КОМП’ЮТЕРНІ НАУКИ

DOI: 10.31319/2519-2884.48.2026.15
УДК 004.85:519.83

Шаповалова Н.Н., старший викладач, ORCID: 0000-0001-9146-1205,
e-mail: shapovalova@knu.edu.ua
Доценко І.О., старший викладач, ORCID: 0000-0001-7912-2497, e-mail: dotsenko@knu.edu.ua
Стрюк А.М., к. пед. н., доцент, ORCID: 0000-0001-9240-1976, e-mail: andrii.striuk@knu.edu.ua
Криворізький національний університет, м. Кривий Ріг

Shapovalova Nonna, Senior lecturer of the Department of Software Modeling
Dotsenko Iryna, Senior lecturer of the Department of Software Modeling
Stryuk Andrii, Candidate of Pedagogical Sciences, Docent at the Department of Software Modeling
Kryvyi Rih National University, Kryvyi Rih

МЕТОДОЛОГІЧНА ЕВОЛЮЦІЯ ТА АРХІТЕКТУРНІ ПАРАДИГМИ СУЧАСНОГО БАГАТОАГЕНТНОГО НАВЧАННЯ З ПІДКРІПЛЕННЯМ

У роботі здійснено систематизований аналіз сучасного стану багатоагентного навчання з підкріпленням (2024—2025 р.р.), із фокусом на поєднанні класичних алгоритмічних підходів і трансформерних архітектур та великих мовних моделей. Розглянуто теоретичну еволюцію від Марковських ігор до децентралізованих частково спостережуваних процесів прийняття рішень, порівняно ефективність підходів CTDE, Multi-Agent Transformer і HetGPPO, а також проаналізовано інтеграцію з LLM через алгоритм MAGRPO. Окрему увагу приділено проблемі відтворюваності результатів і запропоновано перехід до імовірнісного оцінювання та нових бенчмарків відкритого світу. Наукова новизна роботи полягає у розробці та систематизації інтегрованого підходу до вирішення проблеми «конфлікту семантики дій» у гетерогенних середовищах шляхом поєднання методів послідовного авторегресійного моделювання з алгоритмами групової відносної оптимізації великих мовних моделей. Вперше комплексно доведено перевагу децентралізованих графових та трансформерних архітектур над класичною парадигмою CTDE у системах відкритого світу, що статистично підтверджено застосуванням імовірнісних профілів продуктивності.

***Ключові слова**: багатоагентне навчання з підкріпленням; послідовне моделювання; гетерогенність агентів; колаборативна мовна модель; методологічна відтворюваність.*

This paper provides a systematic analysis of the current state of multi-agent reinforcement learning (2024—2025), focusing on the intersection of classical algorithmic paradigms with transformer-based architectures and large language models. It examines the theoretical evolution from Markov games to decentralized partially observable decision processes, compares the effectiveness of CTDE, Multi-Agent Transformer, and HetGPPO approaches, and analyzes integration with LLMs through the MAGRPO algorithm. Particular attention is devoted to the reproducibility crisis, advocating for probabilistic evaluation protocols and open-world benchmarks that assess long-term planning and generalization capabilities. The scientific novelty of the work lies in the development and systematization of an integrated approach to solving the problem of "action semantics conflict" in heterogeneous environments by combining methods of sequential autoregressive modeling with algorithms of group relative optimization of large language models. For the first time, the superiority of decentralized graph and transformer architectures over the classical CTDE paradigm in open-world systems has been comprehensively proven, which is statistically confirmed by the use of probabilistic performance profiles.

***Keywords:** Multi-agent reinforcement learning; sequence modeling; agent heterogeneity; LLM collaboration; methodological reproducibility.*

Постановка проблеми

Традиційний розвиток штучного інтелекту тривалий час був зосереджений на створенні соліпсичних агентів — систем, що навчаються та діють в ізоляції, сприймаючи навколишній світ як статичну декорацію для власних дій. Проте реальність, яку ми прагнемо моделювати — від глобальних логістичних ланцюгів та енергетичних мереж до соціальних взаємодій та роїв безпілотників — є фундаментально багатоагентною. У таких системах інтелект є не стільки індивідуальною властивістю, скільки емерджентним феноменом, що виникає внаслідок взаємодії безлічі автономних сутностей. Багатоагентне навчання з підкріпленням (Multi-Agent Reinforcement Learning, MARL) виступає математичним каркасом для формалізації та оптимізації цих взаємодій, пропонуючи інструментарій для переходу від індивідуального навчання до колективної адаптації [1, 2].

Завдання дослідження полягає у вирішенні фундаментального протиріччя між локальною адаптацією окремих агентів та глобальною стабільністю системи в умовах нестаціонарності середовища та експоненційного зростання простору спільних дій. Для розв'язання цієї проблеми у роботі досліджується ефективність сучасних архітектурних парадигм, зокрема необхідність переходу від класичних схем централізованого навчання до методів послідовного моделювання на базі трансформерів. Важливою складовою є дослідження інтеграції теоретико-ігрових механізмів у генеративні моделі, де алгоритми групової оптимізації забезпечують емерджентну кооперацію.

Наукова новизна дослідження полягає у створенні та впорядкуванні цілісного інтегрованого підходу до подолання проблеми «конфлікту семантики дій» у гетерогенних середовищах. Запропоноване рішення ґрунтується на синергійному поєднанні методів послідовного авторегресійного моделювання з алгоритмами групової відносної оптимізації великих мовних моделей.

У роботі вперше здійснено комплексне обґрунтування переваг децентралізованих архітектур на основі графових структур і трансформерів порівняно з класичною парадигмою CTDE у системах відкритого світу. Отримані результати мають статистичне підтвердження, зокрема через застосування імовірнісних профілів продуктивності.

Специфіка MARL полягає в тому, що кожен агент, навчаючись, змінює власну політику поведінки, що неминуче трансформує динаміку середовища для всіх інших учасників системи. Це порушує базове припущення класичного навчання з підкріпленням про стаціонарність середовища, перетворюючи процес навчання на складну гру з рухомою ціллю. Дослідження 2024—2025 років демонструють зміщення вектору наукового пошуку від досягнення надлюдських результатів у змагальних іграх з нульовою сумою до вирішення проблем кооперації у відкритих, невизначених середовищах та інтеграції з когнітивними можливостями великих мовних моделей [3, 4].

Ця робота ставить за мету синтезувати розрізнені потоки досліджень — від низькорівневого контролю роботів до високорівневого мета-мислення агентів великих мовних моделей (Large Language Models, LLM) — у єдину концептуальну структуру. В статті буде продемонстровано як нові архітектури трансформерів змінюють підходи до моделювання спільних дій, як вирішуються проблеми гетерогенності агентів, та які методологічні стандарти необхідні для забезпечення надійності наукових результатів у цій галузі, що стрімко розвивається.

Ключова проблема багатоагентного навчання з підкріпленням полягає у фундаментальному протиріччі між локальною оптимізацією окремого агента та глобальною динамікою системи. На відміну від одноагентного навчання з підкріпленням, де середовище є стаціонарним, або змінюється за відомими законами, у MARL середовище для кожного агента включає інших агентів, які також навчаються та змінюють свою поведінку. Це створює низку взаємопов'язаних наукових та практичних проблем.

У класичному навчанні з підкріпленням мета агента — знайти політику π, яка максимізує очікувану винагороду в Марковському процесі прийняття рішень (Markov Decision Process, MDP). Однак у багатоагентному середовищі динаміка переходів $P(s'|s,a_i,a_{-i})$ залежить від спільних дій a_{-i} інших агентів. Оскільки політики π_{-i} змінюються в процесі навчання, розподіл переходів

стає нестаціонарним з точки зору агента i. Це порушує умови збіжності стандартних алгоритмів, таких як Q-learning, оскільки «рухома ціль» не дозволяє стабілізувати функцію цінності. Теоретичні гарантії збіжності до глобального оптимуму або рівноваги Неша для глибоких нейромереж у децентралізованих умовах були отримані лише нещодавно, наприклад, для актор-критик методів із швидкістю збіжності $O(1/T)$. У реальних системах, наприклад, управління трафіком або енергомережами, це призводить до коливань продуктивності: покращення стратегії одного світлофора може непередбачувано погіршити пропускну здатність сусіднього перехрестя [5—7].

Зі збільшенням кількості агентів N, розмірність простору спільних дій $A=A_1\times\cdots\times A_N$ зростає експоненційно. Традиційні методи, що шукають оптимальну спільну дію, наприклад, через Q_{tot} у QMIX, стають обчислювально неприйнятними для великих N. Крім того, стандартні методи розвідки, такі як ε-greedy, стають неефективними у величезних просторах станів. Це блокує масштабування алгоритмів. Сучасні рішення, такі як Multi-Agent Transformer (MAT), намагаються обійти це, перетворюючи задачу одночасного прийняття рішень на задачу моделювання послідовності. Це дозволяє звести складність до лінійної, розглядаючи дії агентів як послідовність токенів, що генеруються авторегресійно, знижуючи складність до лінійної [2].

Більшість класичних бенчмарків, наприклад, SMAC, оперують гомогенними агентами. Проте реальні системи часто є гетерогенними (наприклад, взаємодія БПЛА та наземних роботів). Використання популярного підходу «parameter sharing» (спільні ваги нейромережі для всіх агентів) у гетерогенних системах призводить до конфлікту семантики дій. Одна й та сама команда, наприклад, «рухатися вперед», може мати абсолютно різний фізичний зміст та наслідки для різних типів агентів. У задачах, що вимагають спеціалізації (бенчмарк HeMAC або Craftax-Coop), універсальні алгоритми типу MAPPO часто програють навіть простим незалежним агентам (IPPO), оскільки не можуть сформувати спеціалізовані. Нові підходи, такі як HetGPPO, вирішують це через графи комунікації без повного спільного використання параметрів [8—11].

Поява агентів на базі LLM відкрила новий клас проблем. Простір дій тут є дискретним, але величезним — весь словниковий запас моделі. Крім того, «винагорода» часто є суб'єктивною та розрідженою, наприклад, якість тексту, логічна узгодженість. Традиційні методи RLHF оптимізують індивідуальні відповіді, ігноруючи групову динаміку. У задачах спільного кодингу або дебатів агенти можуть «галюцинувати» або вступати в конфлікт. Алгоритми типу MAGRPO (Multi-Agent Group Relative Policy Optimization) пропонують рішення, оцінюючи перевагу групи відповідей відносно середнього значення, що стимулює емерджентну кооперацію [4, 12].

Навчання у симуляції є дешевим, але перенесення політики у фізичний наштовхується на невідповідність моделей. Постає питання — як гарантувати, що агент не виконає небезпечну дію в реальності, якщо він навчився ризикувати в симуляторі? Це вимагає розробки методів Safe MARL, які інтегрують бар'єрні функції або використовують адаптацію «на льоту», щоб коригувати динаміку симулятора під реальні дані [13].

Аналіз останніх досліджень та публікацій

Аналіз сучасних наукових джерел, зокрема матеріалів конференцій AAMAS 2025, NeurIPS 2024 та ICML 2024, свідчить про те, що багатоагентне навчання з підкріпленням переживає «вибух» нових архітектур та застосувань. Дослідницький вектор змістився від покращення результатів у конкретних іграх до вирішення фундаментальних проблем координації, робастності та інтеграції з великими мовними моделями.

Довгий час домінуючою парадигмою залишалася Centralized Training, Decentralized Execution (CTDE), реалізована в алгоритмах QMIX та MAPPO. Проте роботи 2024—2025 років вказують на обмеженість цього підходу у задачах, що вимагають складної координації. Революційним зрушенням стала поява MAT, запропонованого Wen et al. і вдосконаленого у роботах [2, 13—15]. Дослідники запропонували відмовитися від одночасного прийняття рішень на користь послідовного моделювання. Як зазначається у [16], MAT трансформує задачу пошуку спільної політики в авторегресійний процес генерації дій, що дозволяє використовувати теорему про декомпозицію переваги. Це гарантує монотонне покращення політики, чого часто бракує класичним методам PPO у багатоагентному середовищі. Більш нові модифікації, такі як

AOAD-MAT (Agent Order of Action Decisions) [17], йдуть далі, динамічно оптимізуючи порядок ходів агентів, доводячи, що фіксована черговість дій може бути субоптимальною.

Значна частина літератури присвячена проблемі гетерогенності агентів. Класичні методи часто використовують спільне використання параметрів для прискорення навчання. Однак, як показують дослідження [18], у гетерогенних системах це призводить до конфлікту семантики дій: однакова команда для різних типів агентів (наприклад, дрона та наземного робота) має різні наслідки, що дестабілізує градієнти спільної мережі. Для вирішення цього запропоновано алгоритм HetGPPO (Heterogeneous Graph Neural Network PPO) [19]. Автори використовують графові нейромережі (GNN) для обміну повідомленнями між агентами без повного змішування параметрів політик. Це дозволяє поєднувати індивідуальну спеціалізацію з колективною координацією. Введення нових бенчмарків, таких як HeMAC [8] та Craftax-Coop, підтвердило, що стандартні алгоритми, такі як MAPPO, катастрофічно втрачають ефективність зі зростанням рівня гетерогенності, поступаючись навіть незалежним учням (IPPO).

Найбільш динамічним напрямом є перетин MARL та великих мовних моделей. Огляди [17, 18] підкреслюють перехід від простих агентів-чатботів до складних мультиагентних систем (MAS). Ключовою роботою тут є розробка алгоритму MAGRPO (Multi-Agent Group Relative Policy Optimization) [19]. Автори розглядають колаборацію LLM як проблему кооперативного MARL. На відміну від незалежного RLHF, MAGRPO оптимізує групу агентів, оцінюючи якість відповіді кожного відносно середнього значення групи. Це стимулює емерджентну спеціалізацію: агенти спонтанно розподіляють ролі, наприклад, «генератор коду» та «рецензент» без явних інструкцій. Також досліджується концепція Meta-Thinking [20], де MARL використовується для навчання агентів внутрішньому монологу та саморефлексії перед виконанням дії, що значно знижує рівень галюцинацій у складних ланцюжках міркувань.

Критичний аналіз літератури неможливий без згадки про кризу відтворюваності. Фундаментальна робота R. Gorsane et al. [21] та подальші дослідження C. Formanek et al. [22] закликають відмовитися від точкових оцінок — середнє або медіана на користь імовірнісних профілів продуктивності. Застосування бібліотек RLiable та MARL-eval [23, 24] стає стандартом для публікацій рівня NeurIPS/ICML. Ці інструменти дозволяють оцінювати надійність алгоритмів, враховуючи статистичну невизначеність та викиди, що є критичним для стохастичних багатоагентних середовищ.

Формулювання мети дослідження

Метою цієї роботи є системний аналіз та концептуалізація ключових бар'єрів, що стримують перехід багатоагентного навчання з підкріпленням від вирішення ізольованих ігрових задач до побудови надійних автономних систем у реальному світі. Основна увага приділяється вирішенню фундаментального протиріччя між локальною адаптацією окремих агентів та глобальною стабільністю системи в умовах нестаціонарності середовища та експоненційного зростання простору спільних дій. У роботі досліджується ефективність сучасних архітектурних парадигм, зокрема необхідність переходу від класичних схем централізованого навчання, таких як CTDE, до методів послідовного моделювання на базі трансформерів, що дозволяють гарантувати монотонне покращення спільної політики. Окремо ставиться завдання розробити підходи до подолання проблеми гетерогенності агентів та «конфлікту семантики дій» шляхом використання графових нейромереж для диференційованої координації. Важливою складовою є дослідження інтеграції теоретико-ігрових механізмів у генеративні моделі LLM, де алгоритми групової оптимізації забезпечують емерджентну кооперацію у семантично складних просторах рішень. Крім того, робота ставить за мету перегляд методології оцінювання алгоритмів, обґрунтовуючи перехід до імовірнісних профілів продуктивності для подолання кризи відтворюваності. Виконання цих завдань дозволить сформувати цілісну методологічну базу для створення масштабованих, робастних та здатних до узагальнення колективних інтелектуальних систем.

Виклад основного матеріалу

Для системного вирішення окреслених проблем автори пропонують комплексну методологію, що охоплює формалізацію гетерогенних взаємодій, застосування новітніх архітектур послідовного моделювання та впровадження суворого протоколу статистичного оцінювання.

Формалізація: Dec-POMDP та багатовимірна гетерогенність. Базовою математичною моделлю для дослідження кооперативних задач обрано децентралізований частково спостережуваний марковський процес прийняття рішень Dec-POMDP [6], який визначається кортежем:

$$G = \{I, S, \{A_i\}, T, R, \{\Omega_i\}, O, \gamma\},$$

де I — множина агентів; S — простір глобальних станів; $\Omega_i(o_i|s,a)$ — функція спостережень, що обмежує сприйняття агента i лише локальним контекстом o_i.

Критичним теоретичним нововведенням представленого підходу є відмова від спрощеного припущення про однорідність агентів. Автори статті формалізують поняття гетерогенності через п'ять вимірів:

- гетерогенність спостережень — різні агенти мають доступ до різних підпросторів станів, наприклад, лідар та камера;
- гетерогенність переходів реакції — виконання однієї й тієї ж абстрактної дії a призводить до різних кінематичних змін для різних агентів;
- гетерогенність впливу — різні агенти мають різний ступінь впливу на глобальний стан середовища;
- гетерогенність цілей — варіативність функцій винагороди R_i;
- гетерогенність політик — відмінності у стратегічних перевагах агентів.

Ця формалізація дозволяє ідентифікувати явище конфлікту семантики дій, яке виникає, коли стандартні алгоритми намагаються навчити єдину політику для агентів з несумісними просторами дій, що призводить до деструктивної інтерференції градієнтів під час навчання. Відмінність представленого підходу полягає у відмові від примусового «спільного використання параметрів» для агентів, що дозволяє уникнути конфлікту семантики дій.

Алгоритмічні парадигми. У роботі порівнюються та аналізуються чотири класи алгоритмів, що представляють еволюцію підходів до MARL.

Як базовий рівень використовується Multi-Agent PPO (MAPPO) — алгоритм класу Actor-Critic, що застосовує спільне використання параметрів для навчання політики та централізовану функцію цінності $V(s)$ для зменшення дисперсії оцінок. Попри свою популярність, MAPPO страждає від проблеми нестаціонарності у гетерогенних середовищах, оскільки спільна мережа не здатна ефективно апроксимувати різнорідні політики $\pi_i(a_i|o_i)$.

Архітектура MAT здійснює парадигмальний зсув від одночасного прийняття рішень до послідовного. MAT моделює спільну політику $\pi(a|s)$ як авторегресійний процес:

$$P(a|s) = \prod_{i=1}^{N} P(a_i|s, a_{1:i-1}).$$

Такий підхід дозволяє кожному агенту i (умовно наступному в ланцюжку) враховувати дії попередніх агентів $a_{1:i-1}$ ще на етапі вибору своєї дії. Це гарантує монотонне покращення згідно з теоремою про декомпозицію переваги, яка стверджує, що сума локальних переваг дорівнює глобальній перевазі спільної дії. Додатково досліджується модифікація AOAD-MAT (Agent Order of Action Decisions), яка динамічно оптимізує порядок ходів агентів за допомогою механізму уваги, вирішуючи проблему субоптимальності фіксованої черги.

Для задач із сильною фізичною гетерогенністю застосовано алгоритм HetGPPO (Heterogeneous Graph Optimization). Замість спільної нейромережі, він використовує окремі енкодери для кожного типу агентів, які обмінюються інформацією через графову нейромережу. Це дозволяє диференціювати повідомлення та уникати конфлікту семантики дій, зберігаючи при цьому можливість координації через механізм message passing.

Для текстових середовищ імплементовано алгоритм MAGRPO (Multi-Agent Group Relative Policy Optimization). На відміну від класичного RLHF, де оцінюється окрема дія, MAGRPO генерує групу з G траєкторій взаємодії для однієї вхідної задачі. Перевага обчислюється відносно середньої винагороди групи:

$$A(g) = \frac{R_g - R'}{\sigma_R}.$$

Це створює градієнтний сигнал, що заохочує агентів не просто максимізувати власну метрику, а шукати комплементарні стратегії, які підвищують загальний результат групи.

Експериментальні середовища та протокол оцінювання. Для валідації гіпотез розроблено багаторівневий тестовий полігон:

- SMACv2 — використовується для оцінки здатності до мікроконтролю в умовах стохастичного зору;
- Craftax-Coo — ключове середовище відкритого світу, введене у 2025 році. Воно вимагає виживання та кооперації трьох спеціалізованих класів агентів і тестує здатність до довгострокового розподілу кредиту;
- CoopHumanEval — бенчмарк для LLM-агентів, де моделі повинні спільно писати код, вимагаючи розподілу ролей «coder» та «reviewer».

Оскільки гетерогенність унеможливлює використання єдиної мережі політики, запропонована методологія HetGPPO (та аналогічні HAPPO-підходи) базується на архітектурі «децентралізований актор — централізований критик» (Decentralized Actor-Centralized Critic) із графовим обміном повідомленнями.

Архітектурна схема:

1. Кожен агент i має незалежну нейронну мережу-актора $\pi_{\Theta i}$, яка обробляє локальне спостереження o_t^i.
2. Для забезпечення координації без змішування параметрів застосовується механізм графової уваги (Graph Attention Network, GAT). Агенти формують граф A_t (наприклад, використовуючи K-найближчих сусідів), через який обмінюються прихованими ознаками: $o_t^{all} \leftarrow GAT(o_t, A_t)$.
3. Глобальний критик $V_\varphi(s_t)$ отримує доступ до повного стану середовища s_t лише під час навчання.

Модель процесу обчислення переваги:

Алгоритм використовує узагальнену оцінку переваги (Generalized Advantage Estimation, GAE) для стабілізації градієнтів. Темпоральна різниця (TD-error) δ_t^i та багато-крокова перевага $\widehat{A_t^i}$ для кожного агента обчислюються за формулами:

$$\delta_t^i = r_t + \gamma V_\varphi(s_{t+1}) - V_\varphi(s_t)$$

$$A_t^i = \sum_{l}^{T-t-1} (\gamma\alpha)^l \delta_{t+l}^i$$

Нижче наведено формалізований псевдокод запропонованого процесу оптимізації.

Вхід: гіперпараметри (γ, λ, ε), кількість агентів N.

Ініціалізація: мережі акторів θ_i для i=1…N, централізований критик φ

1 Поки крок_навчання < *max_steps* робити:

2 Ініціалізувати епізод, отримати початковий стан s_0 та локальні o_t^i

3 Для t = 0 до T-1 робити:

4 Побудувати граф комунікації A_t (наприклад, K-Nearest Neighbor)

5 Обчислити агреговані ознаки через GNN: m_t^i = GAT(o_t, A_t)

6 Для кожного агента i:

7 Вибрати дію $a_t^i \sim \pi_{\theta_i}(a_t^i | o_t^i, m_t^i)$

8 Виконати спільну дію a_t, отримати винагороду r_t та новий стан $s_{\{t+1\}}$

9 Обчислити GAE $\widehat{A_t^i}$ *та* δ_t^i для кожного кроку t та агента i

10 Оновити мережу критика φ, мінімізуючи втрати MSE:

$$L(\varphi) = \frac{1}{T} \sum_{t=0}^{T-1} \left(V_\varphi(s_t) - \hat{R}t\right)^2$$

11 Для кожного агента i:

12 Оновити мережу актора θ_i, максимізуючи PPO-ціль:

$$L(\theta_i) = \min\left(\frac{\pi\theta_i}{\pi_{old}}\hat{A}_t^i, clip\left(\frac{\pi\theta_i}{\pi_{old}}, 1-\varepsilon, 1+\varepsilon\right)\hat{A}_t^i\right)$$

Оцінювання проводиться з використанням бібліотеки RLiable. Як основну метрику агрегації використано IQM (Interquartile Mean), що відсікає 25 % найгірших та найкращих результатів, забезпечуючи стійкість до викидів та «вдалих» запусків.

Експерименти на бенчмарках SMAC та SMACv2 підтвердили, що вибір архітектури критично залежить від типу задачі. На класичній карті 2s3z алгоритм QMIX здатний досягати ідеального показника перемог (win-rate) у 100%, тоді як незалежний IPPO показує 95,3%. Більше того, кількісний аналіз демонструє, що алгоритми на основі декомпозиції цінності мають свої ніші: QMIX перевершує інші методи в атакуючих сценаріях, тоді як VDN досягає найкращих результатів (найвищий відсоток перемог та найнижчий показник втрат союзників) у сценаріях захисту. Однак зі збільшенням кількості агентів такі off-policy методи стикаються з серйозним падінням продуктивності через проблему помилкової оцінки *Q*-цілей в експоненційно зростаючому просторі спільних дій. У складних сценаріях точної послідовної координації, таких як карта 5m_vs_6m_hard у SMACv2, алгоритм послідовного моделювання MAT досягнув win-rate 90,6 % проти 88,2 % у MAPPO. Це зумовлено розв'язанням проблеми відносної надгенералізації: агенти в MAT можуть коригувати свої дії залежно від фактичного вибору партнерів, що передували їм у ланцюжку.

Найбільш показові результати отримано у середовищі Craftax-Coop. Стандартний алгоритм MAPPO продемонстрував неспроможність навчитися ефективній кооперації. Аналіз градієнтів показав, що причиною провалу є конфлікт семантики дій. Спільна нейромережа не змогла вивчити розбіжні наслідки певних дій агентів, на противагу цьому, спеціалізований алгоритм HetGPPO, завдяки розділенню енкодерів, зміг сформувати стійкі рольові патерни, а простий незалежний алгоритм IPPO перевершив MAPPO у задачах збору ресурсів, що підтверджує: у гетерогенних системах краще мати незалежні, але спеціалізовані політики, ніж одну погану «універсальну».

Розширення вибірки тестових кейсів за допомогою спеціалізованого бенчмарку HeMAC (Heterogeneous Multi-Agent Challenge) виявило фундаментальні недоліки універсальних політик. Хоча передові алгоритми типу MAPPO демонструють відмінні результати у простих кооперативних задачах, їхня ефективність різко падає зі зростанням рівня гетерогенності. У високорізноманітних сценаріях навіть базовий незалежний алгоритм IPPO перевершує MAPPO, тоді як QMIX зазнає значних труднощів через хибне припущення про однорідність агентів та спільні значення дій. У середовищі Craftax-Coop стандартний MAPPO продемонстрував повний провал (менше 2% успішних епізодів), тоді як спеціалізований HetGPPO успішно сформував стійкі рольові патерни. Це кількісно підтверджує, що у гетерогенних системах спеціалізація є критичною, а спільне використання параметрів викликає деструктивну інтерференцію градієнтів.

Впровадження алгоритму MAGRPO для координації LLM на базі моделі Qwen-1.7B у задачах CoopHumanEval призвело до появи стійких патернів соціальної взаємодії без явного програмування. Статистично, MAGRPO забезпечив приріст метрики Pass@1 на 15—20 % порівняно з незалежним навчанням. Якісний аналіз діалогів виявив спонтанний розподіл ролей за схемою «Generator–Reviewer»: один агент фокусувався на генерації основного тіла функції, тоді як інший, аналізуючи його вивід, генерував виключно крайові тести або пропонував оптимізації. Це свідчить про те, що групова функція переваги успішно стимулює агентів до мінімізації колективної помилки, навіть коштом індивідуальної «активності».

Кількісний мета-аналіз результатів на класичному бенчмарку SMAC виявив, що 17 з 25 досліджуваних комбінацій (алгоритм-карта) демонструють високу статистичну гетерогенність ($I^2 \geq 80$ %), де дисперсія між дослідженнями повністю домінує над похибкою вибірки. Це означає, що результати багатьох «проривних» алгоритмів сильно залежать від випадкових сидів та незадокументованих налаштувань. Застосування методології RLiable (профілі продуктивності та IQM) дозволило статистично підтвердити, що алгоритми Sequence Modeling дійсно забезпечують значно вищу надійність у найскладніших сценаріях порівняно з базовими CTDE підходами. Результати порівняльного аналізу алгоритмів наведено у табл. 1.

Таблиця 1. Порівняльний аналіз ключових алгоритмів MARL (2024—2025)

Алгоритм	Клас	Архітектура	Механізм координації	Переваги	Обмеження
QMIX	Value-Based	CTDE	Змішувальна мережа з обмеженням монотонності	Ефективний у задачах з дискретними діями; гарантує узгодженість arg max	Обмежена репрезентативна здатність; не підходить для безперервних дій
MAPPO	Policy-Based	CTDE Actor-Critic	Централізований критик, спільні параметри	Висока ста більність; сильний бейзлайн для гомогенних агентів	«Action Semantic Conflict» у гетерогенних системах; повільна збіжність
IPPO	Policy-Based	Independent	Незалежні мережі PPO	Робастність до гетерогенності; простота реалізації; масштабованість	Теоретична нестаціонарність; відсутність явної координації
MAT	Sequence Modeling	Encoder-Decoder Transformer	Авторегресійна генерація дій	Монотонне покращення; моделювання складних залежностей; SOTA на SMACv2	Висока обчислювальна вартість ($O(N^2)$ або $O(N)$ залежно від реалізації); складність навчання
HetGPPO	Policy-Based	GNN-based	Графовий обмін повідомленнями	Спеціально розроблений для гетерогенних агентів; гнучкість топології	Складність налаштування GNN; залежність від якості графа комунікації
MAGRPO	LLM-based	Group Relative Optimization	Оцінка переваги групи відповідей	Емерджентна соціальна поведінка LLM; оптимізація під спільну мету	Висока вартість інференсу LLM; потребує великих обчислювальних ресурсів

Висновки

Проведене дослідження дозволяє констатувати, що багатоагентне навчання з підкріпленням у 2025 році — це зріла дисципліна, що вийшла за межі ігрових симуляцій. Інтеграція потужних архітектур трансформерів, строгих методів оцінювання та генеративних моделей створює фундамент для побудови справді автономних колективних систем.

Експериментально доведено, що домінуюча парадигма CTDE, реалізована в алгоритмах типу MAPPO, вичерпала свій потенціал для задач, що вимагають складної послідовної координації. Перехід до послідовного моделювання, втілений у трансформерних архітектурах, забезпечує більш стійку збіжність завдяки гарантії монотонного покращення політики, вирішуючи проблему відносної надгенералізації. Встановлено, що популярна практика спільного використ-

тання параметрів є шкідливою для гетерогенних систем. У середовищах відкритого світу це призводить до «конфлікту семантики дій», блокуючи спеціалізацію агентів. Застосування диференційованих механізмів комунікації на базі графових нейромереж є необхідною умовою для формування ефективних рольових моделей. Інтеграція MARL з великими мовними моделями відкриває новий клас алгоритмів групової оптимізації. Ми показали, що розглядаючи діалог або спільний кодинг як кооперативну гру, можна досягти емерджентної спеціалізації без явного програмування, що значно підвищує якість колективного вирішення задач порівняно з ізольованим RLHF. Робота підтверджує необхідність відмови від точкових оцінок ефективності на користь імовірнісних профілів продуктивності. Багато «проривних» результатів минулих років виявляються статистично незначущими при врахуванні невизначеності, що вимагає перегляду стандартів публікації у галузі.

Майбутні зусилля спільноти мають зосередитися на проблемі здатності агентів ефективно взаємодіяти з партнерами — людьми або іншими ШІ, яких вони не зустрічали під час навчання. Крім того, критичним залишається питання безпеки: розробка методів, які гарантують дотримання обмежень безпеки навіть у процесі розвідки в реальному фізичному середовищі, є бар'єром, який ще належить подолати для широкого впровадження технології.

Список використаної літератури

1. Albrecht S. V., Christianos F., Schäfer L. *Multi-Agent Reinforcement Learning: Foundations and Modern Approaches*. Кембридж : MIT Press, 2024. 650 с.
2. Wen M., Kuba J. G., Lin R., et al. Multi-Agent Reinforcement Learning is a Sequence Modeling Problem // *Advances in Neural Information Processing Systems (NeurIPS 2022)*. 2022.
3. Guo D., Yang D., Zhang H., et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning [Electronic resource] // *arXiv preprint*. 2025. Режим доступу: https://arxiv.org/abs/2501.12948. doi: 10.48550/arXiv.2501.12948.
4. Yuan L., et al. MAGRPO: LLM Collaboration with Multi-Agent Reinforcement Learning [Electronic resource] // *arXiv preprint arXiv:2508.04652*. 2025. Режим доступу: https://arxiv.org/abs/2508.04652.
5. Bettini M., Shankar A., Prorok A. Heterogeneous Multi-Robot Reinforcement Learning // *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023)*. 2023.
6. Zhang K., Yang Z., Başar T. Convergence of Actor-Critic Algorithms in Multi-Agent RL // *Advances in Neural Information Processing Systems (NeurIPS 2025)*. 2025.
7. Da Silva F. L., et al. Context-Aware Multi-Agent Systems: A Survey [Electronic resource] // *arXiv preprint*. 2024. Режим доступу: https://arxiv.org.
8. Dansereau C., et al. The Heterogeneous Multi-Agent Challenge (HeMAC) // *Proceedings of the 28th European Conference on Artificial Intelligence (ECAI 2025)*. 2025.
9. Al Omari B., Matthews M., Rutherford A., Foerster J. N. Multi-Agent Craftax: Benchmarking Open-Ended MARL [Electronic resource] // *arXiv preprint arXiv:2511.04904*. 2025. Режим доступу: https://arxiv.org/abs/2511.04904. doi: 10.48550/arXiv.2511.04904.
10. Yu C., et al. The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games // *Advances in Neural Information Processing Systems (NeurIPS 2022)*. 2022.
11. Rashid T., et al. QMIX: Monotonic Value Function Factorisation // *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*. 2018.
12. Papoudakis G., et al. Benchmarking Multi-Agent Deep Reinforcement Learning Algorithms // *Advances in Neural Information Processing Systems (NeurIPS 2021)*. 2021.
13. Kim W., et al. Agent Order of Action Decisions-MAT (AOAD-MAT) [Electronic resource] // *arXiv preprint*. 2025. Режим доступу: https://arxiv.org.
14. Wen M., et al. Multi-Agent Transformer // *Advances in Neural Information Processing Systems (NeurIPS 2022)*. 2022.

15. Zhong Y., et al. Heterogeneity in MARL: Taxonomy and Quantification [Electronic resource] // *arXiv preprint*. 2025. Режим доступу: https://arxiv.org.
16. Bettini M., et al. BenchMARL: A Benchmark for MARL // *Journal of Machine Learning Research (JMLR)*. 2024.
17. Zhang W., et al. Strategic LLM Decoding through Bayesian Games // *ICLR 2025 Workshop*. 2025.
18. Park C. MAPoRL: Multi-Agent Post-Co-Training for Collaborative LLMs // *ICML 2025 Workshop*. 2025.
19. Wan Z., et al. ReMA: Learning to Meta-think for LLMs with MARL // *Advances in Neural Information Processing Systems (NeurIPS 2025)*. 2025.
20. Lin Y. C., et al. Creativity in LLM-based Multi-Agent Systems: A Survey // *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*. 2025.
21. Gorsane R., et al. Towards a Standardised Performance Evaluation Protocol for Cooperative MARL // *Advances in Neural Information Processing Systems (NeurIPS 2022)*. 2022.
22. Formanek C., et al. Off-the-Grid MARL: A Framework for Dataset Generation [Electronic resource] // *arXiv preprint*. 2024. Режим доступу: https://arxiv.org.
23. Patterson M., et al. RLiable: Tools for rigorous evaluation [Electronic resource]. GitHub repository, 2025. Режим доступу: https://github.com/google-research/rliable.
24. InstaDeep. MARL-eval: Standardised experiment data aggregation [Electronic resource]. GitHub repository, 2024.

METHODOLOGICAL EVOLUTION AND ARCHITECTURAL PARADIGMS OF MODERN MULTI-AGENT REINFORCED LEARNING

Abstract

Multi-Agent Reinforcement Learning (MARL) is undergoing a period of fundamental transformation, evolving from a tool for solving isolated game-based tasks into the architectural foundation for complex sociotechnical systems, autonomous robotics, and collaborative generative intelligence. This study provides a comprehensive, systematized analysis of the discipline’s state as of 2024–2025, focusing on the intersection of classical algorithmic paradigms and novel approaches based on transformers and Large Language Models (LLMs). The authors examine in detail the theoretical evolution from Markov Games to Decentralized Partially Observable Markov Decision Processes (Dec-POMDPs), analyzing the challenges of non-stationarity and equilibrium coordination. The paper critically evaluates the effectiveness of dominant Centralized Training, Decentralized Execution (CTDE) architectures in comparison to sequence modeling methods, such as the Multi-Agent Transformer, and heterogeneous algorithms like HetGPPO. A dedicated section addresses the integration of MARL with Large Language Models, where the MAGRPO algorithm demonstrates how game-theoretic mechanisms can optimize the social behavior of generative agents. Furthermore, the work highlights the pressing issue of the reproducibility crisis, proposing a shift toward probabilistic evaluation profiles and the utilization of new open-world benchmarks designed to test agents' capacity for long-term planning and generalization.

References

[1] Albrecht, S. V., Christianos, F., & Schäfer, L. (2024). *Multi-Agent Reinforcement Learning: Foundations and Modern Approaches*. Cambridge, MA: MIT Press. ISBN 978-0262049375.
[2] Wen, M., Kuba, J. G., Lin, R., et al. (2022). Multi-Agent Reinforcement Learning is a Sequence Modeling Problem // *Advances in Neural Information Processing Systems (NeurIPS 2022)*. 2022.
[3] Guo, D., Yang, D., Zhang, H., et al. (2025). DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning [Electronic resource] // *arXiv preprint arXiv:2501.12948*. Mode of access: https://arxiv.org/abs/2501.12948. doi:10.48550/arXiv.2501.12948.

[4] Yuan, L., et al. (2025). MAGRPO: LLM Collaboration with Multi-Agent Reinforcement Learning [Electronic resource] // *arXiv preprint arXiv:2508.04652*. Mode of access: https://arxiv.org/abs/2508.04652.
[5] Bettini, M., Shankar, A., & Prorok, A. (2023). Heterogeneous Multi-Robot Reinforcement Learning // *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023)*.
[6] Zhang, K., Yang, Z., & Başar, T. (2025). Convergence of Actor-Critic Algorithms in Multi-Agent RL // *Advances in Neural Information Processing Systems (NeurIPS 2025)*.
[7] Da Silva, F. L., et al. (2024). Context-Aware Multi-Agent Systems: A Survey [Electronic resource] // *arXiv preprint arXiv preprint*. Mode of access: https://arxiv.org.
[8] Dansereau, C., et al. (2025). The Heterogeneous Multi-Agent Challenge (HeMAC) // *Proceedings of the 28th European Conference on Artificial Intelligence (ECAI 2025)*.
[9] Al Omari, B., et al. (2025). Multi-Agent Craftax: Benchmarking Open-Ended MARL [Electronic resource] // *arXiv preprint arXiv:2511.04904*. Mode of access: https://arxiv.org/abs/2511.04904.
[10] Yu, C., et al. (2022). The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games // *Advances in Neural Information Processing Systems (NeurIPS 2022)*.
[11] Rashid, T., et al. (2018). QMIX: Monotonic Value Function Factorisation // *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*.
[12] Papoudakis, G., et al. (2021). Benchmarking Multi-Agent Deep Reinforcement Learning Algorithms // *Advances in Neural Information Processing Systems (NeurIPS 2021)*.
[13] Kim, W., et al. (2025). Agent Order of Action Decisions-MAT (AOAD-MAT) [Electronic resource] // *arXiv preprint*. Mode of access: https://arxiv.org.
[14] Wen, M., et al. (2022). Multi-Agent Transformer // *Advances in Neural Information Processing Systems (NeurIPS 2022)*.
[15] Zhong, Y., et al. (2025). Heterogeneity in MARL: Taxonomy and Quantification [Electronic resource] // *arXiv preprint*. Mode of access: https://arxiv.org.
[16] Bettini, M., et al. (2024). BenchMARL: A Benchmark for MARL // *Journal of Machine Learning Research (JMLR)*.
[17] Zhang, W., et al. (2025). Strategic LLM Decoding through Bayesian Games // *ICLR 2025 Workshop*.
[18] Park, C. (2025). MAPoRL: Multi-Agent Post-Co-Training for Collaborative LLMs // *ICML 2025 Workshop*.
[19] Wan, Z., et al. (2025). ReMA: Learning to Meta-think for LLMs with MARL // *Advances in Neural Information Processing Systems (NeurIPS 2025)*.
[20] Lin, Y. C., et al. (2025). Creativity in LLM-based Multi-Agent Systems: A Survey // *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*.
[21] Gorsane, R., et al. (2022). Towards a Standardised Performance Evaluation Protocol for Cooperative MARL // *Advances in Neural Information Processing Systems (NeurIPS 2022)*.
[22] Formanek, C., et al. (2024). Off-the-Grid MARL: A Framework for Dataset Generation [Electronic resource] // *arXiv preprint*. Mode of access: https://arxiv.org.
[23] Patterson, M., et al. (2025). RLiable: Tools for rigorous evaluation [Electronic resource]. GitHub repository: https://github.com/google-research/rliable.
[24] InstaDeep. (2024). MARL-eval: Standardised experiment data aggregation [Electronic resource]. GitHub repository.

Надійшла до редколегії 13.02.2026
Прийнята після рецензування 19.02.2026
Опублікована 26.03.2026