УДК 004.8+004.02+004.9
DOI https://doi.org/10.32782/2663-5941/2025.3.2/15

Бочок В.О.
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»

Федорова Н.В.
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»

ОБМІН ЗНАННЯМИ В INDEPENDENT DQN БАГАТОАГЕНТНІЙ СИСТЕМІ ТА ОСОБЛИВОСТІ ЙОГО РЕАЛІЗАЦІЇ

У статті розглянуто проблему підвищення ефективності навчання агентів у багатоагентних системах шляхом впровадження механізму обміну знаннями. Зазначено, що традиційні підходи, засновані на алгоритмах Deep Q-Network (DQN), переважно орієнтовані на використання власного досвіду агентів, що обмежує швидкість збіжності, знижує якість отриманих стратегій та не дозволяє повною мірою використовувати потенціал системи. Водночас у багатоагентних системах існує значний потенціал покращення результатів завдяки обміну знаннями між агентами, що дозволяє ефективніше використовувати наявні ресурси досвіду.

Запропоновано механізм обміну знаннями за моделлю «вчитель–учень», що базується на дистиляції політики. Агент із вищими результатами визначається як вчитель і передає знання агенту-учню з урахуванням різниці в сумарно отриманих за епізод винагородах. Передбачено регулювання інтенсивності навчання на основі співвідношення нагород, що дозволяє зберігати індивідуальність агентів, поступово адаптувати їхні стратегії, а також застосовувати більш інтенсивне навчання для агентів зі значною різницею в отриманих винагородах, яка використовується як об'єктивний показник якості стратегії.

Особливу увагу приділено реалізації асинхронного обміну знаннями без блокування агентів. Запропонований алгоритм дає змогу агентам самостійно накопичувати досвід, публікувати повідомлення з узагальненою інформацією про політики та обирати відповідного вчителя без необхідності синхронізації процесів навчання або зупинки виконання інших агентів. Це забезпечує безперервність навчання, покращує гнучкість системи та підвищує ефективність її роботи в розподілених обчислювальних середовищах.

Експериментальні дослідження, проведені в середовищі OpenAI Gym CartPole-v1 з використанням набору DQN-агентів, підтвердили ефективність запропонованого підходу. Встановлено, що обмін знаннями за моделлю «вчитель–учень» сприяє прискоренню збіжності навчання, підвищенню стабільності стратегій агентів, покращенню загальної продуктивності системи та зменшенню часу досягнення заданих цільових показників якості.

Ключові слова: *Deep Q-Network (DQN), Обмін знаннями, Асинхронна комунікація, Багатоагентні системи, Аналітика даних, Навчання з підкріпленням.*

Постановка проблеми. Багатоагентні системи (Multi-Agent Systems, MAS) є одним із ключових напрямів сучасної інформатики та штучного інтелекту, що знаходять все ширше застосування – від комп'ютерних ігор до розподілених систем управління і робототехніки. Особливістю MAS є взаємодія декількох інтелектуальних агентів, які спільно вирішують завдання, часто недоступні для монолітних підходів [1]. Така парадигма забезпечує низку переваг: декларативність постановки цілей, автономність кожного агента, стійкість системи до відмов окремих компонентів та відсутність централізованого керування [1; 2]. Завдяки децентралізованому характеру MAS є гнучкими та масштабованими: нові агенти або функції можуть додаватися без суттєвої переробки всієї системи, а вихід з ладу окремого агента не призводить до відмови всієї системи [2]. Таким чином, багатоагентний підхід дозволяє створювати розподілені і стійкі до збоїв рішення для складних проблем.

Одним із методів, що надає агентам можливість ефективно навчатися поведінки на основі взаємодії з середовищем, є навчання з підкрі-

пленням (reinforcement learning, RL). У рамках RL агент вивчає оптимальну стратегію дій шляхом багаторазового проб і помилок із отриманням зворотного зв'язку у вигляді винагород чи штрафів за дії [3]. Цей підхід особливо корисний для автономних агентів, оскільки не потребує нагляду або попередньо розмічених даних: агент самостійно вдосконалює свою політику, поступово максимізуючи сумарну винагороду [3]. Завдяки навчанню з підкріпленням інтелектуальні агенти здатні адаптуватися до динаміки середовища та опановувати навички, не закладені явно в їх програмі, що робить цей підхід фундаментальним для розробки багатоагентних систем.

Для розв'язання задач навчання агентів у середовищах з великими або неперервними просторами станів одним із класичних підходів став алгоритм Deep Q-Network (DQN). DQN поєднав метод Q-навчання з глибинними нейронними мережами, що дозволило агентам досягти рівня людини в комп'ютерних іграх, навчаючись безпосередньо за сирими піксельними спостереженнями (наприклад, у середовищі Atari) [4]. Втім, базовий алгоритм DQN має низку обмежень, виявлених подальшими дослідженнями. По-перше, навчання DQN може бути нестабільним: оцінки функції цінності здатні коливатися або навіть дивергувати за неналежних параметрів, тому для стабілізації застосовуються спеціальні прийоми (наприклад, фіксована цільова мережа) [5]. По-друге, збіжність DQN зазвичай вимагає дуже великої кількості взаємодій з середовищем, тобто алгоритм навчається відносно повільно [6]. Частково це зумовлено тим, що DQN неефективно використовує отриманий досвід: кожна взаємодія зі середовищем робить лише обмежений внесок у навчання, тому агенту потрібні численні повторні епізоди для наближення до оптимальної стратегії [6]. У результаті, незважаючи на успіх DQN, його класична версія вважається алгоритмом із низькою стабільністю, повільною збіжністю та недостатнім використанням досвіду [5; 6]. Зазначені проблеми стимулювали появу низки удосконалень, зокрема методів зменшення переоцінювання Q-значень дій (Double DQN) [5] та пріоритетного повторного відтворення досвіду (Prioritized Experience Replay) [6].

Окрім технічних недоліків алгоритму, у середовищах, що змінюються з часом або містять кілька активних агентів, постає проблема адаптації: агентам необхідно безперервно навчатися протягом всього циклу їх функціонування. Система може змінюватися з часом, а отже фіксованої моделі, навченої один раз, може виявитися недостатньо – потрібна здатність алгоритму навчання підлаштовуватися до нових умов і продовжувати навчання без повторного перезапуску. Безперервне навчання агентів дозволяє підтримувати ефективність системи у довгостроковій перспективі та враховувати появу нових ситуацій, відсутніх на етапі початкового тренування. Таким чином, розробка методів, які забезпечують тривале поступове навчання (continual learning) агентів у динамічних багатоагентних середовищах, є актуальним науковим завданням [7].

Більшість існуючих підходів до покращення ефективності DQN-агентів зосереджуються на вдосконаленні процесу навчання на основі власного досвіду. Проте в умовах багатоагентних систем, де одночасно функціонує кілька незалежних агентів з різними траєкторіями навчання, обмін досвідом між агентами може стати додатковим джерелом інформації, що сприяє прискоренню збіжності та покращенню якості навчання. Такий підхід не заперечує наявні модифікації DQN, а доповнює їх, дозволяючи ефективніше використовувати наявні дані в системі.

Особливої уваги потребує реалізація механізмів обміну знаннями у системах, де агенти працюють незалежно та асинхронно. У подібних умовах важливо уникати блокування чи затримок, спричинених очікуванням синхронізації з іншими агентами. Відтак, виникає науково-практичне завдання: розробити ефективний механізм обміну знаннями між агентами, який враховує гетерогенність їх досвіду та забезпечує асинхронність навчального процесу.

Аналіз останніх досліджень і публікацій. Автори [8] розглядали сценарій, у якому агент, натренований на виконання окремої дії, передає свої знання іншому агентові, який має вирішувати багатозадачні проблеми. Передача здійснювалася за допомогою дистиляції політики – методу, який дозволяє переносити стратегію одного агента до іншого через навчання за м'якими мітками. Результати експериментів продемонстрували ефективність такого підходу навіть у випадках, коли агенти не були гетерогенними, що свідчить про широку застосовність методів дистиляції для обміну знаннями у багатоагентних середовищах. Крім того, у роботі коротко обговорено можливість організації процесу «онлайн» дистиляції, тобто безперервної передачі знань під час навчання, однак без поглибленого опрацювання цього питання [8].

Ще одним прикладом сучасного підходу до обміну знаннями є метод KnowSR, запропонований у [9]. У цій роботі було розроблено механізм кооперації однорідних агентів у рамках багатоагентного навчання з підкріпленням. KnowSR базується на концепції дистиляції знань, що дозволяє агентам не обмежуватися лише власним досвідом, а також інтегрувати інформацію, отриману іншими агентами. Це забезпечує скорочення часу, необхідного для досягнення прийнятної політики, та підвищує ефективність навчання загалом. З технічної точки зору, метод був реалізований на базі алгоритму MADDPG (Multi-Agent Deep Deterministic Policy Gradient), який передбачає централізоване навчання із використанням критика, що має доступ до глобальної інформації, та децентралізоване виконання дій кожним агентом окремо. Така архітектура дозволяє ефективно поєднати індивідуальне та колективне навчання, що є важливим у контексті складних багатоагентних середовищ [9].

Таким чином, результати останніх досліджень підтверджують перспективність підходів до обміну знаннями для прискорення та покращення процесів навчання в багатоагентних системах. Надалі важливим завданням є подальший розвиток методів безперервного обміну знаннями в динамічних та нестаціонарних середовищах.

Постановка завдання. Середовище, в якому функціонує багато агентів, часто є нестаціонарним, що значно ускладнює процес навчання та призводить до зниження ефективності класичних DQN-агентів. Це зумовлено тим, що алгоритм DQN спочатку розроблявся для розв'язання завдань в умовах стаціонарних середовищ і не враховує впливу дій інших агентів, які змінюють динаміку середовища та, відповідно, впливають на формування винагород. Проте існують середовища, які залишаються стаціонарними або можуть розглядатися як такі за певних припущень, навіть за наявності декількох агентів. Прикладом таких систем є фінансові торгові середовища, у яких дії окремого агента мають незначний вплив на загальний стан середовища і, в більшості випадків, не змінюють поведінку інших агентів.

У зв'язку з цим, завданням даного дослідження є розробка механізму обміну знаннями між агентами у стаціонарному середовищі, аналіз його впливу на ефективність системи, а також вивчення особливостей програмної реалізації такого підходу в умовах безперервного навчання агентів.

Виклад основного матеріалу. Було проведено серію експериментів з дослідження підходу до обміну знаннями між агентами. Експеримент проводився у класичному середовищі OpenAI Gym CartPole-v1 [10]. Завдання полягає в тому, щоб утримувати жердину (Pole) у вертикальному положенні на рухомій платформі (Cart), балансуючи нею шляхом застосування сил вліво або вправо. Агент отримував нагороду +1 за кожну ітерацію, доки жердина залишалася у межах допустимого кута нахилу (±24 градусів) та діапазону положення (± 4.8 одиниці). Додатково було обмежено кількість кроків у 500 максимум. Для представлення Q-функції використовувалася глибока нейромережа з двома прихованими шарами (24 нейрони, функція активації ReLU, 12 нейронів, функція активації ReLU). Відповідно, вхідний шар – 4 нейрони (відповідають 4-елементному вектору стану середовища), і вихідний шар – 2 нейрони (Q-значення для кожної з двох можливих дій: рух платформи ліворуч або праворуч). Функція втрат – середньоквадратична похибка (MSE), яка використовується для оновлення оцінок Q-функції. Оптимізація параметрів здійснювалася за допомогою алгоритму Adam (Adaptive Moment Estimation) із коефіцієнтом навчання 0.001 [11]. Batch size дорівнює 128. Значення discount factor у рівнянні Беллмана було встановлено на 0.95, що дозволяє агенту враховувати майбутні винагороди при прийнятті рішень [12]. Для розв'язання дилеми дослідження-експлуатації використовувався Adaptive ε-greedy exploration підхід [13], де параметр ε зменшувався з 1 (повністю випадкові дії агента) на 0.5% з кожним епізодом. Для покращення стабільності навчання застосовувався непріорітезований буфер досвіду, що зберігав до 10 тисяч останніх переходів (стан, дія, винагорода, наступний стан, завершення епізоду). У кожній імітації навчалися 3 агенти, що діяли у власних середовищах та обмінювалися досвідом за одним з запропонованих нижче методів.

Обмін знаннями між вчителем і учнем. Запропонований підхід умовно названий «Обмін знаннями між вчителем і учнем» (рис. 1), базується на ідеї дистиляції політики. У межах цього підходу агент, обраний системою як вчитель, передає свої знання агенту-учню, який сприймає їх як еталон.

Основною ідеєю методу є припущення, що наближення розподілу політики учня до політики вчителя позитивно вплине на його показники. При цьому повне узгодження розподілів

не є метою; передбачається часткова дистиляція, яка зберігає індивідуальні відмінності агентів та дозволяє змінювати домінуючу стратегію у різних агентів у процесі навчання.

Регулювання інтенсивності навчання здійснювалося через кількість епох оптимізації. Базове значення становило 3 епохи. Залежно від співвідношення сумарної винагороди вчителя та учня, кількість епох коригувалася шляхом множення на відношення їх нагород, однак не перевищувала 5 епох. Таким чином, якщо агент-учень демонстрував суттєво гірші результати, сила його узгодження із вчителем збільшувалася. Для агентів із близькими показниками вплив навчання залишався мінімальним. Варто зауважити, що обрані параметри є типовими для проведеного експерименту, однак можуть змінюватися залежно від характеристик середовища та архітектури нейронних мереж.

Вибір вчителя здійснювався на основі порівняння сумарних нагород, отриманих агентами за попередній епізод: агент із вищою нагородою вважався краще адаптованим до середовища та визначався вчителем.

Для перевірки ефективності методу «Учень–вчитель» було створено спрощену експериментальну систему, у якій після завершення епізоду кожен агент очікував завершення епізодів усіх інших агентів. Після цього визначалися ролі вчителя та учня, а також сила навчального впливу. За результатами експериментів встановлено, що запропонований підхід сприяє прискоренню процесу навчання агентів (рис. 2).

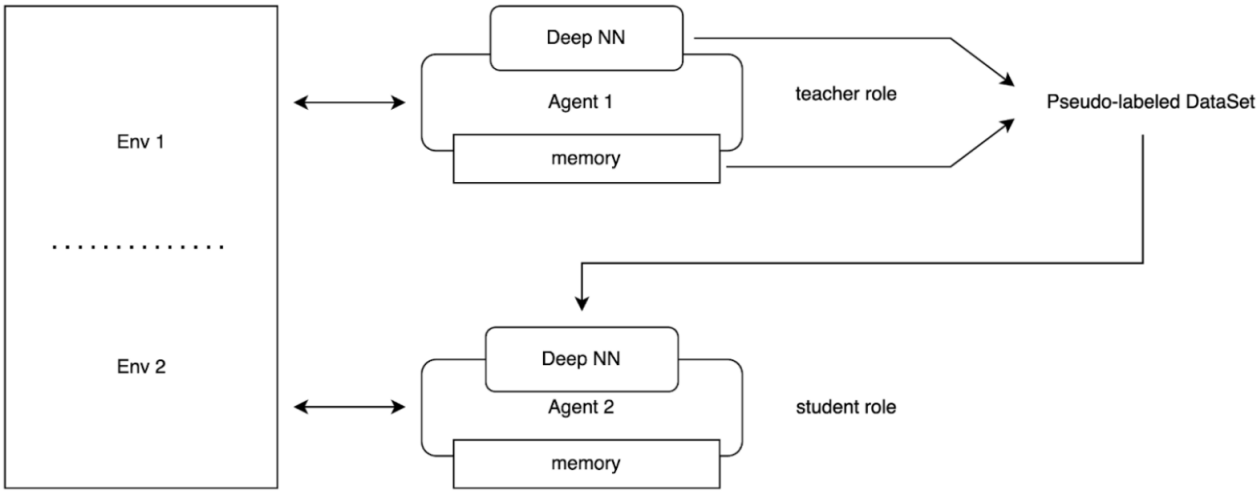

Рис. 1. Схема обміну знаннями між учнем і вчителем

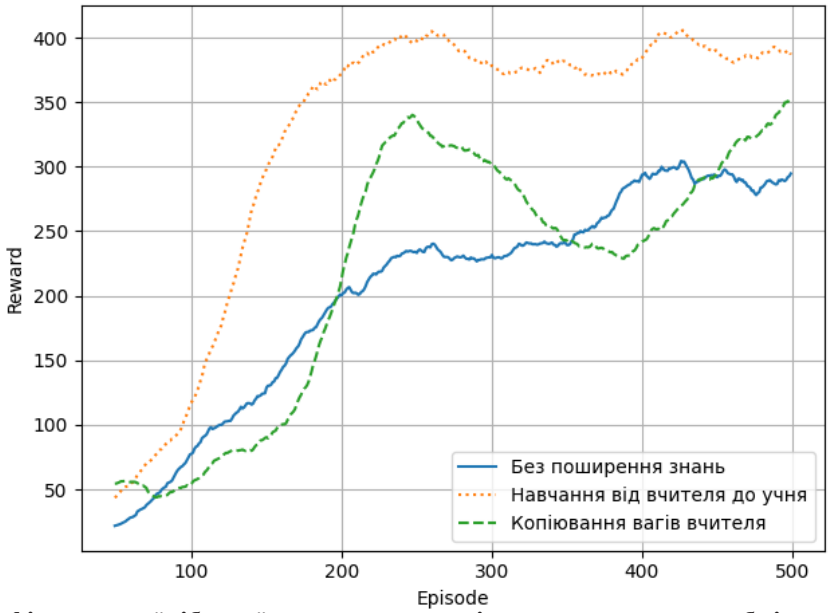

Рис. 2. Графік сумарної зібраної винагороди за епізод з застосуванням обміну знаннями між учнем та вчителем (графік зглажений скользячим середнім з вікном у 50)

Важливо зазначити, що запропонований метод використовує показник зібраної винагороди для визначення більш впливових даних, а отже навчання на власному досвіді має відбуватися після поширення знань, для збереження оціненої моделі. Також, досліди показують, що навчання на власному досвіді має бути так само присутнім, як і обмін знаннями. Більше того, навчання на власному досвіді має бути більше, ніж поширення знань. Інакше, це приводить до сповільненого росту показників, нестабільності або ж деградації.

Особливості неблокуючого обміну знаннями. У реальних багатоагентних системах тривалість епізодів для різних агентів може істотно відрізнятися. Це створює проблему синхронізації: примусове очікування завершення епізодів іншими агентами може призвести до блокування роботи системи, що є неприйнятним в умовах реального часу або розподілених обчислень. Відтак виникає потреба в реалізації алгоритму обміну знаннями, що не вимагає синхронної взаємодії агентів і забезпечує безперервність їх роботи.

Запропонований асинхронний механізм обміну знаннями має наступний алгоритм функціонування:

1. Накопичення досвіду. Під час взаємодії з середовищем агент накопичує дані у власному буфері досвіду. Кожен запис містить інформацію про стан, обрану дію, отриману нагороду, наступний стан та ознаку завершення епізоду.

2. Обробка після завершення епізоду. Після завершення власного епізоду агент виконує наступні дії:

   1. Обчислює сумарну нагороду, отриману за весь епізод.
   2. Випадковим чином обирає певну кількість записів зі свого буфера та обчислює для них передбачення Q-функції.
   3. Формує повідомлення, яке містить обрану вибірку, передбачені Q-значення та сумарну нагороду, і публікує його для інших агентів.

3. Отримання знань від інших агентів. Агент отримує повідомлення, надіслані іншими агентами. Якщо серед цих повідомлень є такі, де сумарна нагорода перевищує власну, агент визначає відповідного агента як вчителя.

4. Навчання. Після визначення вчителя агент додатково навчається на псевдо-розміченому наборі даних, наданому вчителем. Окрім цього, продовжується навчання на власному досвіді, накопиченому у буфері.

Такий підхід (рис. 3) дозволяє організувати процес обміну знаннями без синхронізації агентів, мінімізуючи затримки у навчанні та забезпечуючи безперервність роботи багатоагентної системи.

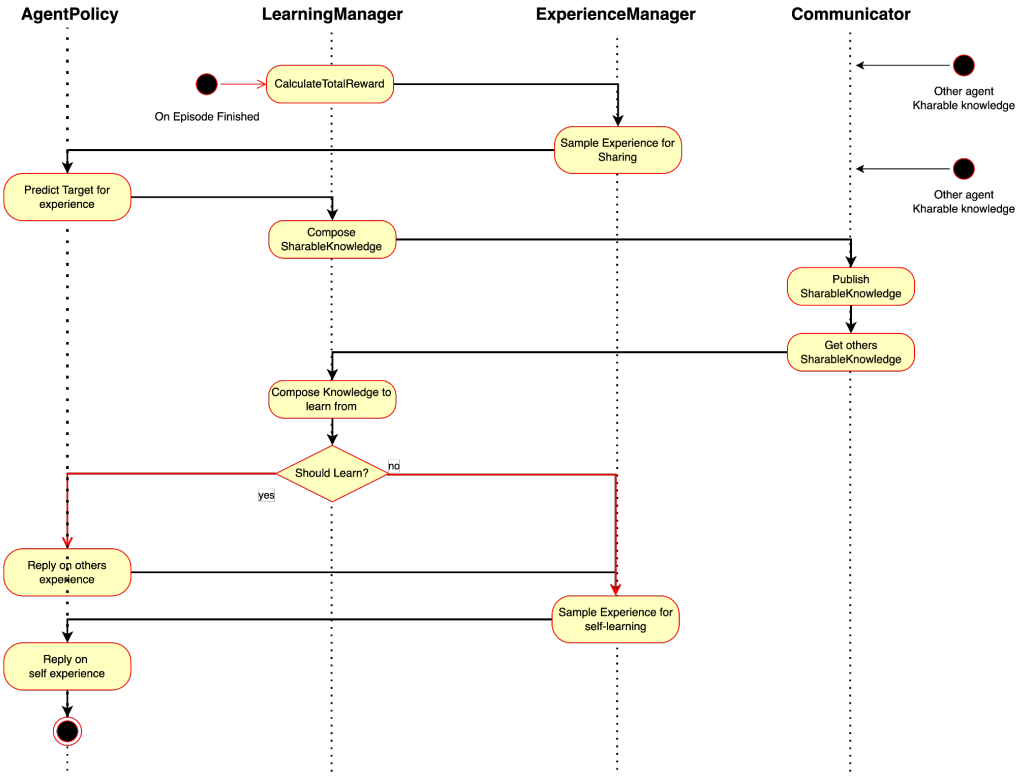

Рис. 3. Схема асинхронного обміну знаннями за схемою Учень-вчитель

Конкретна реалізація механізму обміну повідомленнями залежить від архітектури багатоагентної системи та організації середовища. Існує кілька можливих підходів до побудови комунікатора:

• Комунікатор на рівні середовища. У цьому варіанті роль комунікаційного вузла виконує саме середовище. Воно приймає повідомлення від агентів, зберігає їх і забезпечує доступ для читання іншими агентами. Такий підхід спрощує логіку кожного окремого агента, оскільки обмін знаннями відбувається через централізовану проміжну ланку.

• Спеціалізований агент-комунікатор. Альтернативним варіантом є виділення окремого агента, який відповідає за прийом, зберігання та розповсюдження повідомлень. Інші агенти передають йому записи для публікації та отримують повідомлення у відповідь. Це дозволяє відокремити функції комунікації від основної логіки агентів і забезпечити масштабованість системи.

• Прямий обмін повідомленнями між агентами. У цьому випадку агенти зв›язуються попарно напряму без використання центрального вузла. Кожен агент самостійно приймає, зберігає та оновлює актуальні версії отриманих повідомлень. Під час публікації даних агент самостійно знаходить інших доступних агентів і передає їм свої повідомлення. Можлива також організація передачі за принципом «ланцюга», коли агент передає дані іншому, який у свою чергу ретранслює їх далі.

У всіх описаних варіантах ключовою особливістю є відсутність блокування агентів для очікування синхронізації. Кожен агент продовжує навчання незалежно від інших, що забезпечує асинхронність роботи системи та високу ефективність у реальному часі.

Висновки. У цій роботі було досліджено підходи до підвищення ефективності навчання в багатоагентних системах шляхом обміну знаннями між агентами. Встановлено, що поряд із вдосконаленням індивідуальних алгоритмів навчання (наприклад, модифікацій DQN), впровадження механізмів обміну знаннями є перспективним напрямом для прискорення збіжності та підвищення стабільності навчального процесу.

На основі аналізу існуючих підходів, зокрема методів дистиляції політики та обміну досвідом (Policy Distillation, KnowSR), запропоновано дві модифікації організації навчання:

• Модель «вчитель–учень», у якій обраний агент-вчитель передає знання агенту-учню з контролем сили навчання залежно від відставання в сумарній винагороді. Проведені експерименти продемонстрували, що навіть часткове узгодження розподілів політик сприяє прискоренню навчання.

• Асинхронний механізм обміну повідомленнями, який дозволяє агентам обмінюватися досвідом без необхідності синхронізації завершення епізодів. Це забезпечує безперервність роботи системи та робить підхід придатним для застосування в реальних розподілених середовищах.

Розглянуто також можливі варіанти реалізації системи обміну знаннями: через середовище, через окремого агента-комунікатора та за допомогою прямого обміну повідомленнями між агентами. Кожен варіант має свої переваги та обмеження залежно від архітектури середовища та вимог до масштабованості.

Список літератури:

1. Wooldridge M. An Introduction to MultiAgent Systems. John Wiley & Sons, 2002. 366 с.
2. Tichý P., Staron R.J. Multi-agent technology for fault tolerant and flexible control. In: D. Srinivasan, L.C. Jain (eds.). Innovations in Multi-Agent Systems and Applications – 1. Studies in Computational Intelligence. Vol. 310. Berlin, Heidelberg: Springer, 2010. С. 223–246.
3. Sutton R.S., Barto A.G. Reinforcement Learning: An Introduction. 2nd ed. Cambridge, MA: MIT Press, 2018. 552 с.
4. Mnih V., Kavukcuoglu K., Silver D. et al. Human-level control through deep reinforcement learning. Nature. 2015. Vol. 518. No. 7540. С. 529–533.
5. Van Hasselt H., Guez A., Silver D. Deep reinforcement learning with double Q-learning. Proc. of the 30th AAAI Conf. on Artificial Intelligence. 2016. С. 2094–2100.
6. Schaul T., Quan J., Antonoglou I., Silver D. Prioritized experience replay. Proc. of 4th Int. Conf. on Learning Representations (ICLR). 2016. 9 с.
7. Hernandez-Leal P., Kaisers M., Baarslag T., Muñoz de Cote E. A survey of learning in multiagent environments: dealing with non-stationarity. arXiv preprint arXiv:1707.09183. 2019. 64 с. URL: https://arxiv.org/abs/1707.09183.

8. Rusu A.A., Colmenarejo S.G., Gulcehre C. et al. Policy Distillation [Електронний ресурс]. arXiv preprint arXiv:1511.06295. 2015. URL: https://doi.org/10.48550/arxiv.1511.06295.

9. Gao Z., Xu K., Ding B. et al. KnowSR: Knowledge Sharing Among Homogeneous Agents in Multi-agent Reinforcement Learning. arXiv preprint arXiv:2105.11611. 2021. URL: https://doi.org/10.48550/arxiv.2105.11611.

10. Brockman G., Cheung V., Pettersson L. et al. OpenAI Gym. arXiv preprint arXiv:1606.01540. 2016. URL: https://doi.org/10.48550/arxiv.1606.01540.

11. Kingma D.P., Ba J. Adam: A Method for Stochastic Optimization. arXiv preprint arXiv:1412.6980. 2014. URL: https://doi.org/10.48550/arxiv.1412.6980.

12. Sutton R.S., Barto A.G. Reinforcement Learning: An Introduction. MIT Press, 2018. 552 c.

13. Tokic M. Adaptive ε-greedy exploration in reinforcement learning based on value differences. Annual Conf. on Artificial Intelligence. 2010.

Bochok V.O., Fedorova N.V. KNOWLEDGE SHARING IN INDEPENDENT DQN MULTI-AGENT SYSTEM AND FEATURES OF ITS IMPLEMENTATION

This paper addresses the problem of improving agent learning efficiency in multi-agent systems by introducing a knowledge-sharing mechanism. It is noted that traditional approaches based on Deep Q-Network (DQN) algorithms mainly focus on utilizing each agent's own experience, which limits the convergence rate, reduces the quality of the obtained strategies, and prevents the system from fully realizing its potential. At the same time, multi-agent systems have significant potential for performance improvement through knowledge sharing among agents, enabling more efficient utilization of available experience resources.

A knowledge-sharing mechanism based on the "teacher–student" model is proposed, which utilizes policy distillation techniques. The agent with the highest performance is designated as the teacher and shares its knowledge with the student agent, considering the difference in their cumulative episode rewards. The learning intensity is regulated based on the ratio of obtained rewards, allowing agents to preserve individuality, gradually adapt their strategies, and apply more intensive learning for agents with significant differences in performance, using the reward as an objective quality indicator.

Special attention is given to the implementation of asynchronous knowledge sharing without blocking agents. The proposed algorithm enables agents to independently accumulate experience, publish messages with generalized policy information, and select an appropriate teacher without the need for process synchronization or halting other agents' activities. This ensures continuous learning, improves system flexibility, and enhances the overall efficiency of operations in distributed computational environments.

Experimental studies conducted in the OpenAI Gym CartPole-v1 environment using a set of DQN agents confirmed the effectiveness of the proposed approach. The results demonstrate that knowledge sharing based on the "teacher–student" model accelerates convergence, improves policy stability, increases overall system performance, and reduces the time required to achieve target quality metrics.

Key words: *Deep Q-Network (DQN), Knowledge Sharing, Asynchronous Communication, Multi-Agent Systems, Data Analytics, Reinforcement Learning.*