https://doi.org/10.31891/2219-9365-2024-79-2
УДК 004.8

ВОЛОКИТА Артем
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»
https://orcid.org/0000-0001-9069-5544
e-mail: artem.volokita@kpi.ua
МОРОЗОВ-ЛЕОНОВ Олександр
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»
e-mail: olmorleon@yahoo.com
https://orcid.org/0009-0001-8283-0248

ПІДХОДИ ТА МЕТОДИ ПРИЙНЯТТЯ РІШЕНЬ В СЕРЕДОВИЩАХ ІЗ НЕПОВНОЮ ТА НЕВИЗНАЧЕНОЮ ІНФОРМАЦІЄЮ

В сучасному світі часто виникають задачі, в яких необхідно приймати рішення на основі неповної або невизначеної інформації. У цій статті зроблено огляд деяких сучасних методів та підходів прийняття рішень, наведені їхні сильні та слабкі сторони, особливості застосування та можливості до інтеграції в інші галузі.

Розгляд методів та підходів показав що саме методи, що базуються на навчанні з підкріпленням, є найбільш універсальними, ефективними та мають потенціал для подальшого покращення. Показані також можливі рішення для прийняття рішень із плануванням.

Ключові слова: прийняття рішень, неповна інформація, невизначена інформація, навчання з підкріпленням.

VOLOKYTA Artem, MOROZOV-LEONOV Oleksandr
National Technical University of Ukraine “Ihor Sikorskyi Kyiv Polytechnic Institute”

APPROACHES AND METHODS OF DECISION-MAKING IN ENVIRONMENTS WITH INCOMPLETE AND UNCERTAIN INFORMATION

Artificial intelligence (AI) plays an increasingly important role in decision-making in the modern world, providing effective solutions in various spheres of activity. One of the most important areas of decision-making for the functioning of modern society is the navigation and safety of autonomous vehicles and cyber security.

With the growing complexity and scale of the environments in which these artificial intelligence systems work, new challenges arise: yes, in many cases, autonomous agents must interact with dynamic and multifactorial environments, where an accurate description of all the factors necessary for the agent's work is often impossible. This requires the development and application of special approaches and algorithms capable of adapting to unpredictable conditions and making decisions based on incomplete or inaccurate information.

In the future, the scope and role of using artificial intelligence for decision-making will continue to grow, which will lead to even greater complexity of the environments in which these systems operate. This indicates the great potential of scientific research in this direction.

In the modern world, tasks often arise in which it is necessary to make decisions based on incomplete or uncertain information. This article provides an overview of some modern decision-making methods and approaches, their strengths and weaknesses, application features, and possibilities for integration into other industries.

The review of methods and approaches has shown that methods based on reinforcement learning are the most versatile, effective and have the potential for further improvement. Possible solutions for decision-making with planning are also shown.

The purpose of this paper is to review approaches and methods for solving decision-making problems with incomplete or uncertain information in various fields, to determine the feasibility of using them based on the requirements of the application area, and to analyze their flexibility and versatility.

Keywords: decision making; incomplete information; uncertain information; reinforcement learning.

ПОСТАНОВКА ПРОБЛЕМИ У ЗАГАЛЬНОМУ ВИГЛЯДІ ТА ЇЇ ЗВ’ЯЗОК ІЗ ВАЖЛИВИМИ НАУКОВИМИ ЧИ ПРАКТИЧНИМИ ЗАВДАННЯМИ

Штучний інтелект (ШІ) відіграє все більшу роль у прийнятті рішень у сучасному світі, забезпечує ефективні рішення в різних сферах діяльності. Одними із найбільш важливих сфер прийняття рішень для функціонування сучасного суспільства є навігація та безпека автономних транспортних засобів та кібербезпека.

Зі зростанням складності та масштабів середовищ, у яких працюють ці системи штучного інтелекту, виникають нові виклики: так, в багатьох випадках автономні агенти повинні взаємодіяти із динамічними та багатофакторними середовищами, де точний опис всіх необхідних для роботи агента факторів часто є неможливим. Це вимагає розробки та застосування спеціальних підходів та алгоритмів, здатних адаптуватися до непередбачуваних умов та приймати рішення на основі неповної або неточної інформації.

У майбутньому масштаби та роль використання штучного інтелекту для прийняття рішень будуть зростати і надалі, що призводитиме до ще більшої складності середовищ, у яких діють ці системи. Це свідчить про великий потенціал наукових досліджень у цьому напрямку.

ПОСТАНОВКА ПРОБЛЕМИ

У багатьох реальних ситуаціях, у яких агенти мають приймати рішення, інформація часто буває неповною чи неточною через випадкові збурення у динаміці системи чи через обмежені можливості спостереження за поточним станом або поведінкою інших агентів. Розвиток методів та підходів прийняття рішень в середовищах із таким характером інформації є критично важливим для досягнення надійності та безпеки роботи систем в умовах ризику та невпевненості та мінімізації можливих втрат. У зв'язку із різноманіттям галузей та конкретних рішень існує необхідність в їх огляді, аналізі їх слабких та сильних сторін та особливостей застосування.

АНАЛІЗ ДОСЛІДЖЕНЬ ТА ПУБЛІКАЦІЙ

Методам прийняття рішень із неповною чи невизначеною інформацією присвячені наукові праці [1]-[13]. Стаття [1] приводить класифікацію підходів до вирішення задач прийняття рішень для автономних транспортних засобів, а саме підходи теорії ігор, вірогідніснi підходи, частково спостережувані марківські процеси вирішування (англ. partially observable Markov decision process, POMDP) та підходи, що базуються на навчанні. В роботі [2] розглядається проблематика невизначеності у кібербезпеці. Автори статті [3] розглядають методи наближеного розв'язання задач, сформальованих у POMDP. У статті [4] розглядається адаптація MCTS на неперервні простори дій (англ. continuous action spaces) із станом, що не спостерігається напряму, та поєднання із навчанням з підкріпленням (reinforcement learning, RL) для тактичного прийняття рішень автономними транспортними засобами. У статті [5] розглянуто використання ансамблю нейронних мереж для прийняття рішень в ситуаціях із високою невизначеністю, навіть якщо вони суттєво відрізняються від навчальних даних, та із дотриманням вимог безпеки через перехід до заздалегідь безпечної дії при великій мірі невизначеності.

ВИДІЛЕННЯ НЕДОСЛІДЖЕНИХ ЧАСТИН ЗАГАЛЬНОЇ ПРОБЛЕМИ

В результаті аналізу публікацій показано, що існує потреба у об'єднанні та абстрагуванні методів вирішення задач прийняття рішень від конкретних сфер застосування, виходячи із загальних вимог та обмежень, що притаманні багатьом із них. Існує необхідність в розгляді доцільності використання різних методів та підходів, виходячи із вимог швидкості навчання, швидкодії та точності.

ФОРМУЛЮВАННЯ ЦІЛЕЙ СТАТТІ

Метою роботи є огляд підходів та методів розв'язання задач прийняття рішень при неповній чи невизначеній інформації у різних сферах, визначення доцільності використання виходячи із вимог сфери застосування, та аналіз їх гнучкості і універсальності. Особливу увагу приділено аналізу конкретних покращень та розробок, впроваджених у різних сферах застосування, з метою виявлення універсальних підходів, що можуть бути ефективно адаптовані та використані в інших галузях для підвищення ефективності прийняття рішень в умовах невизначеності.

ВИКЛАД ОСНОВНОГО МАТЕРІАЛУ

Неповнота та невизначеність інформації в середовищах стосується ситуацій, коли агент або система не має доступу до всіх необхідних для прийняття рішення даних, або коли ці дані є неоднозначними або містять суттєві похибки. Наприклад, в системах автопілоту транспортних засобів невідомими є наміри водіїв інших транспортних засобів. Автопілот може адекватно вимірювати положення та швидкість оточуючих транспортних засобів, проте передбачити зміни смуг або різке гальмування інших водіїв він не може. Ще одним прикладом невизначеної інформації можуть бути GPS-дані про положення транспортних засобів. В сфері кібербезпеки прикладами такої інформації можуть бути невідомі вектори атаки та неповна мережева інформація про стан системи, що зазнає кібератаки.

Основним сучасним підходом до формулювання таких задач є застосування математичного апарату POMDP. Існують традиційні методи точного вирішення задач, сформульованих за його допомогою, однак їх застосування не є доцільним для сучасних застосувань штучного інтелекту через обчислювальну складність та конкретні вимоги та обмеження, що визначаються сферою застосування: так, для задач прийняття рішень для автономних транспортних засобів та для захисту від кібератак важливим фактором є швидкість отримання вирішення, хоча б наближеного.

Основною групою методів для розв'язання цих задач є навчання з підкріпленням, зокрема глибоке (deep reinforcement learning, DRL). В ньому агент вчиться приймати рішення шляхом взаємодії із середовищем та отримання зворотнього зв'язку у вигляді винагород. В умовах невизначеності чи неповноти інформації агент поступово набуває досвіду та оптимізує свою стратегію дій, навіть не маючи повної інформації про всі можливі стани середовища або наслідки своїх дій.

За допомогою деяких засобів роботи із неповною та невизначеною інформацією можливе поступове уточнення чи апроксимація початково невідомих факторів середовища. Це дозволяє також використовувати

модифіковані версії більш традиційних методів прийняття рішень у середовищах із повною інформацією у поєднанні із навчанням з підкріпленням.

Приклад такого поєднання наводять автори [4]. В статті розглядається задача тактичного прийняття рішень автономним транспортним засобом на багатополосній трасі серед трафіку. Класична форма алгоритму пошуку по дереву Монте-Карло (англ. Monte Carlo tree search, MCTS) є одним із найбільш часто застосовуваних засобів теорії ігор, проте не підходить для прийняття рішень у середовищі із неповною чи невизначеною інформацією. Автори адаптували метод на неперервні простори дій (англ. continuous action spaces) зі станом, що не спостерігається напряму, та поєднали його із навчанням з підкріпленням, одержаний алгоритм автори назвали MCTS/NN. Пошук по дереву надає можливості для планування, що може бути перерване у будь-який час із приблизним рішенням, а за наявності більшого часу для обчислень результат покращується. Для опрацювання невизначеної інформації щодо станів водіїв оточуючих транспортних засобів використовується частковий фільтр (англ. particle filter), що надає апроксимуючу інформацію про найімовірніші рішення оточуючих водіїв, виходячи із історії їх дій у минулому. Порівняння із класичними методами навігації транспортних засобів на багатополосних трасах та із застосуванням лише компоненту навчання чи компоненту планування ясно вказує на переваги у швидкості переміщення та уникнення аварій при застосуванні розробленого методу. Автори вказують на перевагу MCTS/NN над використання глибоких Q-мереж (Deep Q-Network, DQN) в ефективності використання окремих зразків даних при навчанні завдяки застосуванню пошуку по дереву: для навчання їх потрібно менше, що може бути особливо важливо для застосувань метода у сферах, де неможливо чи недоцільно використовувати комп'ютерну симуляцію, а необхідний саме сбір реальних даних для навчання. Також автори наголошують на гнучності та загальності розробленого методу.

Застосування покращеної версії Double DQN (DDQN) описується в [6]. Автори пропонують новий метод навчання модифікації Double DQN - Rainbow DQN - із більш ефективним використанням навчальних даних, що поєднує прийняття тактичних рішень у динамічному середовищі, дані в якому невизначені та зашумлені, із дотриманням обмежень, що накладаються вимогами безпеки. Отримана реалізація Rainbow DQN перевершує звичайний Double DQN у ефективності, а новітній спосіб використання шару захисту (safety layer) для винагород, названий авторами safety feedback, значно покращує ефективність використання даних при навчанні та продуктивність роботи (рис. 1). Експерименти в симуляції демонструють здатність отриманої реалізації знаходити рішення в ситуаціях, коли необхідне довгострокове планування, та добре справляється із невизначеністю та зашумленістю інформації.

Рис. 1. Ефективність навчання стандартного DDQN та двох версій Rainbow DQN: із safety feedback та без нього
Джерело: [6]

Покращення DQN із використанням методу байєсівського навчання з підкріпленням (Bayesian reinforcement learning) описується в [5]. Пропонується використання ансамблю нейронних мереж із рандомізованими апріорними функціями (randomized prior functions, RPF). Для приблизної оцінки невизначеності використовується коефіцієнт варіації у значеннях Q-функцій, що оцінюються мережами ансамблю. На основі цього введено критерій, що дозволяє визначити ступінь впевненості агента для прийняття певного рішення. В результаті експериментальних порівнянь зі стандартним DQN виявлено, що розроблений метод дозволяє приймати рішення в ситуаціях із високою невизначеністю, навіть якщо вони суттєво відрізняються від навчальних даних. Показано, що розроблений метод добре підходить для вибору безпечних дій в ситуаціях із невизначеністю при проходженні транспортним засобом перехресть (рис. 2). Уникнення зіткнень було одним із важливих завдань при розробці цього методу, і алгоритм обирає апріорно безпечні дії у випадку надто великої невизначеності, що виділяє його серед інших методів і свідчить про перспективність його використання у інших критичних середовищах.

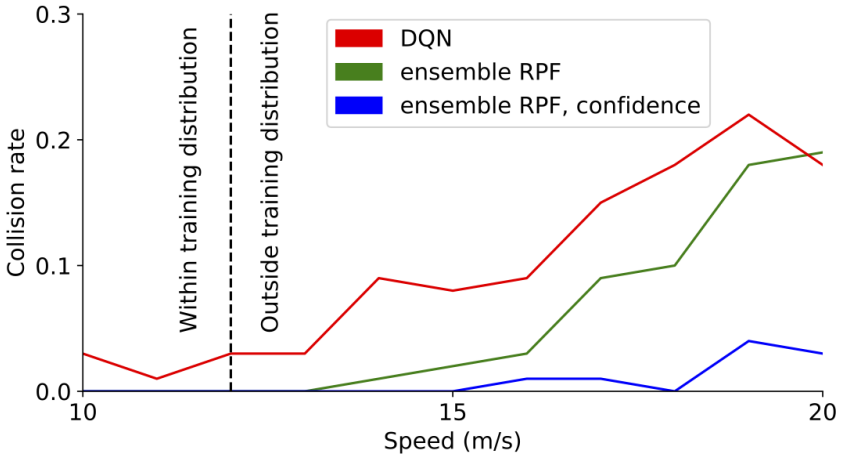

Рис. 2. Частота зіткнень при використанні стандартного DQN, ансамблю мереж із PRF без критерію впевненості та із ним
Джерело: [5]

\Однією із проблем навчання з підкріпленням є дилема exploration-expolitation, автори [7] пропонують новий алгоритм навчання heuristic decaying state entropy (HDSE), що пришвидшує навчання агента RL. Для вирішення проблем невизначеності середовища використано модель future integrated risk assessment model. З поєднанням цієї моделі та long short-term memory (LSTM) для передбачення траєкторій руху транспортних засобів, отримана модель показує кращі результати щодо уникнення зіткнень та покращення ефективності трафіку в модельованих середовищах із великою та малою щільністю трафіку.

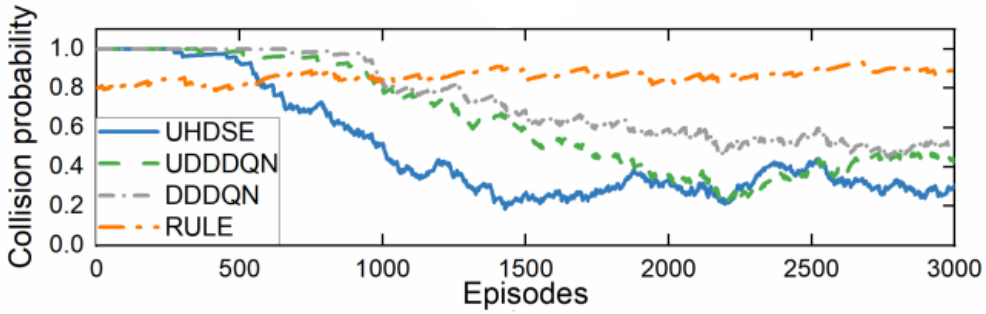

Рис. 3. Частота зіткнень впродовж навчання при використанні агентів RL із новим алгоритмом навчання та future integrated risk assessment model (UHDSE) та зі звичайним навчанням (UDDDQN, DDDQN)
Джерело: [7]

В статті [8] розглядаються задачі прийняття рішень при проходженні транспортним засобом перехрестя. Автори роблять акцент на перевагах POMDP для задач в реальному світі із притаманними йому факторами невизначеності та непередбачуваності, однак також вони звертають увагу на те, що вирішення задач, сформульованих в POMDP, за допомогою RL часто вимагає зберігання великої кількості спостережень, та на те, що така система є обчислювально неефективною при використанні в неперервному просторі дій. Для вирішення цих проблем, пропонується моделювання задачі за допомогою MDP та застосування ієрархічних варіантів (hierarchical options) для навчання з підкріпленням. Такий підхід (Hierarchical Options MDP, HOMDP) дозволяє зберігати за допомогою LSTM лише поточне спостереження. В результаті порівняння із звичайним POMDP виявлено, що розроблений метод показує кращі результати в навігації та уникненні зіткнень (таб. 1).

В сфері безпеки комп'ютерних систем на сьогоднішній день розрізняють одноступеневі та багатоступеневі кібератаки. До одноступеневих атак належать, наприклад, SQL-ін'єкції (SQL injection) чи міжсайтовий скріптинг (Cross-Site Scripting, XSS). Одноступеневі атаки на сьогодні мають вже досить довгу історію супротиву та протидії, та архітектура комп'ютерних систем часто створюється з урахуванням їх загрози: у багатьох випадках одноступенева атака може загрожувати одному окремому компоненту у системі, часто периферійному, і пошкодження такої загрози можуть бути виправлені досить швидко. На відміну від них, багатоступеневі атаки поєднують у собі одноступеневі у послідовніть, що дозволяє зловмисникам заволодіти окремим компонентом у системі і одразу ж продовжити дії, вже виходячи із можливостей скомпрометованого елемента. Такий підхід використовується щоб проникнути до ключових компонентів системи. Для окремих ступенів зловмисники використовують окремі вразливості поточних версій програмного забезпечення, протоколів безпеки тощо. Традиційний підхід - виявлення вразливостей та внесення відповідних виправлень у систему є процесом тривалим та виконується він загалом вручну. Автори [9] розглядають засоби адаптивної кібербезпеки (adaprive cyber defense, ACD) для вирішення цієї

проблеми. Важливим засобом дослідження взаємозалежностей вразливостей та шляху зараження системи є графічні моделі, проте вони носять детерміністичний характер представлення прогресу зловмисника і не підходять для вироблення ефективних шляхів протидії з боку ACD. В розробленому рішенні вводяться байєсівські графи нападу (Bayesian attack graphs, BAGs), що представляють вірогіднісну інформацію про можливе використання тих чи інших шляхів кібератаки. Завдяки цьому, стає можливим використання MDP для формулювання оптимальних стратегій як кібератаки, так і протидії їй, а також використання POMDP у випадках, коли система захисту має доступ лише до підмножини компонентів системи через вади засобів виявлення атаки чи відмову компонентів. Для загальної оцінки захищеності мережі та вибору оптимальних дій для протидії нападу використовується навчання з підкріпленням, а саме Q-навчання. Для вирішення проблеми невідомих вірогідностей переходів між станами у BAG використовується виборка Томпсона (Thompson sampling), що дозволяє оновлювати початкові довільно задані вірогідності із часом та балансує exploration-exploitation навчання з підкріпленням. Було проведено тестування розробленої системи для захисту мережі із 10 комп'ютерів, показано ефективність розробленого методу протидії у симуляціях, що базуються на реальних випадках багатоступеневих кібератак.

Таблиця 1

Порівняння POMDP та HOMDP за метриками частоти успіху, зіткнень, незавершення дії та сукупної винагороди

Задача	Метрика	POMDP	HOMDP
Проїзд прямо	% успіху	97.1	98.3
	% зіткнень	1.7	1.7
	% незавершених	1.2	0.0
	Винагорода	621	873
Правий поворот	% успіху	99.5	99.8
	% зіткнень	0.5	0.2
	% незавершених	0.0	0.0
	Винагорода	892	903
Лівий поворот	% успіху	95.6	97.3
	% зіткнень	2.4	2.6
	% незавершених	2.0	0.1
	Винагорода	213	632

Джерело: таблиця складена на основі даних зі статті [8]

У статті [10] розглядається проблематика прийняття рішень у техніках захисту Moving Target Defense (MTD), що змінюють характеристики програмного забезпечення системи в реальному часі для того, щоб зловмиснику було складніше скористатися можливими вразливостями або для передбачення поведінки зловмисника. Зокрема, зазначається що розробка та впровадження систем із MTD представляють труднощі, що виходять із невизначеності під час роботи систем, а також що існуючі реалізації не враховують факторів невизначеності у параметрах самої моделі та стану системи та обмежені в можливостях до адаптації. Пропонується вирішення проблематики застосування MTD завдяки використанню POMDP та байєсівського навчання (Bayesian learning). Впроваджено байєс-адаптивний POMDP (Bayes-Adaptive POMDP, BA-POMDP), що дозволяє вирішувати проблему невизначеності параметрів моделі. Розроблена реалізація здатна адаптуватись до невизначеності, що випливає із ходу кібератаки та стану компонентів системи, що захищається, та краще координує необхідні засоби та стратегії захисту. Початкові експерименти із застосування розробленої системи показують перспективність цього напрямку, автори вважають, що система може добре масштабуватись та здатна протистояти багатьом векторам атак одночасно, але також зазначається що для достовірних висновків щодо ефективності роботи системи необхідне подальше, більш глибоке експериментальне тестування.

Однією із задач кібербезпеки є введення зловмисника в оману (Cyber Deception, CD) стосовно стану системи, що атакується, для вивчення стратегій, тактик, можливостей та намірів нападника. Часто рішення про першочергові цілі та методі кібератак приймаються зловмисником виходячи із попередніх знань про систему та її інфраструктуру. При застосуванні засобів CD захисник може використати цю особливість для того, щоб змінити вектори атаки на неефективні та затратні за часом, «підкидаючи» нападнику хибні шляхи для атаки. Успішність застосування CD визначається тим, наскільки багато таких шляхів нападник буде змушений виокремлювати та відрізняти від справжніх вразливостей системи. Автори [11] приводять метод такого кібер-захисту для інтернету речей (Internet of Things, IoT). Через динамічну природу мережевих зв'язків саме цей клас комп'ютерних систем може використати CD повною мірою. Розроблена реалізація використовує алгоритм частково спостережуваного планування Монте-Карло (partially observable Monte Carlo planning, POMCP) для представлення невизначеності векторів атаки. Це один із евристичних методів пошуку по дереву для приблизного розв'язання задач, сформульованих в POMDP, що надає задовільні рішення при великому масштабі мережевої структури системи, що захищається. Дії, які виконує система у відповідь на атаки - це створення хибних вузлів системи із заздалегідь заданими характеристиками затримки

та пропускної здатності з'єднання. Розроблена реалізація дозволяє перешкоджати кібератакам, значно зменшуючи їх вплив на систему, та разом із тим надає можливості для вивчення стратегій та засобів зловмисника.

В статті [12] досліджується перспективність застосування підвиду навчання з підкріпленням, табличного Q-навчання (Tabular Q-learning), для створення повністю автономних засобів захисту від кібератак. Цей підвид використовує Q-матрицю, таблицю відповідностей всіх очікуваних винагород та пар «стан-дія», що дозволяє швидко оновлювати значення Q-функції без необхідності в додаткових обчисленнях, проте вимагає додаткового об'єму пам'яті для зберігання. Розглядаючи дилему exploration-explitation, в даному випадку алгоритм віддає перевагу діям, що надають пріорітет пошуку в глибину (exploitation). На відміну від наведених авторами існуючих засобів, що використовують POMDP та враховують невизначеність інформації, цей підхід використовує звичайний MDP. За результатами експериментального порівняння із існуючими реалізаціями систем кібер-захисту, було зроблено висновок про те що розроблене рішення не є оптимальним через відсутність гнучкості: так, окремі конфігурації створеної системи показували результати кращі за деякі аналоги, проте жоден із варіантів конфігурацій не перевершував всі аналоги. Для малих задач POMDP можна отримати приблизне рішення із використанням Tabular Q-Learning, але ефективність цих рішень швидко падає при зростанні складності. Тим не менше, деякі варіанти Tabular Q-Learning можливо використовувати для деяких задач кібер-захисту та специфічних задач прийняття рішень в середовищах із неповною чи невизначеною інформацією в інших галузях.

Автори [13] описують повністю автономний механізм кібер-захисту від динамічних I-DDoS атак великої складності під назвою Horde. Він має розподілену багатоагентну архітектуру для захисту критичних ланок мережі без втручання людини. Цей підхід надає можливості для створення комбінацій обмеження трафіку та перенаправлення в мережі, виходячи із динамічної поведінки зловмисника та спостереження стану мережі. Horde може обчислювати оптимальні послідовності дій для критичного зниження впливу атаки на систему, виявлення стратегій та засобів нападника. Для покращення роботи методів навчання з підкріпленням Horde відкидає неважливі в даний момент секції мережі, тим самим покращуючи ефективність пошуку в ширину (exploration). Для обробки невизначеної іноформації використовуються моделі POMDP, пропонується новітній підхід BRITE Loop, що керує діяльністю захисних агентів так, що вони оновлюють стан власної моделі виходячи зі своїх поточних спостережень та отримують оцінку своїх дій щодо захисту системи. Таким чином, в мережі одночасно знаходиться множина автономних захисних агентів, кожен із яких вивчає власні уявлення про стан системи та вектори атаки та вносить свій вклад в формування загальної стратегії протидії. Розроблена система захисту найкраще придатна до використання у комп'ютерних системах великого масштабу. Недоліком даної системи є слабкість захисту проти атак що використовують досі невідомі вразливості.

Окрім автономних транспортних засобів та кібербезпеки, задачі прийняття рішень у середовищах із неповною та невизначеною інформацією розглядаються та вирішуються і в інших областях застосування (таб. 2).

Таблиця 2

Дослідження прийняття рішень в інших сферах застосування

Сфера застосування	Методи прийняття рішень
Автопілот морських дронів	Deep Deterministic Policy Gradients (DDPG)/DRL ([14]), Proximal Policy Optimization (PPO)/DRL ([15]), POMDP/PPO ([16])
Автопілот літальних апаратів	DDQN ([17]), DRL/MDP ([18]), POMDP/RL ([19])
Організація обслуговування та підтримки систем	POMDP/HMM ([20]), POMDP ([21])
Біржові торги	DRL/Gated Recurrent Unit (GRU) ([22]), Hidden Markov Model (HMM) ([23])

Джерело: розроблено авторами

Отже, POMDP являє собою досить гнучкий та універсальний засіб для опису задач прийняття рішень при неповній чи невизначеній інформації у різноманітних сферах.

Автори [24] пропонують новий швидкий та точний підхід до розв'язання задач POMDP для систем підтримки прийняття рішень, названий авторами Permutable POMDP. Цей підхід вимагає значно менше обчислень для досягнення тієї ж точності та якості вирішення. Тестування показало прискорення обчислень на декілька порядків (рис. 4), що потенційно дозволить розглядати, моделювати та вирішувати задачі із просторами дій значно більшими, ніж ті, що є вирішуваними зі звичайним POMDP. Цей метод вимагає деяких особливих властивостей у структурі POMDP, проте багато задач підтримки прийняття рішень, як вважають автори, можливо апроксимувати як такі, що мають саме таку структуру.

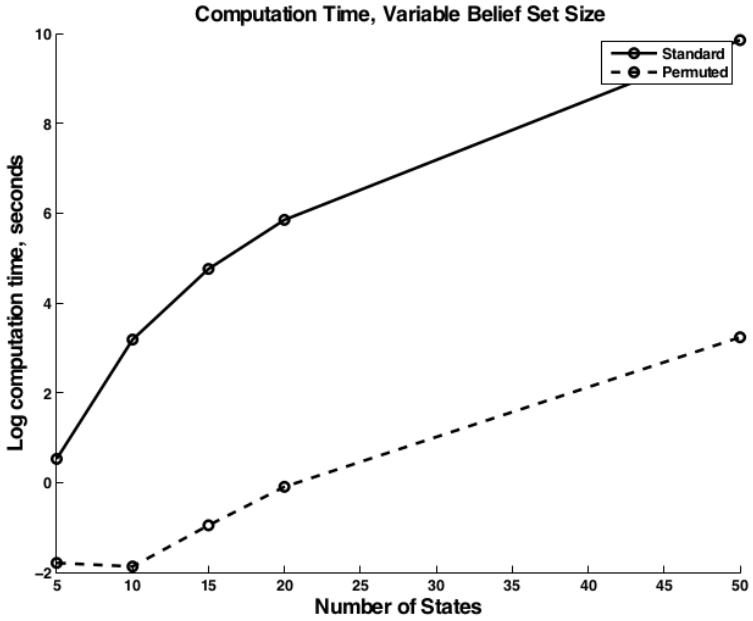

Рис. 4. Залежність логарифму часу обчислень від розміру простору станів для стандартного POMDP та для Permutable POMDP
Джерело: [24]

У статті [25] пропонується новий метод навчання з підкріпленням для розв'язання задач, сформульованих у POMDP, що використовує лише потік винагород та неповні або зашумлені спостереження. Цей підхід, названий авторами deep variational reinforcement learning (DVRL), відрізняється від існуючих методів глибокого навчання з підкріпленням, що використовують рекуррентну нейронну мережу (recurrent neural network, RNN), наприклад, LSTM, для зберігання історії спостережень - таких як deep recurrent Q-network (DRQN). Результати порівняння із ними на деяких комп'ютерних іграх (Mountain Hike, що визначається як неперервна задача контролю із зашумленими спостереженнями, та модифікація ігор Atari із неповною інформацією) показують, що розроблений метод перевершує існуючі методи, що базуються на RNN.

ВИСНОВКИ З ДАНОГО ДОСЛІДЖЕННЯ І ПЕРСПЕКТИВИ ПОДАЛЬШИХ РОЗВІДОК У ДАНОМУ НАПРЯМІ

Огляд досліджень із розробки, розвитку та застосування методів та підходів до прийняття рішень в середовищах із неповною та невизначеною інформацією показав, які методи є ефективними та гнучкими, в якому напрямку варто вести дослідження, та яка проблематика цієї теми.

Навчання з підкріпленням є найпоширенішим підходом до знаходження наближеного розв'язання задачі, сформульованої у POMDP. Розповсюджені методи глибокого навчання, проте звичайний DQN та DDQN поступаються більш новим модифікаціям, що швидше навчаються та показують кращі результати за метриками ефективності та точності. В деяких випадках, коли окрім прийняття рішень у реальному часі доцільне також і планування дій у майбутньому, поєднання навчання з підкріпленням із модифікаціями методів пошуку по дереву демонструє кращі результати. Планування із використанням модифікації MCTS є досить гнучким та добре відповідає вимогам швидкодії. Модифікації POMDP відкривають можливості для розв'язання деяких складних задач, що неможливо зі стандартним POMDP, чи розв'язання тих самих задач із набагато меншою кількістю обчислень, та приймати рішення без зберігання довгої історії минулих спостережень.

Обраний напрямок є перспективним для подальших досліджень, що випливає із активного розвитку автономних транспортних засобів, дронів, розвитку комп'ютерних систем та необхідності у їх захисті. Покращення та нові розробки різних сфер можуть бути застосовані у інших, наприклад, використання планування за допомогою MCTS може застосовуватись для кращого передбачення дій зловмисника при протидії кібератаці. Також варто розглянути, які задачі, що формулюються у POMDP, можливо апроксимувати для застосування Permutable POMDP, що має набагато кращу ефективність.

Література

1. Schwarting W., Alonso-Mora J., Rus D. Planning and Decision-Making for Autonomous Vehicles, Annual Review of Control, Robotics, and Autonomous Systems, Vol. 1 (2018), P. 187-210, 2018. DOI:10.1146/annurev-control-060117-105157

2. Jajodia S., Cybenko G., Liu P., Wang C., Wellman M. Adversarial and Uncertain Reasoning for Adaptive Cyber Defense: Survey. Switzerland: Springer Nature, 2019. 262 c. URL: https://link.springer.com/book/10.1007/978-3-030-30719-6 (дата звернення: 18.08.2024).

3. Bowyer C. M. Approximation Methods for Partially Observed Markov Decision Processes (POMDPs), arXiv:2108.13965, 2021, 59 p. URL: https://arxiv.org/abs/2108.13965 DOI:10.48550/arXiv.2108.13965

4. Hoel C.-J., Driggs-Campbell K., Wolff L., Kochenderfer M. J. Combining Planning and Deep Reinforcement Learning in Tactical Decision Making for Autonomous Driving, IEEE Transactions on Intelligent Vehicles, Vol. 5 no. 2, P. 294-305, 2020. DOI:10.1109/TIV.2019.2955905

5. Hoel C.-J., Tram T., Sjöberg J. Reinforcement Learning with Uncertainty Estimation for Tactical Decision-Making in Intersections, arXiv:2006.09786, 2020, 7 p. URL: https://arxiv.org/abs/2006.09786 DOI:10.48550/arXiv.2006.09786

6. Ugur Yavas M., Kemal Ure N., Kumbasar T. A New Approach for Tactical Decision Making in Lane Changing: Sample Efficient Deep Q Learning with a Safety Feedback Reward, arXiv:2009.11905, 2020, 7 p. URL: https://arxiv.org/abs/2009.11905 DOI:10.48550/arXiv.2009.11905

7. Deng H., Zhao Y., Wang Q., Nguyen A.-T. Deep Reinforcement Learning Based Decision-Making Strategy of Autonomous Vehicle in Highway Uncertain Driving Environments, Automotive Innovation, Vol. 6, P. 438-452, 2023. DOI:10.1007/s42154-023-00231-6

8. Qiao Zh., Mülling K., Dolan J. M., Palanisamy P. POMDP and Hierarchical Options MDP with Continuous Actions for Autonomous Driving at Intersections. In: 2018 IEEE International Conference on Intelligent Transportation Systems (ITSC). IEEE, Maui, USA, p. 2377-2382. DOI:10.1109/ITSC.2018.8569400

9. Hu Zh., Zhu M., Liu P. Adaptive Cyber Defense Against Multi-Stage Attacks Using Learning-Based POMDP, ACM Transactions on Privacy and Security (TOPS), Vol. 24 no. 1, P. 1-25, 2020. DOI:10.1145/3418897

10. Liu R., Tahvildari L. Using POMDP-based Approach to Address Uncertainty-Aware Adaptation for Self-Protecting Software, arXiv:2308.02134, 2023, 7 p. URL: https://arxiv.org/abs/2308.02134 DOI:10.48550/arXiv.2308.02134

11. Al Amin M. A. R., Shetty S., Njilla L. L., Tosh D. K., Kamhoua C. A. Dynamic Cyber Deception Using Partially Observable Monte-Carlo Planning Framework. Modeling and Design of Secure Internet of Things / Edited by Charles A. Kamhoua, Laurent L. Njilla, Alexander Kott and Sachin Shetty. John Wiley & Sons, Inc, 2020. p. 331-355. DOI:10.1002/9781119593386.ch14

12. Applebaum A., Dennler C., Dwyer P., Moskowitz M., Nguyen H., Nichols N., Park N., Rachwalski P., Rau F., Webster A., Wolk M. Bridging Automated to Autonomous Cyber Defense: Foundational Analysis of Tabular Q-Learning. In: AISec'22: Proceedings of the 15th ACM Workshop on Artificial Intelligence and Security. ACM, New York, p. 149-159. DOI:10.1145/3560830.3563732

13. Dutta A. Autonomous Cyber Defense: Formal Models and Applications: дисертація. University of North Carolina, Charlotte, 2021. 178 с.

14. Cui Zh., Guan W., Zhang X. Collision avoidance decision-making strategy for multiple USVs based on Deep Reinforcement Learning algorithm, Ocean Engineering, Vol. 308, 2024. DOI:10.1016/j.oceaneng.2024.118323

15. Guan W., Luo W., Cui Zh. Intelligent decision-making system for multiple marine autonomous surface ships based on deep reinforcement learning, Robotics and Autonomous Systems, Vol. 172, 2024. DOI: 10.1016/j.robot.2023.104587

16. Zheng K., Zhang X., Wang Ch., Zhang M., Cui H. A partially observable multi-ship collision avoidance decision-making model based on deep reinforcement learning, Ocean & Coastal Management, Vol. 242, 2023. DOI:10.1016/j.ocecoaman.2023.106689

17. Han H., Cheng J., Lv M. Interpretable DRL-based Maneuver Decision of UCAV Dogfight, arXiv:2407.01571, 2024, 6 p. URL: https://arxiv.org/abs/2407.01571 DOI:10.48550/arXiv.2407.01571

18. Alvarez L. E., Brittain M. W., Young S. D. Tradeoffs When Considering Deep Reinforcement Learning for Contingency Management in Advanced Air Mobility, arXiv:2407.00197, 2024, 16 p. URL: https://arxiv.org/abs/2407.00197 DOI:10.48550/arXiv.2407.00197

19. Chen Y., Dong Q., Shang X., Wu Zh., Wang J. Multi-UAV Autonomous Path Planning in Reconnaissance Missions Considering Incomplete Information: A Reinforcement Learning Method, Drones, Vol. 7(1) no. 10, 2023. DOI:10.3390/drones7010010

20. Arcieri G., Hoelzl C., Schwery O., Straub D., Papakonstantinou K. G., Chatzi E. Bridging POMDPs and Bayesian decision making for robust maintenance planning under model uncertainty: An application to railway systems, Reliability Engineering & System Safety, Vol. 239, 2023. DOI:10.1016/j.ress.2023.109496

21. Deep A., Zhou Sh., Veeramani D., Chen Y. Partially observable Markov decision process-based optimal maintenance planning with time-dependent observations, European Journal of Operational Research, Vol. 311 no. 2, P. 533-544, 2023. DOI:10.1016/j.ejor.2023.05.022

22. Ansari Y., Yasmin S., Naz Sh., Zaffar H., Ali Z., Moon J., Rho S. A Deep Reinforcement Learning-Based Decision Support System for Automated Stock Market Trading, IEEE Access, Vol. 10, P. 127469-127501, 2022. DOI:10.1109/ACCESS.2022.3226629

23. Zhang L., Li Zh., Xu Y., Li Y. Multi-period mean variance portfolio selection under incomplete information, Applied Stochastic Models in Business and Industry, Vol. 32 no. 6, P. 753-774, 2016. DOI:10.1002/asmb.2191

24. Doshi F., Roy N. The Permutable POMDP: Fast Solutions to POMDPs for Preference Elicitation. In: 7th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2008). Padgham, Parkes, Müller and Parsons (eds.), Estoril, Portugal, p. 493-500. URL: https://groups.csail.mit.edu/rrg/papers/aamas08-fd.pdf (дата звернення: 18.08.2024)

25. Igl M., Zintgraf L., Le T. A., Wood F., Whiteson Sh. Deep Variational Reinforcement Learning for POMDPs. In: Proceedings of the 35th International Conference on Machine Learning. PMLR 80, Stockholm, p. 2117-2126. URL: http://proceedings.mlr.press/v80/igl18a/igl18a.pdf (дата звернення: 18.08.2024)

References

1. W. Schwarting, J. Alonso-Mora, D. Rus, "Planning and Decision-Making for Autonomous Vehicles". *Annual Review of Control, Robotics, and Autonomous Systems*, Vol. 1 (2018), P. 187-210, 2018. DOI:10.1146/annurev-control-060117-105157

2. S. Jajodia, G. Cybenko, P. Liu, C. Wang, M. Wellman, *Adversarial and Uncertain Reasoning for Adaptive Cyber Defense: Survey*. Switzerland: Springer Nature, 2019, 262 p. URL: https://link.springer.com/book/10.1007/978-3-030-30719-6 (Last accessed: 18.08.2024).

3. C. M. Boweyer, "Approximation Methods for Partially Observed Markov Decision Processes (POMDPs)", arXiv:2108.13965, 2021, 59 p. URL: https://arxiv.org/abs/2108.13965 DOI:10.48550/arXiv.2108.13965

4. C.-J. Hoel, K. Driggs-Campbell, L. Wolff, M. J. Kochenderfer, "Combining Planning and Deep Reinforcement Learning in Tactical Decision Making for Autonomous Driving". *IEEE Transactions on Intelligent Vehicles*, Vol. 5 no. 2, P. 294-305, 2020. DOI:10.1109/TIV.2019.2955905

5. C.-J. Hoel, T. Tram, J. Sjöberg, "Reinforcement Learning with Uncertainty Estimation for Tactical Decision-Making in Intersections", arXiv:2006.09786, 2020, 7 p. URL: https://arxiv.org/abs/2006.09786 DOI:10.48550/arXiv.2006.09786

6. M. Ugur Yavas, N. Kemal Ure, T. Kumbasar, "A New Approach for Tactical Decision Making in Lane Changing: Sample Efficient Deep Q Learning with a Safety Feedback Reward", arXiv:2009.11905, 2020, 7 p. URL: https://arxiv.org/abs/2009.11905 DOI:10.48550/arXiv.2009.11905

7. H. Deng, Y. Zhao, Q. Wang, A.-T. Nguyen, "Deep Reinforcement Learning Based Decision-Making Strategy of Autonomous Vehicle in Highway Uncertain Driving Environments". *Automotive Innovation*, Vol. 6, P. 438-452, 2023. DOI:10.1007/s42154-023-00231-6

8. Zh. Qiao, K. Mülling, M. Dolan, P. Palanisamy, "POMDP and Hierarchical Options MDP with Continuous Actions for Autonomous Driving at Intersections", in *2018 IEEE International Conference on Intelligent Transportation Systems (ITSC)*, Maui, 2018, pp. 2377-2382. DOI:10.1109/ITSC.2018.8569400

9. Zh. Hu, M. Zhu, P. Liu, "Adaptive Cyber Defense Against Multi-Stage Attacks Using Learning-Based POMDP", *ACM Transactions on Privacy and Security (TOPS)*, Vol. 24 no. 1, P. 1-25, 2020. DOI:10.1145/3418897

10. R. Liu, L. Tahvildari, "Using POMDP-based Approach to Address Uncertainty-Aware Adaptation for Self-Protecting Software", arXiv:2308.02134, 2023, 7 p. URL: https://arxiv.org/abs/2308.02134 DOI:10.48550/arXiv.2308.02134

11. M. A. R. Al Amin, S. Shetty, L. L. Njilla, D. K. Tosh, C. A. Kamhoua, "Dynamic Cyber Deception Using Partially Observable Monte-Carlo Planning Framework", in *Modeling and Design of Secure Internet of Things*, Charles A. Kamhoua, Laurent L. Njilla, Alexander Kott and Sachin Shetty, Eds. John Wiley & Sons, Inc, 2020. pp. 331-355. DOI:10.1002/9781119593386.ch14

12. A. Applebaum, C. Dennler, P. Dwyer, M. Moskowitz, H. Nguyen, N. Nichols, N. Park, P. Rachwalski, F. Rau, A. Webster, M. Wolk, "Bridging Automated to Autonomous Cyber Defense: Foundational Analysis of Tabular Q-Learning", in *AISec'22: Proceedings of the 15th ACM Workshop on Artificial Intelligence and Security*, New York, pp. 149-149. DOI:10.1145/3560830.3563732

13. A. Dutta, "Autonomous Cyber Defense: Formal Models and Applications", PhD thesis, University of North Carolina, Charlotte, 2021.

14. Zh. Cui, W. Guan, X. Zhang, "Collision avoidance decision-making strategy for multiple USVs based on Deep Reinforcement Learning algorithm". *Ocean Engineering*, Vol. 308, 2024. DOI:10.1016/j.oceaneng.2024.118323

15. W. Guan, W. Luo, Zh. Cui, "Intelligent decision-making system for multiple marine autonomous surface ships based on deep reinforcement learning". *Robotics and Autonomous Systems*, Vol. 172, 2024. DOI: 10.1016/j.robot.2023.104587

16. K. Zheng, X. Zhang, Ch. Wang, M. Zhang, H. Cui, "A partially observable multi-ship collision avoidance decision-making model based on deep reinforcement learning". *Ocean & Coastal Management*, Vol. 242, 2023. DOI:10.1016/j.ocecoaman.2023.106689

17. H. Han, J. Cheng, M. Lv, “Interpretable DRL-based Maneuver Decision of UCAV Dogfight”, arXiv:2407.01571, 2024, 6 p. URL: https://arxiv.org/abs/2407.01571 DOI:10.48550/arXiv.2407.01571

18. L. E. Alvarez, M. W. Brittain, S. D. Young, “Tradeoffs When Considering Deep Reinforcement Learning for Contingency Management in Advanced Air Mobility”, arXiv:2407.00197, 2024, 16 p. URL: https://arxiv.org/abs/2407.00197 DOI:10.48550/arXiv.2407.00197

19. Y. Chen, Q. Dong, X. Shang, Zh. Wu, J. Wang, “Multi-UAV Autonomous Path Planning in Reconnaissance Missions Considering Incomplete Information: A Reinforcement Learning Method”. *Drones,* Vol. 7(1) no. 10, 2023. DOI:10.3390/drones7010010

20. G. Arcieri, C. Hoelzl, O. Schwery, D. Straub, K. G. Papakonstantinou, E. Chatzi, “Bridging POMDPs and Bayesian decision making for robust maintenance planning under model uncertainty: An application to railway systems”. *Reliability Engineering & System Safety*, Vol. 239, 2023. DOI:10.1016/j.ress.2023.109496

21. A. Deep, Sh. Zhou, D. Veeramani, Y. Chen, “Partially observable Markov decision process-based optimal maintenance planning with time-dependent observations”. *European Journal of Operational Research,* Vol. 311 no. 2, P. 533-544, 2023. DOI:10.1016/j.ejor.2023.05.022

22. Y. Ansari, S. Yasmin, Sh. Naz, H. Zaffar, Z. Ali, J. Moon, S. Rho, “A Deep Reinforcement Learning-Based Decision Support System for Automated Stock Market Trading”. *IEEE Access,* Vol. 10, P. 127469-127501, 2022. DOI:10.1109/ACCESS.2022.3226629

23. L. Zhang, Zh. Li, Y. Xu, Y. Li, “Multi-period mean variance portfolio selection under incomplete information”. *Applied Stochastic Models in Business and Industry*, Vol. 32 no. 6, P. 753-774, 2016. DOI:10.1002/asmb.2191

24. F. Doshi, N. Roy, “The Permutable POMDP: Fast Solutions to POMDPs for Preference Elicitation”, in *7th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2008),* Estoril, Portugal, 2008, pp. 493-500. URL: https://groups.csail.mit.edu/rrg/papers/aamas08-fd.pdf (Last accessed: 18.08.2024)

25. M. Igl, L. Zintgraf, T. A. Le, F. Wood, Sh. Whiteson, “Deep Variational Reinforcement Learning for POMDPs”, in *Proceedings of the 35th International Conference on Machine Learning,* Stockholm, 2018, pp. 2117-2126. URL: http://proceedings.mlr.press/v80/igl18a/igl18a.pdf (Last accessed: 18.08.2024)