УДК 004.8:004.896 DOI: 10.31673/2412-4338.2024.011625

Ганенко Людмила Дмитрівна
Центральноукраїнський державний університет ім. В. Винниченка, Кропивницький
ORCID 0000-0003-2219-8196

Жебка Вікторія Вікторівна
Державний університет інформаційно-комунікаційних технологій, Київ
ORCID 0000-0003-4051-1190

ЗАСТОСУВАННЯ МЕТОДІВ НАВЧАННЯ З ПІДКРІПЛЕННЯМ ДЛЯ ПЛАНУВАННЯ ШЛЯХУ МОБІЛЬНИХ РОБОТІВ

Анотація. *Розвиток та впровадження автономних мобільних роботів у різноманітні сфери людського життя стало актуальним завданням сьогодення. Навчання з підкріпленням (Reinforcement Learning — RL) є потужним інструментом для оптимізації навчання та прийняття рішень агентами в реальних умовах. Використання RL стає ключовим аспектом для досягнення ефективності та надійності робототехнічних систем. Навчання з підкріпленням може застосовують у плануванні шляху мобільного робота в складних та динамічних середовищах, для навчання мобільного робота приймати рішення щодо вибору напрямку руху, швидкості та здійснення маневрів на основі показників датчиків, для прийняття рішень щодо ефективного використання енергетичних ресурсів та максимізації часу роботи. Агент може навчатися оптимальним маршрутам, уникати перешкоди та ефективно досягати своїх цілей. В статті розглянуто застосування методів навчання з підкріпленням для оптимізації планування шляху мобільних роботів. Подано класифікацію методів на основі моделі середовища та методів без моделі середовища. Розглянуто методи на основі цінності, на основі політики та методи актора-критика. Зокрема проведено аналіз таких методів навчання з підкріпленням, як Q-learning, Deep Q-Networks (DQN), Double Deep Q-Network (DDQN), алгоритмів актора-критика Advantage Actor-Critic (A2C), Deep Deterministic Policy Gradient (DDPG), Asynchronous Advantage Actor-Critic (A3C), Soft actor-critic (SAC)) та Proximal Policy Optimization (PPO). Дані методи проаналізовано в контексті їх застосування до розв'язання завдань планування шляху мобільного робота в різних середовищах. Досліджено переваги та недоліки використання зазначених методів навчання з підкріпленням в плануванні шляху із врахуванням аспектів ефективності, безпеки та адаптивності. Увага приділяється вирішенню проблем підвищення швидкості та стійкості навчання, ефективної навігації у складних та змінних умовах, де традиційні методи можуть бути неефективними. Запропоновано перспективи для майбутніх досліджень та розвитку даного напрямку в наукових роботах.*

Ключові слова: *машинне навчання, методи навчання з підкріпленням, мобільні роботи, планування шляху, інформаційна технологія, інформаційна система, модель, алгоритм.*

Hanenko Liudmyla
Volodymyr Vynnychenko central ukrainian state university, Kropyvnytskyi
ORCID 0000-0003-2219-8196

Zhebka Viktoriia
State university of information and communication technologies, Kyiv
ORCID 0000-0003-4051-1190

APPLICATION OF REINFORCEMENT LEARNING METHODS FOR PATH PLANNING OF MOBILE ROBOTS

Abstract. *The development and implementation of autonomous mobile robots in various spheres of human life has become an urgent task today. Reinforcement Learning (RL) is a powerful tool for optimizing learning and decision-making by agents in real-world conditions. The use of RL is becoming a key aspect for achieving efficiency and reliability of robotic systems. Reinforcement learning can be used to plan the path of*

a mobile robot in complex and dynamic environments, to teach a mobile robot to make decisions about direction, speed, and maneuvers based on its sensors, to make decisions about the efficient use of energy resources and maximize operating time. The agent can learn optimal routes, avoid obstacles, and effectively achieve its goals. The article discusses the use of reinforcement learning methods to optimize the path planning of mobile robots. A classification of methods based on the environment model and methods without the environment model is presented. Value-based methods, policy-based methods, and actor-critic methods are considered. In particular, the analysis of such reinforcement learning methods as Q-learning, Deep Q-Networks (DQN), Double Deep Q-Network (DDQN), actor-critic algorithms Advantage Actor-Critic (A2C), Deep Deterministic Policy Gradient (DDPG), Asynchronous Advantage Actor-Critic (A3C), Soft actor-critic (SAC) and Proximal Policy Optimization (PPO) is carried out.) These methods were analyzed in the context of their application to solving the problems of planning the path of a mobile robot in different environments. The advantages and disadvantages of using these reinforcement learning methods in path planning are investigated, taking into account the aspects of efficiency, safety, and adaptability. Particular attention is paid to solving the problems of increasing the speed and sustainability of learning, effective navigation in complex and changing conditions where traditional methods may be ineffective. Prospects for future research and development of this area in scientific works are proposed.

Keywords: *machine learning, reinforcement learning methods, mobile robots, path planning, information technology, information system, model, algorithm.*

1. Вступ.

В сучасному світі робототехніки мобільні роботи відіграють важливу роль у багатьох сферах людського життя, від промисловості до медицини. Ефективне планування шляху для мобільних роботів визначає їхню здатність автономно рухатись в різноманітних середовищах. Ефективність традиційних методів планування шляху знижується у складних середовищах. В цьому контексті методи навчання з підкріпленням (Reinforcement Learning — RL) стають перспективним напрямком у розвитку автономної навігації мобільних роботів [1].

Підходи RL базуються на принципах навчання агента через взаємодію з навколишнім середовищем та отриманням винагороди за правильні дії. У порівнянні з традиційними методами, навчання з підкріпленням дозволяє створювати агентів, які здатні адаптуватися до змін у середовищі та вирішувати складні завдання, такі як планування шляху в умовах невизначеності та динамічності.

Проведене дослідження дозволяє виокремити нові напрямки розвитку адаптивності та ефективності автономних мобільних роботів в реальних умовах.

2. Аналіз останніх досліджень і публікацій.

В сучасних наукових дослідженнях використання методів навчання з підкріпленням (RL) в контексті вирішення завдань навігації автономних мобільних роботів є актуальним та перспективним. Порівняно з 2018 роком у 2023 році використання RL у робототехніці зросло у 8 разів [2].

Протягом останніх років спостерігається тенденція до зростання кількості робіт, у яких застосовується глибоке навчання з підкріпленням (DRL) для планування траєкторії руху мобільного робота. У своїх роботах Silver та Mnih використали глибокі нейронні мережі у поєднанні з RL для навчання агентів в складних реальних умовах. Ці дослідження сприяли розвитку нових архітектур та алгоритмів для оптимізації шляхів мобільних роботів [3].

Таі та ін. [4] вперше застосували алгоритм DQN для реалізації планування шляху мобільного робота у віртуальному середовищі [5]. Wang та ін. розробили покращений алгоритм DQN у поєднанні з методами штучного потенційного поля для розробки функцій винагороди, покращуючи ефективність планування шляху мобільного робота [5]. Вдосконалений DQN науковці використовували для навчання патрулюючого робота рухатися по круговій дорозі та уникати перешкод [7], для навчання агента під час руху збирати яблука та уникати лимонів [8].

Застосування RL має такі переваги як відсутність карти, висока здатність до навчання, низька залежність від точності датчиків. Недоліком є довга тривалість процесу навчання.

Незважаючи на значний прогрес у розробці методів навчання з підкріпленням та їхнє використання в робототехніці, розробка нових та вдосконалення існуючих методів залишається актуальним та необхідним.

3. Мета і задачі дослідження.

Метою даного дослідження є аналіз застосування методів навчання з підкріпленням для оптимізації планування шляху мобільних роботів, визначення переваг та недоліків даних методів та виявлення перспективних напрямків майбутніх досліджень. Основна увага приділяється ефективності та адаптивності методів навчання з підкріпленням у різноманітних умовах.

Для досягнення мети поставлено такі завдання:

– проаналізувати алгоритми навчання з підкріпленням та їх застосування в робототехніці;

– визначити переваги і недоліки застосування методів RL у порівнянні з іншими методами планування шляху мобільних роботів та зробити висновки щодо ефективності методів RL у плануванні шляху мобільних роботів;

– запропонувати напрямки для подальших досліджень в цій області.

4. Результати дослідження.

Навчання з підкріпленням є одним із напрямків машинного навчання (ML). На відміну від інших парадигм ML, таких як контрольоване та неконтрольоване навчання, RL реалізується методом спроб та помилок, взаємодіючи із середовищем для максимізації кумулятивної винагороди.

В робототехніці навчання з підкріпленням застосовується в картографуванні середовища, в дослідженні та плануванні шляху мобільних роботів у невідомому середовищі, у відстеженні та взаємодії з перешкодами, у прийнятті рішення щодо напрямку руху, швидкості та маневрів, у прийнятті рішень щодо ефективного використання енергоресурсів [1].

Основними компонентами RL є агент, середовище, стан, дія, нагорода, політика. Агент взаємодіє з оточенням і приймає рішення з метою отримання максимальної винагороди. Середовище надає агенту зворотний зв’язок у вигляді нагород чи штрафів. Станами (s_t) є конфігурації середовища, в якому може перебувати агент. Стани можуть бути відомими або невідомими. Дія (a_t) — це можливий вибір, які агент може здійснити в кожному стані. Нагородою (r_t) є числова величина, яку агент отримує від середовища відповідно до виконаних дій. Політика (π) визначає поведінку агента в середовищі. Оптимальна політика — це політика, яка приносить агенту хорошу винагороду і допомагає агенту досягти мети.

Рис. 1. Модель навчання з підкріпленням

Важливим аспектом RL є баланс між дослідженням та експлуатацією: агенту потрібно ефективно взаємодіяти із середовищем для вивчення нових стратегій і водночас застосовувати вже отримані знання для максимізації винагороди.

Модель навчання з підкріпленням продемонстровано на рис. 1.

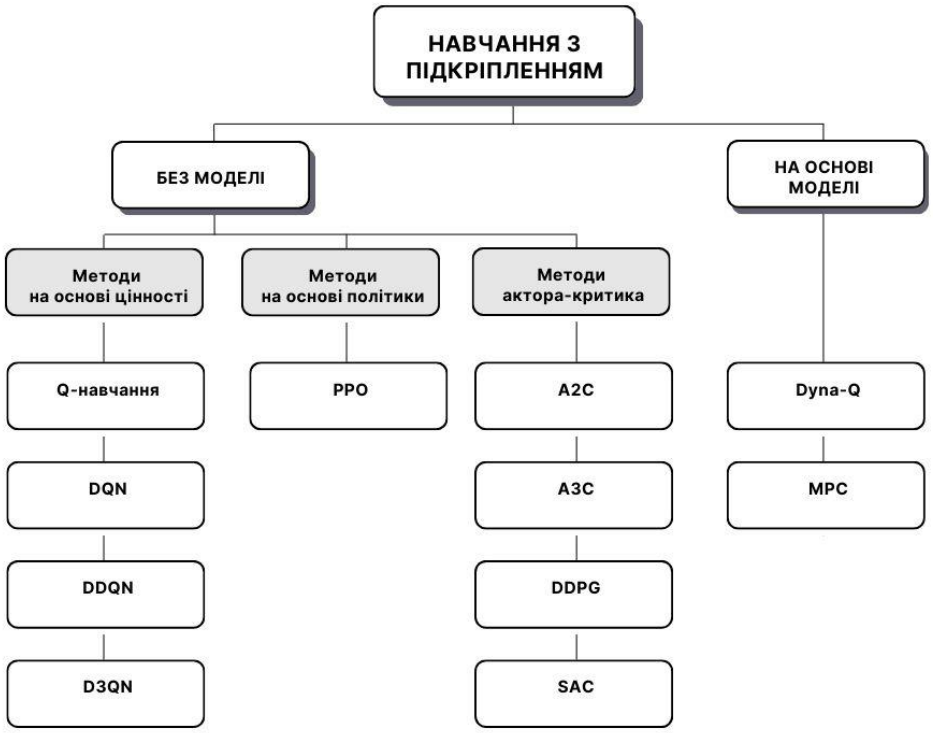

Рис 2. Класифікація методів навчання з підкріпленням

Навчання з підкріпленням може бути або на основі моделі середовища, або без моделі середовища. Алгоритм на основі моделі повинен вивчити (або отримати) усі ймовірності переходу, які описують перехід агента з одного стану в інший. Агенти, які навчаються на моделях, швидко вивчають політику, оскільки наступний стан уже відомий, і майбутню винагороду можна легко передбачити. З іншого боку, вивчення моделі в рази ускладнюється для великих просторів станів. В таких випадках ефективним стає застосування безмодельних методів [3]. У даній роботі досліджуються безмодельні методи навчання з підкріпленням.

Серед безмодельних методів виділяють методи на основі цінності, на основі політики та методи актора-критика. Класифікацію методів RL продемонстровано на рис. 2.

4.1. Методи на основі цінності.

Q-learning є базовим методом навчання з підкріпленням, який застосовують у завданнях із дискретними просторами дій та станів. Агент, який є автономним мобільним роботом (АМР), виконує дію в середовищі та отримує негайну винагороду або штраф за виконану дію. Основна перевага використання Q-learning в АМР — можливість взаємодії з неструктурованим середовищем шляхом самонавчання та реалізація навігації до цільової позиції без зіткнень [9].

Мобільний робот вибирає дію на основі відповідної політики та виконує цю дію. Натомість середовище навігації передає стан (s) та нагороду (r) роботу [10].

У процесі навчання Q-значення кожної пари стан-дія $Q(s,a)$ зберігається та оновлюється у таблиці. Q-значення представляє цінність виконання дії a, коли мобільний робот перебуває в стані s. Q-learning безпосередньо наближається до оптимальної функції дії-цінності, незалежно від поточної політики. Правило оновлення Q- learning обраховується за формулою:

$$Q(s,a) = Q(s,a) + \alpha(r + \gamma \max_{a'} Q(s',a') - Q(s,a))$$

де *r* — винагорода, отримана в епоху t, γ — коефіцієнт дисконтування $0<\gamma<1$, а α —швидкість навчання. Нехай $Q^*(s,a)$ — оптимальне очікуване Q-значення пари стан-дія (s,a). Якщо кожне рішення виконується в кожному стані нескінченно, *Q(s,a)* буде збігатися до *Q*(s,a).*

Алгоритм Q-learning:

1. *Ініціалізація Q-функції Q(s, a) випадковими значеннями.*
2. *Для кожного епізоду:*
    1. *Ініціалізація стану s*
    2. *Для кожного кроку в епізоді:*
        1. *Отримати політику з Q(s, a) і вибрати дію a для виконання в стані s.*
        2. *Виконати дію a, перейти до наступного стану s' і спостерігати за винагородою r*
        3. *Оновити значення Q*

$$Q(s,a) = Q(s,a) + \alpha(r + \gamma \max_{a'} Q(s',a') - Q(s,a)).$$

        4. *Оновити s' (оновити наступний стан s' до поточного стану s)*
        5. *Якщо s не є термінальним станом, повторити кроки з 1 по 5.*

Q-learning має два основні недоліки. По-перше, Q-таблиці значно збільшуються в розмірі під час обробки великої кількості станів і дій, вимагаючи значного часу та ресурсів для пошуку та зберігання, а також схильні до прокляття розмірності. По-друге, оскільки Q-значення збігається лише після того, як кожен стан було відвідано кілька разів, випадкова політика дослідження Q-learning призведе до надмірно повільності збіжності [11].

Незважаючи на те, що Q-learning ефективно використовується в плануванні шляху АМР та уникненні ним перешкод, воно має обмеження. При збільшенні розміру навчального середовища потрібен довший обчислювальний час для оновлення матриць Q-значення та адаптивних матриць пам'яті. Також існує ймовірність невключення правильних ймовірностей у конвергентні матриці пам'яті [8].

Deep Q-Network (DQN). Все більш поширеним стає використання можливостей глибоких нейронних мереж у RL. Deep Q Network (DQN) оцінює Q-функцію $Q(s,a,\theta)$, яка визначає, які дії є оптимальними в конкретних станах та використовує глибоку нейромережу для апроксимації Q-функції. Для обчислення цільового Q-значення використовується цільова Q-функція $Q_{target}(s',a';\theta)$. Функція витрат обчислюється за формулою :

$$L(\theta) = E\left[\left(r + \gamma max_{a'} Q_{target}(s',a';\theta) - Q(s,a;\theta)\right)^2\right],$$

де *E* - математичне сподівання по всіх можливих парах стан-дія, θ − параметр цільової мережі, r – нагорода, γ — коефіцієнт зниження, який використовується для зменшення вагомості майбутніх нагород при обчисленні цільових Q-значень $0<\gamma<1$.

Оновлення параметрів мережі здійснюється за допомогою градієнтного спуску:

$$\theta \leftarrow \theta - \alpha \nabla_\theta L(\theta).$$

Для вибору дій мобільного робота DQN використовує ε-жадібний алгоритм, який вводить елемент випадковості та дозволяє досліджувати нові дії

$$a = argmax_a Q(s,a;\theta).$$

В процесі планування руху робота на основі алгоритму DQN можна виокремити такі недоліки: тенденція до переоціннення значення Q наступної пари стан-дія в цілі та висока кореляція вибірки. Для вирішення проблеми переоціннення значення Q було запропоновано подвійний алгоритм DQN.

Double Deep Q-Network (DDQN). Основна концепція DDQN полягає у відокремленні процесів вибору та оцінки дій. Алгоритм DDQN створює дві копії мережі, а саме мережу прогнозування та цільову мережу. Мережа прогнозування використовує поточний стан і дію, щоб передбачити значення Q з параметром мережі θ, тоді як цільова мережа з параметром мережі θ', використовує наступний стан кортежу даних взаємодії {s, a, r, s'} та прогнозує найкраще значення Q для цього стану, цільове значення Q. Значення Q оновлюється за допомогою наступного рівняння:

$$Q'(s_t, a_t; \theta) = Q(s_t, a_t; \theta) + \alpha\big([r + \gamma maxQ(s_{t+1}, a_{t+1}; \theta')] - Q(s_t, a_t; \theta)\big),$$

де α є гіперпараметром, який називається швидкістю навчання.

Параметри в мережі прогнозування постійно оновлюються, але параметри цільової мережі копіюються з мережі прогнозування після кожних N ітерацій. Для знаходження оптимальних значень використовується функція втрат середньоквадратичної похибки з градієнтним спуском. Функція втрат задається як:

$$L = \frac{1}{n}\sum_{i=1}^{n}([r_i + \gamma maxQ(s_{t+1}, a_{t+1}; \theta')] - Q(s_t, a_t; \theta))^2.$$

Використання окремої мережі для оцінки значень дій дозволяє зменшити збурення та поліпшити стабільність навчання на початкових етапах. DDQN дозволяє мобільному роботу краще пристосовуватися до динаміки змін в середовищі, оскільки мережа-ціль допомагає зберегти стабільніші оцінки значень.

4.2. Методи на основі політики.

Замість вивчення функції значення, градієнт політики безпосередньо намагається оптимізувати функцію політики π . Методи на основі політики не мають цільової мережі і ймовірності дій оновлюються залежно від винагороди, отриманої від виконання цієї дії

$$\pi(a|s, \theta) = P\{A_t = a, S_t = s, \theta_t = \theta\}.$$

Методи на основі політики чутливі до ініціалізації параметрів політики, можуть застрягти в локальних мінімумах і мають високу дисперсію у своїй продуктивності [12].

Proximal Policy Optimization (PPO) [13] — це метод RL на основі політики, який базується на оптимізації політики довірчого регіону (TRPO). Його використовують для роботи з неперервними та дискретними просторами дій. Алгоритм використовує множник важливості для обчислення втрати та оновлення стратегії. Цей множник гарантує, що оновлення буде пропорційним до поточної стратегії, але не дозволяє занадто сильні зміни.

PPO використовується для навчання мобільних роботів управлінню рухом у складних середовищах в режимі реального часу.

Методи, засновані на політиках можуть обробляти як дискретні, так і безперервні простори дій.

4.3. Методи актора-критика.

Методи планування шляху на основі актора-критики (AC) поєднують градієнт політики та функцію цінності. Актор відповідає за прийняття рішень та визначає оптимальну стратегію для вибору дій в конкретних станах. Актор використовує функцію політики для визначення

ймовірностей вибору різних дій у конкретних станах. Критик використовує функцію цінності для обчислення цінності кожного кроку.

Таблиця 1

Порівняльний аналіз методів навчання з підкріпленням

Назва методу	Застосування	Переваги	Недоліки	Простір дій
Q-learning	планування шляху, уникнення перешкод, вивчення оптимальних стратегії в умовах невизначеності	простий в реалізації	висока обчислювальна вартість	дискретний
DQN	взаємодія з об'єктами та перешкодами	ефективний в умовах невизначеності	схильний до переоцінювання Q-значень	дискретний
DDQN	оптимізація планування шляху	стійкість до переоцінювання Q-значень	потребує налаштування гіперпараметрів	дискретний
A3C	навігація та управління рухом	використання асинхронного навчання для навчання кількох агентів	вимагає значних обчислювальних ресурсів, початкові параметри можуть суттєво впливати на результати навчання	дискретний неперервний
A2C	вибір оптимального шляху	простота в реалізації, покращена збіжність	не завжди гарантує оптимальні результати, застрягання в локальних мінімумах	дискретний неперервний
DDPG	уникнення динамічних перешкод	ефективність в задачах з неперервними просторами дій та станів	схильність до нестійкості	неперервний, дискретний
SAC	планування оптимальної траєкторії у складних середовищах	ефективність в реальних умовах, стратегія алгоритму, спрямована на максимізацію ентропії, сприяє стабільності та збалансованості навчання	потребує значних обчислювальних ресурсів, схильність до перенавчання	неперервний дискретний
PPO	вирішення завдань навігації та управління	стабільність навчання	вибір функції винагороди може вплинути на ефективність алгоритму	неперервний дискретний

Deep Deterministic Policy Gradient (DDPG) є ефективним алгоритмом для розв'язання задач із неперервними просторами дій. Тому він застосовується у тих сферах робототехніки, де взаємодія із середовищем є неперервною та потребує гнучкості в прийнятті рішень. Зокрема в праці [14] автори успішно реалізували алгоритм DDPG для відстеження траєкторії колісного робота з ковзним керуванням. RL було застосовано для навчання агента в неконтрольований спосіб. Це дослідження продемонструвало ефективність застосування DDPG у плануванні траєкторій без зіткнень, у зменшенні значення відстані та покращенні часу руху. Даний метод

вимагає ретельного підбору гіперпараметрів, невдало вибрані значення можуть призвести до повільного навчання.

Asynchronous Advantage Actor-Critic (A3C) є розширенням стандартного алгоритму актор-критика, який використовує кілька паралельних потоків актор-критик для одночасного вивчення функцій політики та цінностей.

В A3C параметр стратегії оновлюється як:

$$\theta = \theta + \alpha\nabla_\theta lg\pi_\theta(a_t|s_t)A(s_t, a_t) + \beta\nabla_\theta H(\pi(s_t;\theta)),$$

де H — це ентропія, α і β — параметри мережі.

A3C більш стійкий до налаштування гіперпараметрів, ніж DDPG. Даний метод обмежений середовищами з дискретними просторами дій.

Advantage Actor-Critic (A2C) — метод є удосконаленою версією алгоритму A3C, яка використовує синхронне оновлення моделі. Більшість дослідників віддають перевагу A2C через його швидку конвергенцію та низькі обчислювальні ресурси [15].

Soft actor-critic (SAC) — це алгоритм навчання з підкріпленням, який використовується для вирішення завдань в неперервних просторах дій, заснований на структурі моделі максимальної ентропії [16]. Алгоритм SAC, незважаючи на потужність, має певні обмеження, а саме: труднощі в обробці складної, динамічної інформації про навколишнє середовище; неадекватність обробки довгострокових залежностей у плануванні шляху; відсутність можливостей прогнозування майбутніх станів навколишнього середовища.

В табл. 1 проаналізовано методи навчання з підкріпленням, їхні переваги та недоліки.

5. Висновки.

Вибір алгоритму, який відповідає конкретній постановці задачі, має вирішальне значення для досягнення оптимальних результатів. Дана робота представляє вичерпний огляд методів навчання з підкріпленням, які застосовуються в мобільній робототехніці для планування шляху. Кожен з цих методів має свої унікальні особливості та переваги, які роблять їх ефективними у відповідних сценаріях.

Майбутні дослідження можуть бути зосереджені на розробці гібридних методів, інтеграції сенсорних технологій та забезпеченні надійної навігації в непередбачуваних умовах. Адаптивність даних алгоритмів до динамічного середовища також залишається відкритим питанням.

Список використаної літератури

1. Kober J, Bagnell JA, Peters J. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*. 2013. Vol. 32, 1

2. Rybczak, M.; Popowniak, N.; Lazarowska, A. A Survey of Machine Learning Approaches for Mobile Robot Control. *Robotics.* 2024, Vol.13, №1. P.12-22.

3. Gao, J.; Ye, W.; Guo, J.; Li, Z. Deep Reinforcement Learning for Indoor Mobile Robot Path Planning. *Sensors* 2020. № 20, 5493.

4. Tai L., Paolo G., Liu M., Virtual-to-real deep reinforcement learning: Continuous control of mobile robots for mapless navigation, 2017 *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Vancouver, Canada. 2017, P. 31-36,

5. Zhao Y., Zhang Y., Wang S. A Review of Mobile Robot Path Planning Based on Deep Reinforcement Learning Algorithm. *Journal of Physics: Conference Series*, Vol. 2138, International Conference on Artificial Intelligence and Big Data Applications (ICAIBD 2021) 24-25

6. Wang W., Wu Zh., Luo H., Zhang B. Path Planning Method of Mobile Robot Using Improved Deep Reinforcement Learning. *Journal of Electrical and Computer Engineering*, vol. 2022. P. 7.

7. Zheng, J.; Mao, S.; Wu, Z.; Kong, P.; Qiang, H. Improved Path Planning for Indoor Patrol Robot Based on Deep Reinforcement Learning. *Symmetry* 2022, № 14, 132.

8. Xin J., Zhao H., Liu D., Li M., Application of deep reinforcement learning in mobile robot path planning, *Chinese Automation Congress* (CAC), Jinan, China, 2017, p. 7112-7116.

9. Low Ee S., Ong P., Cheah K. Ch., Solving the optimal path planning of a mobile robot using improved Q-learning. *Robotics and Autonomous Systems*, 2019, Vol. 115, P. 143-161.

10. Jiang Q. Path Planning Method of Mobile Robot Based on Q-learning. *Journal of Physics: Conference Series*, International Symposium on Artificial Intelligence and Intelligent Manufacturing (AIIM 2021) 12/11/2021 - 14/11/2021 Huzhou 2022. Vol. 2181.

11. Khriji L, Touati F, Benhmed K, Al-Yahmedi A. Mobile Robot Navigation Based on Q-Learning Technique. *International Journal of Advanced Robotic Systems*. 2011. Vol.8 №1.

12. Singh, R.; Ren, J.; Lin, X. A Review of Deep Reinforcement Learning Algorithms for Mobile Robot Path Planning. *Vehicles*. 2023. № 5, P. 1423-1451.

13. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. Proximal policy optimization algorithms. 2017. №5. P. 56-67.

14. Srikonda S.; Norris W.R.; Nottage D.; Soylemezoglu A. Deep Reinforcement Learning for Autonomous Dynamic Skid Steer Vehicle Trajectory Tracking. *Robotics*. 2022, № 11, P. 95.

15. Xing X., Ding H., Liang Zh., Li B., Yang Zh., Robot path planner based on deep reinforcement learning and the seeker optimization algorithm, *Mechatronics*, 2022. Vol. 88. P. 102918

16. Zhang, Y.; Chen, P. Path Planning of a Mobile Robot for a Dynamic Indoor Environment Based on an SAC-LSTM Algorithm. *Sensors* 2023. № 23, P. 9802.

17. Ганенко Л. Д., Жебка В.В. Аналітичний огляд питань навігації мобільних роботів в закритих приміщеннях. *Телекомунікаційні та інформаційні технології*. 2023. № 3(80). Ст. 85-98.

18. Malinov V., Zhebka V., Zolotukhina O., Franchuk T., Chubaievskyi V. Biomining as an Effective Mechanism for Utilizing the Bioenergy Potential of Processing Enterprises in the Agricultural Sector. *CEUR Workshop Proceedings*. 2023, 3421, p. 223–230

References

1. Kober J, Bagnell JA, Peters J. Reinforcement learning in robotics: A survey. The International Journal of Robotics Research. 2013. Vol. 32, 1

2. Rybczak, M.; Popowniak, N.; Lazarowska, A. A Survey of Machine Learning Approaches for Mobile Robot Control. Robotics. 2024, Vol. 13, No. 1. P.12-22.

3. Gao, J.; Ye, W.; Guo, J.; Li, Z. Deep Reinforcement Learning for Indoor Mobile Robot Path Planning. Sensors 2020. No. 20, 5493.

4. Tai L., Paolo G., Liu M., Virtual-to-real deep reinforcement learning: Continuous control of mobile robots for mapless navigation, 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, Canada . 2017, P. 31-36,

5. Zhao Y., Zhang Y., Wang S. A Review of Mobile Robot Path Planning Based on Deep Reinforcement Learning Algorithm. Journal of Physics: Conference Series, Vol. 2138, International Conference on Artificial Intelligence and Big Data Applications (ICAIBD 2021) 24-25

6. Wang W., Wu Zh., Luo H., Zhang B. Path Planning Method of Mobile Robot Using Improved Deep Reinforcement Learning. Journal of Electrical and Computer Engineering, vol. 2022. P. 7.

7. Zheng, J.; Mao, S.; Wu, Z.; Kong, P.; Qiang, H. Improved Path Planning for Indoor Patrol Robot Based on Deep Reinforcement Learning. Symmetry 2022, No. 14, 132.

8. Xin J., Zhao H., Liu D., Li M., Application of deep reinforcement learning in mobile robot path planning, Chinese Automation Congress (CAC), Jinan, China, 2017, p. 7112-7116.

9. Low Ee S., Ong P., Cheah K. Ch., Solving the optimal path planning of a mobile robot using improved Q-learning. Robotics and Autonomous Systems, 2019, Vol. 115, P. 143-161.

10. Jiang Q. Path Planning Method of Mobile Robot Based on Q-learning. Journal of Physics: Conference Series, International Symposium on Artificial Intelligence and Intelligent Manufacturing (AIIM 2021) 12/11/2021 - 14/11/2021 Huzhou 2022. Vol. 2181.

11. Khriji L, Touati F, Benhmed K, Al-Yahmedi A. Mobile Robot Navigation Based on Q-Learning Technique. International Journal of Advanced Robotic Systems. 2011. Vol. 8 No. 1.

12. Singh, R.; Ren, J.; Lin, X. A Review of Deep Reinforcement Learning Algorithms for Mobile Robot Path Planning. Vehicles. 2023. No. 5, P. 1423-1451.

13. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. Proximal policy optimization algorithms. 2017. No. 5. P. 56-67.

14. Srikonda S.; Norris W.R.; Nottage D.; Soylemezoglu A. Deep Reinforcement Learning for Autonomous Dynamic Skid Steer Vehicle Trajectory Tracking. Robotics. 2022, No. 11, P. 95.

15. Xing X., Ding H., Liang Zh., Li B., Yang Zh., Robot path planner based on deep reinforcement learning and the seeker optimization algorithm, Mechatronics, 2022. Vol. 88. P. 102918

16. Zhang, Y.; Chen, P. Path Planning of a Mobile Robot for a Dynamic Indoor Environment Based on an SAC-LSTM Algorithm. Sensors 2023. No. 23, P. 9802.

17. Ganenko L. D., Zhebka V. V. Analytical review of issues of navigation of mobile robots in closed spaces. Telecommunications and information technologies. 2023. No. 3(80). Art. 85-98.

18. Malinov V., Zhebka V., Zolotukhina O., Franchuk T., Chubaievskyi V. Biomining as an Effective Mechanism for Utilizing the Bioenergy Potential of Processing Enterprises in the Agricultural Sector. CEUR Workshop Proceedings. 2023, 3421, p. 223–230