистент кафедри інформаційної безпеки та комп'ютерної інженерії
Черкаський державний технологічний університет, м. Черкаси, Україна
ORCID: 0009-0001-2697-0879
roman.tremb@gmail.com

Розломій Інна Олександрівна
кандидат технічних наук, доцент
доцент кафедри інформаційної безпеки та комп'ютерної інженерії
Черкаський державний технологічний університет, м. Черкаси, Україна
ORCID: 0000-0001-5065-9004
inna-roz@ukr.net

ПОРІВНЯЛЬНИЙ АНАЛІЗ МЕТОДІВ POLICY-BASED ТА VALUE-BASED RL ДЛЯ ЛОКАЛЬНОГО ПЛАНУВАННЯ РУХУ БПЛА

Анотація. У роботі проведено порівняльне дослідження ефективності двох фундаментальних архітектур глибокого навчання з підкріпленням, а саме – Policy-based (на прикладі PPO) та Value-based (на прикладі SAC), для вирішення задачі локальної навігації БПЛА в умовах стохастичної невизначеності. Експериментальне моделювання, виконане у середовищі PyFlyt з урахуванням турбулентного вітру, змінної маси дрона та руху цілі, виявило критичні розбіжності у роботі алгоритмів. Встановлено, що алгоритм SAC, попри високу швидкість навчання та ефективність використання даних, формує вразливі політики з низьким показником успішності (63,2%) через помилки оцінки функції цінності в динамічних станах. Натомість метод PPO забезпечив формування стійкої стратегії керування з успішністю 97,1% та генерацію оптимальних траєкторій без осциляцій. Результати підтверджують, що для задач безперервного керування в непередбачуваних середовищах методи прямої оптимізації політики є більш ефективними, оскільки дозволяють уникнути ефекту «Reward Hacking» та краще генералізують фізичні закономірності польоту.

Ключові слова: автономна навігація; БПЛА; навчання з підкріпленням; PPO; SAC; динамічне середовище

ВСТУП

Стрімкий розвиток безпілотних літальних апаратів (БПЛА) відкриває нові перспективи для їх застосування у складних місіях, таких як пошуково-рятувальні операції, моніторинг інфраструктури та логістика в урбанізованому середовищі. Ключовим викликом для автономних систем залишається локальне планування руху в умовах невизначеності, коли дрон стикається з непередбачуваними аеродинамічними збуреннями (вітер), зміною власних інерційних характеристик (змінне корисне навантаження) та наявністю рухомих перешкод або цілей.

Постановка проблеми. Класичні методи керування, такі як ПІД-регулятори або MPC (Model Predictive Control), демонструють високу ефективність у детермінованих умовах, проте їхня залежність від точності математичної моделі робить їх вразливими до реальних стохастичних факторів. У цьому контексті глибинне навчання з підкріпленням (Deep Reinforcement Learning, DRL) пропонує альтернативну парадигму, дозволяючи агенту навчатися оптимальної стратегії керування безпосередньо через взаємодію із середовищем. Однак, вибір конкретної архітектури RL – між методами, що базуються на

оптимізації політики (Policy-based), та методами, що базуються на оцінці цінності (Value-based/Off-policy), – залишається відкритим питанням для задач із високою динамікою.

Аналіз останніх досліджень і публікацій. Застосування методів навчання з підкріпленням для автономної навігації БПЛА стало предметом численних досліджень останніми роками. У систематичному огляді [1] зазначається, що RL-підходи дозволяють долати обмеження класичних методів керування, таких як PID або MPC, особливо коли точна модель динаміки апарата є недоступною або занадто складною для аналітичного опису.

Значна частина сучасних робіт зосереджена на алгоритмах класу Off-policy (таких як Soft Actor-Critic), головною перевагою яких є висока ефективність використання даних (sample efficiency). Наприклад, в роботі [2] запропоновано покращену версію алгоритму SAC для планування маршрутів БПЛА, продемонструвавши його здатність генерувати плавні траєкторії та ефективно уникати перешкод у складних середовищах. Аналогічно, в роботі [3] у своєму огляді успішних застосувань DRL вказують на переваги методів, що базуються на максимізації ентропії (як SAC), для задач, де необхідна гнучка розвідка простору станів.

З іншого боку, алгоритми класу On-policy, зокрема Proximal Policy Optimization, залишаються стандартом де-факто для задач безперервного керування завдяки стабільності процесу навчання. У роботі [4] було показано, що PPO дозволяє агенту самостійно навчатися складним маневрам, таким як плавне гальмування перед ціллю, без явного програмування цих дій. Також в [5] успішно застосували PPO для стабілізації орієнтації БПЛА, підтвердивши його надійність у задачах низькорівневого керування.

Окремий напрям досліджень присвячено проблемі навігації в умовах аеродинамічних збурень. В роботі [6] дослідили використання RL для енергоефективного планування шляху в умовах турбулентного вітру, довівши, що нейромережеві контролери здатні адаптуватися до зовнішніх збурень краще за традиційні методи. Проте, більшість згаданих робіт розглядають або статичні перешкоди, або ізольовані динамічні фактори (тільки вітер).

Як результат аналізу джерел, можна виокремити раніше невирішену частину загальної проблеми: в літературі бракує комплексного порівняльного аналізу, який би зіставляв ефективність архітектур Policy-based (PPO) та Value-based/Off-policy (SAC) в умовах сукупної дії динамічних факторів – одночасної зміни маси апарата, непередбачуваних поривів вітру та наявності рухомих цілей. Залишається нез'ясованим, яка з парадигм забезпечує кращу робастність (стійкість) політики до невизначеності середовища, що і є предметом цього дослідження.

Мета статті. Метою статті є проведення порівняльного аналізу ефективності, стійкості та адаптивності методів навчання з підкріпленням двох різних парадигм – Policy-based (на прикладі PPO) та Value-based/Off-policy (на прикладі SAC) – для задачі локального планування руху БПЛА в умовах комплексної динамічної невизначеності.

МЕТОДИКА ДОСЛІДЖЕННЯ

Для досягнення поставленої мети та верифікації гіпотези про різну стійкість алгоритмів Policy-based та Value-based до динамічних збурень, було розроблено спеціалізоване симуляційне середовище. Методика експерименту базується на принципах відтворюваності та порівняння за інших рівних умов, де єдиною змінною є алгоритм навчання.

Середовище моделювання та динаміка БПЛА. Як програмну основу обрано бібліотеку PyFlyt, яка базується на фізичному рушії PyBullet [7]. Цей інструментарій забезпечує високу точність розрахунку аеродинаміки, гнучкість у налаштуванні завдань та повну сумісність із сучасними фреймворками глибокого навчання. Приклад візуалізації середовища показано на Рис. 1.

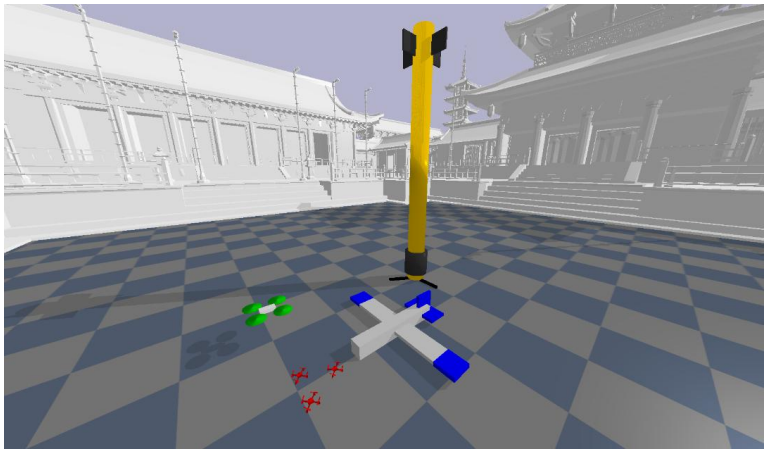


Рис. 1. Приклад візуалізації симуляційного середовища PyFlyt [7]

Об'єктом керування виступає квадрокоптер, динаміка якого описується рівняннями твердого тіла з шістьма ступенями вільності, де рух визначається сумарною тягою пропелерів та зовнішніми силами гравітації й опору. Ключовою особливістю методики є використання режиму керування Acro (Acrobatic mode). У цьому режимі агент безпосередньо контролює кутові швидкості навколо осей крену, тангажу та рискання, а також загальну тягу двигунів, замість використання стабілізованих команд кутів орієнтації. Такий підхід покладає повну відповідальність за стабілізацію апарата на нейромережу, що дозволяє оцінити здатність алгоритмів вивчати складні фізичні залежності без допомоги вбудованих контролерів.

Моделювання стохастичної динаміки. Для створення умов невизначеності, які слугують стрес-тестом для порівнюваних архітектур, реалізовано комплексну систему динамічних збурень.

По-перше, застосовується параметрична рандомізація на початку кожного епізоду навчання. Агент ініціалізується у випадковій точці тривимірного простору зі сферичного сектора, отримує випадкову орієнтацію (кути Ейлера) та початкову лінійну швидкість. Крім того, маса дрона варіюється в діапазоні від 0.01 до 0.04 кг додаткового навантаження, що змінює інерційні характеристики системи. Ціль, до якої має дістатися дрон, також не є статичною, а рухається за випадковою траєкторією, вимагаючи від агента постійної корекції курсу.

По-друге, впроваджено багатокомпонентну модель вітру, що є суперпозицією чотирьох типів атмосферних явищ:

$$V_w(r,t) = V_t + V_s + V_h + V_g(t), \quad (1)$$

де V_t – базова турбулентність, V_s – ефект зсуву вітру, V_h – термальні потоки, та V_g – пориви вітру. Базова турбулентність моделюється за допомогою моделі Драйдена [8], яка генерує стохастичні флуктуації. До неї додається ефект зсуву вітру, де швидкість потоку зростає з висотою за степеневим законом, та термальні потоки, що створюють локальні

висхідні сили. Додаткову непередбачуваність вносять пориви вітру, змодельовані як сума синусоїдальних хвиль з випадковими амплітудами та фазами.

Таке поєднання факторів створює складне нелінійне середовище для тестування стійкості алгоритмів.

Алгоритми та архітектура нейронних мереж. Для порівняльного аналізу обрано два алгоритми, що представляють різні парадигми навчання з підкріпленням і реалізовані за допомогою бібліотеки Stable-Baselines3 [9].

Першим алгоритмом є Proximal Policy Optimization (PPO) – представник класу On-policy методів. Його робота базується на обмеженні кроку оновлення політики за допомогою кліпінгу цільової функції, що забезпечує стабільність навчання та запобігає різкій деградації продуктивності. Другим алгоритмом обрано Soft Actor-Critic (SAC) – метод класу Off-policy, який поєднує архітектуру Actor-Critic із максимізацією ентропії. Використання буфера повторного відтворення дозволяє SAC досягати високої ефективності використання даних, що є характерною рисою Value-based підходів.

Щоб забезпечити коректність порівняння, обидва агенти використовують ідентичну архітектуру нейронної мережі MlpPolicy (Рис. 2).

Рис. 2. MlpPolicy архітектура нейронної мережі

Вона складається з двох повнозв'язних прихованих шарів по 128 нейронів у кожному. Функції активації обрано таким чином: ReLU для першого шару та Tanh для другого. Така конфігурація застосовується як до мережі політики (Actor), так і до мережі цінності (Critic), гарантуючи рівну обчислювальну складність моделей. Відмінність полягає лише у тривалості навчання: 10 мільйонів кроків для PPO проти 1 мільйона для SAC, що відображає різну швидкість збіжності алгоритмів.

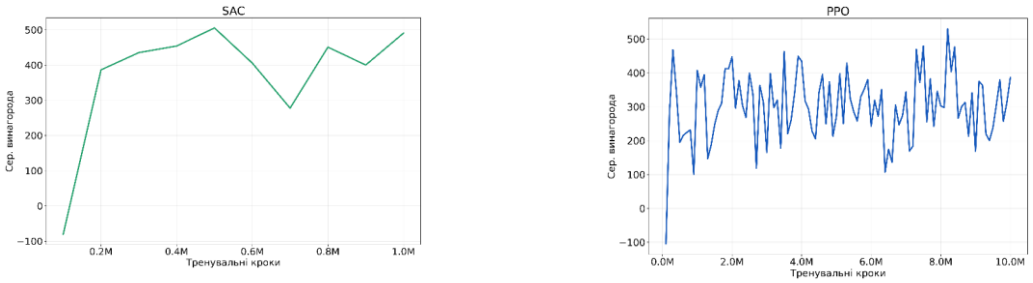
РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Експериментальна частина дослідження складалася з двох етапів: аналізу динаміки навчання агентів та валідації отриманих політик на тестовій вибірці з 1000 епізодів з рандомізованими параметрами середовища.

Аналіз ефективності навчання. На першому етапі порівнювалася здатність алгоритмів до навчання в умовах стохастичної динаміки.

Аналіз кривих навчання виявив фундаментальну відмінність між архітектурами. Агент SAC (Рис. 3-а) продемонстрував значно вищу ефективність використання даних. Для досягнення стабільного рівня винагороди йому знадобилося лише 1 млн кроків, що в 10 разів менше, ніж для PPO (10 млн кроків). Це пояснюється наявністю буфера відтворення, який дозволяє SAC багаторазово навчатися на минулому досвіді.

Натомість, навчання PPO (Рис. 3-б) характеризувалося високою волатильністю (осциляціями) функції винагороди. Однак, незважаючи на "шумний" процес навчання та повільнішу збіжність, PPO не застряг у локальних мінімумах, що часто трапляється з Value-based методами при помилковій оцінці Q-функції.



а б

Рис. 3. Криві навчання моделей

Кількісний аналіз ефективності. Для об'єктивної оцінки стійкості отриманих політик було проведено серію з 1000 випробувальних польотів у рандомізованих динамічних умовах. Отримані результати, зведені в Таблицю 1, демонструють критичний розрив у показниках виконання місії між двома підходами.

Таблиця 1

Порівняльні метрики ефективності агентів PPO та SAC у динамічному середовищі

Метрика	PPO	SAC
Успішність	97.1%	63.2%
Середня винагорода	294	376
Час польоту, фреймів	80	390
Середня довжина шляху	80	300

Агент PPO показав високу надійність, досягнувши показника успішності 97,1%. При цьому він виконував завдання максимально швидко, витрачаючи в середньому менш ніж 80 кадрів симуляції, та рухався за оптимальним маршрутом довжиною 80одиниць. Це свідчить про те, що агент сформував стратегію, орієнтовану на швидке та безпечне досягнення мети.

У випадку з SAC спостерігається деяка аномалія. Цей агент, маючи значно нижчий відсоток успіху (63,2%), отримав суттєво вищу середню винагороду (376) порівняно з PPO. Детальніший аналіз часових метрик пояснює цей парадокс: SAC утримував дрон у повітрі в середньому майже в 5 разів довше (390 кадрів), ніж PPO. Це вказує на явище "Reward Hacking" [10], коли Value-based метод оптимізував не досягнення термінальної цілі, а накопичення покрокових бонусів за стабілізацію польоту, що призвело до формування субоптимальної та пасивної політики.

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

У роботі проведено порівняльний аналіз архітектур глибокого навчання з підкріпленням для задачі локальної навігації БПЛА в умовах стохастичної невизначеності (змінна маса, турбулентний вітер, рухома ціль). За результатами експериментального моделювання методів Policy-based (PPO) та Value-based (SAC) зроблено наступні висновки.

По-перше, встановлено, що для високодинамічних середовищ алгоритми прямої оптимізації політики (PPO) демонструють вищу стійкість. Агент PPO досяг показника успішності 97,1%, формуючи стабільні та оптимальні за часом траєкторії. Це підтверджує здатність Policy-based методів ефективно генералізувати фізичні закономірності польоту навіть за наявності значних зовнішніх збурень.

По-друге, виявлено обмеження Value-based підходів (SAC) у задачах безперервного керування зі змінною динамікою. Попри високу швидкість навчання (збіжність у 10 разів швидша за PPO), сформована політика виявилася схильною до локальних оптимумів, показавши успішність лише 63,2%. Виявлений ефект «Reward Hacking» (висока середня винагорода при низькій успішності) свідчить про неспроможність Critic-мережі коректно оцінювати довгострокову цінність дій в умовах швидкої зміни станів, що призводить до пасивної стратегії утримання висоти замість виконання місії.

По-третє, аналіз кінематики польоту показав, що PPO забезпечує плавність керування без осциляцій, характерних для SAC, що є критичною вимогою для імплементації на реальних фізичних системах. Отримані дані вказують на те, що перевага Off-policy методів у ефективності використання даних нівелюється їхньою чутливістю до похибок апроксимації Q-функції в стохастичних середовищах.

Перспективи подальших досліджень передбачають перевірку отриманих політик на апаратній платформі для оцінки стійкості до сенсорного шуму. Також планується дослідження гібридних архітектур (наприклад, TD3 [11]), спрямованих на зменшення помилки переоцінки функції цінності, та перехід до навігації на основі візуальних сенсорів.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. AlMahamid, F., & Grolinger, K. (2022). Autonomous unmanned aerial vehicle navigation using reinforcement learning: A systematic review. *Engineering Applications of Artificial Intelligence, 115*, 105321. https://doi.org/10.1016/j.engappai.2022.105321
2. Zhou, Y., Shu, J., Zheng, X., Hao, H., & Song, H. (2022). Real-time route planning of unmanned aerial vehicles based on improved soft actor-critic algorithm. *Frontiers in Neurorobotics, 16*, 1025817. https://doi.org/10.3389/fnbot.2022.1025817
3. Tang, C., Abbatematteo, B., Hu, J., Chandra, R., Martín-Martín, R., & Stone, P. (2025). Deep reinforcement learning for robotics: A survey of real-world successes. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 39). https://doi.org/10.1609/aaai.v39i27.35095
4. Oyinlola, S., Subedi, N., & Sarkar, S. (2025). *Reinforcement learning for autonomous point-to-point UAV navigation*. arXiv. https://doi.org/10.48550/arXiv.2509.13943
5. Bălaşa, R.-I., Bîlu, M., & Iordache, C. (2022). A proximal policy optimization reinforcement learning approach to unmanned aerial vehicles attitude control. *Land Forces Academy Review, 27*, 400–410. https://doi.org/10.2478/raft-2022-0049
6. Chen, S., Mo, Y., Wu, X., Xiao, J., & Liu, Q. (2024). Reinforcement learning-based energy-saving path planning for UAVs in turbulent wind. *Electronics, 13*(16), Article 3190. https://doi.org/10.3390/electronics13163190
7. Tai, J. J., Wong, J., Innocente, M., Horri, N., Brusey, J., & Phang, S. K. (2023). *PyFlyt—UAV simulation environments for reinforcement learning research*. arXiv. https://doi.org/10.48550/arXiv.2304.01305
8. Geronel, R. S., Botez, R. M., & Bueno, D. D. (2023). Dynamic responses due to the Dryden gust of an autonomous quadrotor UAV carrying a payload. *The Aeronautical Journal, 127*(1307), 116–138. https://doi.org/10.1017/aer.2022.35
9. Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., & Dormann, N. (2021). Stable-Baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research, 22*(268), 1–8.
10. Skalse, J., Howe, N. H. R., Krasheninnikov, D., & Krueger, D. (2022). Defining and characterizing reward hacking. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS 2022)*.
11. Liu, H., Shen, Y., Zhou, C., Zou, Y., Gao, Z., & Wang, Q. (2024). TD3-based collision-free motion planning for robot navigation. In *Proceedings of the 6th International Conference on Communications, Information System and Computer Engineering (CISCE 2024)* (pp. 247–250). https://doi.org/10.1109/CISCE62493.2024.10653233

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 4 (32), 2026
ISSN 2663 – 4023

Roman Trembovetskyi
Postgraduate
Assistant to the Information Security and Computer Engineering Department
Cherkasy State Technological University, Cherkasy, Ukraine
ORCID: 0009-0001-2697-0879
roman.tremb@gmail.com

Inna Rozlomii
PhD, Associate Professor
Associate Professor of the Information Security and Computer Engineering Department
Cherkasy State Technological University, Cherkasy, Ukraine
ORCID: 0000-0001-5065-9004
inna-roz@ukr.net

COMPARATIVE ANALYSIS OF POLICY-BASED AND VALUE-BASED RL METHODS FOR LOCAL MOTION PLANNING OF UAVS

Abstract. The paper presents a comparative study on the effectiveness of two fundamental Deep Reinforcement Learning architectures – Policy-based (using PPO) and Value-based (using SAC) – for the task of UAV local navigation under stochastic uncertainty. Experimental modeling, conducted in the PyFlyt environment considering turbulent wind, variable drone mass, and target motion, revealed critical discrepancies in algorithm performance. It was established that despite high training speed and sample efficiency, the SAC algorithm forms vulnerable policies with a low mission success rate (63.2%) due to value function estimation errors in dynamic states. In contrast, the PPO method ensured the formation of a robust control strategy with a 97.1% success rate and the generation of optimal trajectories without oscillations. The results confirm that for continuous control tasks in unpredictable environments, direct policy optimization methods are more effective as they avoid the "Reward Hacking" effect and better generalize the physical laws of flight.

Keywords: autonomous navigation, UAV, reinforcement learning, PPO, SAC, stochastic environment, robust control.

REFERENCES (TRANSLATED AND TRANSLITERATED)

1. AlMahamid, F., & Grolinger, K. (2022). Autonomous unmanned aerial vehicle navigation using reinforcement learning: A systematic review. *Engineering Applications of Artificial Intelligence, 115*, 105321. https://doi.org/10.1016/j.engappai.2022.105321
2. Zhou, Y., Shu, J., Zheng, X., Hao, H., & Song, H. (2022). Real-time route planning of unmanned aerial vehicles based on improved soft actor-critic algorithm. *Frontiers in Neurorobotics, 16*, 1025817. https://doi.org/10.3389/fnbot.2022.1025817
3. Tang, C., Abbatematteo, B., Hu, J., Chandra, R., Martín-Martín, R., & Stone, P. (2025). Deep reinforcement learning for robotics: A survey of real-world successes. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 39). https://doi.org/10.1609/aaai.v39i27.35095
4. Oyinlola, S., Subedi, N., & Sarkar, S. (2025). *Reinforcement learning for autonomous point-to-point UAV navigation*. arXiv. https://doi.org/10.48550/arXiv.2509.13943
5. Bălaşa, R.-I., Bîlu, M., & Iordache, C. (2022). A proximal policy optimization reinforcement learning approach to unmanned aerial vehicles attitude control. *Land Forces Academy Review, 27*, 400–410. https://doi.org/10.2478/raft-2022-0049
6. Chen, S., Mo, Y., Wu, X., Xiao, J., & Liu, Q. (2024). Reinforcement learning-based energy-saving path planning for UAVs in turbulent wind. *Electronics, 13*(16), Article 3190. https://doi.org/10.3390/electronics13163190
7. Tai, J. J., Wong, J., Innocente, M., Horri, N., Brusey, J., & Phang, S. K. (2023). *PyFlyt—UAV simulation environments for reinforcement learning research*. arXiv. https://doi.org/10.48550/arXiv.2304.01305
8. Geronel, R. S., Botez, R. M., & Bueno, D. D. (2023). Dynamic responses due to the Dryden gust of an autonomous quadrotor UAV carrying a payload. *The Aeronautical Journal, 127*(1307), 116–138. https://doi.org/10.1017/aer.2022.35

9. Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., & Dormann, N. (2021). Stable-Baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, *22*(268), 1–8.
10. Skalse, J., Howe, N. H. R., Krasheninnikov, D., & Krueger, D. (2022). Defining and characterizing reward hacking. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS 2022)*.
11. Liu, H., Shen, Y., Zhou, C., Zou, Y., Gao, Z., & Wang, Q. (2024). TD3-based collision-free motion planning for robot navigation. In *Proceedings of the 6th International Conference on Communications, Information System and Computer Engineering (CISCE 2024)* (pp. 247–250). https://doi.org/10.1109/CISCE62493.2024.10653233

Отримано редакцією журналу / Received: 27.01.26
Прорецензовано / Revised: 18.02.26
Схвалено до друку / Accepted: 26.03.26