КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 2 (30), 2025
Спеціальний випуск ISSN 2663 – 4023
Special issue
DOI 10.28925/2663-4023.2025.30.994
УДК 004.896

Ганенко Людмила Дмитрівна
аспірантка
Державний університет інформаційно-комунікаційних технологій, Київ, Україна
ORCID: 0000-0003-2219-8196
hanenkoliudmyla@gmail.com

Бушма Олександр Володимирович
доктор технічних наук, професор
професор кафедри ком'ютерних наук
Київський столичних університет імені Бориса Грінченка, Київ, Україна
ORCID: 0000-0003-1604-6129
o.bushma@kubg.edu.ua

МЕТОД НАВЧАННЯ АВТОНОМНИХ МОБІЛЬНИХ РОБОТІВ НА ОСНОВІ DRL ТА CURRICULUM LEARNING

Анотація. Робота присвячена актуальній задачі підвищення ефективності соціально-адаптивної навігації автономних мобільних роботів у динамічних середовищах із присутністю людей. Застосування методів глибокого навчання з підкріпленням (DRL) для вирішення цієї задачі ускладнюється високою розмірністю простору станів, складністю формалізації соціальних норм у функції винагороди та нестабільністю процесу навчання. Для подолання цих викликів запропоновано метод, що інтегрує алгоритм Proximal Policy Optimization (PPO) зі стратегією навчання за програмою Curriculum Learning (CL). Розроблена навчальна програма поєднує поступове ускладнення середовища (від статичних перешкод до середовища із рухомими агентами-людьми) та поетапне формування функції винагороди із додаванням соціальних компонентів. Ключовою особливістю є перехід між етапами, який базується на аналізі стабільності політики. Експериментальне дослідження проведено в розробленому симуляційному середовищі Gazebo із використанням мобільного робота Turtlebot3 Waffle та фреймворку ROS 2 Humble. Поетапне навчання дозволяє автономному мобільному роботу спочатку засвоїти базові навички уникнення статичних перешкод, потім — динамічних, і на завершальному етапі — враховувати соціальні норми взаємодії з людьми. Вхідними даними для системи є дані з LiDAR, стан робота та людей, а також цільова позиція. Результатом методу є оптимізована стохастична політика поведінки, що дозволяє автономному мобільному роботу приймати безпечні, ефективні та соціально прийнятні навігаційні рішення. Проведено порівняльний аналіз запропонованого методу із стандартним алгоритмом PPO. Отримані результати підтверджують, що запропонований метод дозволяє формувати ефективну політику соціально-адаптивної навігації, вирішуючи проблеми нестабільності та повільної збіжності.

Ключові слова: інформаційні технології, методи машинного навчання, методи навчання з підкріпленням, Deep Reinforcement Learning, Curriculum Learning, автономні мобільні роботи, навігація мобільних роботів, ROS 2, Gazebo.

ВСТУП

Ефективна та безпечна навігація автономного мобільного робота в динамічних, насичених людьми середовищах зумовлює появу нових вимог до навігаційних систем. Ключовим викликом стає здатність робота не лише уникати зіткнення з людьми, а й рухатися зрозумілою, передбачуваною та комфортною для людини траєкторією.

Одним з популярних підходів реалізації навігації мобільних роботів є одночасна локалізація та картографування (SLAM) [1]. Однак створення карти є складним процесом

у випадку динамічних середовищ. Точність картографування в таких умовах стає визначальним фактором, який безпосередньо впливає на загальну якість навігації.

Перспективним напрямом подолання цих обмежень є застосування методів машинного навчання, зокрема глибокого навчання з підкріпленням (Deep Reinforcement Learning, DRL). Методи DRL дозволяють агенту автономно формувати складну політику поведінки шляхом безпосередньої взаємодії з середовищем та оптимізації кумулятивної винагороди, вони є оптимальним рішенням для задач з багатофакторними та нелінійними залежностями. Натомість застосування методів глибокого навчання з підкріпленням (DRL) для формування соціально-прийнятної поведінки автономного мобільного робота ускладнюється такими ключовими факторами: високою розмірністю простору станів, складністю формалізації багатогранних аспектів соціальної взаємодії у вигляді функції винагороди та нестабільністю навчання у складному середовищі.

Для подолання зазначених викликів необхідний підхід, що структурує процес навчання, дозволяючи агенту освоювати складну поведінку поступово. Однією з таких стратегій є навчання за програмою (Curriculum Learning, CL). CL дозволяє стабілізувати збіжність, прискорити навчання та знаходити оптимальну політику поведінки.

Аналіз останніх досліджень і публікацій. Центральною ідеєю методів, що структурують процес навчання, стала декомпозиція кінцевої складної задачі на послідовність простіших підзадач. Навчання за програмою дозволяє агенту поступово накопичувати знання та навички. У наукових дослідженнях переважають такі підходи застосування CL: поступове збільшення складності завдань та формування винагороди [2].

Автори [3] запропонували метод CuRLA для навчання агента автономного водіння, в якому реалізували CL як двохетапний процес. Перший етап навчання здійснюється у спрощеному середовищі, агент зосереджується на основних навичках водіння: триматися смуги та підтримувати швидкість. Після 1500 епізодів у симуляції поступово збільшується трафік. Це змушує агента вчитися маневрувати серед інших автомобілів та уникати небезпечних ситуацій. Одночасно із збільшенням трафіку до функції винагороди додається новий компонент — штраф за зіткнення.

В дослідженні [4] автори розглядають проблему DRL у задачах з комплексними, багатокомпонентними функціями винагороди. Автори зазначають, що функції винагороди, які поєднують основну мету та численні обмежувальні умови (дотримання швидкості, слідування маршруту), часто призводять до небажаної поведінки агента. Агент знаходить спосіб максимізувати винагороду за другорядні обмеження, ігноруючи при цьому головну мету. Для вирішення цієї проблеми автори пропонують підхід CL, який реалізовано у вигляді двохетапного CL на основі винагороди. Суть методу полягає у зміні складності цільової функції, а не середовища. На першому етапі агент навчається на спрощеній функції винагороди, яка є підмножиною повної винагороди і включає лише ключові компоненти, які необхідні для освоєння базової навички досягнення мети (винагорода за досягнення цілі, за просування по маршруту та за дотримання швидкості). Це дозволяє агенту уникнути локальних оптимумів і спочатку навчитися вирішувати основні задачі навігації. На другому етапі відбувається перехід до навчання на повній, багатокомпонентній функції винагороди, яка враховує всі обмеження.

В роботі [5] автори реалізували ефективну навігацію агента в щільному натовпі із застосуванням навчальної програми. CL складається з двох основних етапів, які формують навчальну програму від простого до складного: спочатку агент тренується в симуляційному середовищі, де знаходиться лише одна людина, після успішного навчання в простому середовищі, агента переносять у середовище, де знаходиться натовп з п’яти людей. На другому етапі нейронна мережа ініціалізується вагами, отриманими після навчання на першому етапі. Таким чином, агент вже навчився уникати один об’єкт і тепер йому потрібно лише адаптувати ці знання для складнішої ситуації з багатьма людьми.

В праці [6] робота спочатку навчають досягати мети зі стартових станів, які є близькими до заданого цільового стану. Потім, використовуючи ці знання, робота навчають виконувати завдання з дедалі віддаленіших стартових станів.

Незважаючи на значний прогрес застосування CL у навчанні агентів, низка невирішених проблем визначає перспективні напрями для подальших досліджень. Більшість існуючих робіт використовують або ручний перехід між етапами навчання, або покладаються на порогові значення (відсоток успіху, кількість кроків навчання). Актуальною задачею залишається розробка автоматизованих підходів, які б оцінювали стабільність засвоєної навички перед тим, як підвищувати складність завдання.

Мета статті. Метою дослідження є підвищення ефективності соціально-адаптивної навігації мобільного робота в динамічних середовищах із присутністю людей шляхом розробки комбінованого методу навчання.

Для досягнення мети були поставлені наступні завдання:

– проаналізувати існуючі стратегії curriculum learning;

– проаналізувати застосування алгоритмів глибокого навчання з підкріпленням у задачах робототехнічної навігації;

– розробити стратегію curriculum learning;

– створити та налаштувати симуляційне середовище;

провести експериментальне дослідження та виконати порівняльний аналіз ефективності запропонованого методу.

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Розробка автономних навігаційних систем для мобільних роботів є однією з фундаментальних задач сучасної робототехніки. Навчання з підкріпленням (RL) дозволяє роботу навчатися оптимальній поведінці безпосередньо з отриманого досвіду. У процесі навчання, що ґрунтується на методі спроб і помилок, робот виконує дії, такі як зміна швидкості та напрямку, аналізує їхні наслідки у вигляді винагород за досягнення мети та штрафів за зіткнення, і поступово формує навігаційну політику, яка максимізує сукупну винагороду. Цей підхід продемонстрував високі результати у вирішенні простих навігаційних завдань, таких як рух у сіткових світах, де дискретний простір станів та дій дозволяє класичним методам, наприклад Q-навчанню, ефективно знаходити оптимальні шляхи.

Однак, у завданнях з великим простором станів традиційний алгоритм RL стикається з «прокляттям розмірності», коли кількість обчислень різко зростає зі збільшенням кількості вхідних даних. Підхід глибокого навчання (DL) дозволяє апроксимувати нелінійну функцію шляхом навчання глибоких нейронних мереж. Завдяки інтеграції RL з глибокими нейронними мережами було створено DRL (Deep Reinforcement Learning) [7].

DRL можна розділити на методи на основі цінності, методи на основі політики та методи «актора-критика». Методи на основі цінності, такі як DQN, DDQN та D3QN, ефективні для просторів дискретних дій. Вони забезпечують стабільність та масштабованість у завданнях роботизованої навігації з чітко визначеними наборами дій. Методи на основі політики, розроблені для просторів безперервних дій. Такі методи використовують для керування мобільними роботами з диференціальним приводом. Методи «актора-критика» інтегрують переваги обох архітектур: «актор» відповідає за вибір дій (навчання політики), тоді як «критик» оцінює дії (навчання функції цінності).

В задачах автономної навігації метою алгоритмів DRL є пошук оптимальної політики для переміщення робота до цільової позиції шляхом взаємодії з навколишнім

середовищем. Алгоритми DRL успішно використовують для реалізації навігаційної системи автономних мобільних роботів [8].

Алгоритм Proximal Policy Optimization (PPO). У контексті автономної навігації мобільних роботів, алгоритм Proximal Policy Optimization (PPO) набув широкого поширення завдяки поєднанню стабільності навчання, ефективності вибірки та здатності до масштабування на складні динамічні середовища [9]. На відміну від традиційних Q-орієнтованих методів, PPO працює у парадигмі актора-критика, що дозволяє йому безпосередньо оптимізувати політику дій, забезпечуючи плавне оновлення параметрів через механізм обмеженого відсікання політики. Такий підхід зменшує ризик розбалансування навчання та сприяє узгодженості поведінки агента. Порівняно з іншими методами PPO менш чутливий до гіперпараметрів, що робить його ефективним в задачах навігації в складних, непередбачуваних умовах, характерних для взаємодії з людьми.

Метою PPO є максимізація очікуваної винагороди $J(\pi_\theta)$ з оновленням політики, яке не надмірно відхиляється від поточної політики. Мета оптимізації визначається наступним чином:

$$L_{CLIP}(\theta) = E_t[\min(r_t(\theta)A_t, clip(r_t(\theta), 1-\varepsilon, 1+\varepsilon)A_t)],$$

де $r_t(\theta)$– це коефіцієнт ймовірності, який визначається наступним чином:

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$$

A_t – функція переваги, ε – гіперпараметр, який визначає діапазон відсікання, та clip (·) обмежує відношення ймовірностей інтервалом $[1-\varepsilon, 1+\varepsilon]$.

Функція переваги A_t дії a_t у стані s_t обчислюється наступним чином:

$$A_t = Q_\pi(s_t, a_t) - V_\pi(s_t),$$

де $Q_\pi(s_t, a_t)$– функція цінності дії, а $V_\pi(s_t)$ – функція цінності стану.

Алгоритм PPO ґрунтується на оптимізації усіченої сурогатної цільової функції, що обмежує величину кроку оновлення політики на кожній ітерації. Такий підхід запобігає радикальним змінам у поведінці агента, забезпечує стабільність навчального процесу та підтримує ефективний баланс між дослідженням та експлуатацією.

Окрім механізму відсікання, функція загальних втрат у PPO містить втрати функції цінності $L_V(\theta)$ та втрати ентропії $L_E(\theta)$.

$$L(\theta) = L_{CLIP}(\theta) - c_1 L_V(\theta) + c_2 L_E(\theta),$$

де $L_V(\theta)$ – функція втрат, яка апроксимує функцію цінності V(s):

$$L_V(\theta) = E_t[(V_\theta(s_t) - R_t)^2],$$

де $L_E(\theta)$ – втрата ентропії:

$$L_E(\theta) = E_t\left[-\sum_a \pi_\theta(a|s_t)\log\pi_\theta(a|s_t)\right]$$

Коефіцієнти c_1 і c_2 контролюють відносну важливість втрати цінності та втрати ентропії відповідно.

Покроковий псевдокод PPO представлено в алгоритмі 1.

Алгоритм 1. Алгоритм Proximal Policy Optimization (PPO)

Вхідні дані: політика π_θ, мережа цінностей V_θ, параметр відсікання ε, коефіцієнти c_1 і c_2, число епох *K*, розмір батчу *N*, коефіцієнти навчання α_π і α_V.

Вихідні дані: оптимізована політика π_θ

1. **Ініціалізувати** політику π_θ та мережу цінностей V_θ випадковими вагами.
2. **for** кожної ітерації **do**:

зібрати траєкторії $D = \{(s_t, a_t, R_t, s_{t+1})\}$ через запуск політики π_θ в середовищі;

обчислити перевагу A_t для кожної траєкторії, використавши

$$A_t = \sum_{l=0}^{T-t} \gamma^l \delta_{t+1},$$

де $\delta_t = R_t + \gamma V_\theta(s_{t+1}) - V_\theta(s_t)$;

обчислити винагороду

$$R_t = \sum_{l=0}^{T-t} \gamma^l r_{t+1}$$

for *k=1* to *K* **do**:

вибрати міні-батч розміром *N* з *D*;
обчислити відношення ймовірностей

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$$

обчислити усічену наближену цільову функцію

$$L_{CLIP}(\theta) = E_t[\min(r_t(\theta)A_t, clip(r_t(\theta), 1-\varepsilon, 1+\varepsilon)A_t)],$$

обчислити втрати цінності $L_V(\theta) = E_t[(V_\theta(s_t) - R_t)^2]$
обчислити втрати ентропії

$$L_E(\theta) = E_t\left[-\sum_a \pi_\theta(a|s_t)\log\pi_\theta(a|s_t)\right]$$

обчислити загальні втрати $L(\theta) = L_{CLIP}(\theta) - c_1 L_V(\theta) + c_2 L_E(\theta)$
оновити параметри політики $\theta \leftarrow \theta + \alpha_\pi \nabla_\theta L(\theta)$;
оновити параметри мережі цінностей $\theta_v \leftarrow \theta_v + \alpha_v \nabla_\theta L_V(\theta)$

Curriculum learning як метод структурування навчального процесу для складних завдань. У сфері автономної робототехніки навчання з підкріпленням вимагає ефективного вирішення складних завдань у динамічному та часто непередбачуваному середовищі. Проте безпосереднє навчання агентів у складних сценаріях зазвичай призводить до низької швидкості збіжності, локальних мінімумів або повної невдачі навчального процесу. Для подолання цих труднощів дедалі більшого поширення набуває підхід навчання за програмою (Curriculum learning, CL) [10].

CL передбачає організацію навчального процесу у вигляді послідовності завдань, які поступово ускладнюються. Це дозволяє агенту спершу засвоїти базові навички, а потім поступово узагальнювати ці знання для вирішення складніших сценаріїв. У контексті робототехніки та автономної навігації такий підхід сприяє швидшому навчанню, покращенню стабільності та кращій здатності до перенесення знань у нові умови.

Curriculum learning – це стратегія навчання моделі машинного навчання [11], де навчальний процес організований у вигляді послідовності завдань (критеріїв) з різним рівнем складності:

$$C =< Q_1, \dots, Q_t, \dots Q_T >$$

Кожен критерій Q_t формується шляхом вагового зважування цільового розподілу навчання $P(z)$

$$Q_t(z) \propto W_t(z)P(z) \qquad \forall z \in D$$

таким чином, що виконуються наступні три умови:

1) ентропія розподілів поступово зростає

$$H(Q_t) < H(Q_{t+1}).$$

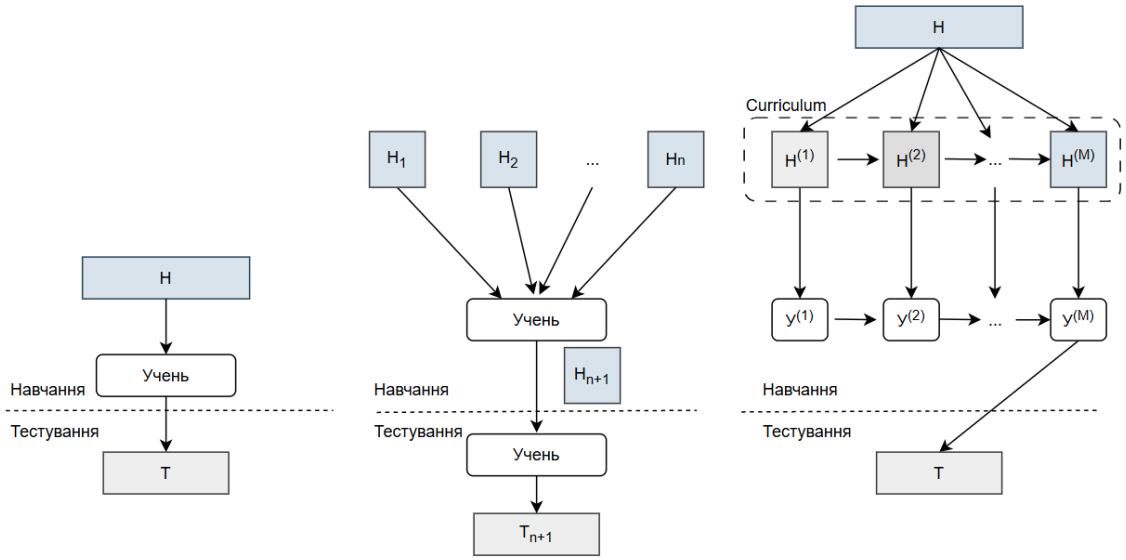
2) вага для будь-якого прикладу зростає

$$W_t(z) \leq W_{t+1}(z) \ \ \forall z \in D.$$

3) в останній ітерації модель навчається на повному наборі даних

$$Q_T(z) = P(z).$$

Фундаментальна відмінність між підходами машинного навчання (традиційне машинне навчання, transfer learning, curriculum learning та ін.) полягає у способі роботи з розподілами тренувальних та тестових даних. Графічну інтерпретацію цих підходів представлено на рис. 1, де Н – навчальні дані, Т – тестові дані, $Н_j$ – навчальні дані для різних завдань, $Н^{(i)}$ – модифікований розподіл на і-му кроці навчання, $У^{(i)}$ – учень на і-му кроці послідовності.



а) б) в)

Рис.1. Схеми навчання: а) традиційного машинного навчання, б) Transfer learning, в) Curriculum learning.

CL дозволяє вирішити проблему неефективних досліджень RL завдяки поступовому підвищенню складності навчального середовища або шляхом модифікації функції винагороди [12, 13].

У підході підвищення складності навчального середовища агент починає навчання у спрощеній версії цільового середовища, складність якого поступово зростає за рахунок зміни одного або кількох параметрів середовища. Навчальний план складається з послідовності середовищ $\{E_1, E_2, ..., E_n\}$, де E_n — це фінальне, цільове середовище.

Складність середовища може змінюватись за рахунок: змін просторових характеристик (розмір карти, кількість та тип перешкод), змін динаміки середовища (швидкість агентів, рівень шуму сенсорів) та змін умови завдання (початкова позиція агента, варіативність цілей).

Підхід формування винагороди передбачає коригування функції винагороди для стимулювання поведінки в напрямку цільового стану, ефективно забезпечуючи підцілі, які ведуть до кінцевої мети. Складність завдання зростає через зміну структури або ваг компонентів функції винагороди. Наприклад, агент, метою якого є рух до цілі, може бути винагороджений за простіше завдання переміщення до цільової точки. Коли підціль буде досягнута, винагорода коригується для стимулювання просування ближче до цільової позиції.

Зворотна навчальна програма передбачає позиціонування агента в цільовому стані і виконання дій для отримання інших допустимих станів поблизу мети. У міру навчання агента ці дії призводять до того, що агент рухається все далі й далі від мети. Водночас, застосування зворотної навчальної програми має недолік у сценаріях з кількома рівноцінними цільовими станами: агент втрачає здатність ефективно досягати альтернативних цілей, оскільки його стратегія стає неоптимальною для інших кінцевих станів.

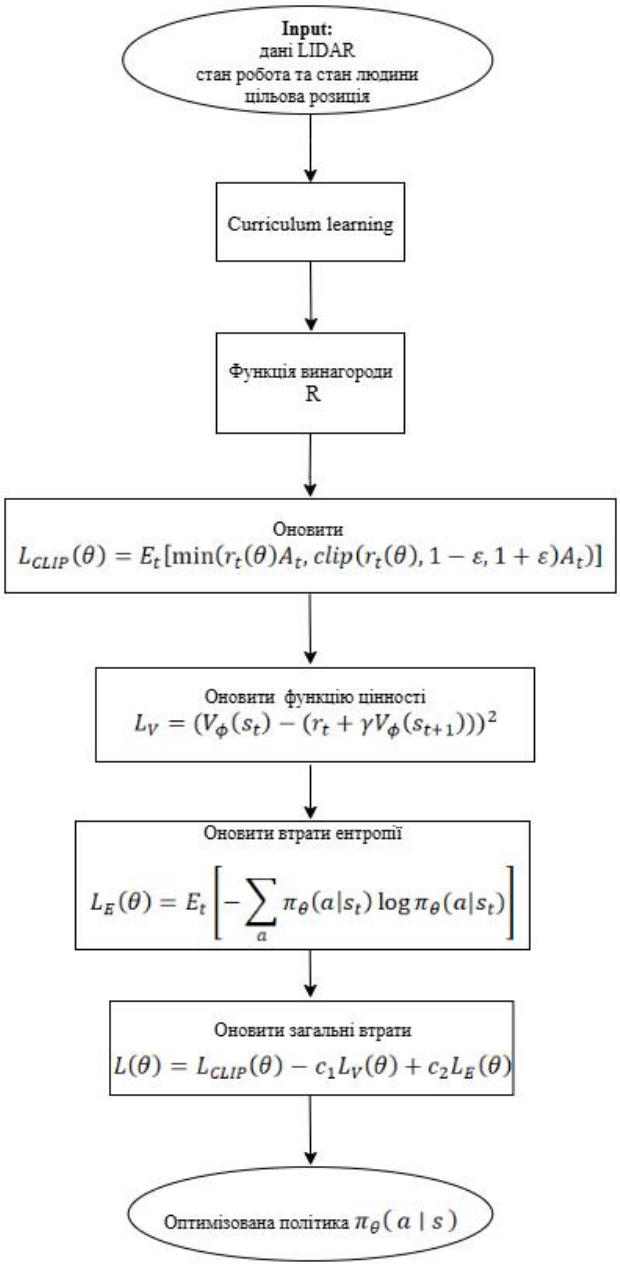


Рис. 2. Блок-схема методу

Метод соціально-адаптивної навігації автономного мобільного робота.

Запропонований метод інтегрує алгоритм PPO та CL і створює єдиний механізм забезпечення соціально-адаптивної навігації автономного мобільного робота. Для створення методу використано побудовану в роботі [14] модель соціально-адаптивної навігації автономного мобільного робота.

Розроблений метод використовує навчальну програму, яка поєднує стратегії поступового ускладнення середовища та формування функції винагороди.

Загальну архітектуру методу представлено на рис.2.

Ключовою особливістю методу є використання модуля CL, який дозволяє агенту поступово засвоювати складні навички навігації. Процес навчання складається з 4 етапів,

кожен з яких відрізняється або середовищем, або функцією винагороди. Перехід до наступного, складнішого етапу, відбувається лише після досягнення стабільної продуктивності на поточному етапі.

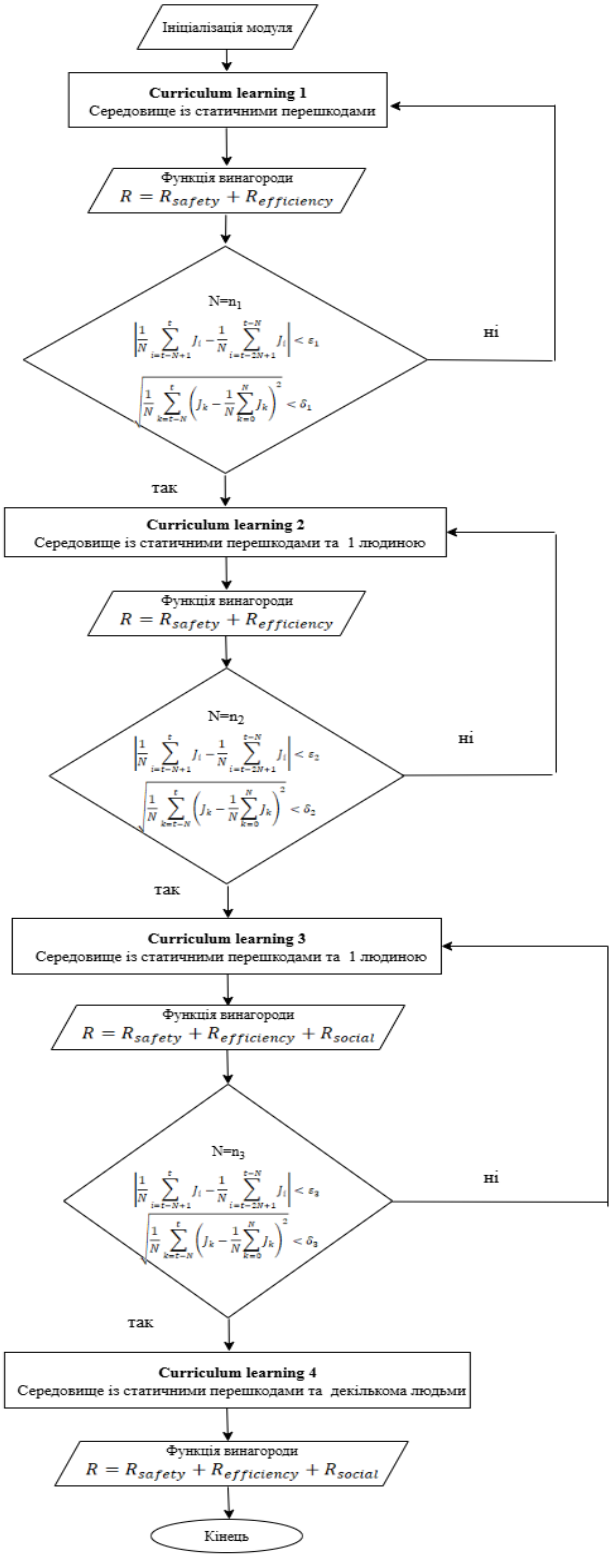


Рис. 3. Блок-схема модуля curriculum learning.

Процес навчання включає наступні етапи:

1) на першому етапі робот навчається досягати цілі, уникаючи лише статичних перешкод. Функція винагороди складається з двох компонентів:

$$R = R_{safety} + R_{efficiency}$$

2) після успішного засвоєння базових навичок на другому етапі у середовище додається динамічна перешкода (агент-людина). Функція винагороди залишається такою ж, як і на першому етапі. Це змушує робота адаптувати свою політику до уникнення рухомих об'єктів.
3) на третьому етапі функція винагороди модифікується для врахування соціальних аспектів:

$$R = R_{safety} + R_{efficiency} + R_{social}$$

4) на четвертому етапі в середовищі збільшується кількість агентів-людей.

Блок-схема модуля CL продемонстрована на рис. 3.

Перехід між етапами здійснюється при одночасному виконанні умов

$$\left| \frac{1}{N} \sum_{i=t-N+1}^{t} J_i - \frac{1}{N} \sum_{i=t-2N+1}^{t-N} J_i \right| < \varepsilon_l$$

$$\sqrt{\frac{1}{N} \sum_{k=t-N}^{t} \left(J_k - \frac{1}{N} \sum_{k=0}^{N} J_k \right)^2} < \delta_l$$

де J_i – винагорода на і епізоді, N – розмір вікна, l приймає значення від 1 до 3.

Система аналізує варіації показника J (середньої винагороди за епізод) за останню N кількість епізодів. Якщо відхилення цього показника від його ковзного середнього є меншим за порогове значення ε_l і стандартне відхилення винагороди менше δ_l, політика стабілізувалася і відбувається перехід на наступний етап.

Результатом процесу навчання є оптимізована політика. Нейронна мережа, здатна в режимі реального часу генерувати дії (лінійну та кутову швидкості) для мобільного робота на основі поточних сенсорних даних, забезпечуючи безпечну, ефективну та соціально-адаптивну навігацію.

Експериментальна реалізація. Даний метод було використано для навчання мобільного робота Turtlebot3 Waffe в симуляційному середовищі Gazebo із використанням ROS 2 Humble [15]. Симуляції проводились у змодельованому середовищі, яке містить 7 статичних перешкод і 3 людини (рис. 4). Цільова точка позначена червоним циліндром.

Для забезпечення узагальнюючої здатності навченої політики та уникнення її перенавчання, початкові позиції мобільного робота та цілі генерувалися випадково.

Результати порівняння ефективності навчання представлені на рис. 5, на якому відображено залежність винагороди від номера епізоду. На графіку порівнюється ефективність алгоритму PPO за двох умов: стандартного тренування в повному цільовому середовищі (крива PPO) та тренування із застосуванням Curriculum learning (крива PPO-CL).

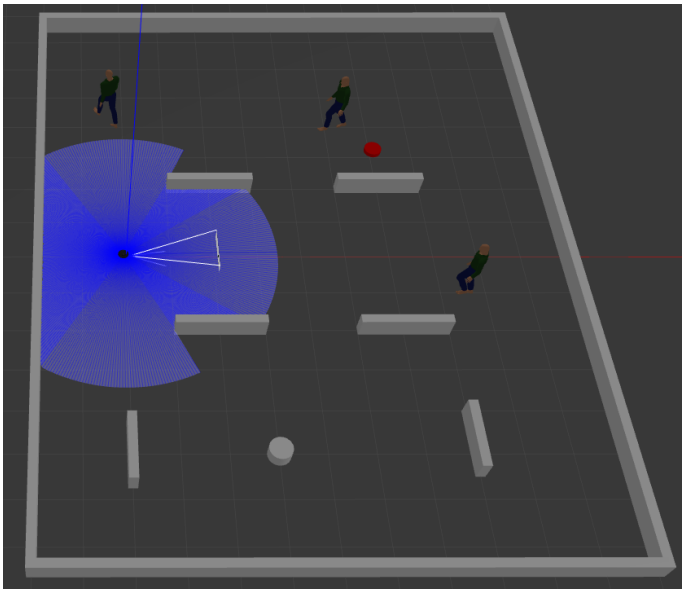


Рис.4. Змодельоване середовище

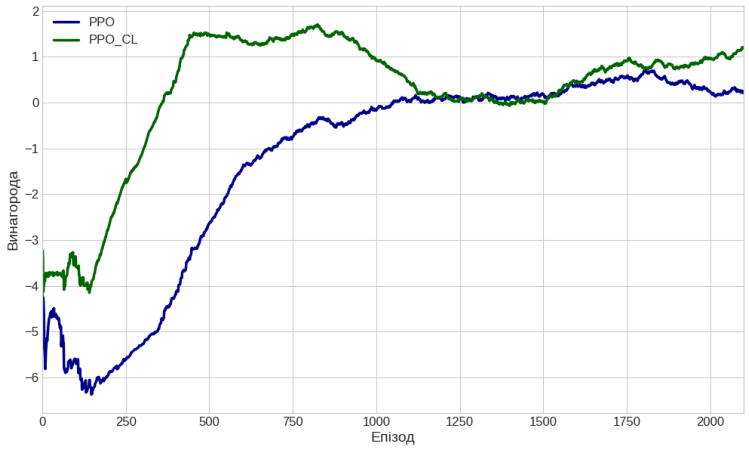


Рис. 5. Графік винагород.

Аналіз отриманого графіку дозволяє зробити висновки щодо динаміки та ефективності обох підходів. На початковому етапі навчання спостерігаються суттєві розбіжності. Мобільний робот, який навчався за традиційним алгоритмом PPO, демонструє низькі початкові показники винагороди (≈ -6.0). Це є результатом того, що мобільний робот врізається у стіни, людей або не може знайти ціль. Процес навчання характеризується повільною збіжністю, що вказує на неефективне дослідження простору станів на ранніх стадіях. Натомість PPO із використанням CL стартує зі значно вищої винагороди (≈ -3.5) і демонструє стрімке зростання.

На середньому етапі навчання агент, тренований за програмою (PPO-CL), демонструє досягнення локального максимуму продуктивності. Подальше зниження кривої винагороди можна інтерпретувати як наслідок переходу до 4 етапу навчальної програми (навчання в середовищі із присутністю 3 людей). Введення нових факторів у середовище, таких як динамічні перешкоди (моделі людей), вимагає від агента суттєвої адаптації раніше засвоєної політики поведінки, що тимчасово знижує його загальну

ефективність. Водночас стандартний агент PPO на цьому інтервалі продовжує демонструвати монотонне, хоча й повільне, зростання.

На завершальному етапі навчання крива PPO-CL демонструє тенденцію до вищих показників. Це свідчить про те, що структурований підхід до навчання дозволив агенту знайти більш ефективну політику поведінки в умовах фінальної складності середовища.

Для порівняння продуктивності обох навчених моделей було проведено тестування. Моделі були протестовані на 20 епізодах із випадково згенерованими позиціями робота та цілі. Метрикою ефективності було визначено показник успішності (Success rate) [16,17], тобто відсоткове співвідношення епізодів, у яких агент досяг цільової позиції без зіткнень із перешкодами або перевищення встановленого часового ліміту.

За результатами тестування, мобільний робот, навчений за стандартним PPO, продемонстрував показник успішності на рівні 70%. Водночас мобільний робот, тренований із застосуванням PPO та CL, досягнув цілі у 80% тестових епізодів. Таким чином, отримані дані емпірично підтверджують, що використання CL сприяє створенню більш ефективної політики навігації.

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

У дослідженні було запропоновано метод, який інтегрує алгоритм глибокого навчання з підкріпленням PPO з комбінованою стратегією навчання за програмою (CL). Розроблено навчальну програму, яка поєднує поетапне ускладнення середовища та формування функції винагороди.

Результати дослідження доводять, що інтеграція PPO з CL є ефективним рішенням для подолання проблем, які властиві DRL у складних задачах робототехнічної навігації. Подальші дослідження можуть бути спрямовані на валідацію розробленого методу на фізичній роботизованій платформі та його адаптацію до більш непередбачуваних соціальних сценаріїв.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Hanenko, L., Storchak, K., Shlianchak, S., Vorokhob, M., & Pitaichuk, M. (2025). SLAM in navigation systems of autonomous mobile robots. *Cybersecurity Providing in Information and Telecommunication Systems 2025*, (3991), 173–182. https://ceur-ws.org/Vol-3991/
2. West, J., Maire, F., Browne, C., & Denman, S. (2020). Improved reinforcement learning with curriculum. *Expert Systems with Applications*, *158*, 113515. https://doi.org/10.1016/j.eswa.2020.113515
3. Uppuluri, B., Patel, A., Mehta, N., Kamath, S., & Chakraborty, P. (2025). CuRLA: Curriculum learning based deep reinforcement learning for autonomous driving. In *17th International Conference on Agents and Artificial Intelligence* (pp. 435–442). SCITEPRESS. https://doi.org/10.5220/0013147000003890
4. Freitag, K., Ceder, K., Laezza, R., Akesson, K., & Haghir Chehreghani, M. (2024). Sample-efficient curriculum reinforcement learning for complex reward functions. *arXiv preprint*. https://doi.org/10.48550/arXiv.2410.16790
5. Li, K., Lu, Y., & Meng, M. Q. H. (2021). Human-aware robot navigation via reinforcement learning with hindsight experience replay and curriculum learning. In *2021 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE. https://doi.org/10.1109/robio54168.2021.9739519
6. Florensa, C., Held, D., Wulfmeier, M., Zhang, M., & Abbeel, P. (2017, October). Reverse curriculum generation for reinforcement learning. In *Conference on Robot Learning* (pp. 482–495). PMLR.
7. Zhu, K., & Zhang, T. (2021). Deep reinforcement learning based mobile robot navigation: A review. *Tsinghua Science and Technology*, *26*(5), 674–691. https://doi.org/10.26599/tst.2021.9010012
8. Gao, J., Ye, W., Guo, J., & Li, Z. (2020). Deep reinforcement learning for indoor mobile robot path planning. *Sensors*, *20*(19), 5493. https://doi.org/10.3390/s20195493

9. Ганенко, Л. Д., & Жебка, В. В. (2024). Застосування методів навчання з підкріпленням для планування шляху мобільних роботів. *Телекомунікаційні та інформаційні технології, 1*, 16–25. https://doi.org/10.31673/2412-4338.2024.011625
10. Soviany, P., Ionescu, R. T., Rota, P., & Sebe, N. (2022). Curriculum learning: A survey. *International Journal of Computer Vision*. https://doi.org/10.1007/s11263-022-01611-x
11. Wang, X., Chen, Y., & Zhu, W. (2021). A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1. https://doi.org/10.1109/tpami.2021.3069908
12. Anca, M., Thomas, J. D., Pedamonti, D., Hansen, M., & Studley, M. (2023). Achieving goals using reward shaping and curriculum learning. In *Lecture Notes in Networks and Systems* (pp. 316–331). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-47454-5_24
13. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint*. https://doi.org/10.48550/arXiv.1707.06347
14. Ганенко, Л., & Жебка, В. (2025). Модель соціально-адаптивної навігації мобільного робота з використанням методів навчання з підкріпленням. *Кібербезпека: освіта, наука, техніка, 1*(29). https://doi.org/10.28925/2663-4023.2025.29.907
15. Ганенко, Л., & Жебка, В. (2025). Створення навігаційної системи автономного мобільного робота засобами ROS 2. *Кібербезпека: освіта, наука, техніка, 4*(28), 498–510. https://doi.org/10.28925/2663-4023.2025.28.824
16. Narvekar, S., Peng, B., Leonetti, M., Sinapov, J., Taylor, M. E., & Stone, P. (2020). Curriculum learning for reinforcement learning domains: A framework and survey. *Journal of Machine Learning Research, 21*(181), 1–50.

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 2 (30), 2025
Спеціальний випуск ISSN 2663 – 4023
Special issue

Hanenko Liudmyla
postgraduate student
State University of Information and Communication Technologies, Kyiv, Ukraine
ORCID: 0000-0003-2219-8196
hanenkoliudmyla@gmail.com

Bushma Oleksandr
Doctor of Technical Sciences, Professor
Professor of the Department of Computer Science
Borys Grinchenko Kyiv Metropolitan University, Kyiv, Ukraine
ORCID: 0000-0003-1604-6129
o.bushma@kubg.edu.ua

METHOD OF LEARNING OF AUTONOMOUS MOBILE ROBOTS BASED ON DRL AND CURRICULUM LEARNING

Abstract: The work is devoted to the urgent task of improving the efficiency of socially adaptive navigation of autonomous mobile robots in dynamic environments with human presence. The application of deep reinforcement learning (DRL) methods to solve this problem is complicated by the high dimensionality of the state space, the complexity of formalizing social norms in the reward function, and the instability of the learning process. To overcome these challenges, a method is proposed that integrates the Proximal Policy Optimization (PPO) algorithm with the Curriculum Learning (CL) training strategy. The developed training program combines a gradual increase in the complexity of the environment (from static obstacles to an environment with moving human agents) and the phased formation of the reward function with the addition of social components. A key feature is the transition between stages, which is based on policy stability analysis. The experimental study was conducted in the developed Gazebo simulation environment using the Turtlebot3 Waffle mobile robot and the ROS 2 Humble framework. Step-by-step training allows an autonomous mobile robot to first learn basic skills for avoiding static obstacles, then dynamic ones, and finally, at the final stage, to take into account social norms of interaction with people. The input data for the system is data from LiDAR, the status of the robot and people, and the target position. The result of the method is an optimized stochastic behavior policy that allows an autonomous mobile robot to make safe, efficient, and socially acceptable navigation decisions. A comparative analysis of the proposed method with the standard PPO algorithm was performed. The results confirm that the proposed method allows the formation of an effective policy of socially adaptive navigation, solving the problems of instability and slow convergence.

Keywords: information technology; machine learning methods; reinforcement learning methods; deep reinforcement learning; curriculum learning; autonomous mobile robots; mobile robot navigation; ROS 2; Gazebo.

REFERENCES

1. Hanenko, L., Storchak, K., Shlianchak, S., Vorokhob, M., & Pitaichuk, M. (2025). SLAM in navigation systems of autonomous mobile robots. *Cybersecurity Providing in Information and Telecommunication Systems 2025*, (3991), 173–182. https://ceur-ws.org/Vol-3991/
2. West, J., Maire, F., Browne, C., & Denman, S. (2020). Improved reinforcement learning with curriculum. *Expert Systems with Applications*, *158*, 113515. https://doi.org/10.1016/j.eswa.2020.113515
3. Uppuluri, B., Patel, A., Mehta, N., Kamath, S., & Chakraborty, P. (2025). CuRLA: Curriculum learning based deep reinforcement learning for autonomous driving. In *17th International Conference on Agents and Artificial Intelligence* (pp. 435–442). SCITEPRESS. https://doi.org/10.5220/0013147000003890
4. Freitag, K., Ceder, K., Laezza, R., Akesson, K., & Haghir Chehreghani, M. (2024). Sample-efficient curriculum reinforcement learning for complex reward functions. *arXiv preprint.* https://doi.org/10.48550/arXiv.2410.16790

5. Li, K., Lu, Y., & Meng, M. Q. H. (2021). Human-aware robot navigation via reinforcement learning with hindsight experience replay and curriculum learning. In *2021 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE. https://doi.org/10.1109/robio54168.2021.9739519
6. Florensa, C., Held, D., Wulfmeier, M., Zhang, M., & Abbeel, P. (2017, October). Reverse curriculum generation for reinforcement learning. In *Conference on Robot Learning* (pp. 482–495). PMLR.
7. Zhu, K., & Zhang, T. (2021). Deep reinforcement learning based mobile robot navigation: A review. *Tsinghua Science and Technology*, *26*(5), 674–691. https://doi.org/10.26599/tst.2021.9010012
8. Gao, J., Ye, W., Guo, J., & Li, Z. (2020). Deep reinforcement learning for indoor mobile robot path planning. *Sensors*, *20*(19), 5493. https://doi.org/10.3390/s20195493
9. Hanenko, L. D., & Zhebka, V. V. (2024). Application of reinforcement learning methods for mobile robot path planning. *Telecommunication and Information Technologies*, *1*, 16–25. https://doi.org/10.31673/2412-4338.2024.011625
10. Soviany, P., Ionescu, R. T., Rota, P., & Sebe, N. (2022). Curriculum learning: A survey. *International Journal of Computer Vision*. https://doi.org/10.1007/s11263-022-01611-x
11. Wang, X., Chen, Y., & Zhu, W. (2021). A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *1*. https://doi.org/10.1109/tpami.2021.3069908
12. Anca, M., Thomas, J. D., Pedamonti, D., Hansen, M., & Studley, M. (2023). Achieving goals using reward shaping and curriculum learning. In *Lecture Notes in Networks and Systems* (pp. 316–331). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-47454-5_24
13. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint.* https://doi.org/10.48550/arXiv.1707.06347
14. Hanenko, L., & Zhebka, V. (2025). Model of socially adaptive navigation of a mobile robot using reinforcement learning methods. *Cybersecurity: Education, Science, Technology, 1*(29). https://doi.org/10.28925/2663-4023.2025.29.907
15. Hanenko, L., & Zhebka, V. (2025). Development of a navigation system for an autonomous mobile robot using ROS 2. *Cybersecurity: Education, Science, Technology*, *4*(28), 498–510. https://doi.org/10.28925/2663-4023.2025.28.824
16. Narvekar, S., Peng, B., Leonetti, M., Sinapov, J., Taylor, M. E., & Stone, P. (2020). Curriculum learning for reinforcement learning domains: A framework and survey. *Journal of Machine Learning Research*, *21*(181), 1–50.