УДК 004.896 DOI: 10.31673/2412-4338.2026.019003

Ганенко Людмила Дмитрівна
аспірантка 4 курсу
Державний університет інформаційно-комунікаційних технологій, Київ, Україна
ORCID 0000-0003-2219-8196
hanenkoliudmyla@gmail.com

МЕТОД АДАПТИВНОГО ФОРМУВАННЯ ВИНАГОРОДИ ЗА УМОВ НЕВИЗНАЧЕНОСТІ ДИНАМІЧНИХ ОБ'ЄКТІВ

Анотація. *У дослідженні обґрунтовано метод адаптивного формування винагороди для навігації автономних мобільних роботів у динамічних соціальних середовищах. Запропонований підхід дозволяє ефективно моделювати поведінку робота в умовах високої невизначеності, створеної непередбачуваним рухом агентів-людей. Актуальність дослідження зумовлена необхідністю безпечної інтеграції автономних мобільних роботів у людський простір. В таких середовищах робот повинен діяти не лише ефективно, а й соціально прийнятно.*

Обмеженням існуючих підходів на основі глибокого навчання з підкріпленням (DRL), є використання функцій винагороди з фіксованими ваговими коефіцієнтами. Такий підхід не дозволяє роботу гнучко адаптуватися до змін середовища. Налаштування на досягнення цілі призводить до підвищеного ризику зіткнень, тоді як пріоритет безпеки часто спричиняє проблему «замороженого робота» та поведінку, яка знижує загальну ефективність системи.

Запропонований підхід розв'язує зазначену проблему шляхом інтеграції архітектури проксимальної оптимізації політики (PPO) з модулем імовірнісного прогнозування. Модуль імовірнісного прогнозування побудовано на основі рекурентної нейронної мережі LSTM, яка кодує часові залежності руху агентів, та мережі суміші густин (MDN), яка дозволяє моделювати мультимодальність людської поведінки. Вихідний шар MDN генерує параметри суміші нормальних розподілів.

Запропоновано механізм динамічно-адаптивного зважування компонентів функції винагороди. Система автоматично регулює баланс між конкурентними цілями. У ситуаціях з високою невизначеністю прогнозу поведінки агентів-людей вагові коефіцієнти безпеки та соціального комфорту нелінійно зростають, змушуючи агента діяти обережніше. І навпаки, коли наміри агентів-людей є більш передбачуваними, система підвищує пріоритет ефективності руху.

Експериментальні дослідження методу підтвердили ефективність запропонованої архітектури.

Ключові слова: *інформаційні технології, моделювання, методи машинного навчання, методи навчання з підкріпленням, автономні мобільні роботи, навігація мобільних роботів.*

Hanenko Liudmyla
4th year postgraduate student
State University of Information and Communication Technologies, Kyiv, Ukraine
ORCID 0000-0003-2219-8196
hanenkoliudmyla@gmail.com

ADAPTIVE REWARD SHAPING METHOD UNDER DYNAMIC OBJECT UNCERTAINTY

Abstract. *The study proposes and substantiates a method of adaptive reward formation for the navigation of autonomous mobile robots in complex dynamic social environments, where the presence of people creates a high level of uncertainty in the socio-dynamic environment. The relevance of the study is determined by the need for the safe integration of autonomous mobile robots into human space, where they must act not only effectively but also in a socially acceptable manner.*

The disadvantage of existing approaches based on deep reinforcement learning (DRL) is the use of reward functions with fixed weight coefficients. This approach does not allow the robot to adapt flexibly to changes in the environment: focusing on achieving the goal leads to an increased risk of collisions, while prioritising safety often causes the problem of a 'frozen robot' and overly conservative behaviour, which reduces the overall efficiency of the system.

The proposed approach solves this problem by integrating the proximal policy optimisation (PPO) architecture with a probabilistic trajectory prediction module. The probabilistic prediction module is based on a recurrent LSTM neural network, which encodes the temporal dependencies of agent movements, and a mixture of density networks (MDN), which allows modelling the multimodality of human behaviour. The output layer of the MDN directly generates the parameters of the mixture of normal distributions.

The proposed mechanism of dynamic adaptive weighting of reward function components. The system automatically adjusts the balance between competing goals: in situations with high uncertainty in predicting the behaviour of human agents, the weighting coefficients of safety and social comfort increase non-linearly, forcing the agent to act more

cautiously. Conversely, when the intentions of human agents are predictable, the system increases the priority of movement efficiency.

Experimental validation of the method confirmed the effectiveness of the proposed architecture.

Keywords: *information technology, machine learning methods, reinforcement learning methods, autonomous mobile robots, mobile robot navigation.*

1. Вступ

Активне впровадження автономних мобільних роботів (АМР) у людський простір є однією з пріоритетних задач сучасної робототехніки та потребує розробки надійних систем навігації. Для успішної інтеграції автономних мобільних роботів у соціальний простір вони повинні не лише результативно виконувати навігаційні завдання, але й демонструвати соціально прийнятну поведінку. Системи керування повинні функціонувати в умовах постійних динамічних змін та миттєво реагувати на непередбачуваність дій людини. Класичні підходи до планування шляху АМР виявляються недостатньо ефективними. Вони здебільшого розглядають людей як рухомі перешкоди, ігноруючи складну стохастичну природу людського руху.

2. Постановка проблеми

Перспективним напрямом розв'язання завдань навігації АМР в соціальному середовищі є застосування методів глибокого навчання з підкріпленням (DRL). Їхня ключова перевага полягає у здатності гнучко адаптуватися до непередбачуваних ситуацій у стохастичному середовищі. Завдяки оптимізації функції винагороди DRL-алгоритми дозволяють агенту засвоювати складні патерни соціальної поведінки.

Постійна зміна намірів людей у реальних соціальних середовищах є причиною високого рівня просторово-часової невизначеності. Використання фіксованих параметрів винагороди у стохастичних умовах призводить до системного конфлікту між безпекою та ефективністю руху АМР. Наприклад, якщо політика керування налаштована на максимально швидке досягнення цілі, робот починає ігнорувати соціальний дискомфорт людини. Натомість надання пріоритету налаштуванням безпеки спричиняє відому проблему «замороженого робота», коли у випадку невизначеності траєкторій людей робот повністю зупиняється.

Таким чином, обмеженням існуючих систем є відсутність механізмів, які б дозволяли DRL-агенту динамічно адаптувати свої дії відповідно до поточного рівня невизначеності середовища. Це зумовлює необхідність розробки методів адаптивного формування винагороди, які дадуть змогу системі в режимі реального часу знаходити оптимальний компроміс між швидкістю переміщення та безпекою людей.

3. Аналіз останніх досліджень і публікацій

Традиційні підходи до планування шляху автономного мобільного робота характеризуються обмеженою ефективністю у складних динамічних сценаріях. Поведінка людини є стохастичною та мультимодальною. З однієї і тієї ж початкової позиції людина може обрати один із декількох різних, але однаково ймовірних шляхів. Наприклад, обхід перешкоди з різних боків. Враховуючи ці обмеження, сучасні дослідження зосереджуються на розробці методів, які здатні моделювати розподіл ймовірностей майбутніх траєкторій людини.

Використання генеративних моделей продемонструвало результативність у вирішенні даної проблеми. Зокрема, в роботі [1] запропоновано архітектуру Social-GAN для генерації набору реалістичних і соціально прийнятних майбутніх траєкторій. Схожий підхід, реалізований на основі умовних варіаційних автоенкодерів (Social-CVAE) [2], моделює невизначеність через прихований простір змінних. Шляхом семплування з цього простору модель генерує траєкторії та ефективно відтворює мультимодальність людських намірів.

Для моделювання складних взаємодій людини та робота активно використовуються графові нейронні мережі як, наприклад, у моделі Social-BiGAT [3]. У моделі розглядають сцену як соціальний граф, де вузли – це люди, а ребра – їхні взаємовпливи. Це дозволяє поширювати соціальну невизначеність майбутніх станів через граф. Рух однієї людини є невизначеним у просторі і часі і впливає на ймовірнісний прогноз для всіх, хто з нею взаємодіє.

На противагу підходам, які зосереджені на явному моделюванні ймовірнісних прогнозів, розвинувся напрямок, який базується на використанні глибокого навчання з підкріпленням (DRL) для формування навігаційної політики. У методі SARL [4] інтегровано механізми соціальної уваги безпосередньо в архітектуру нейронної мережі. Такий підхід дозволив агенту неявно моделювати просторово-часову невизначеність переміщень шляхом динамічного розподілу «уваги». Система

автоматично надає вищий пріоритет тим пішоходам, чия непередбачувана поведінка становить найбільший ризик зіткнення або критично впливає на найближчі навігаційні рішення робота.

Однак ефективне розпізнавання джерел невизначеності середовища є лише першим етапом, наступним етапом є формування оптимальної політики реагування на ці ризики. Саме на цьому етапі важливого значення набувають методи адаптивного формування функції винагороди. Сучасні архітектури здатні в режимі реального часу трансформувати оцінку зовнішнього середовища в адаптивні сигнали керування для робота.

Зокрема, у роботі [5] запропоновано гібридний підхід до навігації робота, в якому процес зважування компонентів функції вартості делегований окремій нейронній мережі, навченій за допомогою навчання з підкріпленням. Замість використання фіксованих, наперед заданих вагових коефіцієнтів для різних цілей система використовує RL-агента як «менеджера пріоритетів». На кожному кроці агент аналізує поточний стан середовища (карти перешкод та пішоходів) і генерує вектор вагових коефіцієнтів, який передається до планувальника траєкторій. Таким чином, робот вчиться динамічно змінювати свою поведінку. Він підвищує ваговий коефіцієнт безпеки при наближенні до людей і ваговий коефіцієнт ефективності, коли шлях вільний.

У дослідженні [6] адаптацію функції винагороди реалізовано через масштабування штрафів на основі кількісної оцінки ризику в реальному часі. Сукупний показник ризику використовується як динамічний множник для безпосереднього посилення штрафу за зіткнення, роблячи його суворішим у небезпечних ситуаціях. Окрім цього, вводиться додатковий штраф за небезпечну зону, який також залежить від показника ризику. Такий підхід дозволяє системі автоматично пріоритезувати безпеку в соціальних середовищах, нелінійно посилюючи відповідні штрафи.

Автори [7], в запропонованому методі AQDR, динамічно змінюють функцію винагороди шляхом масштабування заохочень та штрафів залежно від відстані до цілі.

Незважаючи на значний прогрес у зазначеному дослідницькому напрямі, малодослідженим залишається питання динамічної інтеграції міри невизначеності поведінки агентів-людей безпосередньо в процеси прийняття рішень автономних мобільних роботів (АМР). Це створює об’єктивну потребу в розробці методів, здатних трансформувати ймовірнісні прогнози мультимодальних траєкторій у кількісні параметри адаптивної функції винагороди.

4. Мета і задачі дослідження

Метою дослідження є розробка методу соціально-адаптивної навігації на основі глибокого навчання з підкріпленням, який використовує імовірнісне прогнозування траєкторій людей для динамічної адаптації поведінки робота залежно від рівня просторової невизначеності в середовищі шляхом коригування функції винагороди.

Задачі дослідження: розробити математичну модель оцінки невизначеності майбутніх траєкторій руху агентів-людей та створити метод адаптивного формування функції винагороди, який забезпечує динамічне масштабування компонентів функції винагороди відповідно до міри невизначеності руху агентів-людей.

5. Результати дослідження

Аналіз методів моделювання стохастичної невизначеності траєкторій

Проблема прогнозування руху людей у соціальному середовищі ускладнюється стохастичною природою людської поведінки, оскільки з однієї початкової позиції можлива реалізація кількох вірогідних траєкторій. Для ефективного використання в системі глибокого навчання з підкріпленням (DRL) модуль прогнозування повинен не лише генерувати майбутні координати, але й надавати кількісну оцінку невизначеності для адаптації функції винагороди.

Одним із поширених підходів до вирішення цієї задачі є використання генеративних змагальних мереж (GAN), зокрема архітектури Social-GAN. У даному підході генератор формує правдоподібні траєкторії, а дискримінатор оцінює їх відповідність реальним соціальним нормам поведінки. Перевагою GAN є здатність вирішувати проблему «усереднення» прогнозів, генеруючи чіткі мультимодальні траєкторії. Головним недоліком є збільшене обчислювальне навантаження через необхідність генерації великої кількості семплів для оцінки невизначеності. Крім того, GAN схильні до нестабільності навчання.

Архітектури на основі умовних варіаційних автоенкодерів CVAE, такі як Social-CVAE, моделюють невизначеність через прихований (латентний) простір змінних. Прогноз формується шляхом семплування з латентного розподілу за умови поточного стану спостереження. CVAE забезпечують стабільніше навчання порівняно з GAN і дозволяють явно моделювати розподіл

латентних змінних. Однак для обчислення міри невизначеності потрібне багаторазове застосування декодера з метою генерації набору траєкторій та оцінки їхньої дисперсії. Складний обчислювальний процес може стати причиною виникнення затримок керування роботом у режимі реального часу.

Методи, зокрема Social-BiGAT та підходи на базі механізмів уваги, моделюють сцену як граф взаємодій. Вони демонструють високу ефективність в оцінюванні впливу оточення на невизначеність руху окремого агента. Головною перевагою таких алгоритмів є точність навігації у натовпі, яка досягається завдяки аналізу топології соціальних зв'язків. Водночас їхня значна обчислювальна складність є недоліком для обмежених апаратних ресурсів автономного мобільного робота.

З огляду на зазначені обмеження, для практичної реалізації соціально-адаптивної навігації виникає потреба у забезпеченні високої обчислювальної ефективності. Оскільки DRL-агент функціонує в режимі реального часу, він вимагає безперервного обчислення стану середовища та оновлення функції винагороди з мінімальною затримкою. Використання архітектур, які потребують ітеративного виведення або генерації великої кількості вибірок, призводить до затримок керування і дестабілізує поведінку робота.

На відміну від проаналізованих підходів, архітектура MDN здійснює апроксимацію умовного розподілу ймовірностей у вигляді суміші гауссових розподілів. Головною перевагою цієї мережі є здатність безпосередньо генерувати повний набір параметрів розподілу за один прямий прохід [8]. Така архітектурна особливість здатна забезпечити можливість аналітичного обчислення міри просторової невизначеності U та усунути потребу в ресурсоємному стохастичному семплуванні.

Порівняльний аналіз розглянутих методів моделювання невизначеності для навігації роботів наведено в Таблиці 1.

Таблиця 1

Порівняльний аналіз методів моделювання невизначеності для навігації роботів

Метод	Механізм моделювання невизначеності	Спосіб обчислення метрики невизначеності (U)	Обчислювальна ефективність	Придатність для адаптивної функції винагороди
MDN	Суміш гауссових розподілів (GMM)	Аналітичний	Висока (один прямий прохід мережі)	Висока
Social-GAN	Генеративно-змагальна мережа (GAN)	Емпіричний	Низька (потребує генерації N семплів)	Низька
Social-CVAE	Умовний варіаційний автоенкодер (CVAE)	Емпіричний	Обмежена (залежить від розміру вибірки N)	Середня
Social-BiGAT	Графова мережа уваги (GAT) у поєднанні з GAN	Емпіричний	Низька (складні матричні обчислення на графі)	Низька

Враховуючи необхідність мінімізації затримок та отримання диференційованого сигналу невизначеності для функції винагороди, архітектура MDN є найбільш доцільним вибором для задачі адаптивного керування навігацією.

Метод адаптивного формування винагороди на основі прогнозу невизначеності

Запропонований метод базується на гібридній архітектурі, яка є інтеграцією механізмів оцінювання невизначеності та глибокого навчання з підкріпленням. Структурно система поділяється на два взаємопов'язані модулі: модуль імовірнісного прогнозування на основі архітектури LSTM-MDN та модуль адаптивного планування руху.

Функціонування даної архітектури реалізується у вигляді багатоетапного обчислювального процесу, першим етапом якого є аналіз поточного стану динамічного соціального середовища з подальшою кількісною оцінкою рівня невизначеності руху пішоходів. Реалізація цього завдання забезпечується модулем імовірнісного прогнозування.

Для моделювання стохастичної природи руху людей використано рекурентну нейронну мережу LSTM у поєднанні з мережею змішаної щільності (MDN).

Вхідними даними для моделі є позиції (x, y) просторового розташування та швидкості (V_x, V_y) переміщення агентів-людей у середовищі за фіксований проміжок часу $T_{history}$. Оскільки рух пішоходів

у соціальному середовищі містить високий рівень невизначеності та передбачає наявність кількох варіантів руху, пряме перетворення цих історичних даних на точні координати є неефективним.

На відміну від детермінованих підходів, вихідний шар мережі прогнозує не єдину траєкторію, а параметри суміші K двовимірних нормальних розподілів для кожного часового кроку t у майбутньому горизонті T_p.

Навчання моделі здійснено шляхом мінімізації функції негативної логарифмічної правдоподібності (NLL), яка дозволяє мережі апроксимувати мультимодальність руху агентів-людей

$$L_{NLL} = -\frac{1}{T}\sum_{t=1}^{T} ln\left(\sum_{k=1}^{K} \pi_{k,t} N\left(\mu_{k,t}, \sigma_{k,t}, \rho_{k,t}\right)\right) \quad (1)$$

Структурна схема модуля імовірнісного прогнозування представлена на рис. 1.

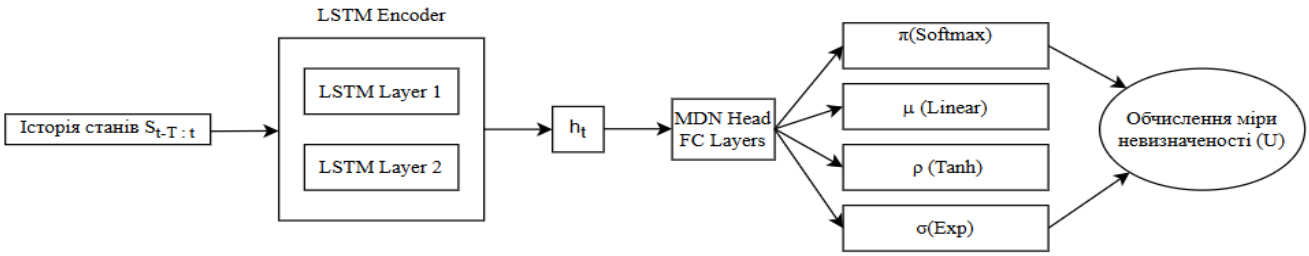

Рис. 1. Структурна схема модуля імовірнісного прогнозування

Визначальною особливістю запропонованого методу є використання вихідних параметрів MDN для розрахунку міри невизначеності середовища в реальному часі.

На кожному кроці для N_{max} агентів розраховується показник U_t за формулою

$$U_t = \frac{1}{N_{max}} \sum_{n=1}^{N_{max}} \sum_{k=1}^{K} \pi_{n,k} \sqrt{(\sigma_{x,n,k})^2 + (\sigma_{y,n,k})^2} \quad (2)$$

Отримане значення U_t є мірою невизначеності сцени: високі значення свідчать про хаотичну поведінку людини, тоді як низькі вказують на високу впевненість моделі у прогнозі. Відповідно, ця метрика використовується як керуючий параметр для процесу прийняття рішень AMP.

Модуль адаптивного планування руху функціонує як DRL-агент і безпосередньо формує навігаційну політику автономного робота. Агент отримує на вхід розширений вектор стану, який формується на основі поточних сенсорних даних. Адаптація до середовища відбувається шляхом динамічного формування функції винагороди під час навчання.

Для вирішення проблеми конфліктуючих цілей безпеки чи ефективності розроблено механізм динамічного балансування вагових коефіцієнтів функції винагороди. Вагові коефіцієнти компонентів винагороди визначаються, залежно від рівня невизначеності U_t, за допомогою формули

$$\omega_{type} = \omega_{base} + \left(\frac{2S}{1 + e^{-k(U_t - U_{mid})}} - S\right), \quad (3)$$

де ω_{base} – базове значення вагового коефіцієнта компонента; S – коефіцієнт масштабу, який визначає діапазон зміни вагового коефіцієнта; k – параметр, який регулює чутливість до змін невизначеності середовищ; U_{mid} – порогове значення невизначеності майбутніх просторових позицій.

Вибір сигмоїдальної функції обґрунтований потребою у нелінійному моделюванні реакції системи. Властивість асимптотичного насичення даної функції на краях області визначення дозволяє мінімізувати вплив екстремальних відхилень, локалізуючи максимальну чутливість моделі в околі точки перегину U_{mid}. Крім того, обмеженість області значень гарантує, що вагові коефіцієнти залишатимуться у заданому цільовому інтервалі.

У реалізованій системі для вагових коефіцієнтів безпеки ω_{safe} та соціального комфорту ω_{social} використовується позитивний коефіцієнт масштабування $S > 0$, який автоматично посилює штрафи у зонах високої невизначеності. Використання значення $S < 0$ для вагового коефіцієнта ω_{eff} знижує пріоритет швидкості руху в небезпечних ситуаціях, змушуючи робота діяти обережніше.

Загальна функція винагороди R розраховується за формулою

$$R = \omega_{eff} \cdot R_{progress} + \omega_{safe} \cdot R_{obstacle} + \omega_{social} \cdot R_{prox} \quad (4)$$

Таким чином, штрафи за порушення соціальної дистанції масштабуються пропорційно до рівня невизначеності руху оточуючих людей. Це стимулює робота приймати соціально безпечні рішення у стохастичних ситуаціях, генеруючи на виході безперервний вектор дій керування.

Вибір запропонованого математичного апарату для обчислення міри невизначеності та адаптивного формування винагороди зумовлений необхідністю вирішення компромісу між обчислювальною ефективністю системи навігації реального часу та стабільністю процесу глибокого навчання з підкріпленням.

Загальну структурну схему системи навігації з інтегрованим оцінюванням невизначеності представлено на рис. 2.

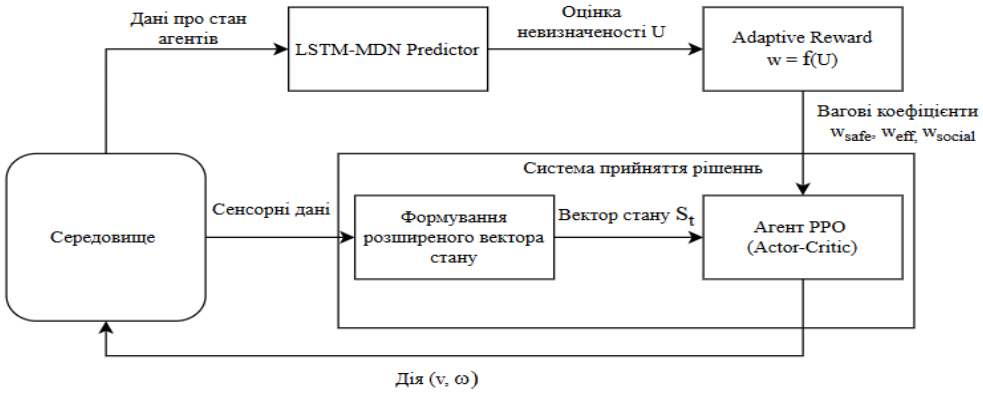

Рис. 2. Загальна структурна схема системи навігації з інтегрованим оцінюванням невизначеності

Експериментальні дослідження. Для перевірки ефективності запропонованого методу проведено серію експериментів у віртуальному середовищі. Мета тестування – кількісне порівняння ефективності та соціальної прийнятності навігаційної політики запропонованого методу з базовим алгоритмом навчання з підкріпленням.

Симуляційна платформа була розгорнута на основі 2D-середовища розробленого з використанням фреймворку Gymnasium. В середовище інтегровано динамічну поведінку агентів-людей, яку було згенеровано на основі моделі соціальних сил (SFM). Було створено імітацію неперервного руху агентів-людей, які здатні активно уникати зіткнень один з одним та зі статичними перешкодами.

В основу розробленого методу покладено раніше запропоновану авторами модель соціально-адаптивної навігації [9]. Для оптимізації процесу навчання застосовано стратегію Curriculum Learning, обґрунтовану авторами в дослідженні [10].

З метою валідації результатів та кількісної оцінки переваг розробленого методу над базовим алгоритмом PPO порівняння моделей здійснювалося на основі просторово-часових та соціальних метрик:

1. Success Rate – відсоток епізодів, в яких мобільний робот успішно досягнув цільової точки;
2. Collisions – відсоток фізичних зіткнень робота з перешкодами;
3. Timeout – кількість епізодів, перерваних через перевищення ліміту часу на виконання завдання;
4. SCS (Social Compliance Score) – комплексна оцінка дотримання соціальних норм. Оцінку SCS задано формулою

$$SCS = \left(0, \left(1 - \frac{W_I \cdot t_{int} + W_P \cdot t_{per} + W_F \cdot t_{frt}}{t_{total}}\right) \cdot 100\right), \quad (5)$$

де t_{int} – час перебування робота в інтимному просторі людини (відстань менше 0.5 м); t_{per} – час перебування робота в особистому просторі людини (відстань 0.5 - 1.0 м); t_{frt} – час руху фронтального зближення; t_{total} – загальний час виконання епізоду; W_I, W_P, W_F – вагові коефіцієнти.

5. Time– середній час руху за епізод;
6. Path – середня довжина шляху за епізод.

Для порівняльного аналізу із запропонованим методом (PPO-Adaptive) було використано базову модель PPO. Зведені результати тестування наведено у Таблиці 2.

Таблиця 2

Зведені результати тестування навігаційних моделей

Метрика	PPO	PPO-Adaptive	Динаміка змін
Success Rate, %	95	98	+3%
Collisions, %	1	1	0
Timeout, %	4	1	-3%
SCS, %	72,6	77,3	+4,7%
Time, с	16,78	8,92	-7,86 с
Path, м	5,88	6,01	+0,13 м

Отримані експериментальні дані свідчать про перевагу гібридної архітектури над базовим алгоритмом. Загальний показник успішності досягнення цілі для PPO-Adaptive становить 98,0%, порівняно з 95,0 % у базовій моделі. Показник метрики соціальної взаємодії SCS збільшився з 72,6% до 77,3%. Крім того, зменшився середній час руху з 16,78 с до 8,92 с. Це свідчить про те, що робот приймає більш оптимальні рішення щодо об'їзду перешкод, не сповільнюючись і не зупиняючись без необхідності.

Незначне збільшення середньої довжини траєкторії руху пояснюється превентивним уникненням динамічних перешкод. Завдяки інтеграції прогнозної моделі LSTM-MDN мобільний робот завчасно ідентифікує ймовірні напрямки руху людей і формує довшу траєкторію. Тим самим AMP уникає прямолінійного наближення до агентів-людей та потенційного зіткнення.

6. Висновки та перспективи подальших досліджень

У дослідженні представлено метод адаптивного формування винагороди для навігації мобільного робота, який базується на імовірнісному прогнозуванні траєкторій людей та глибокому навчанні з підкріпленням. В основі запропонованої архітектури лежить гібридна модель, яка поєднує рекурентну нейронну мережу (LSTM) та мережу суміші густин (MDN). Оцінку просторової невизначеності майбутніх траєкторій безпосередньо інтегровано у функцію винагороди. Політику керування робота оптимізовано за допомогою алгоритму проксимальної оптимізації політики (PPO). Запропонований підхід стимулює мобільного робота діяти безпечніше в умовах високої невизначеності руху людей та ефективніше маневрувати за умови надійності прогнозів.

Експериментальні результати підтверджують, що запропонований метод дозволяє отримати вищі показники безпечної та соціально прийнятної навігації порівняно з методами, які використовують функцію винагороди із фіксованими ваговими коефіцієнтами.

Перспективою подальших досліджень є тестування методу на реальному роботі TurtleBot3 у фізичному середовищі.

Декларація про штучний інтелект

Автор не використовував штучний інтелект при створенні матеріалів статті.

Конфлікт інтересів

Автор заявляє про відсутність конфлікту інтересів та підтверджує, що під час підготовки цієї роботи не існувало жодних комерційних, фінансових чи інших взаємовідносин, які могли б бути розцінені як такі, що здатні вплинути на результати дослідження або їх інтерпретацію. Робота виконана відповідно до принципів академічної доброчесності, етичних норм проведення наукових досліджень та вимог редакційної політики щодо запобігання конфлікту інтересів.

Список використаної літератури

1. Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., & Alahi, A. (2018). Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks. У *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. https://doi.org/10.1109/cvpr.2018.00240

2. Xiang, W., YIN, H., Wang, H., & Jin, X. (2024). SocialCVAE: Predicting Pedestrian Trajectory via Interaction Conditioned Latents. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(6), 6216–6224. https://doi.org/10.1609/aaai.v38i6.28439

3. Kosaraju, V., Sadeghian, A., Martín-Martín, R., Reid, I., Rezatofighi, H., & Savarese, S. (2019). Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. *Advances in neural information processing systems*, *32*. https://proceedings.neurips.cc/paper/2019/file/d09bf41544a3365a46c9077ebb5e35c3-Paper.pdf

4. Li, K., Xu, Y., Wang, J., & Meng, M. Q. H. (2019). SARL*: Deep Reinforcement Learning based Human-Aware Navigation for Mobile Robot in Indoor Environments. У *2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE. https://doi.org/10.1109/robio49542.2019.8961764

5. Cao, M., Xu, X., Yang, Y., Li, J., Jin, T., Wang, P., Hung, T.-Y., Lin, G., & Xie, L. (2025). Learning Dynamic Weight Adjustment for Spatial-Temporal Trajectory Planning in Crowd Navigation. У *2025 IEEE International Conference on Robotics and Automation (ICRA)* (с. 8196–8202). IEEE. https://doi.org/10.1109/icra55743.2025.11128766

6. He, J., Zhao, D., Liu, T., Zou, Q., & Xie, J. (2025). Research on Adaptive Reward Optimization Method for Robot Navigation in Complex Dynamic Environment. *Computers, Materials & Continua*, 1–10. https://doi.org/10.32604/cmc.2025.065205

7. Alshammari, A. B. (2025). Dynamic Rewards in Reinforcement Learning for Robotic Navigation. *Engineering, Technology & Applied Science Research*, *15*(4), 25766–25771. https://doi.org/10.48084/etasr.11986

8. Choi, S., Lee, K., Lim, S., & Oh, S. (2018). Uncertainty-aware learning from demonstration using mixture density networks with sampling-free variance modeling. *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 6915–6922. https://doi.org/10.48550/arXiv.1709.02249

9. Ганенко, Л., & Жебка, В. (2025). Модель соціально-адаптивної навігації мобільного робота з використанням методів навчання з підкріпленням. *Електронне фахове наукове видання «Кібербезпека: освіта, наука, техніка»*, 1(29), 559-570. https://doi.org/10.28925/2663-4023.2025.29.907

10. Ганенко, Л. & Бушма, О. (2025). Метод навчання автономних мобільних роботів на основі DRL та Curriculum Learning. *Електронне фахове наукове видання «Кібербезпека: освіта, наука, техніка»*, *30*(2), 568-582. https://doi.org/10.28925/2663-4023.2025.30.994

References

1. Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., & Alahi, A. (2018). Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks. У *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. https://doi.org/10.1109/cvpr.2018.00240

2. Xiang, W., YIN, H., Wang, H., & Jin, X. (2024). SocialCVAE: Predicting Pedestrian Trajectory via Interaction Conditioned Latents. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(6), 6216–6224. https://doi.org/10.1609/aaai.v38i6.28439

3. Kosaraju, V., Sadeghian, A., Martín-Martín, R., Reid, I., Rezatofighi, H., & Savarese, S. (2019). Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. *Advances in neural information processing systems*, *32*. https://proceedings.neurips.cc/paper/2019/file/d09bf41544a3365a46c9077ebb5e35c3-Paper.pdf

4. Li, K., Xu, Y., Wang, J., & Meng, M. Q. H. (2019). SARL∗: Deep Reinforcement Learning based Human-Aware Navigation for Mobile Robot in Indoor Environments. У *2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE. https://doi.org/10.1109/robio49542.2019.8961764

5. Cao, M., Xu, X., Yang, Y., Li, J., Jin, T., Wang, P., Hung, T.-Y., Lin, G., & Xie, L. (2025). Learning Dynamic Weight Adjustment for Spatial-Temporal Trajectory Planning in Crowd Navigation. У *2025 IEEE International Conference on Robotics and Automation (ICRA)* (с. 8196–8202). IEEE. https://doi.org/10.1109/icra55743.2025.11128766

6. He, J., Zhao, D., Liu, T., Zou, Q., & Xie, J. (2025). Research on Adaptive Reward Optimization Method for Robot Navigation in Complex Dynamic Environment. *Computers, Materials & Continua*, 1–10. https://doi.org/10.32604/cmc.2025.065205

7. Alshammari, A. B. (2025). Dynamic Rewards in Reinforcement Learning for Robotic Navigation. *Engineering, Technology & Applied Science Research*, *15*(4), 25766–25771. https://doi.org/10.48084/etasr.11986

8. Choi, S., Lee, K., Lim, S., & Oh, S. (2018). Uncertainty-aware learning from demonstration using mixture density networks with sampling-free variance modeling. *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 6915–6922. https://doi.org/10.48550/arXiv.1709.02249

9. Hanenko, L., & Zhebka, V. (2025). Model of social-adaptive navigation of mobile robot using reinforcement learning methods. *Electronic professional scientific publication "Cybersecurity: education, science, technology"* , *1* (29), 559–570. https://doi.org/10.28925/2663-4023.2025.29.907

10. Hanenko, L., & Bushma, O. (2025). Method of training autonomous mobile robots based on drl and curriculum learning. *Electronic professional scientific publication "Cybersecurity: education, science, technology"*, *2* (30), 568–582. https://doi.org/10.28925/2663-4023.2025.30.994

Надійшла до редакції: 18.11.25
Прийнята до друку: 17.03.26
Опубліковано: 30.03.26