НАЦІОНАЛЬНА АКАДЕМІЯ НАУК УКРАЇНИ

ІНСТИТУТ ПРОБЛЕМ МОДЕЛЮВАННЯ В ЕНЕРГЕТИЦІ ІМ. Г.Є. ПУХОВА

ЗБІРНИК МАТЕРІАЛІВ

НАУКОВО-ПРАКТИЧНОЇ КОНФЕРЕНЦІЇ «ШТУЧНИЙ ІНТЕЛЕКТ І БЕЗПЕКА»

04 грудня 2025 року

Київ–2025

УДК 004(8+056+413.4)
ББК 32.813
Ш-94

Рекомендовано до друку Вченою радою Інституту проблем моделювання в енергетиці ім. Г.Є. Пухова НАН України (протокол № 15 від 27 листопада 2025 р.)

Ш-94 **Штучний інтелект і безпека,** науково-практична конференція Інституту проблем моделювання в енергетиці ім. Г.Є. Пухова Національної академії наук України, Інституту проблем реєстрації інформації Національної академії наук України : матеріали, 04 грудня 2025 р. Київ : ІПМЕ ім. Г.Є.Пухова НАН України. 271 с.

SH-94 **Artificial intelligence and security**, scientific-practical conference of the G.E. Pukhov Institute for Modeling in Energy Engineering National Academy of Sciences of Ukraine, Institute for Information Recording of the National Academy of Sciences of Ukraine : materials, December 04, 2025. Kyiv: PIMEE NAS of Ukraine, 2025. 271 p.

ОРГАНІЗАТОРИ КОНФЕРЕНЦІЇ

Інституті проблем моделювання в енергетиці ім. Г.Є. Пухова НАН України
(м. Київ)

ПРОГРАМНИЙ КОМІТЕТ

Мохор Володимир Володимирович
член-кореспондент НАН України, доктор технічних наук, професор,
директор Інституту, голова програмного комітету

Чемерис Олександр Анатолійович
доктор технічних наук, старший науковий співробітник
заступник директора з наукової роботи

Чьочь Вікторія Володимирівна
кандидат технічних наук,
заступник директора з науково-технічної роботи

Артемчук Володимир Олександрвич
доктор технічних наук, старший науковий співробітник
заступник директора з науково-організаційної роботи

ОРГАНІЗАЦІЙНИЙ КОМІТЕТ

Артемчук Володимир Олександрвич
доктор технічних наук, старший науковий співробітник
заступник директора з науково-організаційної роботи

Клименко Тетяна Михайлівна
Завідувачка науково-організаційного відділу

Цуркан Оксана Володимирівна
молодший науковий співробітник

В.М. Горбачук, А.О. Камуз, П.В. Ламонов, Д.І. Ніколенко, С.П. Осипенко

НОВА ЕНЕРГЕТИКА ШТУЧНОГО ІНТЕЛЕКТУ

Штучний інтелект (ШІ) видозмінює (трансформує) ринок праці США (обсяг понад 9,4 трлн дол, 151 млн працівників, 32 тис людських навичок станом на 2025 р.) із системними ефектами за межами видимих технологічних секторів [1]. Коли ШІ автоматизує контроль якості на автомобільних заводах, то наслідки такої автоматизації поширюються через логістичні мережі, ланцюги поставок і місцеві економіки послуг [2]. Проте традиційні метрики робочої сили не можуть охоплювати подібні каскадні ефекти, оскільки вимірюють наслідки впровадження ШІ для фактичної зайнятості, а не перетини можливостей ШІ та людей під час такого впровадження [3]. Тому MIT спільно з ORNL започаткував Проєкт Айсберг (Project Iceberg) з метою координації економіки людина-ШІ (Human-AI) за допомогою наукових методологій та основаного на даних (data-driven) аналізу. Інфраструктура Iceberg допомагає координувати взаємодію популяцій людей та ШІ, який сьогодні здатний замість людей читати, писати, творити музику, складати вірші та пісні, робити покупки, виконувати державні функції тощо, тим самим замінюючи людей на ринку праці.

Коли мільйони розумних (smart) штучних інтелектів працюють разом, інтелект постає не від індивідуальних ШІ-агентів, а від протоколів, які їх координують [4]. Проєкт Iceberg вивчає, як загалом ШІ-агенти координують свої дії між собою, а також з людьми. Iceberg імітує Агентні США (Agentic US) − популяцію людина-ШІ, де понад 151 млн модельованих (синтезованих) людей координують свої дії з тисячами ШІ-агентів.

ORNL є світовим лідером у розробці ПЗ і даних для моделювання та симуляції ядерної енергетики, застосуванні спеціально розроблених (custom-developed) і комерційних засобів для міждисциплінарних прикладних сценаріїв, у передових підходах до валідації результатів аналізу даних, у проектуванні експериментів, у регуляторній підтримці та навчанні, особливо в таких галузях, як дані ядерної фізики, нейтроніка, транспорт та екранування (shielding) радіації, теплогідравліка, структурна механіка, продуктивність (performance) палива, безпека реакторів, системна динаміка, приладобудування (instrumentation) та управління, ядерна безпека, відпрацьоване ядерне паливо, інтерфейси користувача [5−7]. Для зазначеного імітування (через симуляційну віртуалізовану скриньку (sandbox)) Агентних США застосовуються суперкомп’ютери ORNL (найпотужніші вітчизняні суперкомп’ютери обслуговує Інститут кібернетики імені В.М. Глушкова НАН України), щоб проаналізувати ключові сценарії: розробка протоколів (Iceberg) для координації ШІ-команд між собою при виконанні конкретних завдань; симуляція каскадних ефектів ШІ-спроможностей поміж тисяч людських навичок в економіці; вимірювання впливу людських навичок і спільнот [8−11]. Уваги заслуговують питання координації роботи команд ШІ-агентів і команд людей при виконанні важливих

соціально-економічних і державних завдань. Очевидно, що подібна комплексна координована робота дозволяє ефективніше виконувати подібні завдання.

Проєкт Iceberg з'ясовує зазначені питання, використовуючи великі моделі населення (Large Population Models (LPMs), аналогічно до LLMs) для імітування ринку праці ШІ-людина, що представляють 151 млн працівників через автономних агентів, які виконують понад 32 тис навичок у 3 тис округах (counties) і взаємодіють з тисячами ШІ-інструментів. Цей проєкт запроваджує Індекс Айсберга (Iceberg Index) − орієнтовану на навички (skills-centered, skills-based) метрику, що вимірює вартість навичок у заробітній платі, які ШІ-системи можуть виконувати в межах кожної професії. Цей індекс охоплює технічний вплив, де ШІ може виконувати професійні завдання, але не охоплює наслідки витіснення (displacement) робочих місць людей чи терміни впровадження ШІ.

Аналіз показує, що видиме впровадження (adoption) ШІ, зосереджене в обчислювальній техніці та технологіях (2,2% від вартості заробітної плати, тобто понад 9400×0,022=207 млрд дол), є лише верхівкою айсберга [12]. Технічна спроможність сягає набагато глибше через когнітивну автоматизацію, що охоплює адміністративні, фінансові та професійні послуги (11,7% від вартості заробітної плати, тобто понад 9400×0,117=1110 млрд дол), в 5,32 рази перевищуючи видиме впровадження та поширюючись від прибрежних хабів (hubs) вглиб усіх штатів. Водночас традиційні індикатори (ВВП, доходи, безробіття) пояснюють на два порядки менше, ніж пропонований Iceberg Index, − менше 5% такої орієнтованої на навички мінливості, підкреслюючи потребу в нових індексах для охоплення впливу ШІ в сучасній економіці інформаційної ери. Імітуючи можливості поширення спроможностей за альтернативних сценаріїв, проєкт Iceberg дозволяє політикам і бізнес-лідерам ідентифікувати точки доступу (hotspots) впливу, пріоритетизувати інвестиції в інфраструктуру та відповідну підготовку кадрів, тестувати втручання перед фактичним фінансуванням впровадження ШІ-інструментів. Проєкт Iceberg побудований за допомогою фреймворку AgentTorch (https://agenttorch.github.io/AgentTorch/).

Федеральний уряд США та уряди окремих штатів виділили і продовжують виділяти значні кошти на підготовку до економіки ШІ [13]. Федеральний план дій з ШІ окреслює 90 політичних позицій і запускає Національний дослідницький хаб робочої сили з ШІ (AI Workforce Research Hub) [14]. Міністерство енергетики США (Department of Energy (DOE)) визначило 16 федеральних місць (sites) для розвитку центрів обробки даних [15]. Окремі штати реагують масштабними ініціативами [16].

4 червня 2025 р. компанія Amazon оголосила про плани інвестувати приблизно 10 млрд дол у Північній Каролині для розширення інфраструктури своїх центрів обробки даних з підтримкою технологій ШІ та хмарних обчислень. Очікується, що ця знакова інвестиція створить щонайменше 500 нових висококваліфікованих робочих місць, а також підтримає тисячі інших робочих місць у ланцюгу поставок центрів обробки даних Amazon Web Services (AWS): одне подібне робоче місце потребує 200 млн дол інвестицій. Генеративний ШІ

стимулює зростання попиту на передову хмарну інфраструктуру й обчислювальну потужність, а ці інвестиції підтримають майбутнє ШІ від центрів обробки даних AWS. Це розгортання передової інфраструктури хмарних обчислень зміцнить інноваційні позиції штату.

18 серпня 2025 р. компанія Google оголосила про перше розгортання вдосконаленого ядерного реактора Kairos Power (електростанція Hermes 2 в Оук-Рідж біля ORNL, штат Теннессі) через нову угоду купівлі електроенергії (power purchase agreement (PPA)) між Kairos Power та комунальною компанією Tennessee Valley Authority (TVA). Компанія Kairos Power була заснована у 2016 р. командою інженерів і вчених для сектору ядерної енергетики малих модульних реакторів (small modular reactors (SMRs)). Поточний проект Kairos Power – це розробка високотемпературного SMR Hermes із фторидним солевим охолодженням (fluoride salt-cooled high-temperature (FHR)). У 2021 р. Kairos Power подала заявку про отримання дозволу на будівництво в Теннессі до Комісії з ядерного регулювання (Nuclear Regulatory Commission (NRC)) США для свого демонстраційного FHR Hermes. Заявку було схвалено у грудні 2023 р.: це схвалення стало першим за 50 років для неводяної (non-water-cooled) АЕС в США. Kairos Power не зверталася до NRC за схваленням проекту для своїх SMRs. У жовтні 2024 р. Google розпочала довгострокову співпрацю з Kairos Power, щоб розблокувати до 500 МВт ядерної енергії для енергосистеми США шляхом багаторазових розгортань SMRs від Kairos Power. Вироблена енергія буде використовуватися для живлення центрів обробки даних Google з ШІ. Google замовила 6−7 SMRs для розгортання протягом 2030−2035 рр. Ці плани очікують схвалення NRC, а також місцевих агентств.

Згадана угода 2025 р. означає першу покупку електроенергії від вдосконаленого реактора GEN IV комунальною компанією США, забезпечуючи 50 МВт ядерної енергії мережею TVA, яка живить центри обробки даних Google в округах Монтгомері (Теннессі) та Джексон (Алабама). Отже, постає тристороннє рішення, де споживачі енергії, комунальні підприємства, розробники технологій співпрацюють над просуванням таких технологій, які здатні допомогти задовольняти зростаючі енергетичні потреби світу за допомогою надійної та доступної потужності.

1. Chopra, A., Bhattacharya, S., Salvador, DA, Paul, A., Wright, T., Garg, A., Ahmad, F., Schwarze, A.C., Raskar, R., Balaprakash, P. (2025). The Iceberg Index: Measuring skills-centered exposure in the AI economy, *Project Iceberg*. Cambridge, MA: MIT, 21 p. https://iceberg.mit.edu.
2. Горбачук, В., Дунаєвський, М., Сулейманов, С.-Б. (2025). Теорія технологічних та економічних трансформацій М. Туган-Барановського. Н.А.Супрун (ред.), *Наукова спадщина Михайла Туган-Барановського як концептуальне підґрунття суспільного розвитку* (29 січня 2025 р., Київ, Україна). Київ: ДУ Інститут економіки та пронозування НАН України. ISBN 978-617-95404-4-8.
3. Горбачук, В.М., Рибачок, Д.О. (2025). Від комп'ютеризації, інформатизації, цифровізації до платформизації освіти і науки, досліджень і розробок. *Інноваційні*

ідеї в економічній науці: пошуки вирішення сучасних проблем (17 листопада 2025 р., Київ, Україна). Київ: НаУКМА,.

4. Chopra, A., Sharma, A., Ahmad, F., Muscariello, L., Pandey, V., Raskar, R. (2025). Ripple Effect Protocol: Coordinating agent populations, *Project Iceberg*. *arXiv*. 2025, October 18, 13 p. https://arxiv.org/pdf/2510.16572.
5. Горбачук, В.М., Бардадим, Т.О., Дунаєвський, М.С., Годлю,к В.В., Голоцукова, Т.Г., Рибачок, Д.О. (2025). Програмне забезпечення для цифрових систем ядерних реакторів. *Наукова конференція Інституту ядерних досліджень НАН України за підсумками 2024 року* (26−30 травня 2025 р., Київ, Україна) (с. 102−103). Київ: Інститут ядерних досліджень НАН України.
6. Gaivoronski, A., Gorbachuk, V., Bardadym, T., Dunaievskyi, M. (2025). Digital transformations of modern energy sciences. In A. Doroshenko, N. Kussul (eds.), 1st Workshop on Software Engineering and Semantic Technologies (SEST) 2025, co-located with the 15th International Scientific and Practical Programming Conference UkrPROG'2025 (May 13–14, 2025, Kyiv, Ukraine). *CEUR Workshop Proceedings*, 4053, pp. 61–71.
7. Kussaiyn-Murat, A. Kadyroldina, A., Krasavin, A., Tolykbayeva, M., Orazova, A., Nazenova, G., Krak, I., Haidegger, T., Alontseva, D. (2025). Application of discrete exterior calculus methods for the path planning of a manipulator performing thermal plasma spraying of coatings, *Sensors*, 25 (3), 708.
8. Gorbachuk, V., Bardadym, T., Dunaievskyi, M. (2025). Digitalization of modern energy in the information society. In A. Ostenda, O. Nestorenko (eds.), *Education, Economy, and AI: Multidisciplinary Perspectives for a Digital Future* (pp. 513−527). Katowice, Poland: University of Technology in Katowice. ISBN 978-83-68422-04-7.
9. Rybachok, D., Hodliuk, V., Kushnir, O. (2025). Nuclear informatics: key technologies and modeling. In A. Ostenda, O. Nestorenko (eds.), *Education, Economy, and AI: Multidisciplinary Perspectives for a Digital Future* (pp. 540−551). Katowice, Poland: University of Technology in Katowice. ISBN 978-83-68422-04-7.
10. Горбачук, В.М., Бардадим, Т.О., Дунаєвський, М.С., Сулейманов, С.-Б., Рибачок, Д.О. (2025). Конкурентоспроможність ядерної енергетики. *Перспективи впровадження інновацій у атомну енергетику* (25–26 вересня 2025 р., Київ, Україна) (с. 76−79). Київ: Українське ядерне товариство. ISBN 978-617-8189-49-5.
11. орбачук, В.М., Бардадим, Т.О., Дунаєвський, М.С., Сулейманов, С.-Б. (2025). Розробка віртуальної АЕС та її застосування в ядерній безпеці. С.В.Литовченко (ред.), *Проблеми сучасної ядерної енергетики* (16–18 квітня 2025 р., м. Харків) (с. 62). Київ: Українське ядерне товариство. ISBN 978-617-8189-40-2.
12. Горбачук, В.М., Батіг, Л.О., Симонов, Д.І. (2021). Інноваційна поведінка на цифрових ринках. О.Л.Гальцова (ред.), *Цифрова економіка як фактор економічного зростання держави* (с. 219–238). Запоріжжя: Класичний приватний університет; Херсон: Гельветика.
13. *Removing barriers to American leadership in Artificial Intelligence* (2025). Executive Order 14179 of January 23, 2025.
14. *Winning the Race. America's AI Action Plan* (2025). Washington, DC: White House; Executive Office of the President of the United States, 2025, 23 p.
15. *Request for information on Artificial Intelligence infrastructure on DOE lands* (2025). Washington, DC: Office of Policy; Department of Energy, 42 p.
16. *Unleashing American energy* (2025). Executive Order 14154 of January 20, 2025.

А.Є. Діденко, А.О. Олійник, С.О. Субботін

ЗАСТОСУВАННЯ МЕТОДІВ ПОКРАЩЕННЯ ЯКОСТІ ЗОБРАЖЕНЬ В СИСТЕМАХ БЕЗПЕКИ ТА ВІДЕОСПОСТЕРЕЖЕННЯ

Системи безпеки та відеоспостереження часто страждають від низької якості зображень та відео через недостатню роздільну здатність камер, шум, погодні умови та ін. Це ускладнює процес аналізу отриманого матеріалу, що в свою чергу критично обмежує здатність системи розпізнавати обличчя, номерні знаки і підозрілу активність. Використання методів відновлення зображень потенційно може значно поліпшити точність систем безпеки.

Методи покращення якості зображень покривають широкий спектр задач, але для систем відеоспостереження можна виділити декілька релевантних сфер: збільшення роздільної здатності, знешумлення, підвищення різкості, підвищення освітленості, а також видаленя артефактів, спричинених погодними умовами. Крім того, ці методи застосовуються не тільки для кольорових зображень, а і для зображень інфрачервоного спектру (для темної пори доби).

Нейромережий підхід є основним способом реалізації методів покращення якості зображень. На сьогоднішній день існує велика кількість різноманітних нейромережевих архітектур (в основному згорткових нейромереж і візуальних трансформерів), створених для вирішення задач, перерахованих вище. Ці підходи можна поділити на дві групи: підходи, які використовують одну нейромережу для однієї задачі [1], і підходи, які використовують одну нейромережу для вирішення спектру задач [2]. Перший підхід дозволяє підвищити якість конкретної задачі, але потребує значної кількості ресурсів при мультизадачності. Другий підхід потребує менше ресурсів і є більш універсальним, але тренування однієї нейромережі під групу задач значно ускладнюється.

Серед переваг застосування методів покращення якості в системах безпеки можна виділити:

- **покращення візуальної якості для подальшої обробки операторами.** Покращуючи параметри зображення (яскравість, роздільну здатність і т.п.) можна значно поліпшити роботу операторів для аналізу даних у реальному часі або на записі;
- **покращення автоматичного розпізнавання.** Багато систем безпеки використовують автоматичні алгоритми для ропзізнавання облич, номерних знаків, об’єктів. Такі алгоритми погано працюють на пошкоджених зображеннях. Це в свою чергу дає підстави на використання методів покращення якості як механізму передобробки;
- **розширення сфер використання.** Методи покращення якості зображень дозволяють використовувати системи безпеки

(відеоспостереження зокрема) в більшому діапазоні ситуацій, наприклад, під час дощу, туману або в темну пору доби [3], без необхідності використання спеціалізованого обладнання;

- **поліпшення прийняття рішень в реальному часі.** Прийняття рішень в реальному часі сильно залежить від якості даних. Але таке поліпшення може призводити до компромісу між якістю і швидкістю.

Основними проблемами програмної реалізації і тренування таких методів є:

- **генералізація методів під реальні умови.** Часто методи покращення якості зображень тренерують на синтетичних датасетах, де погіршене зображення створюється за допомогою додавання випадкового шуму, розмиття, балансу освітлення і т.п. В результаті, модель може добре працювати в "лабораторних умовах", але мати низьку точність на камерах з реальними артефактами;
- **обмежена кількість реальних даних для навчання.** У відкритому доступі наявна дуже обмежена кількість датасетів, зібраних із реальних зображень та відео, через складність отримати пару зображень "оригінальне-пошкоджене". Крім того, такі датасети зазвичай містять невелику кількість даних. Це не дає змогу нейромережам досягти бажаної точності, що змушує дослідників використовувати синтетичні дані;
- **оцінка якості.** Ще одне обмеження полягає в оцінці відновленого зображення. Традиційні метрики, такі як SSIM і PSNR використовують для порівняння пари "ground truth - predicted", але в реальних задачах ground truth зображення часто невідоме. Можна використовувати no-reference метрики, такі як NIQE [4], але ці метрики не говорять про те, як відновлене зображення допомогло при аналізі контенту;
- **швидкодія.** Сучасні state-of-the-art моделі для покращення якості зображень, особливо трансформери, складаються з великої кількості параметрів, що збільшує використання пам'яті та обчислювальних ресурсів. Враховуючи те, що зазвичай системи безпеки повинні працювати в реальному часі, необхідно застосовувати невеликі і легковісні моделі, через що виникає компроміс між швидкістю і якістю.

Також, не менш важливими є ризики використання нейромережевих методів підвищення якості зображення в системах безпеки:

- **артефакти та галюцинації.** Фундаментальна проблема всіх нейромережевих методів підвищення якості зображень (особливо генеративних) – це галюцинації. Наприклад, модель для збільшення роздільної здатності може видати чітке зображення обличчя, але деякі риси (очі, ніс та ін.) можуть бути артефактами із датасету і належати іншій людині. Те саме можна сказати і про номерні знаки автомобіля. Таким чином, надмірне використання покращених зображень може призвести до неправильної ідентифікації [5];

- **юридична відповідальність і правове використання.** Використання зображень, поліпшених нейромережами, особливо якщо воно сильно модифіковане, у якості доказів ускладнюється можливими артефактами.

Отже, методи покращення якості зображень на практиці підтверджують свою ефективність у різноманітних сферах. Очевидно, що ці методи відіграватимуть дедалі важливішу роль і у системах безпеки і відеоспостереження, збільшуючи попит на них. Але, враховуючи юридичні особливості і практичні перешкоди, реалізація і подальше використання таких підходів в системах безпеки потребують підвищеної уваги до точності і правових норм.

1. Su, J., Xu, B., & Yin, H. (2022). A survey of deep learning approaches to image restoration. Neurocomputing, 487, 46–65. doi:10.1016/j.neucom. 2022.02.046.
2. J. Jiang, Z. Zuo, G. Wu, K. Jiang, & X. Liu. (2025). A Survey on All-in-One Image Restoration: Taxonomy, Evaluation and Future Trends. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(12), 11892–11911. doi:10.1109/TPAMI.2025.3598132.
3. Guo, J., Ma, J., García-Fernández, Á. F., Zhang, Y., & Liang, H. (2023). A survey on image enhancement for Low-light images. Heliyon, 9(4), e14558. doi:10.1016/j.heliyon. 2023.e14558.
4. A. Mittal, R. Soundararajan, & A. C. Bovik. (2013). Making a “Completely Blind” Image Quality Analyzer. IEEE Signal Processing Letters, 20(3), 209–212. doi:10.1109/LSP.2012.2227726.
5. Velan, E., Fontani, M., Carrato, S., & Jerian, M. (2022). Does Deep Learning-Based Super-Resolution Help Humans With Face Recognition? Frontiers in Signal Processing, 2–2022. doi:10.3389/frsip.2022.854737.

Р.І. Драгунцов, В.Ю. Зубок

ЗАСТОСУВАННЯ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ ДЛЯ АВТОМАТИЧНОЇ ІДЕНТИФІКАЦІЇ ОКРЕМИХ АТРИБУТІВ ПОВЕРХНІ АТАКИ АКТИВІВ В КІБЕРБЕЗПЕЦІ

В умовах воєнного часу дані щодо динамізму кіберактивів та властивих для них загроз мають важливе значення для підрозділів кібербезпеки в розрізі пріоритезації сповіщень, планування резервування та підтримання процесів інформаційної безпеки інфраструктури за умов погіршення спостережності. Водночас обчислення цих атрибутів для великої кількості різнорідних активів потребує значних людських ресурсів або складної, жорстко формалізованої логіки класичної автоматизації, яка погано масштабується та важко підтримується в динамічному середовищі.

У моделі процесу управління безпековими активами (Security Asset Management, SAM) кожний актив розглядається як кортеж з унікальним ідентифікатором, класом (тип комп'ютерної системи, сервісу чи інформаційного об'єкта) та часово залежним набором атрибутів. Динамізм активу визначається як інтегральний показник чутливості його технічних і мережевих характеристик до втрати електроживлення, каналів зв'язку та топологічних змін; підвищений динамізм корелює зі збільшенням латентності спостереження, ризиком "зникнення" активу з поля зору та зростанням ймовірності пропущених інцидентів [1]. Поняття динамізму можливо формалізувати до (1):

$$D(a) = \frac{1}{m}\sum_{i=1}^{m} d_i(a) \tag{1}$$

де m - кількість атрибутів активу, d - нормований до [0,1] динамізм кожного атрибуту.

Дескриптор загроз є синтетичним атрибутом, що відображає роль активу, його експозицію в мережі, належність до бізнес-сервісів та відображення на тактиках і техніках фреймворку MITRE ATT&CK, і використовується для ризик-орієнтованої пріоритезації подій та планування контролів, що може бути формалізований як (2):

$$\Theta(a) = \big(R(a), E(a), K(a), T(a)\big) \tag{2}$$

де R - роль активу, E - експозиція (ступінь доступності для порушника), K - критичність, T - бінарний вектор релевантності технік з моделі Mitre Att&ck. Асоціація динамізму активу з пов'язаним ризиком будується на основі твердження про зворотню залежність спостережності активів під час

глобальних перебоїв електроживлення та зв’язку від рівня динамізму активу, що відповідно впливає на процеси виявлення та реагування на інциденти.

Для обчислення динамізму та дескрипторів загроз необхідна консолідація інформації з декількох шарів спостереження: мережевого (дані мережевих сканерів, телеметрія про досяжність,історія змін IP-адрес, сегментації), конфігураційного (Configuration Management Database, CMDB, дані про роль, життєвий цикл, відповідальних), шару ідентичностей та облікових записів (служби каталогів, системи централізованої автентифікації), а також спеціалізованих систем інвентаризації хмарних компонентів. Додатково використовуються результати управління вразливостями, журнал змін, географічне положення та зовнішні бази знань щодо загроз і розвідданих (каталоги технік атак, типові ланцюжки компрометації для певних класів активів). Подібні підходи до інтеграції інвентаризаційної інформації є типовими для процесу Asset Management в SOC [2]. Ручна побудова правил, що інтегрують усі ці джерела, є трудомісткою, а зміна інфраструктури вимагає постійного перегляду логіки.

Використання LLM (Large Language Model) в якості засобу інтеграції інвентаризаційної інформації вже знаходило відображення в профільних джерелах [3][4], однак без чіткої орієнтації на обрахунок специфічних, трудомістких атрибутів. Пропонується метод використання великої мовної моделі як центрального інтелектуального компонента, і, на відміну від вже описаних, на основі багатоджерельної інвентаризаційної інформації формує оцінки динамізму та дескрипторів загроз для кожного активу Технологічний підхід реалізує конвеєр із чотирьох основних стадій: збору й нормалізації даних, побудови контексту активу, семантичної інтерпретації за допомогою LLM та постпроцесінгу результатів.

На стадії збору й нормалізації здійснюється автоматизоване злиття даних з мережевих сканерів, CMDB, служб каталогів, систем управління вразливостями та CNAPP у єдину модель активу. Для кожного активу формується структурований запис, що включає: історію змін мережевих реквізитів та стану досяжності, інформацію про рівні резервування електроживлення та зв’язку, часові мітки останньої ідентифікації та останньої успішної ідентифікації, належність до бізнес-сервісу, класифікацію за роллю (критичний сервер, робоча станція користувача, елемент технологічної мережі тощо), типові сценарії використання та пов’язані облікові записи з їхніми правами.

На стадії побудови контексту цей структурований запис перетворюється на стандартизоване текстове або напівструктуроване подання, придатне для інтерпретації LLM. Контекст містить стислий опис інфраструктурної ролі активу, часовий профіль його появи та зникнення в різних джерелах спостереження, дані про географічне розташування та резервування, перелік основних програмних компонентів, а також посилання на релевантні

елементи зовнішньої бази знань щодо загроз (наприклад, типові техніки доступу до подібних систем, відомі ланцюжки атак для даних технологій). Додатково контекст може включати агреговані статистичні показники, такі як частота змін атрибутів чи середня тривалість недоступності.

Роль LLM полягає у семантичній інтерпретації цього контексту з урахуванням зовнішньої бази знань. Модель отримує у підказці (prompt) опис активу, короткі інструкції щодо ранжування рівня динамізму (наприклад, низький, середній, високий, або відповідні кількісні оцінки) та структури дескриптора загроз (перелік найвірогідніших тактик і технік, очікуваний вектор атак, роль активу у ланцюжку компрометації), а також приклади вже верифікованих оцінок для типових активів. LLM оцінює, наскільки актив схильний до частих змін стану та топології, і які класи атак для нього є найбільш релевантними, аналізуючи відповідний наданий контекст та доступ до референтних джерел інформації за допомогою Retrieval Augmented Generation (RAG). Результатом роботи моделі є структурована відповідь із вказанням категорії динамізму, переліку релевантних технік за MITRE ATT&CK, коротким поясненням (раціоналізацією) та допоміжними ознаками, що вплинули на рішення.

На стадії постпроцесінгу відповідь LLM перетворюється на формалізовані атрибути, прийнятні для подальшої автоматизованої обробки та використання в інвентаризації. Для динамізму активу застосовується проекція на заздалегідь визначену шкалу з фіксованими значеннями, що може доповнюватись числовим індексом, обчисленим, наприклад, як функція частоти змін і тривалості відмов. Для дескриптора загроз формується кортеж, до якого входять: перелік технік, ключові вектори атак, рівень очікуваного впливу, а також агрегований індикатор критичності профілю загроз. На цьому етапі застосовується стандартна автоматизація на рівні детермінованої логіки та програмного коду, накладаються обмеження (наприклад, допустимі значення шкал, максимальна кількість технік), що мінімізують варіативність відповідей моделі.

Запропонований метод передбачає ряд контрольних заходів, спрямованих на обмеження похибок, властивих великим мовним моделям [5]. Усі атрибути, обраховані LLM зберігаються разом із даними, що використовувались для формування контексту.Це дозволяє здійснювати вибірковий експертний перегляд та визначення коректності роботи моделі. Експертний перегляд здійснюється на регулярній основі для невеликих кількостей активів, обраних рандомізовано. Для контролю якості ідентифікації дескриптора загроз застосовується евристична модель - порогові правила за частотою виявлення, наявністю резервування, IT-сервісу - яка виконує функцію референсної лінії для виявлення аномальних відхилень. Аналіз таких відхилень дозволяє як виявити очевидні помилки в роботі LLM, так і навпаки, ідентифікувати найбільш ризикові активи.

До основних обмежень методу належать залежність якості оцінок від повноти та актуальності даних інвентаризації, ризик некоректної генерації у разі нечітких або суперечливих вхідних атрибутів [6]. Практична реалізація методу потребує можливості швидкого ручного коригування атрибутів під час оперативної роботи SOC.

1. Drahuntsov, R., & Zubok, V. (2025). DYNAMIC INFRASTRUCTURE COMPONENTS AND SYSTEM VISIBILITY DURING CYBERSECURITY INCIDENT RESPONSE. Cybersecurity: Education, Science, Technique, 1(29), 493–507. https://doi.org/10.28925/2663-4023.2025.29.891.
2. Coppolino, L., Iannaccone, A., Nardone, R., & Petruolo, A. (2025). Asset Discovery in Critical Infrastructures: An LLM-Based Approach. Electronics, 14(16), 3267. https://doi.org/10.3390/electronics14163267.
3. Hasanov, I., Virtanen, S., Hakkala, A., & Isoaho, J. (2024). Application of Large Language Models in Cybersecurity: a Systematic Literature Review. IEEE Access, 1. https://doi.org/10.1109/access.2024.3505983.
4. Jaffal, N. O., Alkhanafseh, M., & Mohaisen, D. (2025). Large Language Models in Cybersecurity: A Survey of Applications, Vulnerabilities, and Defense Techniques. AI, 6(9), 216. https://doi.org/10.3390/ai6090216.
5. King, R., & Sellers, T. (2024, 25 липня). AI in CAASM: The Risks of LLM data in security-critical workflows. https://www.runzero.com/blog/ai-caasm-llm-security. https://www.runzero.com/blog/ai-caasm-llm-security.

R.Yu. Ivanova

ARTIFICIAL INTELLIGENCE IN FINANCIAL LAW: LEGAL SECURITY AND INTERNATIONAL COOPERATION

AI's incorporation into virtually every sector of global governance, policy, and the economy has changed the world, and the concept of automation and algorithmic decision-making has moved to the forefront of attention. In the context of financial law, the use of AI to detect transactional anomalies, forecast credit risk, automate the generation of regulatory reports, and manage compliance with the AML and CTF regulations has become common practice. These benefits, however, come with additional problems, such as the need to address concerns regarding personal data protection, algorithmic transparency and accountability, fault liability, and the general perils of the cyberspace.

In the financial sector, this paper considers the implications of AI as both a technology and a branch of law. The author believes that as more AI technology becomes integrated into financial law and regulations, the traditional legal concepts of responsibility, fault, and intent need to be rethought. The challenge arises with the fact that damage, be it financial or reputational, is done by autonomous systems that function without the immediate control of a human.

Attention is given to the international legal cooperation in the area of AI-based financial governance. The International Monetary Fund (IMF) Transdigitalization advocates the digitalization of financial supervision and the TransMonetary Resilience frameworks. The Financial Action Task Force (FATF) sets the standards for the application of AI to identify money laundering. The Organisation for Economic Co-operation and Development (OECD) developed principles for Responsible Artificial Intelligence which are centered on human fairness and transparency. Together, these frameworks are the foundation for protecting the financial system on the global scale within the digital ecosystem.

Within Europe, the Artificial Intelligence Act (2024) proposes a risk-based approach to regulation, which involves categorizing AI systems by the risk they pose for the infringement of an individual’s safety and rights. One of the examples of a high-risk system is credit scoring. Such systems must satisfy certain standards regarding quality of data, human control, and record-keeping. From the perspective of building the EU single market, trust of investors in Ukraine is largely determined by the convergence of its legislation with this regulation.

The financial sector of Ukraine, especially the Reconstruction after the War, is beginning to utilize algorithmic technologies to automate the monitoring of finance, customs, and the growing digital banking services. The author states that the application of AI in these technologies must be balanced, and fences must be put in place to mitigate potential ethical abuses and discrimination, such as having institutions propose Ethics Independent Oversight and Transparent Audits. The primary national regulators, the National Bank of Ukraine and the State Financial

Monitoring Service, need to propose and discuss the policy parameters and draft model documents to be bulwark in the development of AI technologies.

The application of Artificial Intelligence in the field of financial legislation and public administration is very promising, but it needs a design coherent from the legal point of view the determines the liability and the protection against the misuse. The control of AI applications in financial legislation should be based on: The Consistency of national laws with International legal (soft law) standards; The Collaborativeness of institutions (Regulators, Financial entities, and Technological providers); The Ethics of control (ensure human agency, and social control). The application of these principles will help Ukraine strengthen the digital sovereignty, develop confidence in financial institutions and create a secure human-centric global financial system.

1. European Commission. (2024). Artificial Intelligence Act: Regulation (EU) 2024/1689. Official Journal of the European Union.
2. Financial Action Task Force (FATF). (2023). Opportunities and Challenges of New Technologies for AML/CFT Compliance. Paris: FATF/OECD.
3. Organisation for Economic Co-operation and Development (OECD). (2019). Recommendation of the Council on Artificial Intelligence. OECD Legal Instruments, OECD/LEGAL/0449.
4. World Bank. (2022). Harnessing Artificial Intelligence for Financial Inclusion: Policy and Regulatory Perspectives. Washington, DC: World Bank Group.
5. UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence. Paris: UNESCO Publishing.
6. Khairullina, N., & Pakhomov, A. (2024). AI Ethics and Legal Accountability in Financial Technology. Journal of Digital Law and Governance, 12(2), 45–59. https://doi.org/10.5281/zenodo.11457892.

Л.О. Митько

ДЕЯКІ АСПЕКТИ ЗАЛУЧЕННЯ ШТУЧНОГО ІНТЕЛЕКТУ ДЛЯ СТАБІЛЬНОЇ РОБОТИ ЕНЕРГЕТИЧНОЇ ІНФРАСТРУКТУРИ УКРАЇНИ

Військові дії в Україні вимагають дієвих рішень в енергетичній галузі України. Різке зменшення виробництва енергії на теплових електростанціях, ускладнення процесів у розподілених мережах і зростання інтеграції відновлюваних джерел енергії невеликої потужності вимагає нових підходів до управління енергосистемами. Важливим для енергосистем є прогнозування навантаження в кожний момент часу, оскільки правильно передбачений попит електроенергії дозволяє оптимально запланувати генерацію, уникнути надлишкового резерву та знизити ризик відключень.

З кожним днем в енергетичній галузі розширюється заміна традиційних методів розрахунків та планувань на інтелектуальні алгоритми. Такі алгоритми надають можливість здійснювати аналіз величезних масивів даних та формувати рішення за більш короткий час і з більшою точністю в порівнянні з минулим. Коли використовуються інтелектуальні алгоритми , то в цьому випадку йде мова про застосування штучного інтелекту (ШІ) [1]. За допомогою інтелектуальних алгоритмів можна проводити аналіз значних обсягів даних, враховувати споживання та місцезнаходження об'єктів. При цьому з’являється новий комплексний підхід, який полягає у створенні та використанні кількох сценаріїв у разі непередбачених подій, в тому числі, пов’язаних з військовими діями. Використовуючи ШІ, а саме задіянні методу аналізу великих даних можна побачити відхилення від нормального функціонування об’єктів, такі як коливання напруги або незвичні відхилення у роботі устаткування. ШІ може здійснювати постійний моніторинг та управління енергомережами. При цьому вся система повинна працювати на зразок команди роботів, де окремий елемент (трансформатор, батарея, сонячна панель, вітряк, тощо) приймає самостійні рішення, узгоджуючи їх з іншими елементами. Такий підхід нагадує схожість з функціонуванням розумного будинку, в якому за допомогою пристроїв проходить управління роботою всієї мережі будинку з метою уникнення аварійних ситуацій та економного використання ресурсів. В майбутньому використання ШІ в енергетичній галузі повинно забезпечити створення розумних мереж, які будуть вирішувати проблеми, що виникають при функціонуванні енергосистем, дозволити економити ресурси та бути менш залежними від дій персоналу.

Ситуація, що склалась в Україні в зв’язку з військовими діями вимагає постійних змін в процесі функціонування енергосистем. Замість сценаріїв, що прораховуються наявними засобами потрібні генеративні моделі, тобто потрібне задіяння таких типів алгоритмів машинного навчання, які здатні створювати нові, оригінальні контенти, що подібні до того контенту, на якому проходило навчання і при цьому великі нейромережі матимуть змогу перебирати тисячі варіантів. На сьогодні компанія IBM пропонує концепцію

GridFM - foundation-модель для енергомережі, яка навчена на сотнях тисяч рішень оптимізаційних завдань енергосистеми [2]. Наявні версії моделі GridFM спрямовані на розв'язування задач оптимального навантажувального потоку для різних мереж, щоб можна було швидко оцінювати стан мережі та рекомендувати керуючі дії. Результатом використання таких моделей та ШІ-асистентів буде допомога операторам швидше прийняти рішення на основі більш повної інформації. На сьогодні існують прототипи систем, які в режимі реального часу пропонують диспетчеру оптимальні перемикання чи перестановки генерації на основі прогнозу стану мережі на кілька годин наперед.

Важливою складовою енергетичних систем стають безпілотні літальні апарати (БпЛА), що являють собою як інструмент підвищення безпеки в енергетиці так і загрозу її функціонування. Їх залучають для проведення аерофотозйомки та тепловізійного контролю високовольтних ліній, для збору даних з метою аналізу пошкоджень та складанню звітів, що сприяє підвищенню ефективності роботи об'єктів енергетики. Життєво необхідним для української енергосистеми є захист енергетичних об'єктів від БпЛА. Сучасні БпЛА, в тому числі і FPV-платформи, мають можливість атакувати невеликі об'єкти, здійснювати маневри в обмежених просторах та міняти вектори підльотів. Така ситуація з БпЛА вимагає перехоплення траєкторії фізичними методами, такими як: перекриття ймовірних коридорів, екранування вузлів та створення двох контурів захисту там, де потрібний максимальний захист. Отже для захисту енергетичних об'єктів від загроз, що створюються БпЛА, потрібно об'єднання декількох складових, в саме: наявність зенітно-ракетних комплексів, мобільних вогневих груп та інших засобів ППО навколо критичних об'єктів для фізичного знищення повітряних цілей; наявність систем для виявлення, чатування і знешкодження каналів зв'язку та навігації БпЛА (GPS-глушіння), що примушує їх відхилятися від енергетичних об'єктів або падати; наявність інженерних загороджень, наприклад, сітки, аеростати загородження, тощо, які є фізичними перешкодами на шляху БпЛА; наявність інтегрованих платформ, що мають можливість об'єднувати дані з різних датчиків (радарів, тепловізорів, акустичних систем) з метою раннього виявлення та ідентифікації БпЛА; наявність БпЛА-перехоплювачів, що можуть нейтралізувати ворожі безпілотники ще у повітрі. Ефективне управління енергосистемами та їх захист від БпЛА вимагає постійного вдосконалення технологій та адаптації до нових загроз, яке неможливо здійснювати без залучення ШІ, тому що методи атак, варіанти фізичних впливів на енергетичні об'єкти зі сторони противника мають тенденцію до розширення та еволюції.

1. Митько Л.О. Кібербезпека в енергетиці на тлі швидкого розвитку штучного інтелекту./ Електронне моделювання. 2024. Т. 46. № 1. С. 70-77. https://doi.org/10.15407/emodel.46.01.070.
2. https://arxiv.org/abs/2407.09434 (date of access: 19.11.2025).

Д.В. Глушкова, А.А. Рось

ШТУЧНИЙ ІНТЕЛЕКТ У ПЕРІОД ВІЙНИ: ІНТЕЛЕКТУАЛЬНІ МОДЕЛІ ПРОГНОЗУВАННЯ ЗАГРОЗ ТА СИСТЕМИ ПІДТРИМКИ РІШЕНЬ ДЛЯ СЕКТОРА НАЦІОНАЛЬНОЇ БЕЗПЕКИ

У XXI столітті війна перетворилася на багатовимірне протистояння, у якому інформація, швидкість обробки даних і точність прогнозування відіграють не меншу роль, ніж кількість техніки чи особового складу. Як зазначено у вихідному документі, сучасні конфлікти стали «війною алгоритмів», і саме штучний інтелект (ШІ) дедалі більше визначає характер ведення бойових дій, прийняття рішень та стратегічне планування. Умови повномасштабної агресії Російської Федерації проти України спричинили безпрецедентне прискорення цифрової трансформації сектору безпеки та оборони. Україна стала одним зі світових лідерів у впровадженні алгоритмів ШІ для розвідки, кіберзахисту, аналізу даних, керування операціями, автоматизації роботи з великими масивами інформації та прогнозування ворожих дій. Це не просто технологічна еволюція — це поява нового рівня мислення у військовій сфері, коли людські рішення доповнюються потужністю машинної аналітики, здатної обробити те, що раніше було неможливим.

Починаючи з 2014 року, Україна зіткнулася з необхідністю швидко аналізувати великі обсяги розвідувальної інформації: відео з дронів, супутникові знімки, радіоперехоплення, дані з камер спостереження, повідомлення підрозділів, телеметрію артилерійських систем, інформацію про пересування техніки та живої сили противника. У перші роки війни ці дані оброблялися переважно вручну, що обмежувало швидкість, створювало ризики помилки й не відповідало масштабу загроз. Після 2022 року ситуація змінилася кардинально. Створення цифрової інфраструктури «Армія дронів», інтеграція програм «Дія.Цифрова армія», поява високотехнологічних проєктів у співпраці з державою, ІТ-бізнесом і волонтерськими фондами дали можливість перейти від реагування до випередження. Саме тоді ШІ почав виконувати ключові функції: автоматичне розпізнавання техніки, класифікацію цілей, аналіз траєкторій руху, прогнозування артилерійських ударів, оптимізацію логістики й ухвалення оперативних рішень — у режимі майже реального часу 1.

Одним із перших успішних українських рішень стали алгоритми, здатні аналізувати відео з безпілотників і розпізнавати російську бронетехніку за характерними ознаками корпусу, тепловим слідом та поведінкою на місцевості. Такі системи використовуються вже з 2023 року на передових напрямках, забезпечуючи навідників артилерії актуальними даними про ціль. Інший приклад — системи прогнозування артилерійських обстрілів, що на основі навчання на реальних бойових даних можуть передбачати ймовірність

удару, його напрямок і приблизний час. В окремих випадках точність прогнозу перевищує 80 %, що дозволяє своєчасно евакуювати цивільних або зміцнювати критичні ділянки оборони.

Крім того, українські інженери та військові аналітики створили системи, що автоматично аналізують радіоелектронний простір. Алгоритми виявляють активність станцій РЕБ, визначають місця роботи ворожих радарів, аналізують зміни інтенсивності сигналів і попереджають підрозділи про зміну стратегії противника. З 2024 року подібні інструменти інтегровані у структури Сил оборони та використовуються під час планування наступальних і оборонних дій. Завдяки цьому командири отримують не просто розрізнені дані, а комплексну картину операції, побудовану на базі машинного аналізу, що значно скорочує час ухвалення рішень 2.

Важливою складовою розвитку воєнного ШІ стала співпраця держави, ІТ-компаній і волонтерських фондів. Саме волонтери першими створили прототипи автоматизованих систем розпізнавання цілей, а українські технологічні компанії запропонували інноваційні рішення зі сфер комп'ютерного зору, аналізу великих даних та побудови нейронних моделей для прогнозів. У 2023–2025 роках такі інструменти почали інтегруватися у державні системи управління військовими операціями, створюючи єдиний цифровий простір безпеки.

На стратегічному рівні ШІ забезпечує колосальні можливості для моделювання майбутніх загроз. Аналітичні моделі можуть враховувати сотні факторів: інтенсивність боїв, логістику, погодні умови, темпи виробництва озброєнь противником, особливості його тактики, маршрути пересування техніки, супутникові знімки та сигнали розвідки. Подібні системи дозволяють прогнозувати сценарії розвитку подій: від локальних атак до масштабних операцій. Наприклад, у 2024–2025 роках українські фахівці розробили моделі, які допомагають визначити, чи може противник здійснити наступ у конкретному районі, з якою швидкістю він здатен перекидати резерви та де саме існує ризик прориву. Такий підхід дозволяє Збройним Силам діяти на випередження, концентрувати ресурси в потрібних місцях і мінімізувати втрати 3.

Кібербезпека — ще одна сфера, де штучний інтелект став критично важливим. Масштабні кібератаки, спрямовані на українські енергетичні компанії, державні реєстри, фінансову систему та військові мережі, вимагають систем, здатних виявляти вторгнення за лічені секунди. Алгоритми, які аналізують мільйони подій у мережі та визначають аномальні патерни, стали основою кібероборони країни. Такі системи можуть передбачити атаку ще до того, як вона фактично розпочнеться, блокуючи її або зменшуючи збитки. У 2023–2025 роках за допомогою ШІ було відбито сотні спроб проникнення у державні бази даних і системи критичної інфраструктури 4.

Окремої уваги заслуговує етичний та правовий вимір застосування ШІ під час війни. Виникають питання відповідальності за рішення, ухвалені на основі алгоритмів, визначення меж автономності бойових систем, дотримання міжнародного гуманітарного права й запобігання неконтрольованому використанню автономних ударних платформ. Україна, інтегруючи технології ШІ, водночас бере до уваги міжнародні стандарти та ризики, розробляючи внутрішні протоколи використання інтелектуальних систем у бойових умовах.

Сучасна війна показала, що людський фактор залишається ключовим, але роль людини зміщується: від ручного управління — до стратегічного контролю, від обробки даних — до ухвалення рішень, від рутинних процесів — до творчого мислення. ШІ не замінює військового, а підсилює його можливості, дозволяючи зосередитися на тому, що машини зробити не можуть — інтуїції, моральній оцінці, стратегічному баченні та відповідальності.

Таким чином, розвиток і застосування алгоритмів штучного інтелекту у секторі безпеки та оборони України стали одним із ключових факторів стійкості держави у період війни. Завдяки інтеграції воєнних алгоритмів, систем прогнозування загроз, інструментів аналізу великих даних та платформ підтримки рішень, Україна змогла створити новий тип цифрової обороноздатності — швидкий, адаптивний, точний і здатний до самооновлення. Надалі роль ШІ у війні лише зростатиме, визначаючи майбутні стандарти оборони, безпеки та стратегічного управління, а досвід України може стати основою для переосмислення глобальних підходів до ведення воєнних дій у XXI столітті.

1. Національна академія Національної гвардії України. Збірник матеріалів науково-практичного форуму МНППК 03.2025. Харків: НАНГУ, 2025. URL: https://nangu.edu.ua/uploads/files/documenty/Naukova%20diyalnist/naukovu%20forumy/2025/Збірник%20тез%20МНПК%2027_03_2025%20НАНГУ.pdf (дата звернення 23.11.2025).
2. Розробили сенсори для виявлення артилерії: як вони працюють. Hromadske, 22.02.2025. URL: https://hromadske.radio/news/2025/02/22/v-ukraini-rozrobyly-sensory-dlia-vyiavlennia-artylerii-iak-vony-pratsiuiut (дата звернення 23.11.2025).
3. Міністерство оборони України. National Cyber Security Digest. Вересень 2025. Київ: МОУ, 2025. URL: https://www.rnbo.gov.ua/files/2025/NATIONAL_CYBER_SCC/20251014_Digest%20for%20September%202025/niss_cadep_digest_september_2025.pdf (дата звернення 23.11.2025).
4. Інститут проблем прогнозування НАН України. Збірник наукових праць, вип. 2, 2024. Київ: ІППІ НАНУ, 2024. URL: https://ippi.org.ua/sites/default/files/2024-2.pdf (дата звернення 24.11.2025).

В.О. Пархоменко

ВИКОРИСТАННЯ ШІ ДЛЯ АНАЛІЗУ СОЦІАЛЬНИХ МЕРЕЖ З МЕТОЮ ВИЯВЛЕННЯ ФЕЙКОВИХ НОВИН

У наш час соціальні мережі стали середовищем не тільки для поширення правдивої інформації, а і джерелом розповсюдження фейкових новин і провокативних постів. Поява та прийняття цифрових технологій спричинили інформаційне забруднення та інфодемію в онлайн-соціальних мережах, блогах та онлайн-сайтах. Зловмисне поширення незаконного, небажаного та оманливого контенту викликає зміни в поведінці та соціальні заворушення, впливає на економічне зростання та національну безпеку, а також загрожує безпеці користувачів [1]. Відсутність гарантій правдивості інформації, низький безпековий бар'єр і необов'язкова авторизація користувачів, підсилюють вплив маніпуляції в онлайн просторі, частіше за все використовуючи штучний інтелект.

Деякі з інструментів ШІ використовують декілька завдань одночасно для перевірки постів та коментарів на прозорість інформації: виявлення та класифікація повідомлень за рівнем достовірності, розпізнавання патернів поширення фейків у соціальній мережі, виявлення ботів, а також оцінка емоційного тону та семантичного контексту публікацій. Для цього системи використовують ознаки різних рівнів - лінгвістичні (лексика, стилістика, запити фактів), поведінкові (шаблони ретвітів, частота публікацій, часові ряди активності), мережеві (структура соцграфа, центральність вузлів, кластери) та мультимодальні (поєднання тексту з зображеннями й відео) [2]. Комбінація цих ознак дозволяє підвищити рівень класифікації, по зрівнянні з іншими методами, які зазвичай спираються на один тип даних.

Сучасні моделі на основі трансформерів (такі як BERT і RoBERTa) мають в собі ключову перевагу: такі моделі мають змогу витягати контекст з представленого тексту [4]. Це робить систему більш чутливою до нюансів брехні чи маніпуляцій. Особливо ефективним є їх можливість аналізувати поведінкові і мереживні характеристики, це дозволяє виявити дію ботів-мереж і штучні патерни поширення інформації. Дослідження підтверджують, що якість виявлення фейків/ботів значно зростає після удосконалення моделей-трансформерів на спеціальних базах даних з соцмереж про перепоширення і коментарі [3]. Однак навіть такі моделі можуть бути обманутими, якщо в дописі/коментарі будуть використовуватись певні стилі мовлення, або так звані атаки генеративних ботів на соцмережі, які здатні створювати дуже переконливий фальшивий контент.

Для протидії з дезінформацією потрібна певна комплексна стратегія, яка буде поєднувати технології, освіту та певні інституційні заходи. Технологічний захист повинен включати розробку моделей детекторів і систем відстеження інформації та її походження. Ключову роль відіграє

інституційний компонент, який спрямований на підвищення медіаграмотності та критичного мислення» та навчання людей розпізнавати фейкові новини [4]. Інституційна основа несе за собою підтримку фактчекерів, які використовують журналістські методи для перевірки фактів. Важливість цих зусиль підкреслює їх спрямованість на протидію пропаганді та дезінформації по всьому світу, завдяки об'єднанню журналістів, дослідників, експертів та волонтерів [4].

Для того, аби створити ефективні системи виявлення фейків треба подолати ключові проблеми, наприклад дисбаланс у наявних даних, відсутність високоякісно анотованих даних та складнощі у представленні ознак [1]. Моделі глибинного навчання можуть автоматично навчатися складним поданням даних із мультимодальної інформації, підвищуючи ефективність виявлення фейкових новин порівняно з традиційними підходами [2]. Для досягнення результатів треба забезпечити якісну підготовку великою кількістю даних, використовувати методи аугментації для того, аби забезпечити стійкість моделі, впроваджувати механізми для пояснень рішень і проводити оцінку якості моделі до стійкості до атак і її можливість їх запобігання.

Але існують і технологічні обмеження. Деякі з таких моделей можуть демонструвати явище «overfitting» до конкретних корпусів і погано переносяться на нові теми або локальні мовні варіанти [2], такі ситуації вимагають кроку назад, надання моделі нових даних, її перенавчання і досягання бажаного результату. Крім цього, інструменти, які використовуються для відстеження походження контенту в мережах поки не часто використовуються в навчанні моделей, а їх впровадження ускладнене різноманітністю технологічних особливостей платформ. Такі складнощі наочно демонструють потребу спільної роботи спеціалістів різних галузей. Ефективна автоматична детекція фейкових новин потребує інтеграції різних підходів, адаптивного оновлення моделей та міждисциплінарної співпраці для врахування локальних особливостей і нових форматів контенту [5].

Підсумовуючи, застосування ШІ для аналізу соціальних мереж, інформаційних просторів і виявлення фейкових новин має великий потенціал і хоча вже демонструє успіхи в фільтруванні підозрілого контенту, він не може впоратися з усім напливом дезінформації. Найкращого результату ми доб'ємося лише використовуючи комплексний підхід до вирішення цієї проблеми, який включає багато факторів. Для успіху критично важливо залучати експертів різних галузей, таких як: машинне навчання, соціологія, право, журналістика. Лише прозорі методи оцінки, постійне оновлення і вдосконалення моделей, залучення більшої кількості баз даних і люди, допоможуть моделям стати чудовим інструментом фільтрації інформації в просторі інтернету.

1. Harris S. Fake News Detection Revisited: An Extensive Review of Approaches and Datasets / S. Harris // Technologies. — 2024. — Vol. 12, No.11. — P. … (article). — Available at: https://www.mdpi.com/2227-7080/12/11/222.
2. Lv J. Multi-modal fake news detection: A comprehensive survey on deep learning / J. Lv // Multimedia Tools and Applications (Springer). — 2025. — Available at: https://link.springer.com/article/10.1007/s44443-025-00317-7.
3. Wang Z. Implementing BERT and fine-tuned RoBERTa to detect AI generated news / Z. Wang // arXiv preprint. — 2023. — Available at: https://arxiv.org/pdf/2306.07401.pdf (дата звернення: 24.11.2025).
4. StopFake: про проект / StopFake.org. — Kyiv, 2014–. — Available at: https://www.stopfake.org/en/about-us/.
5. Emil RȘ. A Review of Automatic Fake News Detection / RȘ Emil // Future Internet (MDPI). — 2025. — Available at: https://www.mdpi.com/1999-5903/17/10/435 .

М.В. Богачов

ГЛИБОКЕ НАВЧАННЯ ДЛЯ ДЕТЕКЦІЇ ТА КЛАСИФІКАЦІЇ СИГНАЛІВ БПЛА В РАДІОЕФІРІ В РЕАЛЬНОМУ ЧАСІ

У сучасних умовах задача інтелектуального виявлення джерел специфічного радіовипромінювання в радіоефірі в реальному часі є критично важливою. Основними труднощами виступають високий рівень шуму та інтерференційні перешкоди в ефірі, які ускладнюють класифікацію сигналів стандартними методами цифрової обробки.

У ході дослідження було проаналізовано масив IQ-даних [1], отриманих за допомогою програмно-визначеного радіо (SDR). Експерименти з використанням методу пікових значень не продемонстрували достатньої ефективності у виявленні та розпізнаванні радіосигналів: хоча в окремих випадках метод коректно класифікував цільові сигнали, значний рівень завад та специфіка випромінювань не дозволили досягти необхідної селективності.

Для вирішення цієї проблеми запропоновано застосувати методи глибокого навчання, зокрема архітектуру LSTM (Long Short-Term Memory) [2]. Такий вибір зумовлений тим, що дані, з якими ведеться робота, мають вигляд числового ряду, і для коректної детекції та класифікації пристроїв за їх сигнатурами в радіоефірі критично важливо зберігати часовий контекст сигналу.

Ключовим етапом роботи, що виконується наразі, є розроблення алгоритму розмітки навчальної вибірки. Алгоритм має очищати вхідний потік даних, мінімізуючи шум і сторонні перешкоди; формувати сегменти потенційної активності; а також розширювати їх залежно від характеру сигналу (наприклад, циклічної появи й зникнення, що вказує на постійну активність пристрою впродовж відповідного періоду).

На поточному етапі виконано нормалізацію даних шляхом віднімання медіани, що дало змогу зменшити кількість викидів та рівень шумів. Розроблено базовий алгоритм розмітки на основі порогового значення потужності сигналу: якщо у певний момент часу максимальна потужність перевищує встановлений поріг N dB, цей інтервал фіксується як позитивна мітка. Запроваджено правило продовження позитивних міток: якщо впродовж 100 мс фіксується щонайменше 60% позитивних міток - увесь проміжок необхідно вважати позитивним.

Отримані результати забезпечують можливість переходу до етапу проєктування та тренування моделі LSTM на якісно розмічених даних. Передбачається навчання моделі на «вікнах» тривалістю N мс із подальшим застосуванням прогнозу для кожного часового кроку.

1. Lyons, R. (2008, November). Quadrature signals: Complex, but not complicated. https://www.ieee.li/pdf/essay/quadrature_signals.pdf.
2. Staudemeyer, R. C. (2019). Understanding LSTM—A tutorial into Long Short-Term Memory Recurrent Neural Networks. arXiv. https://arxiv.org/abs/1909.09586.

К.С. Кандагура, Д.О. Кулай

ЕТИКА РІЗНИХ КУЛЬТУР У ВИКОРИСТАННІ ШТУЧНОГО ІНТЕЛЕКТУ В БІЗНЕСІ

Стрімке впровадження технологій штучного інтелекту (ШІ) у бізнес-процеси змінює підходи до прийняття рішень, управління даними та корпоративної відповідальності. Актуальність теми зумовлена глобальною різницею в культурних, правових та етичних нормах, що визначають допустимість, ризики та межі використання ШІ у різних країнах [1]. У таких умовах компанії стикаються з необхідністю адаптувати технологічні рішення до місцевих стандартів, зберігаючи при цьому міжнародну конкурентоспроможність.

Мета дослідження - проаналізувати культурні відмінності в етичному сприйнятті та регулюванні штучного інтелекту в бізнесі, а також визначити їхній вплив на глобальні корпоративні практики.

Етичні підходи до використання ШІ тісно пов'язані з культурними моделями, які формують уявлення суспільства про приватність, контроль, відповідальність і ступінь довіри до технологій. Дослідження Гофстеде показують, що культури з високим рівнем індивідуалізму приділяють більше уваги захисту особистих прав та мінімізації втручання у приватне життя, тоді як колективістські суспільства більше орієнтуються на суспільний добробут та безпеку [5].

Такі відмінності впливають на те, які саме дані дозволяється збирати, у яких сферах можна застосовувати автоматизовані рішення, та наскільки суспільство довіряє алгоритмам. Наприклад, у країнах з високою дистанцією влади громадяни частіше позитивно ставляться до централізованого контролю за цифровими технологіями, тоді як у державах із низькою дистанцією влади існує тенденція вимагати від бізнесу прозорості та підзвітності ШІ-систем [3].

Європейські країни спираються на цінності захисту людської гідності, фундаментальних прав та приватності. У цьому контексті регуляторна політика ЄС у сфері штучного інтелекту спрямована на забезпечення прозорості, запобігання дискримінації та мінімізацію ризиків для користувачів [2].

Основні характеристики європейського підходу: **принцип «безпечного і надійного ШІ» (trustworthy AI)** як базовий стандарт для бізнесу; чіткий поділ застосування ШІ на рівні ризику: неприйнятний, високий, обмежений та мінімальний; жорсткі вимоги до обробки персональних даних відповідно до GDPR; зобов'язання компаній розкривати критерії роботи алгоритмів у критичних сферах (кредитування, рекрутинг, управління персоналом).

У європейській бізнес-практиці велике значення мають інструменти етичного аудиту, проведення оцінювання впливу алгоритмів (AI impact

assessment) та залучення мультидисциплінарних команд для перевірки упередженості моделей [1].

Американська модель ділової культури ґрунтується на поєднанні високої інноваційності, підприємницької ініціативи та максимальної ринкової свободи. Важливим елементом американського підходу є **пріоритет ефективності та швидкості прийняття рішень**. Американський стиль вирізняється відкритістю до експериментів та толерантністю до помилок, що розглядаються як частина інноваційного процесу. У культурі, де важлива підприємницька свобода, заохочується ініціатива, нестандартні підходи та швидке масштабування нових ідей. Це формує специфічний етичний контекст: важливо дотримуватися законодавства та чесних правил гри, але інші аспекти - наприклад, жорсткий переговорний стиль або агресивна конкурентна стратегія - часто сприймаються як нормальні елементи ринкової боротьби [5].

Ще однією особливістю є прагматичний підхід до партнерства. Взаємини в американському бізнесі більшою мірою будуються на взаємній вигоді та професіоналізмі, ніж на довгостроковій соціальній лояльності. За моделлю Тромпенаарса, США тяжіють до **універсалізму**, де правила мають однакову силу для всіх, що формує передбачувані та стандартизовані етичні практики у сфері контрактів і регуляцій [5].

Азійські країни демонструють інше бачення етичності, що зумовлено високою довірою до державних інституцій та колективістськими цінностями. Наприклад, у Китаї фокус робиться на соціальній стабільності та користі для суспільства, а регулювання штучного інтелекту має централізований характер. Етичні норми дозволяють ширше використання даних, що в західних культурах може сприйматися як порушення приватності [4].

Культурний контекст також впливає на готовність компаній до впровадження етичних стандартів. У скандинавських країнах активно підтримується концепція «етичного дизайну» (ethical-by-design), тоді як у країнах, що розвиваються, етичні питання часто відходять на другий план через економічний тиск та нерозвиненість нормативних систем. Крос-культурні відмінності у сприйнятті ризиків і відповідальності створюють виклики для міжнародних корпорацій, що працюють на різних ринках та змушені адаптувати свої AI-політики під локальні етичні вимоги [5].

З огляду на це, застосування універсальних етичних принципів штучного інтелекту в глобальному бізнесі потребує врахування культурних моделей поведінки, традицій правового регулювання та очікувань споживачів.

Дослідження етичних підходів різних культур до використання штучного інтелекту в бізнесі демонструє, що культурні цінності, моделі поведінки та інституційні особливості суттєво впливають на формування норм відповідального застосування цифрових технологій. Американська модель підкреслює інноваційність, підприємницьку свободу та домінування

правових механізмів, що сприяє швидкому впровадженню AI, але потребує посилення контролю за ризиками для суспільства. Європейський підхід, навпаки, орієнтований на захист прав людини, прозорість і превентивне регулювання, формуючи етичні стандарти через принципи сталого розвитку та відповідальності. Азійські країни переважно керуються колективістськими цінностями, пріоритетом стабільності та вагомим державним втручанням, що забезпечує широке масштабування технологій за умови збереження соціальної гармонії.

Порівняльний аналіз свідчить про те, що етика використання штучного інтелекту не може бути універсальною, оскільки кожна культурна система формує власні уявлення про межі допустимого, роль технологій у суспільстві та баланс між свободою та контролем. Одночасно простежується тенденція до зближення підходів через розвиток міжнародних стандартів, рекомендацій і рамкових документів, що спрямовані на забезпечення безпеки, прозорості та справедливості у використанні AI. Це створює умови для формування глобальної етичної екосистеми, яка враховує культурні відмінності, але водночас прагне до уніфікації ключових принципів.

Практична значущість результатів полягає у можливості враховувати культурні особливості під час розроблення корпоративних політик, міжнародних партнерств та освітніх програм у сфері штучного інтелекту. Усвідомлення цих відмінностей дозволяє мінімізувати ризики непорозумінь, підвищити ефективність взаємодії між компаніями різних культур та забезпечити етичне й відповідальне використання AI у глобальному бізнес-середовищі.

1. Floridi L. The Ethics of Artificial Intelligence. Oxford University Press, 2023.
2. European Commission. Ethics Guidelines for Trustworthy AI. Brussels, 2021.
3. Bryson J. The Past Decade and Future of AI Ethics. *AI & Society*, 2022.
4. Roberts H., Cowls J. The Chinese Approach to AI Governance. *Nature Machine Intelligence*, 2021.
5. Hofstede G. Culture’s Consequences: Comparing Values, Behaviors, Institutions and Organizations Across Nations. Sage, 2010.

С.А. Смирнов, В.І. Полуциганова

ЗАСТОСУВАННЯ ШТУЧНОГО ІНТЕЛЕКТУ ДЛЯ ПЕРЕВІРКИ УЗГОДЖЕНОСТІ ЕКСПЕРТНИХ ОЦІНОК

Анотація

У роботі представлено новий підхід до контролю якості експертної інформації в задачах підтримки прийняття рішень. На відміну від традиційних підходів застосовано інструменти штучного інтелекту (ШІ) для виявлення структурних патернів неузгодженості, оцінювання стабільності групових рішень та прогнозування впливу помилок окремих експертних суджень на підсумковий висновок.

Ключові слова: матриця попарних порівнянь, ШІ-агент, підтримка прийняття рішень

Вступ

Системи підтримки прийняття рішень (СППР) активно використовують експертні ранжування та матриці попарних порівнянь для формування діагностичних і прогностичних моделей. Однак ключовою проблемою залишається оцінювання якісності експертної інформації, оскільки суперечливі або неточні судження можуть значно знизити достовірність отриманих рекомендацій. Класичні методи оцінювання – індекс Сааті [2-3], ентропійні підходи спираються на спрощене розуміння складності, тому не ефективні при нелінійних залежностях та високовимірної структурі даних. Метою роботи є розробка підходу до перевірки узгодженості експертних оцінок з використанням інструментів штучного інтелекту [4].

Методика ШІ-оцінки узгодженості матриці парних порівнянь

Для забезпечення логічної цілісності експертних оцінок Сааті запропонував індекс узгодженості (CI) та відношення узгодженості (CR), порівнюючи CI з випадковим індексом (RI). Якщо CR < 0.1, матриця вважається «прийнятно узгодженою».

Однак на практиці цей підхід виявляється недостатнім. Навіть при CR < 0.1 можуть існувати внутрішні логічні суперечності, такі як:

- *Цикли переваги*: A краще за B, B краще за C, але A лише трохи краще за C (або навіть гірше), що порушує транзитивність.

- *Непослідовність у шкалі*: експерт використовує одні й ті самі числові оцінки для різних за суттю пар, що призводить до «виправленої», але беззмістовної матриці.

Ці недоліки стають особливо критичними при роботі з великими ієрархіями, груповими експертними оцінками або нечіткими судженнями.

Тому виникає необхідність у більш гнучкому та інформаційно насиченому підході до оцінки якості матриць попарних порівнянь. Одним із

перспективних напрямків є використання інструментів штучного інтелекту, які дозволяють оцінити ступінь хаосу або впорядкованості у розподілі експертних оцінок, незалежно від розмірності матриці.

Архітектура моделі системи прийняття рішень з підтримкою ШІ-агента

Пропонується архітектура, що складається з експерта або наборів експертів, які будуть заповнювати матрицю попарних порівнянь (МПП) виходячи з поставленої задачі. В роботі [1] розглядається випадок, що за неможливості визначити всю матрицю попарних порівнянь, є необхідний і достатній набір елементів матриці, з якого його можна відновити повністю. В силу великої кількості значень, експерт може помилятися і можуть виникати цикли, такі що варіант А кращий ніж Б, варіант Б кращий ніж В, але варіант В кращий ніж А. Ще породжує неузгодженість оцінок і математично виглядає так:

$$a_{АБ} * a_{ББ} * a_{ВА} \gg 1 \quad (1)$$

де a_ij- оцінки в скільки разів і більше ніж j. Це дає порушення транзитивності в порядковій шкалі і породжує неузгодженість, яку даних припущенням можливо однозначно визначити.

При цьому виходячи з [1] випливає, що згаданий вище мінімальний набір може бути відновлений якщо дана характеристика (1) буде близька до 1. Якщо був заданий мінімальний набір значень, то за допомогою ШІ-агента ми можемо відшукати всі можливі цикли і перевірити їх на порушення транзитивності. Якщо транзитивність зберігається і близька до 1, то є можливість корегувати відповідні елементи МПП для строгої рівності одиниці. Якщо ж дані цикли далекі від 1, тоді маємо повернути матрицю експертові для корегування оцінок. Роль ШІ-агента тут висуває в ролі контролера відповідності матиці заданим вище критеріям та менеджером для підказки експертові в не точностях оцінювання [4].

Інтегральне використання даного підходу до заповнення матриці попарних порівнянь з використання штучного інтелекту для попередження неузгодженості дає можливість створення надійних систем прийняття рішень в будь-якій області знань. До інших переваг методу відносяться: працює для будь-якої розмірності n; підтримує неповні матриці; дозволяє локалізувати конкретні пари чи цикли, що порушують узгодженість; забезпечує адаптивність до поведінки експертів.

Приклад застосування даного підходу

Для прикладу було запропоновано експертам оцінити 20 альтернатив, хоча це більше ніж може оцінити людина одночасно, але дає можливість оцінити адекватність опрацювання ШІ-агента. В ході опрацювання було виявлено, що виникає два цикли. А1-А12 – одна зв’язка альтернатив, яка дає перший цикл,

A13-A20 – друга зв’язка альтернатив. За допомогою ШІ-агента було виявлено типове відхилення від нормованого значення рис.1.

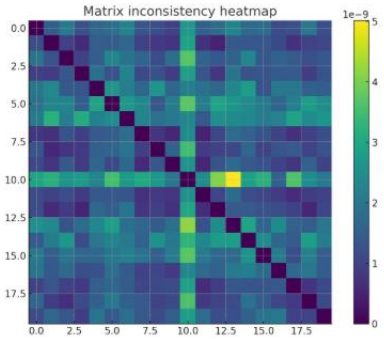

Рисунок 1 – Теплова карта відхилення від нормованого значення

Чим світліший колір - тим більша локальна неузгодженість, темніший - ідеальна консистентність. Бачимо повністю синю діагональ - це логічно, бо $a_{ii} = 1$ дає нульове відхилення.

Висновки

Представлений підхід демонструє перспективність використання штучного інтелекту для формалізації та контролю узгодженості експертних оцінок. Такий підхід виходить за межі використання традиційних інформаційних критеріїв та забезпечує вищий рівень адаптивності в контексті підтримки прийняття рішень. Метод забезпечує перевірку форм транзитивності оцінок експертів, гарантуючи логічну цілісність судження, а також виявляє помилки структурно, а не просто «приблизно». Метод працює для високої розмірності даних, включно з дуже великими МПП (>100×100), де класичні підходи проблематичні.

1. Polutsyhanova, V. I., & Smyrnov, S. A. (2025). Method for refining weights in multi-criteria utility function in MAUT. KPI Science News, 140(3). https://doi.org/10.20535/kpisn.2025.3.328644.
2. Keeney, R. L., & Raiffa, H. (2014). Decisions with multiple objectives: Preferences and value trade-offs. Cambridge University Press.
3. Zgurovsky, M. Z., Pavlov, A. A., & Shtankevych, A. S. Sh. (2010). Modified method of hierarchy analysis. System Research and Information Technologies, (1), 7–25.
4. Дриньов, Д. М., Войтех, К. Р., & Тимошенко, Р. Р. (2023). Штучний інтелект в процесі прийняття та реалізації управлінських рішень. Таврійський науковий вісник. Серія: Економіка, (18), 74–79. https://doi.org/10.32782/2708-0366/2023.18.7.

А.Ю. Жигадло

ГІБРИДНИЙ ПІДХІД ДО ВИЯВЛЕННЯ ШАХРАЙСЬКИХ BITCOIN-ТРАНЗАКЦІЙ НА ОСНОВІ RANDOM FOREST ТА GRAPH ATTENTION NETWORKS

Стрімке зростання обсягу криптовалютних транзакцій супроводжується збільшенням масштабів фінансового шахрайства. За оцінками, 2-4% операцій у мережі Bitcoin пов'язані з нелегальною діяльністю [1]. Основні виклики у виявленні таких транзакцій включають відсутність централізованих органів контролю, анонімність користувачів та використання складних схем відмивання коштів через mixing services і layering-техніки [2]. Традиційні KYC/AML процедури виявляються неефективними в децентралізованому середовищі, що вимагає розробки нових технологічних підходів.

Метою дослідження є розробка та аналіз гібридної моделі машинного навчання, що поєднує переваги класичних алгоритмів (Random Forest) та графових нейронних мереж (Graph Attention Networks) для підвищення точності детекції нелегальних Bitcoin-транзакцій.

Об'єктом дослідження є датасет, що містить 203,769 транзакцій загальною вартістю понад $6 млрд, з яких 46,564 розмічені як легальні (42,019) або нелегальні (4,545) [1].

Класичні алгоритми машинного навчання (Random Forest, XGBoost) демонструють високу ефективність на табличних даних, проте ігнорують топологічну структуру блокчейну [3]. Tree-based моделі не враховують складні зв'язки між транзакціями, що критично важливо для виявлення layering-схем та транзакцій через міксери.

Graph Convolutional Networks (GCN) та Graph Attention Networks (GAT) спеціалізовані на моделюванні графових структур і показують кращу здатність виявляти складні топологічні патерни [4]. Однак GCN страждають від проблеми "over-smoothing" на розріджених графах з роз'єднаними компонентами, типових для часових зрізів Bitcoin-транзакцій [3]. GAT частково вирішують цю проблему через механізм attention, що дозволяє зважувати важливість різних сусідів.

Ансамблеві методи використовують переваги обох підходів: стійкість і інтерпретованість Random Forest та здатність GAT моделювати контекстні зв'язки [5, 6]. Практичні дослідження підтверджують, що такі гібриди досягають кращих показників точності порівняно з окремими методами.

Датасет розділено за часовим принципом (temporal split) для запобігання витоку даних [7]. Такий підхід відповідає реальному сценарію прогнозування та уникає look-ahead bias.

Архітектура Random Forest : 100 дерев рішень з параметром class_weight='balanced' для компенсації дисбалансу класів (співвідношення 1:9.2); використання 165 табличних ознак (94 локальні + 71 агреговані).

Graph Attention Network : двошарова архітектура: 8 attention heads; activation function: ELU; dropout: 0.4; loss function: CrossEntropyLoss з вагами [1.0, 3.0] для класів.

Гібридний ансамбль:

$$P\_final(illicit) = \alpha \times P\_RF(illicit) + (1-\alpha) \times P\_GAT(illicit)$$

де α – ваговий коефіцієнт, підібраний через grid search на валідаційній вибірці, P_final – кінцевий прогноз, P_RF – ймовірність класу від Random Forest, P_GAT – ймовірність від Graph Attention Network.

З огляду на критичність виявлення всіх шахрайських операцій в AML-системах [8], основний фокус зроблено на максимізації Recall при збереженні прийнятного рівня Precision. Додатково розраховано F1-score для комплексної оцінки.

Таблиця 1 – Порівняльний аналіз моделей

Модель	Precision	Recall	F1-score
Random Forest	0,99	0,68	0,8083
GAT	0,33	0,70	0,4498
Гібрид (α=0.85)	0,97	0,71	0,8186

Random Forest продемонстрував найвищу точність (Precision=0.99), проте пропускав 32% шахрайських транзакцій. GAT збільшив повноту виявлення до 70%, але ціною різкого падіння Precision до 0.33 через проблеми з класифікацією на розріджених підграфах.

Проведено grid search по сітці параметрів: $\alpha \in [0, 1]$ (крок 0.05), threshold $\in [0.1, 0.9]$ (крок 0.05). Виявлено два оптимуми:

1. Максимум F1-score: Precision=0.97, Recall=0.71 (α=0.85, threshold=0.45).
2. Максимум Recall: Precision=0.64, Recall=0.75 (α=0.75, threshold=0.25).

Вибір α=0.85 обґрунтовується стабільністю Random Forest на табличних ознаках, натомість 15% внеску GAT достатньо для врахування графового контексту.

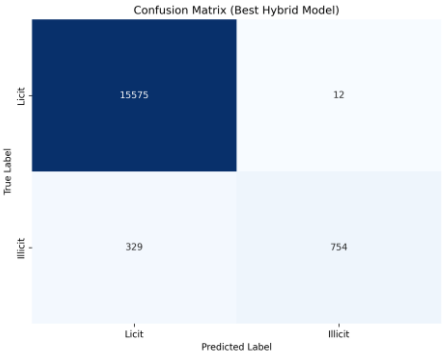

Рисунок 1 – Confusion Matrix для оптимальної конфігурації

Покращення F1-score на 1% та Recall на 3% порівняно з baseline-моделлю підтверджує ефективність гібридного підходу. Ключовим фактором успіху стала здатність GAT виявляти складні транзакційні ланцюжки, типові для layering-схем, які Random Forest класифікував як легальні через відсутність підозрілих індивідуальних ознак.

1. Elliptic Data Set. https://www.kaggle.com/datasets/ellipticco/elliptic-data-set/data.
2. Alarab, I., Prakoonwit, S., & Nacer, M. I. (2020). Competence of graph convolutional networks for anti-money laundering in bitcoin blockchain. У ICMLT 2020: 2020 5th international conference on machine learning technologies. ACM. https://doi.org/10.1145/3409073.3409080.
3. Ahmed, M. J., Mozo, A., & Karamchandani, A. (2025). A survey on graph neural networks, machine learning and deep learning techniques for time series applications in industry. *PeerJ Computer Science*, *11*, Стаття e3097. https://doi.org/10.7717/peerj-cs.3097.
4. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2017). Graph attention networks. arXiv Preprint arXiv:1710.10903. https://doi.org/10.48550/arXIV.1710.10903.
5. Salsabil, F., & Rahman, A. S. M. M. (2025). Intelligent detection of e-wallet transaction fraud using hybrid deep learning and ensemble machine learning models. International Journal of Science and Research Archive, 17(1), 1156–1166. https://doi.org/10.30574/ijsra.2025.17.1.2923.
6. Sohony, I., Pratap, R., & Nambiar, U. (2018). Ensemble learning for credit card fraud detection. У CoDS-COMAD '18: The ACM india joint international conference on data science & management of data. ACM. https://doi.org/10.1145/3152494.3156815
7. Park, M., et al. (2025). HyPV-LEAD: Proactive early-warning of cryptocurrency anomalies through data-driven structural-temporal modeling. arXiv Preprint arXiv:2509.03260. https://doi.org/10.48550/arXiv.2509.03260.
8. Financial action task force (FATF). (2019). У Secretary-General's report to ministers 2019 (с. 126). OECD. https://doi.org/10.1787/337ab3ff-en.

К.О. Бегунець

ШТУЧНИЙ ІНТЕЛЕКТ І ФОРМАЛЬНІ МЕТОДИ В ЗАДАЧІ ДЕОБФУСКАЦІЇ КОДУ ПРОГРАМНИХ ЗАСОБІВ РЕАЛІЗАЦІЇ КІБЕРЗАГРОЗ

Штучний інтелект (ШІ) активно впроваджується у системи безпеки: для виявлення аномалій, класифікації зловмисного програмного забезпечення, аналізу журналів подій та підтримки прийняття рішень у центрах моніторингу безпеки [1]. Водночас ускладнюється й програмна складова кіберзагроз: зловмисне програмне забезпечення подвійного використання, експлойти, скрипти наприклад VBS, PowerShell тощо, які широко застосовують обфускацію коду, віртуальні машини, шифрування рядків і конфігурацій, змішані булево-арифметичні вирази [2].

Для цифрової криміналістики, аудиту безпеки та реагування на інциденти важливо не лише зафіксувати факт атаки, а й зрозуміти реальну логіку конкретного фрагмента коду. Для цього необхідна деобфускація - перетворення заплутаного коду в простішу, але семантично еквівалентну форму коду, що придатна для аналізу людиною й інструментами. Актуальним питанням є: чи може ШІ стати основним засобом такої деобфускації, чи для цієї задачі потрібні спеціалізовані формальні методи?

Метою роботи є дослідження використання штучного інтелекту в задачах деобфускації коду програмних засобів реалізації кіберзагроз.

Використання штучного інтелекту для розв’язання цієї задачі має задовольняти низку вимог:

- Коректність: спрощений код не повинен змінювати поведінку програми: відкидання або додавання логіки є неприйнятним, особливо у цифровій криміналістиці.
- Повторюваність та детермінованість: за однакових вхідних умов інструмент повинен давати однаковий результат, що важливо для відтворюваності аналізу.
- Прозорість: аналітик повинен мати змогу зрозуміти, які саме перетворення було виконано і чому.
- Стійкість до протидії: зловмисник може цілеспрямовано модифікувати обфускацію, щоби ускладнити роботу інструментів аналізу.

У даному дослідженні деобфускація розглядається як системний процес, для якого в майбутньому планується розробити архітектуронезалежну модель проміжного представлення коду та модифікований метод деобфускації на її основі.

Потенційні ролі ШІ у деобфускації коду

Методи ШІ вже показали ефективність у виявленні зловмисного програмного забезпечення та його класифікації [1, 3]. У задачі деобфускації можна виділити кілька потенційних застосувань таких методів:

● Класифікація та пріоритизація фрагментів коду. Моделі машинного навчання можуть визначати, які частини великого виконуваного файлу чи скрипта з більшою ймовірністю містять обфускацію або небезпечну логіку.

● Розпізнавання відомих патернів обфускації. Наприклад у наступній роботі [3], демонструється, що гібрид machine learning (ML) в поєднанні з статичним аналізом може з високою точністю (до 98 %) виявляти непрозорі предикати (opaque predicates) та усувати їх у статичному режимі.

● Підтримка аналітика-реверсера. Великі мовні моделі й нейронні деобфускатори можуть допомагати формувати гіпотези щодо призначення коду, відновлювати імена змінних та структур кращої читабельності [4].

Отже, ШІ може виступати корисним інструментом для пошуку, класифікації та інтерпретації у задачах аналізу коду.

Обмеження ШІ як базового механізму деобфускації

Водночас використання ШІ як ядра процесу деобфускації має низку фундаментальних обмежень:

● Статистичний характер рішень. Навіть моделі з дуже високою точністю на тестових вибірках (наприклад, 98 % для класифікації opaque predicates) [3] залишають ненульову ймовірність помилки. У задачі деобфускації така помилка може означати втрату важливого фрагмента логіки або появу "вигаданої" поведінки, якої не було в оригінальній програмі.

● Залежність від навчальних даних та еволюції атак. Як показують огляди застосування глибокого навчання для виявлення зловмисного програмного забезпечення, доступні датасети й бенчмарки є неповними та швидко застарівають [1]. Нові обфускаційні техніки, власні віртуальні машини, змішані сценарії (скрипт взаємодіє з бінарним файлом) часто виходять за межі того, на чому навчалась модель.

● Вразливість до adversarial-атак (змагальні атаки). Роботи з adversarial malware показують, що ML-детектори можна обійти, змінюючи лише незначну частину байтів у виконуваному файлі [5]. Аналогічно, деобфускатор на основі ШІ може бути навмисно "зламаний" ретельно сконструйованою обфускацією.

● Чорна скринька (black box) та відсутність строгих доказів. Наявні наукові роботи, в яких досліджується інтерпретованість ML-декторів зловмисного програмного забезпечення підкреслюють, що більшість моделей залишаються непрозорими; їх рішення важко пояснити та формально обґрунтувати. Для деобфускації це критично: ми не можемо гарантувати, що два фрагменти коду дійсно еквівалентні для всіх можливих вхідних даних.

● Обмеження LLM у деобфускації складно обфускованого коду. Нещодавнє дослідження LLM для деобфускації асемблерного коду показало, що хоча моделі іноді успішні на простих обфускаціях, вони систематично зазнають невдачі на комбінованих техніках (flattening з instruction substitution тощо), демонструючи типові помилки у відновленні логіки й структури програми [6].

Отже, ШІ має значний потенціал як допоміжний інструмент, але не відповідає вимогам до базового механізму деобфускації, який потребує строгих гарантій коректності й стійкості до протидії.

Формально-орієнтований підхід на основі моделі проміжного представлення

Як альтернативу автор обирає формально-орієнтований підхід, у якому ядром методу є модель проміжного представлення коду. Ідея полягає в тому, щоби:

- перетворити різнорідний код, наприклад x86/x64, VBS тощо на уніфіковане проміжне представлення;
- чітко визначити, як кожна інструкція цієї мови змінює стан програми;
- подавати програми у формі з єдиним статичним присвоєнням та графом керування.

На цьому рівні можна використовувати:

- символьне виконання та SMT-розв'язувачі (satisfiability modulo theories) для перевірки досяжності гілок, виявлення opaque predicates та "мертвих" шляхів виконання;
- алгебраїчне переписування змішаних булево-арифметичних виразів (інструмент GAMBA для їх деобфускації) [2];
- формальні семантики ISA (instruction set architecture) для побудови більш точних і розширюваних інструментів аналізу бінарного коду, як це продемонстровано на прикладі виявлення помилок в існуючих засобах символьного виконання, побудованих на основі проміжного представлення;

Більш докладна реалізація ідеї даного підходу буде наведено у наступних роботах.

Роль ШІ у гібридній системі деобфускації

Попри наведені обмеження, ШІ залишається корисним у гібридній архітектурі:

- над формальним ядром - для виявлення підозрілих ділянок у великих кодових базах, кластеризації вже деобфускованих фрагментів, напівавтоматичного пояснення поведінки коду;
- поруч із формальними методами - як засіб побудови евристичних методів і пріоритизації, не замінюючи при цьому строгі докази еквівалентності.

Підсумовуючи можна зробити висновок, що оптимальною видається архітектура, де формально перевірювані перетворення (проміжне представлення, символьний аналіз, SMT) забезпечують коректність і прозорість, а ШІ використовується як засіб підвищення продуктивності та зручності роботи експертів із кібербезпеки. Це відповідає загальній тенденції у сучасних дослідженнях, де ШІ розглядається не як заміна класичних методів безпеки, а як інструмент їх посилення.

1. Bensaoud, A., Kalita, J., & Bensaoud, M. (2024). A survey of malware detection using deep learning. *Machine Learning with Applications*, 16, 100546. https://doi.org/10.1016/j.mlwa.2024.100546.
2. Reichenwallner, B., & Meerwald-Stadler, P. (2023). Simplification of general mixed Boolean-arithmetic expressions: GAMBA. In *2023 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)* (pp. 427–438). IEEE. https://doi.org/10.1109/EuroSPW59978.2023.00053.
3. Tofighi-Shirazi, R., Asăvoae, I.-M., Elbaz-Vincent, P., & Le, T.-H. (2019). Defeating opaque predicates statically through machine learning and binary analysis. In *Proceedings of the 3rd ACM Workshop on Software Protection (SPRO'19)* (pp. 3–14). Association for Computing Machinery. https://doi.org/10.1145/3338503.3357719.
4. Hosseini, I., & Dolan-Gavitt, B. (2022). Beyond the C: Retargetable decompilation using neural machine translation. In *NDSS Workshop on Binary Analysis Research (BAR 2022)*. Internet Society. https://doi.org/10.14722/bar.2022.23009.
5. Demetrio, L., Coull, S. E., Biggio, B., Lagorio, G., Armando, A., & Roli, F. (2021). Adversarial EXEmples: A survey and experimental evaluation of practical attacks on machine learning for Windows malware detection. *ACM Transactions on Privacy and Security*, 24(4), 1–31. https://doi.org/10.1145/3473039.
6. Tkachenko, A., Suskevic, D., & Adolphi, B. (2025). Deconstructing obfuscation: A four-dimensional framework for evaluating large language models' assembly code deobfuscation capabilities. *arXiv*. https://doi.org/10.48550/arXiv.2505.19887.

О.В. Мельничук

ПОРІВНЯННЯ УКРАЇНСЬКИХ МАЛИХ МОВНИХ МЕРЕЖ У КОНТЕКСТІ ДЕТЕКЦІЇ МАНІПУЛЯЦІЙ У СОЦІАЛЬНИХ МЕРЕЖАХ

Актуальність

Стрімкий розвиток штучного інтелекту та великих мовних моделей (LLM) докорінно змінює інформаційний ландшафт, створюючи безпрецедентні виклики для верифікації контенту в цифровому просторі. AI-керована дезінформація була визнана Світовим економічним форумом у Global Risk Report як Глобальний ризик №1 у 2024 році [1]. Масштаб проблеми досягає критичних значень: швидкість поширення неправдивої інформації в соціальних мережах перевищує правдиву у шість разів [2], створюючи системні ризики для демократичних процесів, громадського здоров'я та національної безпеки.

Рік 2024, який називали "супервиборчим роком" з виборами у понад 60 країнах, показав, що контент, який поширює дезінформацію у соціальних мережах, став характерною рисою більшості виборчих кампаній у світі [3]. Актуальність цієї проблеми особливо підкреслюється в контексті України, де за даними дослідження USAID-Internews, 74% населення споживають новини через соціальні мережі, з яких 60% надають перевагу Telegram [4].

Комерційні LLM (GPT-5, Gemini) забезпечують високу якість аналізу, проте мають критичні недоліки для практичного застосування: висока вартість інференсу, культурний баєс через тренування переважно на англомовних даних, залежність від хмарних API. Традиційні ML-методи стикаються з новими викликами: обмежене розуміння контексту, сильна залежність від датасету та критична вразливість до LLM-генерованого контенту [5], що робить детектори майже неефективними на коротких текстах, типових для соціальних мереж.

Ця праця представляє ефективну мультиагентну систему верифікації інформації, яка досягає високої точності при економії вартості порівняно з великими мовними моделями шляхом поєднання малих локальних моделей, що пройшли культурно-специфічну валідацію для українського контексту у єдину систему. Ключові переваги українських локальних мовних моделей включають повний контроль над даними та приватність, краще розуміння українського контексту, топонімів та власних назв та відсутність залежності від зовнішніх API.

Концептуальна модель

Розроблена система базується на мультиагентній архітектурі, яка декомпозує складний процес верифікації інформації на спеціалізовані компоненти. Подібний підхід до декомпозиції життєвого циклу дезінформації продемонстровано у роботі [6], проте представлена система фокусується на детекції маніпулятивних технік із культурно-специфічною

адаптацією для українського контексту. На відміну від монолітних підходів, мультиагентна архітектура забезпечує модульність, спеціалізацію та можливість незалежної оптимізації кожного компонента. Система реалізує повний життєвий цикл обробки: виявлення маніпулятивних технік, витягування наративів, верифікацію фактів та синтез фінального висновку.

Архітектура базується на принципах спеціалізації агентів, технік "prompt chaining" та паралелізму обробки. Кожен агент має вузьку спеціалізацію та оптимізований для конкретного типу завдань, що дозволяє використовувати різні моделі та підходи для кожного етапу.

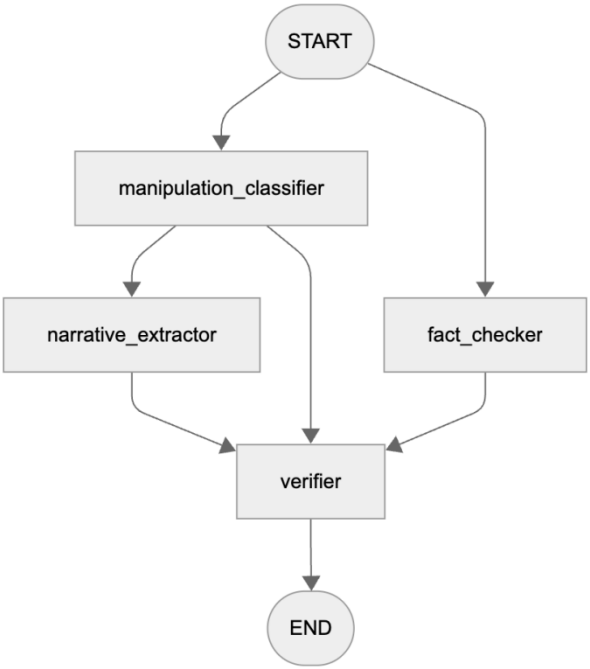

Рисунок 1– Етапи мультиагнетної системи

Компонентна архітектура

Manipulation Classifier Agent: Перший компонент у ланцюзі обробки, який аналізує вхідний контент на предмет маніпулятивних технік. Агент використовує запит із структурованими шаблонами для виявлення 5 категорій маніпуляцій:

• Емоційна маніпуляція — використовує експресивну мову з сильним емоційним забарвленням або ейфорійний тон для підняття бойового духу та впливу на думку;

• Апеляції до страху — грає на страхах, стереотипах чи упередженнях, включає тактики страху, невизначеності та сумнівів (FUD);

• Ефект натовпу — використовує загальні позитивні концепції або заклики до мас («всі так думають») для заохочення згоди;

• Селективна правда — використовує логічні помилки, такі як вибірковий відбір фактів, whataboutism для відвернення критики, або створення опудальних аргументів;

• Думко-припиняючі кліше — використовує формульні фрази, розроблені для припинення критичного мислення та завершення дискусії.

Агент повертає ймовірність маніпуляції у діапазоні 0-1 та список виявлених технік з використанням порогового значення.

Narrative Extractor Agent: Працює послідовно після Manipulation Classifier і спеціалізується на витягуванні ключових наративів з контенту з урахуванням виявлених маніпулятивних технік. Використовує контекстуальні промпти, що включають інформацію про ймовірність маніпуляції та конкретні техніки для більш точного витягування наративу. Цей компонент є критично важливим для подальшого визначення маніпуляції, оскільки забезпечує структурований опис ключових наративів.

Fact Checker Agent: Найбільш ресурсоємний компонент, який здійснює верифікацію фактів через зовнішні джерела. Працює паралельно з Manipulation Classifier для зниження загальної тривалості аналізу. Процес включає два етапи: генерацію пошукових запитів (LLM аналізує контент для створення запитів, що б перевіряли конкретні факти) та веб-пошук через Perplexity API з фільтрацією ненадійних джерел.

Verifier Agent: Фінальний компонент, який синтезує результати всіх попередніх агентів у структурований висновок. Отримує оригінальний контент, ймовірність маніпуляції, виявлені техніки, витягнуті наративи та результати факт-чекінгу. Вихідний формат включає бульове значення присутності маніпуляцій, список технік, елементи дезінформації та структуроване пояснення висновку з використанням Structured Output для подальшого використання.

Експериментальні результати

Методологія

Для валідації ефективності розробленої системи проведено експериментальне дослідження на реальному датасеті маніпулятивного контенту, що містить дані українською та російською мовами. Експеримент спрямований на оцінку точності детекції маніпулятивного контенту та порівняння продуктивності різних мовних моделей.

Експеримент проводився на підготовленому тестовому датасеті з 40 україномовних та російськомовних повідомлень з соціальних мереж та інформаційних джерел [7]. Датасет збалансований за двома критеріями: 50% маніпулятивний контент / 50% достовірний контент; 50% українською мовою / 50% російською мовою. Датасет сформований на основі корпус постів Telegram, що були анотовані медіа-експертами за узгодженою таксономією 10 маніпулятивних технік, сформованою на основі локальної експертизи та фокус-груп для UNLP 2025[9].

Апаратна конфігурація: Apple M1 Pro з 32 ГБ оперативної пам'яті та 24.96 ГБ відеопам'яті. Всі локальні моделі запускалися через LM Studio. Для факт-чекінгу використовувалась Perplexity API як пошукова система, яка забезпечує можливість паралельного пошуку за декількома різними search_queries. Система генерує до 3 пошукових запитів для кожного факту, що має бути перевіреним.

Протестовані моделі: Gemma 3 4B [10] (базова модель Google у повній версії fp16), MamayLM Gemma-3-4B-it-v1.0 [11], MamayLM Gemma-3-4B квантизована версія Q4_K_S, Gemini 2.5 Flash [12] (базова модель Google з доступом через AI Studio API).

Метрики оцінювання

Для оцінки ефективності детекції маніпулятивного контенту використовуються класичні метрики бінарної класифікації, що базуються на матриці помилок (confusion matrix):

TP (True Positives) — правильно визначені маніпулятивні повідомлення

TN (True Negatives) — правильно визначені достовірні повідомлення

FP (False Positives) — помилково визначені як маніпулятивні

FN (False Negatives) — пропущені маніпулятивні повідомлення

На основі цих значень обчислюються наступні метрики:

Accuracy (Точність) — частка правильно класифікованих повідомлень:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

Precision (Прецизійність) — частка справжніх маніпуляцій серед усіх, що були класифіковані як маніпулятивні:

$$Precision = TP / (TP + FP) \quad (2)$$

Recall (Повнота) — частка виявлених маніпуляцій серед усіх справжніх маніпулятивних повідомлень:

$$Recall = TP / (TP + FN) \quad (3)$$

F1-Score — гармонічне середнє між Precision та Recall, що забезпечує збалансовану оцінку якості:

$$F_1 = (2 \times Precision \times Recall) / (Precision + Recall) \quad (4)$$

F1-Score обрано як основну метрику порівняння моделей, оскільки вона враховує як точність детекції, так і повноту покриття маніпулятивного контенту, що є критично важливим для практичного застосування системи.

Результати та порівняння

Експериментальне тестування сфокусовано на порівнянні великої хмарної моделі (Gemini Flash) з малими локальними LLM (MamayLM) у двох варіантах квантизації. Основна гіпотеза: малі локальні моделі можуть досягати порівнянних результатів з великими хмарними рішеннями при значно меншій вартості та кращій приватності.

Таблиця 1 – Порівняння ефективності моделей

Модель	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Gemini 2.5 Flash (хмара)	0.750	0.667	1.000	0.800
MamayLM повна Q8_0	0.625	0.576	0.950	0.717
Gemma 3 4B fp16	0.625	0.581	0.900	0.706
MamayLM Q4_K_S	0.500	0.500	0.900	0.643

Порівнянна ефективність моделей: Локальні моделі MamayLM (F1-score 0.717) та базова Gemma 3 (F1-score 0.706) показують результати на 10-12% нижче хмарної моделі Gemini Flash (F1-score 0.800), підтверджуючи можливість використання локальних рішень. Українська адаптація MamayLM показує дещо кращі результати порівняно з базовою Gemma 3.

Вплив квантизації: Q4_K_S забезпечує 75% економію пам'яті при помірному зниженні якості (F1-score 0.643 проти 0.717 для повної версії). Recall залишається високим (0.900), що важливо для детекції.

Культурна адаптація: Українська модель MamayLM краще обробляє культурно-специфічний контекст порівняно з базовою Gemma 3, підтверджуючи важливість локальних адаптацій.

Результати експериментів збережені на Github [8].

Висновки

Представлена мультиагентна система демонструє можливість ефективної детекції маніпулятивного контенту з використанням локальних моделей. Система успішно поєднує спеціалізовані агенти в єдину

архітектуру, що забезпечує економію вартості порівняно з використанням виключно комерційних LLM при збереженні високої точності аналізу.

Культурна адаптація через інтеграцію української мовної моделі (MamayLM) забезпечила культурно-специфічну валідацію, критично важливу для ефективної детекції маніпуляцій у локальному контексті. Експериментальна валідація на реальних даних підтвердила гіпотезу, що малі мовні моделі можуть ефективно справлятися з конкретними задачами на рівні, близькому до комерційних LLM. Локальна модель MamayLM показує F1-score 0.717, що лише на 10% нижче хмарної моделі Gemini Flash (0.800).

Довгострокова перспектива розвитку включає Parameter Efficient Fine-Tuning через LoRA адаптери [13], використання спеціалізованих класифікаторів на відповідних етапах та загальну оркестрацію з early stopping механізмами для економії ресурсів з мінімальною втратою точності.

1. World Economic Forum. Global Risks Report 2024. Geneva: World Economic Forum, 2024. URL: https://www.weforum.org/publications/global-risks-report-2024/
2. Vosoughi S., Roy D., Aral S. The spread of true and false news online. Science. 2018. Vol. 359, No. 6380. P. 1146-1151.
3. Allen K., Nehring C. AI-Generated Disinformation in Europe and Africa: Use Cases, Solutions and Transnational Learning. Johannesburg: Konrad-Adenauer-Stiftung Media Programme Sub-Saharan Africa, 2025.
4. USAID-Internews. Media consumption and trust in Ukraine. 2024. URL: https://www.internews.org/.
5. Sadasivan V. S., Kumar A., Balasubramanian S., Wang W., Feizi S. Can AI-Generated Text be Reliably Detected? arXiv preprint arXiv:2303.11156. 2024.
6. Gautam A. Multi-agent Systems for Misinformation Lifecycle: Detection, Correction And Source Identification. Workshop Proceedings of the 19th International AAAI Conference on Web and Social Media. Copenhagen, Denmark, 2025. 7 p.
7. UNLP-2025 Shared Task Dataset. URL: https://github.com/unlp-workshop/unlp-2025-shared-task/tree/main/data.
8. Melnychuk O. VerifAI PhD Research Repository. URL: https://github.com/olehmell/verifai-phd.
9. Kyslyi, R., Romanyshyn, N., & Sydorskyi, V. (2025, July). The unlp 2025 shared task on detecting social media manipulation. In Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025) (pp. 105-111).
10. Google DeepMind. Gemma 3: Open Models Based on Gemini Research and Technology. Technical Report. 2025. URL: https://ai.google.dev/gemma.
11. Lang-UK. MamayLM: Ukrainian Language Models. 2025. URL: https://github.com/lang-uk/mamaylm.
12. Google. Gemini 2.5 Flash: Technical Documentation. 2025. URL: https://ai.google.dev/gemini-api.
13. Hu E. J., Shen Y., Wallis P., Allen-Zhu Z., Li Y., Wang S., Wang L., Chen W. LoRA: Low-Rank Adaptation of Large Language Models. arXiv preprint arXiv:2106.09685. 2021.

H. Doroshuk

NATIONAL AI STRATEGY OF UKRAINE: DEVELOPMENT VECTORS, SECURITY, AND INTEGRATION INTO THE GLOBAL CONTEXT

Artificial Intelligence (AI) is considered a crucial socio-technical institution capable of fundamentally transforming the economy, politics, and social order, while improving societal well-being. At the same time, the uncertainty surrounding the future of technologies, including AI, necessitates proactive leadership from the state, businesses, and academic institutions to determine AI development trajectories. Summarizing [1-4], it can be said that a national AI strategy is an official state plan or document that defines objectives, priorities, and measures for the development, implementation, and regulation of AI technologies, taking into account economic, social, scientific, and ethical aspects, as well as national security and international cooperation.

It outlines technological policies, a country's strategic positioning, and the envisioned public and private benefits of AI. Such strategies set the framework for investments, regulatory measures, and coordination between the state, business, and society. In the context of international competition and the threat of exogenous shocks—such as war, cyber threats, or environmental disasters—AI is seen as a key tool for ensuring national security and economic prosperity.

Against the backdrop of growing global AI competition and increasing external challenges, including the pandemic and the war in Ukraine, analyzing how national strategies adapt to extreme circumstances and what new forms of technological development they generate becomes particularly relevant. This leads to the key research question: How does Ukraine's national AI strategy reflect global AI development trends while simultaneously addressing the specific challenges and tasks of AI technology evolution under the impact of war as an external shock?

According to the *AI Index Report* (Stanford), the first national AI strategies were officially published in 2017 by Canada, China, and Finland [5]. The uniqueness of each national AI strategy largely reflects the political, economic, and cultural context of the respective country. Global leadership in AI belongs to the United States and China, with the U.S. emphasizing innovation, business development, and national leadership [6], while China prioritizes social order and regulatory oversight [7]. Germany and France—the largest economies and technological leaders in Europe—develop AI strategies emphasizing ethics, sustainability, efficiency, and security. Germany focuses more on industrial applications and social standards [8], whereas France emphasizes medical and environmental technologies [9].

Ukraine's first document outlining AI development directions was the 2020 AI Concept [10]. The Concept highlighted general directions without specifying

sectoral priorities or addressing Ukraine-specific challenges. Comparing this Concept with the draft national AI strategy presented at the WinWin Summit in November 2025 shows that the new strategy reflects real use cases and ongoing developments, which serve as a basis for future directions. The updated strategy primarily aims to establish Ukraine as a leader in AI, ensuring national sovereignty. Priority sectors include the public sector, defense, and education. By 2030, the strategy aims to achieve the following KPIs: 75% of companies using AI, 90% of the population regularly using AI, 50,000 AI developers, 100% of public services in Diia delivered by AI agents, a sovereign AI infrastructure, and 500 AI companies active in the global market [11].

Case 1. AI Ecosystem for Public Administration in Ukraine

Diia.AI was developed with the support of the "Digitalisation for Growth, Integrity and Transparency" project (UK DIGIT), implemented by the Eurasia Foundation and funded by UK Dev, as well as the Swiss-Ukrainian EGAP program, implemented by the Eastern Europe Foundation [12]. Diia.AI was recognized as the first national AI assistant for public services, launched in September 2025. The emergence of such an AI assistant not only signals the digital transformation but also the AI transformation of the state and defense sector, marking the transition from a "Digital State" to an "Agentic State," where AI is integrated into public services. Consequently, Ukrainian citizens can access public services quickly and easily through a single query.

Case 2. AI Infrastructure in Ukraine

Building sovereign AI is a matter of national security and data protection, especially during wartime [13]. To launch a national AI in Ukraine, the Ministry of Digital Transformation collaborates with NVIDIA, a global leader in computational infrastructure for AI. NVIDIA provides GPUs for training modern AI models, making it critical for AI development. The project aims to create a sovereign AI state with its own infrastructure and strong talent pool. Cooperation will focus on: supporting the creation of national AI infrastructure based on NVIDIA technology; talent development and AI education; joint R&D projects; and support for the startup ecosystem. The first joint project will be the development of Diia AI LLM — a sovereign language model adapted to Ukrainian legislation, public services, and citizens' needs. All AI services in the Diia ecosystem will operate based on Diia AI LLM, including the AI assistant on the portal, the future mobile assistant, and internal government solutions. The next step in Ukraine's Digital Development will be the creation of an AI factory, a specialized infrastructure covering the entire AI lifecycle: data ingestion, model training, and inference.

Case 3. AI Ecosystem for Education in Ukraine

While online education worldwide has diminished after the COVID-19 pandemic, in Ukraine it has continued for five years due to mobility restrictions initially caused by the pandemic and later by the war. In response, the Ministry of Digital Transformation and the Ministry of Education and Science developed "Mriya," a state digital education ecosystem [14].

The Mriya project, implemented with the support of EGAP and the Eastern Europe Foundation, received a $1.5 million investment from Google to integrate AI and personalize learning. The goal is to unite all participants of the educational process — students, teachers, and parents — within a single digital space.

Case 4. AI Ecosystem for Defense in Ukraine

The war has accelerated the development of military AI, which may lead to complete replacement of humans in certain military tasks and change conceptions of tactics and strategy. One of the first projects in this domain was the U.S. Department of Defense "Maven" project, which integrated big data analysis into combat operations [15].

Full-scale war in Ukraine, however, has accelerated global military innovation [16]. Key technologies have rapidly developed:

- Situational awareness and data analytics: The Delta cloud system integrates data from UAVs, satellites, and sensors into a digital map where algorithms prioritize and predict target movements.
- Facial recognition and social profiling: Clearview AI compares photos of deceased or captured individuals with images from social media and open databases to assist in identification.
- Cybersecurity and digital countermeasures: Microsoft Sentinel and Google Mandiant Threat Intelligence use machine learning to analyze terabytes of data, detect anomalies, and block threats automatically. The Ukrainian startup SOC Prime creates Sigma rules daily to quickly identify cyberattacks.
- Autonomous drones and robotic platforms: The MAGURA V5 naval drone independently plans routes, avoids obstacles, and executes strikes. Initiatives like Brave1 and the startup Swarmer develop software for FPV drone swarms that collectively detect and attack targets even with GPS interference.

Studying national AI strategies allows identifying each country's specialization, priorities, and unique experiences, forming the basis for international collaboration, technology exchange, and joint development. Diverse approaches to AI policy foster both constructive competition and mutual enrichment of knowledge, increasing the resilience of the global AI ecosystem. Ukraine receives significant international support, including access to project funding, specialized software stacks for AI development, technical expertise to accelerate model training and reduce costs, and consulting and engineering support from international experts. Simultaneously, Ukraine aims to create a hybrid infrastructure for testing and piloting world-class AI solutions. Developing such a high-tech environment opens opportunities for Ukraine to become a pilot site for large-scale R&D projects and international innovation initiatives.

1. Bareis, J., & Katzenbach, C. (2021). Talking AI into Being: The Narratives and Imaginaries of National AI Strategies and Their Performative Politics. Science, Technology, & Human Values, 47(5), 855-881. https://doi.org/10.1177/01622439211030007 (HIIG).
2. Department for Science, Innovation & Technology; Office for Artificial Intelligence; Department for Digital, Culture, Media & Sport. (2021). National AI Strategy. GOV.UK.

https://assets.publishing.service.gov.uk/media/614db4d1e90e077a2cbdf3c4/National_AI_Strategy_-_PDF_version.pdf.
3. Federal Government of Germany. (2020). Artificial Intelligence Strategy of the German Federal Government: 2020 update. https://www.bmwi.de/Redaktion/EN/Publikationen/Artificial-Intelligence/ai-strategy.pdf.
4. Organisation for Economic Co-operation and Development. (2021). An overview of national AI strategies and policies. OECD. https://www.oecd.org/en/publications/an-overview-of-national-ai-strategies-and-policies_c05140d9-en.html.
5. Organization for Economic Co-operation and Development. (2021, March 15). AI Index Report 2021. https://wp.oecd.ai/app/uploads/2021/03/2021-AI-Index-Report.pdf (oecd.ai).
6. The White House. (2018, May 10). Summary report of the 2018 White House Summit on AI for American Industry. https://trumpwhitehouse.archives.gov/wp-content/uploads/2018/05/Summary-Report-of-White-House-AI-Summit.pdf.
7. State Council of China. (2017, July 8). A Next Generation Artificial Intelligence Development Plan. http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm.
8. Bundesregierung. (2020). National AI Strategy update: Fortschreibung 2020. https://www.bundeswirtschaftsministerium.de/Redaktion/DE/Publikationen/Technologie/strategie-kuenstliche-intelligenz-fortschreibung-2020.pdf.
9. Villani, C. (2018). For a meaningful artificial intelligence: National and European strategy [Report]. Mission Intelligence Artificielle. https://www.jaist.ac.jp/~bao/AI/OtherAIstrategies/MissionVillani_Report_ENG-VF.pdf.
10. Cabinet of Ministers of Ukraine. (2020). Concept for the Development of Artificial Intelligence in Ukraine (Resolution No. 1556-r). https://zakon.rada.gov.ua/laws/show/1556-2020-p#Text.
11. Ministry of Digital Transformation of Ukraine. (2025, November 5). AI Strategy Presentation UA [Video]. YouTube. https://www.youtube.com/watch?v=MYQV-S4wmb4.
12. The Digital. (n.d.). Ukraine sets world record: DiiaAI recognized as the first national AI assistant for public services. https://thedigital.gov.ua/news/progress/ukrayina-vstanovyla-svitovyy-rekord-diiaai-vyznaly-pershym-natsionalnym-shi-asystentom-iz-derzavnykh-posluh.
13. The Digital. (n.d.). Starting work with NVIDIA to develop sovereign AI in Ukraine. https://thedigital.gov.ua/news/progress/pochynayemo-pratsiuvaty-z-nvidia-dlia-rozbudovy-suverennoho-shi-v-ukrayini.
14. The Digital. (n.d.). AI in “Mriya”: How artificial intelligence will build individual educational trajectories for children. https://thedigital.gov.ua/news/education/ai-u-mrii-yak-shtuchniy-intelekt-buduvatime-individualni-osvitni-traektorii-dlya-ditey.
15. MarketsandMarkets. (n.d.). Artificial intelligence in military market. https://www.marketsandmarkets.com/Market-Reports/artificial-intelligence-military-market-41793495.html.
16. The Economist. (2024, April 8). How Ukraine is using AI to fight Russia. https://www.economist.com/science-and-technology/2024/04/08/how-ukraine-is-using-ai-to-fight-russia.

Н.В. Тахтай, Я.В. Левадний, В.Ю. П'ятаченко

АНАЛІЗ ПОВЕДІНКИ LSTM-МОДЕЛІ У ЗАДАЧІ ВИЯВЛЕННЯ АТАК ПІДМІНИ КООРДИНАТ

У сучасних мережевих системах проблема виявлення атак набуває критичного значення, оскільки традиційні сигнатурні методи, на яких будуються класичні системи виявлення вторгнень, виявляються малоефективними у випадках нових або модифікованих загроз. Атаки “нульового дня” та аномальні шаблони трафіку, що не відповідають заздалегідь відомим сигнатурам, залишаються поза можливостями таких систем. Саме тому методи машинного навчання, здатні аналізувати часові закономірності та виявляти відхилення у поведінці трафіку, постають перспективним інструментом підвищення кіберстійкості[1]. У цій роботі досліджується підхід, заснований на поділі трафіку на послідовні часові вікна, з яких екстрагуються ознаки, а далі модель оцінює ймовірність наявності атаки для кожного вікна. Метою є не лише побудова такого детектора, а й оцінка його стійкості, здатності до узагальнення та характеру його помилок у динамічному мережевому середовищі. Для навчання та верифікації гібридної моделі використано AV-GPS-Dataset[2] – відкритий набір навігаційних даних, зібраний дослідницькою групою Autonomic Computing Lab Університету Аризони (США).

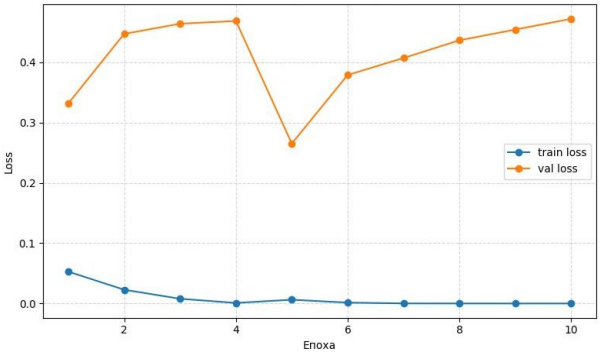

Рисунок 1 – Графік втрат машинного навчання

Поведінку моделі під час навчання ілюструє графік зміни функції втрат (рис.1), який виявляє суттєву різницю між роботою на тренувальних і валідаційних даних. Тренувальна крива стрімко наближається до нуля вже після четвертої епохи, що демонструє швидку здатність моделі запам’ятовувати навчальні приклади. Водночас валідаційні втрати зростають у перших епохах, досягаючи значень понад 0.45, що вказує на початок перенавчання. Короткочасне падіння val loss на п’ятій епосі виявляється

скоріше винятком, оскільки вже з шостої епохи крива знову піднімається, формуючи стійке розходження між train loss і val loss. Така динаміка свідчить про нестійкість процесу оптимізації та про те, що модель надмірно адаптується до тренувальних даних, втрачаючи здатність коректно відтворювати поведінку на нових зразках.

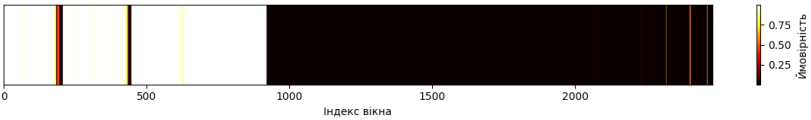

Рисунок 2 – Теплова карта ймовірностей атак

Особливості роботи моделі в умовах реального трафіку наочно демонструє теплова карта ймовірностей атак(рис.2). Вона виявляє чітко окреслений та тривалий інтервал високих значень ймовірності — приблизно від 900 до 2500 вікон. Цей темний сегмент, близький до насичення, підтверджує, що модель впевнено фіксує довготривалу атаку та зберігає стабільність оцінок протягом усього періоду шкідливої активності. Разом із тим на початку послідовності, у діапазоні від 0 до 500 вікон, з'являються поодинокі вертикальні "спалахи", які свідчать про короткочасні аномалії. Подібні вузькі смуги наприкінці (після 2200 вікон) вказують на окремі хибнопозитивні піки, які не утворюють суцільної області атаки, але демонструють підвищену чутливість моделі до рідкісних закономірностей у легітимному трафіку.

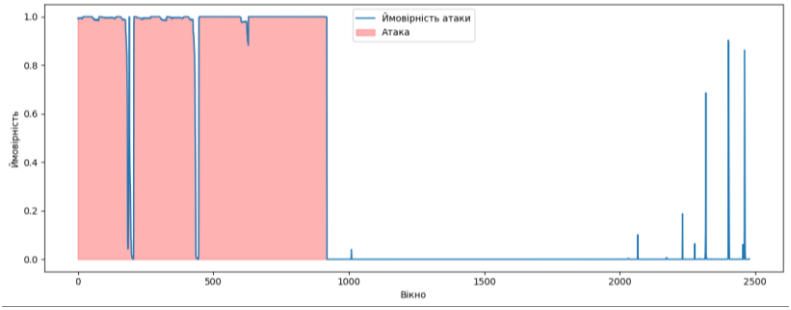

Рисунок 3 – Графік зміни імовірності атаки у часі

Це підтверджує і графік зміни імовірності атаки у часі, який накладає оцінки моделі на фактичний інтервал атаки. Рожева область окреслює реальну атаку, і протягом усього її тривалого періоду модель підтримує ймовірність близьку до одиниці. Така поведінка свідчить про високу повноту виявлення. Проте всередині цього інтервалу спостерігаються кілька різких спадів до нульового рівня, розташованих приблизно біля 250 та 450 вікон. Ці "провали" є пропусками атаки — ситуаціями, коли модель на тлі стабільного шкідливого трафіку все ж класифікує його як нормальний. Поза межами

фактичної атаки крива ймовірності довго утримується на нульових значеннях, однак після двох тисяч вікон з’являються ізольовані піки до 0.8–0.9. Саме ці поодинокі, але інтенсивні коливання свідчать про хибнопозитивні спрацювання, які потенційно ускладнюють експлуатацію детектора в реальних умовах.

Об’єднаний аналіз навчальних кривих, теплової карти та часової динаміки ймовірності показує, що модель демонструє високу здатність до виявлення довготривалих атак, проте її робота характеризується перенавчанням, пороговою нестабільністю та епізодичними пропусками виявлення. Такі особливості є ознаками структурних обмежень моделі в її нинішньому вигляді. У практичному контексті це означає, що модель буде ефективною лише у випадках чітко виражених та інтенсивних атак, але може втрачати якість у ситуаціях зі змінною статистикою трафіку або складнішими типами загроз. Саме тому подальше вдосконалення має включати гібридні архітектури, покращення механізмів екстракції ознак та впровадження адаптивних стратегій навчання, що дозволять підвищити стійкість та знизити.

1. Pyatachenko, V., & Serhieiev, V. (2025). Hybrid model approach with temporal feature extraction for UAV coordinate spoofing attack detection. Transactions of Kremenchuk Mykhailo Ostrohradskyi National University, (3(152)), 90–96. https://doi.org/10.32782/1995-0519.2025.3.10.
2. Abrar, M., Youssef, A., Islam, R., Satam, S., Latibari, B., Hariri, S., Shao, S., Salehi, S., & Satam, P. (2024). GPS-IDS: An Anomaly-based GPS Spoofing Attack Detection Framework for Autonomous Vehicles. https://arxiv.org/pdf/2405.08359v2.

Л.Б. Десятнюк, В.В. Ковальчук, Д.В. Протасов

КОНФІДЕНЦІЙНІСТЬ У ЦИФРОВОМУ СВІТІ: ЗАГРОЗИ ДЛЯ МЕДИЦИНИ ТА СУСПІЛЬСТВА

Цифровізація стала фундаментальною частиною розвитку сучасного суспільства. Щодня величезні масиви даних створюються, зберігаються та обробляються у різних сферах — від медицини й освіти до бізнесу, державного управління та соціальних мереж. На тлі цього зростає і кількість ризиків, пов'язаних із конфіденційністю даних, можливістю дискримінації та все більшим поширенням технологій нагляду. Ці явища впливають на базові права людини, формуючи нові етичні та правові виклики. Конфіденційність — це здатність особи контролювати доступ до особистої інформації. У цифровому середовищі це поняття набуває нового змісту, оскільки дані більше не зберігаються в паперових архівах, а перетворюються на цифрові профілі, які можуть бути скопійовані, передані або неправомірно використані за лічені секунди. Серед основних джерел ризиків — соціальні мережі, мобільні додатки, хмарні сервіси, електронні медичні картки, системи розумного міста та штучний інтелект.

Порушення конфіденційності може призвести до широкого спектра негативних наслідків. Найпоширеніші загрози включають витік персональних даних, хакерські атаки, вторинний продаж даних рекламним компаніям, також використання приватної інформації з політичною або комерційною метою. У сфері охорони здоров'я це може бути особливо небезпечним: інформація про хвороби, генетичні фактори чи психологічний стан пацієнта може бути використана для впливу на страхові тарифи, прийняття на роботу або навіть для шантажування.

Окрему загрозу становлять так звані цифрові сліди — інформація, яку людина залишає несвідомо. Алгоритми можуть аналізувати пошукові запити, пересування за GPS, історію покупок, використання додатків та робити висновки про спосіб життя, здоров'я та фінансове становище людини. Таким чином, конфіденційність стає не лише юридичним, а й технологічним викликом.

Штучний інтелект та алгоритми часто позиціонуються як об'єктивні інструменти, проте вони можуть мати вбудовані упередження. Це пов'язано з тим, що алгоритми навчаються на даних, які містять соціальні нерівності, стереотипи або нерепрезентативні вибірки. У результаті системи прийняття рішень можуть відтворювати або навіть посилювати дискримінацію.

Приклади дискримінації можуть включати:

- алгоритми, які занижують шанси певних груп на отримання кредитів;
- системи розпізнавання обличчя;
- аналіз медичних даних.

У медицині ризик алгоритмічної дискримінації є дуже високим. Якщо система оцінює шанси пацієнта на успішність лікування на основі неповних або викривлених даних, це може призвести до хибних клінічних рекомендацій або нерівного доступу до медичних послуг.

Сучасні технології нагляду охоплюють широкий спектр інструментів: від камер відеоспостереження до комплексних систем обробки даних. Нагляд може бути державним, корпоративним або приватним. Розвиток штучного інтелекту значно підвищує точність і масштаби таких систем.

Основні види цифрового нагляду:

- масове відеоспостереження з розпізнаванням обличчя;
- моніторинг соціальних мереж;
- відстеження геолокації через смартфони;
- аналіз онлайн-поведінки для створення психологічних профілів;
- контроль за електронною поштою та корпоративною активністю працівників.

Та надмірний нагляд може мати негативні психологічні та соціальні наслідки. Людина, яка знає, що за нею стежать, змінює свою поведінку — це називається «ефектом охолоджування». Він може обмежувати свободу висловлювання, креативність та суспільну активність. Для захисту прав людини та мінімізації зловживань створені міжнародні та національні правові механізми. Одним із найважливіших є GDPR — Загальний регламент захисту даних ЄС. Він встановлює правила збирання, зберігання та використання персональних даних, включно з медичними. Етичні аспекти регулювання цифрових технологій включають принципи поінформованої згоди, мінімізації обробки даних, прозорості алгоритмів, відповідальності розробників та пріоритету прав людини. Медична етика підкреслює необхідність конфіденційності як ключового елементу відносин «лікар–пацієнт».

У стоматології неналежний захист таких даних може спричинити порушення приватності пацієнтів та втрату довіри до медичних закладів. Ризики також стосуються дискримінації: якщо алгоритм, що аналізує рентген-знімки або КТ, має низьку точність для певних груп пацієнтів, це може призвести до неправильних діагнозів. Стоматологи повинні розуміти обмеження автоматизованих систем та контролювати їх роботу.

Для зменшення ризиків, пов’язаних із конфіденційністю, дискримінацією та наглядом, необхідно:

- впроваджувати сучасні стандарти кібербезпеки;
- забезпечувати прозорість алгоритмічних систем;
- проводити аудит штучного інтелекту на наявність упереджень;
- навчати медичних працівників цифровій грамотності;
- інформувати пацієнтів про їхні права;

Проблеми конфіденційності, дискримінації та технологічного нагляду є ключовими викликами цифрової епохи. Вони впливають не лише на приватне життя людини, але й на розвиток систем охорони здоров’я. Для

створення безпечного цифрового середовища необхідний комплексний підхід, який включає правові, технологічні та етичні механізми. Гармонійне поєднання цих факторів може забезпечити захист прав людини та ефективність цифрових технологій.

1. *General Data Protection Regulation (GDPR).* (n.d.). URL: https://gdpr-info.eu/.
2. Хаджирадєва, С., Безверхнюк, Т., Назаренко, О., Базика, С., & Доценко, Т. (2024). Захист персональних даних: між дотриманням прав людини та національною безпекою. *Scientific and Analytical Legal Studies*, 7(3), 245–256. https://doi.org/10.32518/sals3.2024.245. URL: https://sls-journal.com.ua/uk/journals/tom-7-3-2024/zakhist-personalnikh-danikh-mizh-dotrimannyam-prav-lyudini-ta-natsionalnoyu-bezpekoyu.
3. Верховна Рада України. (2010). Закон України “Про захист персональних даних”. URL: https://zakon.rada.gov.ua/laws/show/2297-17.

S.V. Levshchanov

DEMONSTRATION OF PRACTICAL APPLICATIONS OF ARTIFICIAL INTELLIGENCE IN ENERGY AND BANKING

The rapid development of artificial intelligence technologies is having a significant impact on various industries and areas of information technology, including cybersecurity and cloud technologies.

Cloud infrastructure is a key platform for deploying AI systems and requires integration with modern cybersecurity tools in both the energy and banking sectors [1].

A general diagram of AI integration with the cloud and cybersecurity using AWS services as an example [2]:

Stages:

1. Data collection and storage → 2. Processing and cleaning → 3. Training AI models → 4. Threat detection/prediction → 5. Response and process optimization → 6. Monitoring and updating models.

AWS services in the chain:

S3, Redshift → storage

IoT Core, Kinesis, Glue → collection and processing

SageMaker, Fraud Detector → training and predictions

GuardDuty, Macie → security

QuickSight → visualization and decision making

Security breaches and data leaks from cloud storage can significantly affect any organization in both the energy and banking sectors [3].

Traditionally, cybersecurity and the implementation of AI with cloud computing are of great importance for the transformation of the banking sector [4].

The integration of AI with cloud technologies optimizes critical business operations in various areas and also strengthens security measures and regulatory compliance in a dynamic digital environment, increasing efficiency and automation, as well as organizing data-driven decision-making [5].

The main cyber threats to digital banks are phishing attacks, social engineering, ransomware, internal threats, and DDoS attacks. Machine learning algorithms such as SVM, RNN, HMM, and LOF are effective for anomaly detection, event classification, and early attack detection. This creates a constant link between cybersecurity protocols and machine learning methods in digital banks. It also makes it mandatory to include machine learning in banks' cybersecurity policies [6].

Various types of AI are capable of classifying and detecting a huge number of cyberattacks affecting modern industrial systems in the era of Industry 4.0. Thanks to their adaptability and predictability, they effectively classify and identify hidden patterns with increased accuracy. This, in turn, promotes their implementation in modern business strategies and production processes in the energy sector. The following machine learning and deep learning models and algorithms are widely used: RF, LR, KNN, LSTM, CyRes (GC-LSTM), XGB [7].

Machine learning (ML), like other types of artificial intelligence (AI), is currently a critical tool for the effective detection of cyber threats in smart grids,

which includes both the introduction of false data and the avoidance of serious operational problems, but it also has ongoing problems such as a lack of high-quality data for training, limited model interpretability, and vulnerability to attacks by malicious actors [8].

Types of AI in cybersecurity provide automation, anomaly detection, threat analysis, and, as a result, improve cyber defense. But they also have a constant need for high-quality data. In practical terms, this means improving the security of equipment and infrastructure by optimizing and monitoring the data centers, servers, and processors responsible for this protection. This includes monitoring aspects such as equipment temperature, cooling systems, power consumption, and backup power, as well as analyzing this data and comparing it with historical information, which in turn improves equipment performance, overall infrastructure efficiency, and minimizes the financial burden of equipment and infrastructure maintenance costs necessary to protect the organization.

Types of AI automate processes by analyzing and predicting threats, identifying vulnerabilities, and helping to make decisions that overall improve the reliability and resilience of system security. However, there are also disadvantages, which include constant data requirements, the need for qualified specialists, increased equipment and infrastructure requirements, and implementation difficulties [9].

1. Liang, X., & Xu, Y. (2025). A novel framework to identify cybersecurity challenges and opportunities for organizational digital transformation in the cloud. Computers & Security, 151, 104339. https://doi.org/10.1016/j.cose.2025.104339.
2. Amazon Web Services. (n.d.). AWS documentation. https://docs.aws.amazon.com/?nc2=h_ql_doc_do.
3. Bhardwaj, A., & Kaushik, K. (2022). Predictive analytics-based cybersecurity framework for cloud infrastructure. International Journal of Cloud Applications and Computing, 12(1). https://doi.org/10.4018/IJCAC.297106.
4. Wang, S., Asif, M., Shahzad, M. F., & Ashfaq, M. (2024). Data privacy and cybersecurity challenges in the digital transformation of the banking sector. Computers & Security, 147, 104051. https://doi.org/10.1016/j.cose.2024.104051.
5. Singh, B., & Dutta, P. K. (2025). Cloud native engineering for AI-driven ERP systems: Enhancing security, compliance and data privacy. Advances in Computers. https://doi.org/10.1016/bs.adcom.2025.06.009.
6. Asmar, M., & Tuqan, A. (2024). Integrating machine learning for sustaining cybersecurity in digital banks. Heliyon, 10(17), e37571. https://doi.org/10.1016/j.heliyon.2024.e37571.
7. Lezzi, M., Montefusco, P., Lazoi, M., & Corallo, A. (2025). AI-based cybersecurity for a sustainable digital industry: Systematic literature review and future research directions. Journal of Industrial Information Integration, 48, 100980. https://doi.org/10.1016/j.jii.2025.100980.
8. Abbasi, A. R. (2025). Safeguarding the energy transition: A review of cybersecurity strategies for vulnerability management in smart grids. Energy Conversion and Management: X, 101419. https://doi.org/10.1016/j.ecmx.2025.101419.
9. Jada, I., & Mayayise, T. O. (2024). The impact of artificial intelligence on organisational cyber security: An outcome of a systematic literature review. Data and Information Management, 8(2), 100063. https://doi.org/10.1016/j.dim.2023.100063.

О.С. Ткаченко, А.В. Ільєнко

АРХІТЕКТУРНЕ ПРОТИСТОЯННЯ У СИСТЕМАХ ДЕТОКСИКАЦІЇ: BART ПРОТИ LLM

Сучасна цифрова індустрія переживає фундаментальну зміну парадигми модерації контенту, відмовляючись від традиційних методів блокування та видалення на користь автоматичної детоксикації тексту. Цей процес розглядається як складна задача перенесення стилю, де головною метою є трансформація токсичного висловлювання у нейтральне зі збереженням первинного семантичного змісту [1]. На відміну від бінарної класифікації, детоксикація вимагає від нейронних мереж глибокого розуміння контексту та нюансів, що призвело до технічного протистояння між двома домінуючими архітектурами: спеціалізованими Sequence-to-Sequence моделями (BART) та універсальними декодерними моделями (LLM). Видалення коментарів часто сприймається як цензура, тоді як детоксикація дозволяє зберегти нитку дискусії, уникаючи конфліктів [2].

Метою дослідження є здійснення порівняльного аналізу архітектур Encoder–Decoder (на прикладі BART) та Decoder-Only моделей (LLM) у задачі автоматичної детоксикації тексту, з оцінкою їхньої здатності зберігати семантику, точності трансформації, стійкості до мовних особливостей та економічної ефективності застосування.

Архітектура BART, що являє собою шумопоглинаючий автоенкодер, демонструє значні переваги в задачах, де критично важливим є збереження структури та змісту повідомлення. Завдяки поєднанню двоспрямованого кодувальника та авторегресійного декодувальника, ця модель здатна аналізувати все речення одночасно, виявляючи складні залежності між віддаленими токенами, та ефективно відновлювати «пошкоджений» токсичністю текст [3]. Натомість сучасні декодерні моделі (GPT-4, LLaMA), які використовують механізм каузальної уваги, схильні до «галюцинацій» та надмірного переписання тексту. Хоча вони забезпечують високу плинність мови, їхнє застосування часто призводить до втрати первинного сенсу повідомлення заради досягнення формальної ввічливості, що робить їх менш надійними для автоматичних систем без нагляду людини [4].

Особливої гостроти проблема вибору архітектури набуває при роботі з морфологічно багатими мовами, такими як українська. Результати змагання PAN 2024 засвідчили, що просте збільшення кількості параметрів у LLM не гарантує якості: переможцем стала компактна модель mT0-XL (архітектура, подібна до BART), яка перевершила GPT-4. Успіх спеціалізованих моделей пояснюється кращою адаптацією токенізаторів до кирилиці та здатністю Encoder-Decoder архітектури утримувати граматичні зв'язки (узгодження роду, числа, відмінка) у довгих реченнях, де декодерні моделі часто втрачають когерентність [5]. Додатковою перевагою стала інтеграція методу прямої оптимізації преференцій (Odds Ratio Preference Optimization [9]), що дозволило ефективно налаштувати модель навіть на обмеженому наборі даних [6].

Таблиця 1 – Порівняння архітектурних підходів

Характеристика	BART (Encoder-Decoder)	LLM (Decoder-Only)
Принцип роботи	Відновлення пошкодженого тексту (Denoising)	Передбачення наступного токена (Next Token Prediction)
Увага (Attention)	Двоспрямована (бачить контекст цілком)	Каузальна (бачить лише минуле)
Збереження змісту	Високе (жорстка прив'язка через cross-attention)	Помірне (ризик зміни сенсу при генерації)
Налаштування	Fine-Tuning на паралельних корпусах	In-Context Learning, RLHF, Prompting

Економічний аспект впровадження систем детоксикації також свідчить на користь спеціалізованих архітектур через явище, відоме як «податок на токени» (Token Tax). Для коректної роботи LLM вимагає включення до кожного запиту розгорнутих інструкцій та прикладів (few-shot), що призводить до ситуації, коли для обробки короткого речення система витрачає ресурси на сотні службових токенів [7]. Це робить використання LLM у реальному часі економічно неефективним порівняно з моделями BART, де всі інструкції фактично закладені у ваги під час навчання, забезпечуючи високу швидкість інференсу на доступному обладнанні [8].

Таблиця 2 – Економічна ефективність моделей

Параметр	BART-Large (Fine-tuned)	LLaMA-3 / GPT-4 (Prompting)
Використання токенів	100% корисної обробки	~5–10% корисної обробки (решта — промпт)
Обладнання	Доступні GPU (NVIDIA T4)	Потужні кластери (NVIDIA A100/H100)
Затримка (Latency)	Мілісекунди (можливе кешування)	Висока (через обробку довгого контексту)

Подальший розвиток технологій детоксикації рухається у напрямку гібридизації методів навчання. Відхід від класичного навчання з учителем (SFT), яке вимагає дорогих паралельних корпусів, на користь навчання з підкріпленням (RLHF) та новітніх методів прямої оптимізації (DPO/ORPO) дозволяє створювати ефективні системи з меншими витратами ресурсів. Найбільш перспективним сценарієм є використання потужних LLM виключно для генерації синтетичних даних та оцінки якості, тоді як фінальним виконавчим механізмом у продакшені залишаються швидкі та точні моделі архітектури Encoder-Decoder, дистильовані з урахуванням специфіки конкретної мови.

Висновки. Проведений аналіз показав, що в задачах автоматичної детоксикації тексту архітектури типу Encoder–Decoder (BART) забезпечують

кращу якість збереження семантичного змісту, нижчу схильність до генеративних викривлень та економічно ефективніший інференс порівняно з LLM. Декодерні моделі демонструють високу плинність мовлення, однак їх використання потребує значних обчислювальних ресурсів і часто призводить до надмірних трансформацій тексту. Оптимальною стратегією є поєднання можливостей великих мовних моделей для створення навчальних даних із застосуванням компактних Encoder–Decoder моделей у виробничих системах.

1. Dementieva, D., et al. (2024). SmurfCat at PAN 2024 TextDetox: Alignment of Multilingual Transformers for Text Detoxification. CEUR Workshop Proceedings. https://ceur-ws.org/Vol-3740/paper-276.pdf.
2. Faggioli, G., et al. (2025). GemDetox at TextDetox CLEF 2025: Enhancing a Massively Multilingual Model for Text Detoxification on Low-resource Languages. University of Padua. https://www.dei.unipd.it/~faggioli/temp/clef2025/paper_288.pdf.
3. Hosseini, P., et al. (2024). DetoxLLM: A Framework for Detoxification with Explanations. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. https://aclanthology.org/2024.emnlp-main.1066.pdf.
4. Lewis, M., et al. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. arXiv preprint arXiv:1910.13461. https://arxiv.org/abs/1910.13461.
5. Logacheva, V., et al. (2025). LLM in the Loop: Creating the ParaDeHate Dataset for Hate Speech Detoxification. arXiv preprint arXiv:2506.01484. https://arxiv.org/html/2506.01484v2.
6. NVIDIA. (n.d.). Mastering LLM Techniques: Inference Optimization. NVIDIA Technical Blog. https://developer.nvidia.com/blog/mastering-llm-techniques-inference-optimization/.
7. Protasov, Vitaly & Babakov, Nikolay & Dementieva, Daryna & Panchenko, Alexander. (2025). Evaluating Text Style Transfer: A Nine-Language Benchmark for Text Detoxification. 10.48550/arXiv.2507.15557.
8. Pu, X., et al. (2024). GPT-DETOX: An In-Context Learning-Based Paraphraser for Text Detoxification. arXiv preprint arXiv:2404.03052. https://arxiv.org/pdf/2404.03052.
9. AI Engineering Academy. (n.d.). ORPO. https://aiengineering.academy/LLM/TheoryBehindFinetuning/ORPO/.

I. Dorohyi, V. Kravchuk, V. Tsurkan

GENERATIVE AI FOR CRITICAL INFORMATION INFRASTRUCTURE PROTECTION

Abstract. Critical information infrastructures (CIIs) – such as energy grids, water systems, and transportation networks – are increasingly targeted by sophisticated cyber-attacks. Generative artificial intelligence (AI) techniques (e.g. large language models, GANs, diffusion models) offer new approaches to detect and mitigate these threats. This report reviews recent literature on applying generative AI to CII protection and discusses applications in threat detection, incident response automation, deception (e.g. honeypots), and resilience. We conclude that generative AI holds great promise for enhancing CII security but faces challenges including data scarcity, model trustworthiness, and adversarial risks. Future research should focus on domain-specific model training, robust evaluation, and integration with digital twins and resilient architectures.

Keywords: Critical Infrastructure Protection, Generative AI, Large Language Models, Generative Adversarial Networks, Cybersecurity.

Introduction. Critical information infrastructures (CIIs) – including energy, water, transportation, and communication networks – are essential to societal functioning but face escalating cyber threats. Recent years have seen sharp rises in targeted attacks on infrastructure systems, causing physical disruptions and economic damage. Traditional security measures (IDS, firewalls, etc.) are proving insufficient in the face of complex, AI-driven threats. Meanwhile, generative AI – such as large language models (LLMs) and generative adversarial networks (GANs) – is maturing rapidly. These models can autonomously generate realistic content and have begun to yield novel cybersecurity tools. For example, CII operators now experiment with AI-driven threat-hunting and automated analysis tools, and there is growing interest in using generative techniques for proactive defense [1, 2]. This paper explores how generative AI can protect CIIs by improving threat detection, accelerating response, enabling deception (e.g. synthetic honeypots), and building resilience. We first survey key literature, then discuss technical approaches and examples in each area, and finally synthesize insights on future directions.

Literature Review. Recent studies have begun to chart generative AI's role in cybersecurity and CII protection. Yigit et al. (2025) provide a comprehensive review of AI-driven approaches to critical infrastructure protection. They emphasize benchmarks for evaluating LLMs on CII tasks and discuss trust, privacy, and resilience issues. Importantly, they highlight the emerging use of generative AI/LLMs in CIP (Critical Infrastructure Protection) for proactive defenses and outline agentic AI methods [1]. Similarly, Crichton et al. (2024) note that "new generative AI techniques have become more capable and offer novel

opportunities for CI operators", such as enhanced threat intelligence synthesis and AI agents for automation. They caution, however, that generative AI also brings difficult new risks and requires careful governance [2].

Generative adversarial networks have been studied for cybersecurity defense. Ndayipfukamiye et al. (2025) systematically review GAN-based methods for threat detection. They find that GANs can significantly improve detection accuracy and robustness across domains (network intrusion, malware, IoT) by augmenting training data and simulating adversarial scenarios. For example, conditional GANs can produce realistic malicious samples to bolster anomaly detectors. However, GAN defenses face challenges like training instability, lack of standard benchmarks, and limited interpretability [3].

On deception strategies, Ahmed et al. (2025) demonstrate how LLMs can automate adaptive cyber-deception. Their SPADE framework uses structured prompt engineering to generate scaled deception ploys (fake credentials, decoy documents) for diverse malware scenarios. In experiments, advanced LLMs (e.g. ChatGPT-4o) outperformed smaller models in creating coherent, contextually relevant deceptive content [4]. This work shows generative AI can greatly enhance honeypot realism and engagement.

Other recent work discusses broad implications of generative AI in cybersecurity resilience. Radanliev et al. (2025) use a PRISMA review to outline how generative AI is employed for automated threat detection, adversarial simulation, and defense acceleration. They also warn that these systems create new attack vectors (deepfakes, malware-as-code) and that governance frameworks are lagging [6]. Overall, the literature suggests promising applications of generative models for CII protection but emphasizes careful evaluation of risks, ethics, and data governance.

Generative AI in Threat Detection and Analysis. Generative models enhance threat detection in CIIs by augmenting data and intelligence analysis. LLMs can process vast unstructured information (log files, threat reports, forums) to identify patterns indicative of attacks. For instance, researchers describe using GPT-4 to sift through online hacker forums and system logs to "predict and identify potential cyber threats, including phishing attacks or malware aimed at energy grid systems" [7]. By summarizing reports and highlighting anomalies, LLMs help analysts detect emerging threats more quickly.

Meanwhile, GAN-based systems improve anomaly detection by generating synthetic malicious traffic. GANs can simulate rare attack scenarios that are underrepresented in real-world logs [3]. For example, a GAN trained on network flow data can produce stealthy intrusion traces to challenge an IDS. Studies report that integrating GAN-generated samples into training datasets raises detection accuracy and robustness in network intrusion and IoT sensors. A recent digital-twin study used a hybrid model (LSTM+GAN) to create realistic multivariate network flows for EV charging infrastructures. This allows validating IDS under rare yet critical attacks (e.g. charging profile manipulation) by preserving temporal

sequencing in the synthetic data [5]. In summary, generative AI can produce representative attack and normal data to train better detectors and enrich threat intelligence.

Incident Response and Automation. Generative AI can streamline incident response and mitigation. LLMs like BERT or GPT variants excel at reading technical texts and extracting relevant details. One approach uses BERT to rapidly parse incident reports and security logs after a breach. For example, a public transit authority could apply BERT to multiple incident reports following a cyber attack on its railway network to identify common vulnerabilities or indicators of compromise [7]. By summarizing key information, the model enables faster situational awareness and prioritization.

Beyond analysis, LLMs can draft response content. Given a summary of an attack, a generative model could suggest mitigation steps, compliance documentation, or forensics templates. Prompt-engineered GPT models might produce post-incident press releases or technical write-ups compliant with regulatory language (e.g. automating SEC breach notifications). In one hypothetical scenario, a water treatment plant uses an LLM to generate a complete audit report after detecting unauthorized access – the model ingests log excerpts and regulations to ensure accuracy. These applications increase response speed and reduce manual workload.

However, ensuring trust in AI-generated guidance is crucial. Models must be robust to hallucinations or adversarial prompts. Techniques like reinforcement learning from human feedback (RLHF) and domain-specific fine-tuning are needed so that LLM outputs (e.g. an incident-response checklist) are reliable and aligned with experts' intent. Integration with decision workflows and human oversight will be key to safely adopting generative response automation.

Deception and Honeypots. A powerful use of generative AI is active cyber-deception. Unlike static honeypots, LLM-driven deception can adapt in real time. The SPADE framework illustrates this by having LLMs autonomously create deceptive content tailored to ongoing attacks [4]. For example, if reconnaissance activity is detected on a SCADA network, a deployed generative agent could fabricate additional fake sensor readings and dummy control commands to lure the attacker deeper. Or, upon spotting credential-phishing attempts, an LLM could generate convincing decoy documents and fake login portals that mimic the real control system but are isolated.

Such adaptive deception scales beyond what human operators can configure manually. Ahmed et al. report that ChatGPT-4o achieved 93% engagement (as measured by expert assessment) in generating context-relevant malware lures [4]. This suggests state-of-the-art LLMs can craft believable text (fake emails, policy memos, chat transcripts) that draw attackers into honey-environments. In practice, a smart-grid operator might deploy generative honeypots producing realistic feeder status logs and simulated intrusion artifacts. When attackers interact with these systems, their tactics are revealed without endangering real assets.

Deception via generative models also includes data masking and encryption. GANs or diffusion networks could generate synthetic versions of critical data streams (e.g. decoy chemical plant sensor outputs) to feed to potential intruders. The goal is to introduce uncertainty into the attacker's model of the system, forcing them to reveal themselves or waste effort. Overall, generative deception bolsters resilience by actively countering adversaries with intelligent traps and counterfeit information.

Resilience and Simulation. Generative AI supports CII resilience by simulating attack scenarios and alternative operations. Digital twins of infrastructure can incorporate generative data to test defenses under stress. For instance, a power grid twin might use a GAN to create time-series of demand surge patterns combined with cyber events, evaluating automated load-shedding responses. The hybrid LSTM/GAN framework mentioned earlier enabled accurate simulation of EV charging attacks for evaluating intrusion detection [5]. Such synthetic scenarios help planners identify vulnerabilities and refine contingency protocols.

Generative models can also aid resilience by augmenting rare event data. Many high-impact CII attacks are low-frequency, so GANs can enlarge these datasets to improve statistical models of risk. Moreover, diffusion models might be explored for anomaly-driven forecasting (e.g. generating possible sensor output sequences under hypothesized failures). Preliminary research suggests diffusion-based methods are promising for capturing complex temporal behaviors (though this is an emerging area).

Additionally, LLMs and generative tools contribute to resilience by enhancing human preparedness. AI-driven "what-if" analysis can automatically produce reports on potential cascading failures (e.g. "If subsea cable is compromised, how would data traffic reroute?"). These narrative simulations inform decision-makers. Generative AI can also optimize routine tasks (e.g. translating technical manuals, updating recovery playbooks), keeping staff focused on critical resilience tasks.

Nonetheless, data availability is a constraint. Effective training of domain-specific generative models requires realistic CII datasets, which are often proprietary or sparse [7]. Collaboration between operators and model developers is needed to share sanitized data. Privacy-preserving techniques (like federated learning or on-site model deployment) can mitigate data sensitivity during generative training.

Conclusion. Generative AI techniques offer powerful new tools for protecting critical information infrastructure. They enhance threat detection by enabling sophisticated data augmentation and large-scale intelligence analysis. They speed and automate response actions by generating reports and extracting insights from vast logs. They revolutionize deception by creating adaptive, realistic honeypots and decoys. And they aid resilience by simulating attacks in digital twins and broadening scenario coverage.

However, realizing these benefits requires overcoming challenges. Critical infrastructure data is often scarce or sensitive, so building and fine-tuning generative models can be difficult. Models must be highly reliable to avoid introducing false assurances or new vulnerabilities (e.g. AI could hallucinate or produce exploitable outputs). Adversaries also use generative AI to enhance attacks (e.g. automated phishing or exploit generation), raising the stakes. Mitigation will demand robust validation (benchmarking CIP-specific tasks), adversarial training, and human-in-the-loop oversight.

Future research should pursue sector-specific generative models (e.g. specialized LLMs for power grid or water systems), standardized evaluation frameworks (e.g. simulated red-team exercises with generative attack tools), and integration with digital twin platforms. Hybrid approaches combining generative and traditional AI (e.g. GANs guiding symbolic security rules, LLMs plus rule-checkers) may yield practical gains. Policy and governance are also critical: as one review notes, generative systems are advancing faster than institutional safeguards. Building transparent, ethically aligned CI AI requires cross-domain collaboration among cybersecurity experts, AI researchers, and regulators. In sum, generative AI has significant potential to strengthen CI resilience, but its power must be harnessed with rigorous safety, ethical, and operational frameworks.

1. Yigit Y., Ferrag M.A., Ghanem M.C. та ін. Generative AI and LLMs for Critical Infrastructure Protection: Evaluation Benchmarks, Agentic AI, Challenges, and Opportunities. Sensors. 2025. Vol. 25, No. 6. P. 1666. DOI: 10.3390/s25061666.
2. Crichton K., Ji J., Miller K. та ін. Securing Critical Infrastructure in the Age of AI. Center for Security and Emerging Technology (CSET) Analysis. 2024 (жовт.). URL: https://surl.li/uwnikx (дата звернення: 05.11.2025).
3. Ndayipfukamiye T., Ding J., Sebastian S. та ін. Adversarial Defense in Cybersecurity: A Systematic Review of GANs for Threat Detection and Mitigation. 2025. arXiv: 2509.20411.
4. Ahmed S., Rahman M., Alam M. та ін. SPADE: Enhancing Adaptive Cyber Deception Strategies with Generative AI and Structured Prompt Engineering. 2025. arXiv: 2501.00940.
5. Iturbe E., Arcas J., Gaminde G. та ін. Hybrid AI-Based Framework for Generating Realistic Attack-Related Network Flow Data for Cybersecurity Digital Twins. Applied Sciences. 2025. Vol. 15, No. 21. P. 11574. DOI: 10.3390/app152111574.
6. Radanliev P., Santos O., Ani U.D. Generative AI Cybersecurity and Resilience. Frontiers in Artificial Intelligence. 2025. Vol. 8. P. 1568360. DOI: 10.3389/frai.2025.1568360.
7. Yigit Y., Ferrag M.A., Sarker I.H. та ін. Critical Infrastructure Protection: Generative AI, Challenges, and Opportunities. 2024. arXiv: 2405.04874. URL: https://surl.li/muvenf (дата звернення: 05.11.2025).

С.О. Кашкевич, О.Л. Стокіпний

ЗАСТОСУВАННЯ ШІ ДЛЯ ПІДВИЩЕННЯ СТІЙКОСТІ ІНФОРМАЦІЙНИХ СИСТЕМ ДО КІБЕРЗАГРОЗ

Розвиток сучасних інформаційних технологій призвів до істотного ускладнення структури кіберпростору та появи багаторівневих загроз. Традиційні засоби кіберзахисту, побудовані на сигнатурному аналізі або жорстко визначених правилах, втрачають ефективність у динамічних середовищах. Тому актуальним є створення інтелектуальних адаптивних систем, здатних самостійно виявляти нові типи аномалій та модифікувати власні алгоритми в процесі роботи.

Теоретичною основою є концепція інтелектуального аналізу даних (Data Mining), яка передбачає багатокрокову обробку потоків даних: збір, очищення, трансформацію, класифікацію та візуалізацію. У межах машинного навчання виділяють дві групи методів: з учителем (supervised) і без учителя (unsupervised). Для задач кіберзахисту доцільно поєднувати їх у гібридній моделі, яка дозволяє одночасно навчатися на позначених вибірках і виявляти нові невідомі загрози.

Методологія базується на гібридній архітектурі, що поєднує методи класифікації та кластеризації. Для формалізації процесу оцінки аномалій використано функцію ризику:

$$Ri = \alpha \cdot P(Ai) + \beta \cdot \Delta Di, \quad (1)$$

де Ri – ризик для *i*-го вузла, $P(Ai)$ – ймовірність аномальної активності, ΔD_i – відхилення параметрів від нормального профіл, α, β –вагові коефіцієнти.

Для перевірки ефективності було проведено моделювання на тестовому наборі даних із 10 000 записів мережевих сесій. Порівняльні результати подано в таблиці 1.

Таблиця 1 – Результати дослідження

Метод виявлення	Точність, %	Час реакції, с
Сигнатурний IDS	78,4	2,5
Машинне навчання (SVM)	89,7	1,4
Гібридна модель (запропонована)	93,1	1,1

Отримані результати демонструють покращення точності виявлення вторгнень на 14,7% та скорочення часу реакції майже удвічі порівняно з традиційними підходами. Розроблений метод може бути основою для

побудови адаптивних систем кіберзахисту, здатних до самооновлення та навчання на нових загрозах.

З теоретичного погляду перспективним напрямом подальших досліджень є поєднання підходів нейронних мереж і нечіткої логіки (fuzzy-systems) для побудови систем, що здатні працювати за умов невизначеності та приймати рішення з урахуванням лінгвістичних правил. Такі системи можуть реалізовувати принцип самоорганізації, коли модель змінює власну структуру відповідно до характеру вхідних даних. Це дозволить забезпечити глибшу адаптивність і стійкість інформаційних систем майбутнього.

Висновки. Використання алгоритмів штучного інтелекту для аналізу потоків даних у реальному часі забезпечує створення ефективних адаптивних систем кіберзахисту. Інтеграція методів машинного навчання, нечіткої логіки та самоорганізації відкриває нові можливості для побудови інтелектуальних інформаційних систем, здатних до самонавчання та прогнозування нових типів загроз.

1. Kashkevich, S. (Ed.) (2025). Decision support systems: mathematical support. Kharkiv: TECHNOLOGY CENTER PC, 202. https://doi.org/10.15587/978-617-8360-13-9.
2. Tamer, K. A., Sova, O., Shaposhnikova, O., Yashchenok, V., Stanovska, I., Shostak, S., Rudenko, O., Petruk, S., Matsyi, O., & Kashkevich, S. Development of a solution search method using a combined bio-inspired algorithm. EasternEuropean Journal of Enterprise Technologies. 2024, Vol. 1, No. 4 (127), pp. 6–13. https://doi.org/10.15587/1729-4061.2024.298205.
3. Development of a methodological approach for assessing the condition of complex organizational and technical systems. Mohammed, B. A., Stanovska,I., Kashkevich, S., Lebedynskyi, A., Vakulenko, Y., Protas, N., Klyuchak, O., Lastivka, O., Semeniuk, A., Kivshar, O. Eastern-European Journal of Enterprise Technologies, 2/4 (134) 2025, 47 –53. https://doi.org/10.15587/1729-4061.2025.326468.

K.O. Manolova

THE USE OF ARTIFICIAL INTELLIGENCE IN ENTERPRISE MANAGEMENT

In today's dynamic business environment, strategic growth of enterprises is impossible without deep integration of artificial intelligence (AI) into management processes. AI technologies offer a transition from traditional management to intelligent, predictive and adaptive management, providing competitive advantages through real-time data analysis and informed decision-making [1].

Artificial intelligence extends beyond standard automation to transform the basis of management decision-making. According to data from Precedence Research, the generative artificial intelligence market is experiencing explosive growth, from $10.79 billion in 2023 to a projected $1.3 trillion by 2032, representing an annual growth rate of 42% over the next decade [2]. This macrotrend is confirmed by the large-scale investments from technology leaders, who allocated about $400 billion in capital expenditures in 2024, mainly related to the development of AI.

A key aspect of successful use of AI is understanding the specific tools and their impact on business processes. According to our research, machine learning finds its application in operations management to forecast demand and optimize supply chains, and to create personalized offers in marketing. It should be noted that natural language processing tools are implemented through AI assistants and chatbots, increasing the efficiency of customer service and automating reporting. Computer vision technology opens up new opportunities for quality control and safety, conducting predictive maintenance of equipment.

A good example of successful implementation is the experience of Siemens, which, due to predictive analysis algorithms, was able to reduce equipment downtime by 30% and increase production efficiency by 15% [3]. Such solutions demonstrate that even basic integration of machine learning algorithms can bring significant economic benefits.

The journey of digital transformation is not devoid of serious challenges. One of the most important issues, according to scientists, remains data quality, since the effectiveness of any AI system is governed by the "garbage in, garbage out" principle. In addition, the cost and complexity of integrating such solutions into the existing IT infrastructure can be very high. It should be emphasized that in addition to the technical aspects, ethical issues arise, in particular the algorithmic bias that can inherit and amplify social stereotypes, leading to biased decisions in the areas such as staff recruitment or lending [4].

The analysis of scientific sources has shown that effective implementation of artificial intelligence requires consistent actions and strategic planning. Therefore, the most effective approach is considered to be the one that involves phased implementation with a focus on quick and measurable results. In particular, we will define the following stages:

1. Diagnostics and prioritization of the organization's activities. The key tasks of this stage are: audit of business processes, identification of "bottlenecks", selection of 1-2 tasks for piloting the project.

2. Resource assessment and solution selection. The key tasks of this stage are: data quality assessment; analysis of available expertise and budget; choice between SaaS solution and custom development.

3. Scaling and integration. The key tasks of this stage are: analyzing POC results, developing a scaling plan and integrating the solution into related processes.

Regarding the expected results of the implementation of the above-mentioned stages, it should be noted that:

- at the first stage, we need to obtain a clearly formulated list of goals for the pilot project with defined priorities;
- at the second stage, a specific tool or service provider for the pilot project is established.
- at the third stage, a strategy for full-scale implementation of AI was formed and substantiated.

The proposed phased approach allows for a systematic process of transformation, starting from diagnostics and prioritization of tasks, moving through resource assessment and launching a pilot project, to scaling and full solution integration. This approach makes it possible to minimize risks and ensure maximum return on investment in artificial intelligence technologies [5].

Therefore, the integration of artificial intelligence into managerial processes is not a short-term trend, but a strategic necessity to increase the enterprise competitiveness. Effective implementation of AI requires not only technical investments, but also a change in corporate mindset, the development of a digital culture and continuous staff training. Successful companies of the future are those that will be able to combine technological innovations with human potential, creating truly intelligent organizations.

1. Anti-Crisis Business Scaling: Experience of the "Nova" Group of Companies Entering the European Markets. URL: https://surl.li/ynywju (Accessed: 20.10.2025).
2. Siemens is Releasing its New AI-based Assistant. URL: https://surl.li/y n ywju.
3. Myths and facts about AI: What Every Manager Should Know. URL: https://surl.li/nwkenr (Accessed: 19.10.2025).
4. AI for Business: Where to Start Implementing Artificial Intelligence in Your Own Business. URL: https://surl.lu/lmusei (Accessed: 21.10.2025).
5. How AI is Changing Business. URL: https://surl.li/soqncy (Accessed: 10.10.2025).

Д.О. Персунов, Н.В. Апенько

ІНТЕЛЕКТУАЛЬНІ АГЕНТИ В СИСТЕМАХ КІБЕРБЕЗПЕКИ: КОНЦЕПЦІЯ АДАПТИВНОГО ЗАХИСТУ НА ОСНОВІ ШІ

У міру зростання складності інформаційних систем та обсягів цифрових потоків класичні підходи до кіберзахисту втрачають ефективність. Традиційні системи виявлення загроз зазвичай працюють у реактивному режимі – реагують після фіксації інциденту, не маючи змоги передбачити подію. Такий підхід не відповідає сучасним вимогам, адже кібератаки стали багатоступеневими, автоматизованими та часто самонавчальними. У цих умовах виникає потреба у впровадженні інтелектуальних агентів – автономних програмних модулів, здатних до самонавчання, адаптації та колективної взаємодії у процесі захисту інформаційного середовища.

Поняття інтелектуального агента (intelligent agent) базується на принципах автономності, реактивності, проактивності та соціальності. Агент може самостійно збирати дані про стан системи, аналізувати загрози, приймати рішення та обмінюватися інформацією з іншими агентами. Теоретичною основою агентних систем є мультиагентна архітектура (MAS – Multi-Agent Systems), у якій безпека досягається не централізованим контролем, а колективною взаємодією. Такий підхід дозволяє моделювати поведінку складних систем, де кожен агент виконує власну функцію: спостереження, аналіз, прогнозування, реагування чи відновлення.

Інтелектуальні агенти інтегрують у собі механізми машинного навчання, нейронних мереж, нечіткої логіки та еволюційних алгоритмів, що забезпечує здатність до самоорганізації. Вони формують адаптивне середовище, яке не просто протидіє загрозам, а й еволюціонує разом із ними.

Структура агентної системи безпеки: агенти-монітори, які збирають дані про мережеву активність і стан вузлів; аналітичні агенти, що здійснюють класифікацію подій і виявлення аномалій; агенти-рішень, які визначають рівень ризику та запускають сценарії реагування; координаційні агенти, що узгоджують дії між різними вузлами системи; агенти-навчання, які оновлюють моделі поведінки на основі зворотного зв’язку.

У центрі системи розташований модуль інтелектуального ядра, який координує обмін даними між агентами та підтримує спільну базу знань. Таке середовище є самоадаптивним: якщо один з агентів виявляє невідомий тип загрози, інформація негайно поширюється до інших, і вся система оновлює власні параметри без участі оператора.

У центрі системи розташований модуль інтелектуального ядра, який координує обмін даними між агентами та підтримує спільну базу знань. Таке середовище є самоадаптивним: якщо один з агентів виявляє невідомий тип загрози, інформація негайно поширюється до інших, і вся система оновлює власні параметри без участі оператора.

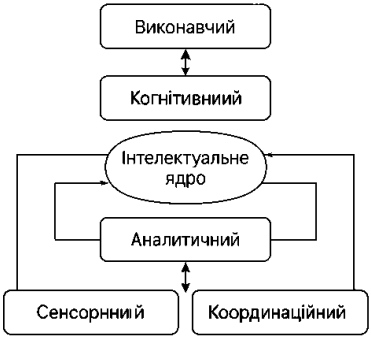

Рисунок 1 – Концептуальна схема мультиагентної системи кібербезпеки

З позиції теорії складних систем, мультиагентна модель реалізує принцип самоорганізації – здатності структури змінювати власні параметри відповідно до зовнішніх умов. Це забезпечує: адаптивність, тобто підлаштування до нових типів атак; стійкість, завдяки відсутності єдиної точки відмови; масштабованість, оскільки агенти можуть додаватися без змін у структурі; розподілену обробку даних, що скорочує час реакції; колективне навчання, коли досвід окремого агента стає спільним знанням системи.

Агентні підходи є синтезом кількох парадигм – системного аналізу, штучного інтелекту та кібернетики. Вони створюють основу для реалізації концепції «Active Cyber Defense» – активного захисту. Інтелектуальні агентні системи безпеки – це новий етап розвитку технологій кіберзахисту. Вони поєднують автономність, адаптивність і колективну взаємодію, дозволяючи системі не лише виявляти атаки, а й прогнозувати їх виникнення. Подальший розвиток цього напряму пов’язаний із інтеграцією мультиагентних середовищ із хмарними платформами, інтернетом речей та квантовими обчисленнями, що забезпечить формування самонавчальних, масштабованих і повністю розподілених систем кібербезпеки.

1. Nechyporuk O., Nechyporuk V., Kashkevich I-F., Poburko O., Suprun O., Apenko N. Identification of combinations of faults in multilevel information systems // The perspective technologies and methods in MEMS Design (MEMSTECH), IEEE 2020. – Львів, 2020. – 76-81 с.
2. Stepanenko, A., Oliinyk, A., Deineha, L., & Zaiko, T. (2018). Development of the method for decomposition of superpositions of unknown pulsed signals using the second-order adaptive spectral analysis. Eastern-European Journal of Enterprise Technologies, 2(9 (92), 48–54. https://doi.org/10.15587/1729-4061.2018.126578.
3. Gupta N., Jolly S. Enhancing Data Quality at ETL Stage of Data Warehousing // International Journal of Data Warehousing and Mining (IJDWM). 2021. Vol. 17, Issue 1. P. 74-91.

О.І. Ластівка, О.П. Нечипорук

ГІБРИДНІ СИСТЕМИ ВИЯВЛЕННЯ КІБЕРАТАК ІЗ ВИКОРИСТАННЯМ ШТУЧНОГО ІНТЕЛЕКТУ

Сучасні тенденції розвитку інформаційних технологій призвели до значного ускладнення архітектури інформаційних систем, що своєю чергою створює нові виклики у сфері кібербезпеки. Умови постійного зростання кількості атак і різноманітності їхніх форм зумовлюють необхідність переходу від традиційних сигнатурних систем до інтелектуальних, які поєднують методи аналізу даних, прогнозування та самоорганізації. Одним із перспективних напрямів розвитку таких технологій є гібридні системи виявлення кібератак, у яких штучний інтелект використовується для аналізу, класифікації та оцінювання ризиків у режимі реального часу.

Традиційні системи виявлення вторгнень (IDS) працюють за принципом зіставлення вхідних подій із базою відомих шаблонів атак. Хоча цей підхід ефективний для вже відомих загроз, він є малоефективним у випадках нових або модифікованих атак. Натомість системи, побудовані на основі штучного інтелекту (ШІ), мають здатність до самонавчання, що дозволяє їм виявляти невідомі загрози шляхом аналізу поведінкових аномалій у даних. Використання алгоритмів машинного навчання забезпечує динамічне оновлення моделей без участі оператора та суттєво знижує час реагування на інциденти. Гібридні системи поєднують переваги сигнатурного та аномального підходів. У них використовуються як бази знань, що містять відомі шаблони атак, так і модулі інтелектуального аналізу, які виявляють нові відхилення від норми. Теоретичною основою побудови таких систем є комбінація кількох методів: класифікаційних алгоритмів (Decision Tree, Random Forest, Gradient Boosting); кластеризаційних підходів (K-Means, DBSCAN); нейронних мереж, які формують поведінкові профілі користувачів; методів нечіткої логіки.

Сутність гібридного підходу полягає в тому, що на першому етапі система проводить фільтрацію даних за сигнатурними правилами, а на другому – передає потенційно підозрілі події до модуля інтелектуального аналізу. Така архітектура зменшує навантаження на систему, водночас підвищуючи точність виявлення та знижуючи кількість хибних спрацювань.

Одним із ключових елементів ефективності є правильний вибір ознак для навчання моделей. До них можуть належати кількість пакетів, обсяг переданих даних, інтервали між запитами, використання портів, частота звернень до серверів тощо. Аналіз цих характеристик дозволяє створити поведінкові профілі нормальної активності користувачів і системних процесів. Виявлення істотних відхилень від цих профілів сигналізує про можливу атаку.

Теоретичні дослідження свідчать, що гібридні системи можуть реалізовувати принцип самоорганізації, коли підсистеми адаптуються до поточних умов функціонування та змінюють свої параметри без зовнішнього втручання. Це наближає їх до концепції інтелектуальних багатошарових систем, де кожен рівень відповідає за певний аспект безпеки – від моніторингу до прийняття рішень. Нижче наведено узагальнену таблицю з порівнянням основних типів гібридних моделей.

Таблиця 1 – Порівняння систем

Тип гібридної системи	Основні методи	Точність	Адаптивність	Обчислювальна складність
ML + Statistical	Random Forest + аналіз частот	90–92	Середня	Низька
ML + Fuzzy Logic	Decision Tree + нечітка логіка	93–95	Висока	Середня
Deep Learning + Rules	CNN/LSTM + експертні правила	95–97	Дуже висока	Висока
ML + Blockchain	Random Forest + розподілений журнал	92–94	Висока	Висока
Комплексна H-IDS	ML + Fuzzy Logic + DL	96–98	Дуже висока	Висока

Як видно з таблиці, найвищі показники точності (до 98%) демонструють комплексні моделі, що поєднують глибинне навчання з нечіткою логікою. Такі системи характеризуються високою адаптивністю, проте вимагають значних обчислювальних ресурсів. Їх доцільно використовувати в критичних інформаційних інфраструктурах або хмарних середовищах із динамічним навантаженням. Гібридні системи виявлення кібератак із використанням штучного інтелекту є ключовим напрямом розвитку сучасної кібербезпеки. Вони забезпечують баланс між точністю, швидкодією та адаптивністю, поєднуючи переваги класичних та інтелектуальних методів аналізу.

1. Shknai, O., Sova, O., Nechyporuk, O., Nalapko, O., Buyalo, O., & Lyashenko, A. (2025). Chapter 3. A set of methods for enhancing the efficiency of information processing in intelligent decision support systems. In S. Kashkevich (Ed.), DECISION SUPPORT SYSTEMS: MATHEMATICAL SUPPORT (pp. 62–94). Kharkiv: TECHNOLOGY CENTER PC. https://doi.org/10.15587/978-617-8360-13-9.ch3.
2. Stallings, W. (2023). Cryptography and network security: Principles and practice (8th ed.). Pearson.

К.І. Гоцуляк

ШТУЧНИЙ ІНТЕЛЕКТ І ЦИФРОВА ЕТИКА У ВИХОВАННІ ПІДРОСТАЮЧОГО ПОКОЛІННЯ

Розвиток штучного інтелекту кардинально змінює не лише економіку та суспільство, а й освітній простір. Сучасні діти взаємодіють із ШІ ще до школи - через ігри, соціальні мережі, навчальні платформи. Це формує нові ціннісні орієнтири, способи сприйняття інформації та соціальної взаємодії. Тому виникає нагальна потреба у вихованні цифрової етики - сукупності норм, що визначають відповідальне, безпечне та моральне використання технологій.

Визначити роль штучного інтелекту у процесі формування цифрової етики та обґрунтувати педагогічні підходи до її розвитку в учнівської молоді.

У контексті глобальних викликів - від упереджень алгоритмів до порушень приватності - освітня взаємодія зі ШІ має супроводжуватися етичним супроводом. Наприклад, UNESCO в документі «Recommendation on the Ethics of Artificial Intelligence» зазначає, що впровадження ШІ повинно базуватись на захисті прав людини, недопущенні дискримінації, забезпеченні прозорості та участі всієї спільноти [2].

У педагогічному середовищі це означає не просто використання технологій, а формування культури відповідального користування ними.

Штучний інтелект як освітній інструмент може сприяти персоналізації навчання, розвитку критичного мислення й цифрової грамотності, однак потребує чітких етичних рамок використання.

Так, наприклад, World Economic Forum виділяє сім принципів відповідального використання ШІ в освіті - серед них: спрямованість на освітню мету, інклюзивність, справедливість, захист приватності. [3].

Цифрова етика — це не лише знання про безпечне користування технологіями, а й формування моральної відповідальності за дії в цифровому середовищі. Педагог має виховувати в учнях свідоме ставлення до власних дій онлайн, розуміння, як алгоритми можуть впливати на їхнє оточення, і як уникати цифрових пасток (дезінформація, маніпуляція, вторгнення в приватність).

Роль педагога полягає у вихованні критичного ставлення до інформації, розвитку медіаграмотності, розумінні питань авторства, приватності та академічної доброчесності при використанні ШІ. Педагог також виконує функцію фасилітатора - створює умови для діалогу, рефлексії й обговорення етичних аспектів.

Практичні кроки:

впровадження уроків або модулів із цифрової етики, присвячених темі ШІ;

використання кейсів, що демонструють моральні дилеми, пов’язані з ШІ (наприклад: «Чи можна довіряти штучному алгоритму для оцінювання учнів?»);

розвиток у школярів навичок етичного використання технологій і співпраці з алгоритмами - наприклад, навчання формулювати питання до ШІ, перевіряти результат, критикувати можливі помилки;

забезпечення інклюзивності - педагог має звертати увагу на те, щоб усі учні мали доступ до навчання про цифрову етику, незалежно від соціально-економічного стану або регіону [4].

Виховання цифрової етики є одним із ключових завдань сучасної освіти, адже саме від ціннісних орієнтирів молоді залежить, як штучний інтелект буде впливати на суспільство у майбутньому. Педагог має не лише навчати користуванню цифровими інструментами, а й формувати у здобувачів освіти розуміння меж їх застосування, відповідальності за власні дії в онлайн-просторі та усвідомлення моральних наслідків рішень, прийнятих за допомогою ШІ.

Забезпечення балансу між технологічним прогресом і гуманістичними цінностями має стати пріоритетом освітньої політики, спрямованої на розвиток покоління, здатного жити та діяти у світі, де людина і штучний інтелект співіснують у взаємоповазі й етичній гармонії [1; 5].

Крім того, важливо підкреслити, що цифрова етика має розвиватися як процес, а не лише як набір правил. Учні повинні мати змогу рефлексувати над власним досвідом, обговорювати помилки, аналізувати інциденти й спільно з педагогом створювати рекомендації щодо безпечного й відповідального підходу до ШІ. Освітня середа має бути відкритою до змін, адже технології швидко розвиваються - і лише постійне оновлення знань, навичок і цінностей дозволить виховувати покоління, готове до майбутнього.

Також слід враховувати, що навчальний процес із акцентом на цифрову етику допомагає не лише уникнути ризиків, але й використовувати потенціал ШІ для навчання як збагачуючого, трансформаційного досвіду. Учні, озброєні етичними підходами, стають не просто користувачами технологій, а активними учасниками цифрової спільноти, здатними критично мислити, вибирати й впливати на майбутнє технологічного світу.

1. Біотехнічний університет. (2025). Цифрова дидактика: навчально-методичний посібник. Біотехнічний університет. https://biotechuniv.edu.ua/wp-content/uploads/2025/02/015b-24rv-ok8nmp.pdf.
2. Запорізький національний університет. (2024). Інтеграція штучного інтелекту в освіту: збірник науково-методичних доповідей. Запорізький національний університет. https://files.znu.edu.ua/files/Bibliobooks/Inshi82/0061842.pdf.
3. Іванов, С. (2024). Цифровий етикет: правила поведінки в онлайн-середовищі. Академія цифрових технологій. https://docs.academia.vn.ua/handle/123456789/1888.
4. Інститут інформаційних технологій та освіти. (2022). Освіта для цифрової трансформації суспільства. LIB IITT. https://surl.li/dxsvrz.
5. Центр освіти та розвитку штучного інтелекту. (2023). Штучний інтелект у вищій освіті: ризики та перспективи *інтеграції*. CUESC. https://cuesc.org.ua/images/informlist/%D0%9C%D0%B0%D0%BA%D0%B5%D1%82%20advanced_training_OLA.pdf.

В.С. Волошин, О.В. Клєнін

МЕХАНІЗМИ ДВОМИСЛЕННЯ, ЯК ПРИЧИНА НЕБЕЗПЕКИ ДЛЯ СИСТЕМ ШТУЧНОГО ІНТЕЛЕКТУ

Тема охоплює складні процеси створення, поширення та розпізнавання дезінформації в цифровому середовищі, до якого вже належать системи штучного інтелекту, які здатні отримати початкову інформацію заявно низької якості і тим самим бути інструментарієм для дезінформації в людському житті. Нижче наведені дані про оцінку обсягу інформації в мережі Інтернет, відомі частки первинної, вторинної та неправдивої інформації. механізми та послідовність формування неправдивої інформації, властивості обману, ознаки її впізнання,

Поточні дані про кількість світової інформації в мережі Інтернет мають тенденцію бути дуже розпливчастими, як кількісно, так і якісно. Відомі оцінки кількості та структури інформації ґрунтуються на прогнозах і можуть змінюватися залежно від методології. Неправдиву інформацію в цьому обсязі важко точно оцінити через її суб'єктивний характер і якісні приховані форми (наприклад, часткова правда). Тим не менш, її аналіз для розпізнавання і оцінки є актуальним і складним завданням для сучасних інтелектуальних штучних систем.

Найбільш об'єктивну інформацію про вихідні дані для оцінки обсягу інформації в мережі Інтернет, можна отримати в інтегральному вигляді з наступних джерел: глобальний обсяг створених даних (Global Datasphere); обсяг реально збережених даних (Storage Footprint); глобальний IP-трафік (Internet Traffic), обсяг даних, що передаються мережею за одиницю часу (оцінки Cisco); обсяг публічного вебу, включаючи сукупність даних, доступних через відкриті ресурси (Common Crawl, Internet Archive) а також кількість веб-сайтів і сторінок (статистика від Netcraft, Internet Live Stats, Siteefy). Ключові кількісні оцінки на 2025 рік надані в таблиці 1.

Таблиця 1 – Вихідні дані про обсяги дієвої інформації в глобальних інформаційних мережах

Показник	Орієнтовне значення	Джерело
Загальний обсяг створених даних (Global Datasphere)	≈ 163–181 зетабайт (ZB)	IDC *Data Age 2025*
Глобальний IP-трафік	≈ 300–420 екзабайт/місяць або ~4.8 ZB/рік	Cisco *Annual Internet Report*
Обсяг публічного вебу (архіви, корпуси)	Десятки–сотні петабайт (ПБ)	Common Crawl, Internet Archive
Кількість сайтів	~1.1–1.5 млрд загалом, з них 190–210 млн активних	Netcraft, Internet Live Stats

Нас цікавить проблема дезінформації, хибної інформації в системах штучного інтелекту (ШІ) і її вплив на користувача. Згідно останніх досліджень наука не має об'єктивної методики оцінки кількості такої інформації. Але її якісний вплив на системи користування ШІ має суттєве значення для користувачів и має бути відомим. Зокрема, репрезентативні емпіричні орієнтири (proxy-метрики) у дослідженнях економістів і соціологів показують, що серед новинних статей, якими користуються в системах ШІ, і які активно поширюються в соцмережах, частка навмисно сфабрикованих «fake news» зафіксована як відносно мала частина всіх новинних реплік — в окремих вибірках це було порядку не менше 8% від усіх поширень новин [1]. Це стосується категорії «верифікованої фальші» у межах політичних та новинних стрічок, в інших, вже наукових, матеріалах. З іншого боку, експозиція людей до неправдивої інформації невимірно висока. Опитування показують, що велика частка аудиторії зустрічалась із дезінформацією (рейтинги експозиції різняться за країнами — десятки відсотків і вище) і проявляє негативне відношення. Reuters Institute та Digital News Report документують поширеність занепокоєнь і високу частоту зустрічі з «misinformation» у новинах. Поширення синтетичного AI-контенту (ризик дезінформації), за даними OECD та інших досліджень, фіксують значне зростання неправдивої інформацій. Це вже десятки відсотків річного приросту (наприклад, ~55% зростання синтетичного контенту на mainstream-сайтах у їх вибірці). Це значно підвищує ризик масового поширення невірної або маніпулятивної інформації, навіть якщо частка прямо «фальшивого» контенту лишається невеликою за абсолютними числами.

Механізм формування неправдивої інформації в мережі Інтернет пов'язаний з навмисним або ненавмисним спотворенням фактів, що поширюються через цифрові платформи. Неправда може виникати через навмисні маніпуляції, помилки або недостатню перевірку даних. Ініціювання неправдивої інформації створюється окремою особою, групою або автоматизованою системою (ботами). Мотиви цих груп різні - від політичної пропаганди, псевдонаукової або комерційної вигоди до соціального впливу, тролінгу або жартів. Як приклад посилимось на створення фейкових новин про вигадані події для залучення трафіку на сайт [2].

Дезінформація маскується під достовірність за допомогою підроблених джерел, маніпулятивних заголовків, фейкових зображень або відео. Це підвищує його переконливість [1].

Методи поширення неправдивої інформацій дуже широкі. Вона присутня на платформах (соціальні мережі, форуми, новини, сайти), поширюється за допомогою репостів, лайків, алгоритмічних рекомендацій тощо. Соціальні мережі, такі як Facebook або Twitter, підсилюють ефект за рахунок поширення вірусу. Алгоритми з платформ часто віддають перевагу сенсаційному контенту, сприяючи швидкому охопленню. Вони включають послідовність: «Створення → форматування для достовірності → публікації → поширення вірусу → поширення через вторинні джерела → закріплення в суспільній свідомості».

Рисунок 1 – Властивості дезінформації, що відрізняють її від правдивої інформації (ознаки правди та неправди)

Неправдива інформація в Інтернеті має особливі властивості, які відрізняють її від правдивої інформації (рис. 1) За метою, неправда створюється заради введення в оману чи маніпуляції користувачами, суперечуючись дійсності, ігноруючи докази. Правдива інформація не має такої мети та, опираючись на відомі факти, емпіричні підтвердження, прагне до об'єктивності [3].

Однією з важливих властивостей фейків є емоційне забарвлення, підхід до невиправданої сенсаційності, в супереч правдивій інформації, яка несе на собі фактор можливої нейтральності та опір на логіку та попередні факти.

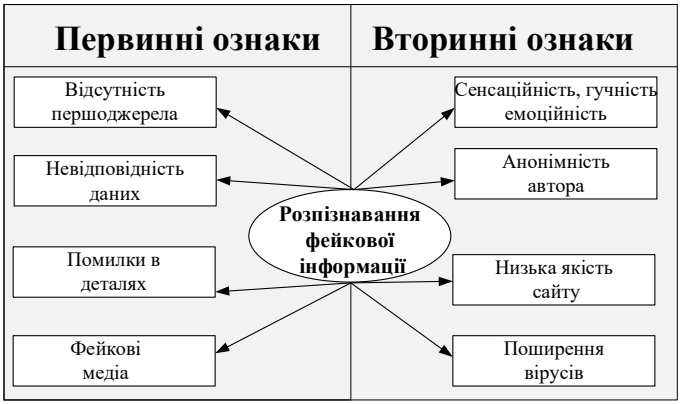

Рисунок 1 – Первинні та вторинні ознаки розпізнавання інформаційної неправди

Окрім того, важливим опосередкованим фактором є швидкість поширення інформації: для дезінформації вона має бути великою завдяки сенсаційності та алгоритмам соціальних мереж [4]. Щоб неправда буда прийнятою як істина, вона має бути посилена за рахунок підтримки користувачами, блогерами, носіями інформації, яку ані хто не перевіряє. Боти та тролі штучно, і навіть не навмисно, підвищують обіг аудиторії. Навіть після одкровення, це може впливати на сприйняття [5]. Правда має властивість повільного поширення, як менш емоційна система.

Не складає зусиль означити існуючі головні ознаки, за якими можна відрізнити достовірність інформації (рис. 1). Наприклад, первинні ознаки, які, як правило, безпосередньо вказують на брехню [2]. Або, якщо немає посилань на оригінальні дані, документи чи свідків, чи факти в повідомленні не узгоджуються з відомими даними або логікою, використовуються неправильні дати, імена, топоніми та ін.

Існують і вторинні ознаки фейковості, непрямі, які вимагають аналізу [6]. Це псевдосенсаційні та гучні заголовки, апеляції до емоцій, що викликають страх, гнів. Або наявність анонімності автора, відсутність ідентифікованого автора або джерела. Опосередковано на фейковість інформації можу вказувати, наприклад, поганий дизайн сайту, орфографічні помилки, підозрілі домени, а також швидке збільшення охоплення без сканування, часто через ботів. Прикладом, коли фейкові новини про «чудодійний медичний препарат» можуть мати як первинну ознаку (відсутність наукових досліджень), так і вторинну (декларована універсальність дії та емоційні заголовки на кшталт «Ліки від усіх хвороб!»).

Достатньо широкий спектр небезпечностей щодо користування системами штучного інтелекту призводять до переліку причин та методів уникнення хибних маніпуляцій в цих системах. До них можна віднести наступне.

1.Контомінація навчальних даних призводить до помилкових, системно спотворюючих моделей ШІ. Як правило, великі мовні моделі та інші нейронні мережі обучаються на величезних базах відкритих даних, і якщо їх корпус містить багато неправильної, маніпулятивної або синтетичної інформації, модель ШІ засвоює її шаблони і починає видавати «впевнені» неправдиві відповіді (феномен «hallucinations», biased outputs).

2. Атаки на дані (data-poisoning) призводять до цілеспрямованої корекції поведінки моделі ШІ. Шкідливі лінгвістичні композиції можуть вставляти спеціально підготовлені приклади в корпуси (або переднавчальні набори), щоб викликати помилки, встановлювати backdoor-триггеры або змушувати модель генерувати неправду в потрібних контекстах. Це реально задокументована загроза для будь-яких систем, які використовують зовнішні/скопійовані дані [6].

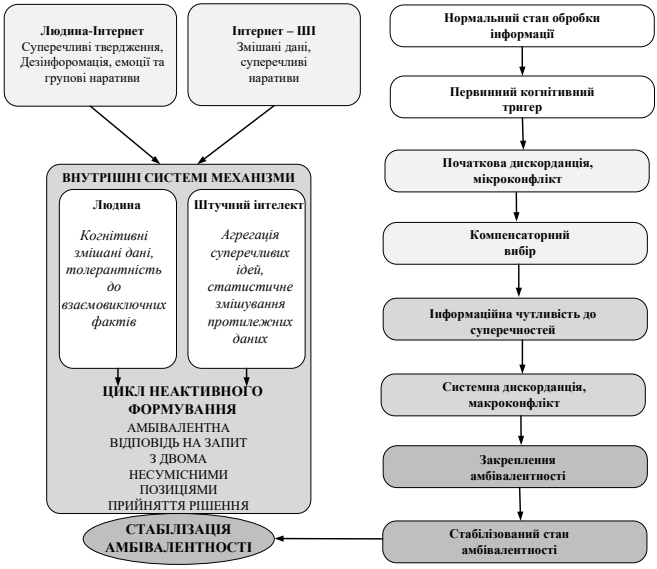

Рисунок 3 – Алгоритм негативної взаємодії «Людина-Штучний інтелект» (а) і механізм переходу людини в стан двомислення (б)

3. Медійні синтетичні методи представляють високо реалістичні доказові «фейки», в яких навіть ідеально підготовлена модель (або людина) не завжди відрізняє правду від реалістичного фейку без інструментарія верифікації. Такі матеріали використовуються для фінансових махінацій, наукової та політичної дезінформації, психологічного впливу.

4. Процедури експлуатації інтерфейсу моделі tromp-injection (jailbreaks), в які шкідливий контент (веб-сторінки, файли, вставлені підказки) можуть «впроваджувати» інструкції, які модель ШІ буде виконувати, порушуючи політику очікуваної об'єктивної поведінки (видача персональної інформації, виконання шкідливих інструкцій). Це робить інтеграцію LLM у додатки вразливою.

Таким чином, дезінформація впливає на модель ШІ на трьох рівнях: *навчання* (довгострокове), *залучення та попередня обробка даних* (середньостроковий), *інтерактивний інтерфейс* (короткостроковий, миттєвий)).

Ми звернемося до актуальної теми, що спонукає співвідношення інформації та людини в межах штучного інтелекту. Це поняття «двомислення» — орвеллівський термін, що означає здатність одночасно приймати та утримувати дві взаємовиключні установки як «істину». У соціальному й психологічному контексті — це механізм раціоналізації, когнітивної адаптації та маніпуляції пам'яттю та сприйняттям, на підставі якого хибні судження приймають зміст об'єктивних.

Ключові механізми перетинання цих явищ в системах штучного інтелекту наступні (рис. 3).

1. Зіткнення когнітивної схильності та алгоритмічної подачі контенту. Коли люди, схильні до двомислення, легше приймають суперечливі наративи, якщо платформа одночасно підтверджує та спростовує певну позицію. Це створює середовище, де протилежні установки співіснують і закріплюються.

2. Наративна багатовекторність. Джерело може навмисно поширювати як відверто неправдиві, так і частково правдиві повідомлення, формуючи у системі ШІ «подвійну правду» — одночасне прийняття «факту» і «контрафакту». Мета — заплутати критичне мислення та знизити довіру до будь-якої конкретної версії.

3. Соціальне підтвердження та нормалізація суперечностей. Масове повторення суперечливих тверджень різними джерелами створює ілюзію суспільної підтримки обох позицій послаблюючи внутрішні критерії істинності на користь стійкості двомислення.

4. Технологічні інструменти: боти, дипфейки, алгоритми рекомендацій в інтернеті, автоматизовані боти та синтетичний контент підсилює зовнішньо індуковане двомислення

Спільні структурні фактори (платформи та середовище), які посилюють зв'язок «двомислення ↔ дезінформація», це: орієнтація платформ на метрики залучення, а не на точність контенту; надмірна персоналізація, яка звужує інформаційну бульбашку та підсилює існуючі упередження; низька прозорість модераційних політик і часті зміни правил поведінки; можливість створення та масштабування контенту без людської участі (бот-мережі, генеративні моделі); ослаблення інституцій незалежного фактчекінгу.

Висновки. Таким чином, дезінформація та цілеспрямовані атаки на дані є серйозною та задокументованою загрозою надійності систем сучасного штучного інтелекту: вони погіршують якість будь-яких інформаційних моделей (hallucinations), перешкоджають серйозній аналітиці, зокрема інформаційній, дозволяють здійснювати шахрайство (діпфейки), відкривають вектори експлуатації через оперативне вливання та отруєння даних. Для систем штучного інтелекту найнебезпечніші комбінації реалістичних синтетичних медіа та скоординованих кампаній, а також оперативне впровадження в інтегровані додатки мають швидкі та масштабні наслідки як для користувачів, так і для корпоративних систем.

Дезінформаційні екосистеми, і зокрема «двомислення», активно використовують суперечливі повідомлення як інструмент зниження довіри та руйнування когнітивної стабільності. Системи ШІ чутливо вразливі до суперечливих даних і потребують суворої фільтрації, багато джерельної перевірки та ентропійно-консистентного аналізу інформаційних корпусів. Ефективна протидія можлива лише через міждисциплінарні сукупні підходи: технології, соціальний захист, спеціальна комунікація та освітні інтервенції.

1. Allcott H., Gentzkow M. Social Media and Fake News in the 2016 Election.The Journal of Economic Perspectives, Vol. 31, № 2, 2017. pp. 211-235.

2. Wardle C., Derakhshan H., Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making, 2017, Strasbourg, France, Council of Europe, p. 20.
3. Pennycook, G., Rand, D. G., The Psychology of Fake News. Cambridge, UK, Trends in Cognitive Sciences, Vol. 25, № 5, 2021. pp. 388–389.
4. Vosoughi S., Roy D., Aral S. The Spread of True and False News Online. Science, 359, 2018. pp.1146-1151.
5. Wardle C., Derakshan H. Information Disorder: Toward an interdisciplinary framework for research and policy making, Council of Europe. 2017. p.217.
6. Pennycook G., McPhetres J., Zhang Y., Lu J. G., Rand D. G.. Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention". Psychological science, 31(7), 2020. pp.770–780.

G.P. Kostenko, A.O. Zaporozhets

DIGITAL TWIN FOR SECOND-LIFE EV BATTERY OPERATION IN ENERGY SYSTEMS

The rapid expansion of electric mobility and renewable energy integration has created a new operational landscape in which energy storage systems play a central role in ensuring flexibility, stability, and resilience of the power grid [1]. At the same time, the global market is generating a growing stream of retired electric vehicle batteries that remain technically viable for reuse in stationary applications. These Second-Life Batteries (SLBs) represent a valuable resource for supporting distributed energy systems; however, their non-uniform technical condition, undocumented degradation history, and high stochasticity of ageing processes introduce substantial uncertainty into their operation. In this context, Digital Twin (DT) technologies are emerging as a key enabler for predicting, managing, and optimizing SLB performance under real-world conditions.

A digital twin of a battery can be defined as a continuously updated virtual replica of the physical asset that synchronizes in real time with sensor data and operational measurements [2]. For SLBs, such synchronization is particularly important because their state of health (SOH) and degradation trajectory depend strongly on their previous usage in electric vehicles, which is often only partially known or entirely unavailable. DTs compensate for incomplete historical data by integrating two layers of information: real-time telemetry (current, voltage, temperature, SOC) and model-based representations of degradation, electrical behaviour, and thermal dynamics. Through this integration, the digital twin becomes the core predictive component capable of estimating SOH, forecasting remaining useful life (RUL), tracking degradation indices, and supporting KPI-based decision making [3].

The functional roles of digital twins vary across different layers of the energy system, from cell-level diagnostics to grid-level coordination. Tab. 1 summarizes these functions in the context of the SLB lifecycle and their integration into energy networks.

Table 1 - Digital Twin Functions Across SLB Lifecycle and Grid Levels

Grid/Operational Layer	DT Functions for SLB	Key Outputs	Value for SLB Integration
Cell / Module Level	SOH tracking, thermal estimation, degradation modelling	SOH, SOC, RUL	Safe operation, failure prevention
Pack Level	Balancing optimization, thermal control, anomaly detection	Optimal balancing, fault flags	Extends SLB lifespan

System-Level (BESS)	Dispatch optimization, degradation-aware power allocation	Optimal P/E ratio use	Higher efficiency & lower LCOS
Microgrid Level	PV smoothing, peak shaving, emergency backup	Load profiles, optimal charging	Reliability & resilience
Grid-Level / VPP	Aggregated forecasting, fleet management, uncertainty modelling	Fleet RUL, degradation cost	Market participation & reliability

Depending on the modelling approach, digital twins for battery systems can be categorized as physics-based, data-driven, and hybrid (Tab.2). Physics-based DTs rely on electrochemical or equivalent-circuit models that capture internal mechanisms such as SEI growth, charge-transfer resistance, and thermal evolution. Their main advantage is physical interpretability and high diagnostic accuracy, making them suitable for failure assessment and safety analysis. However, they require precise parameterization and may be computationally demanding, which limits their ability to support fast decision-making in distributed environments.

Table 2 - Comparative Evaluation of Digital Twin Architectures for Second-Life Batteries

Architecture Type	Strengths	Limitations	Best Use Cases
Physics-Based DT	High interpretability; physically consistent degradation and thermal dynamics; good for diagnostics and safety assessment.	Requires accurate parameterization; computationally expensive; sensitive to unknown SLB history.	Safety validation, capacity fade modelling, detailed thermal analysis, certification-related studies.
Data-Driven DT	Adaptive to real-world data; scalable to fleets; fast deployment; effective with incomplete historical information.	Limited extrapolation ability; low interpretability; requires diverse datasets.	Operational forecasting, fast SOH/RUL estimation, fleet-wide SLB monitoring, anomaly detection.
Hybrid DT	Optimal balance between physics and learning; high accuracy; robust under uncertainty; suitable for heterogeneous SLBs.	More complex integration; requires both domain knowledge and ML expertise.	Predictive control, uncertainty-aware RUL forecasting, SLB deployment in microgrids and VPPs.

Cloud-Based Deployment	High computational power; scalable analytics; suitable for large datasets and long-horizon predictions.	Latency; dependency on connectivity; cybersecurity considerations.	Aggregators, VPPs, multi-site SLB coordination, regulatory analysis.
Edge-Based Deployment	Low latency; real-time decision making; autonomy; low communication requirements.	Limited computational resources; reduced model complexity.	Frequency regulation, peak shaving, PV smoothing, microgrid operation.
Cloud–Edge Hybrid	Combines scalability with responsiveness; optimal partitioning of tasks; supports adaptive and health-aware control.	Requires careful synchronization and architecture design.	Distributed SLB systems with real-time constraints and long-horizon optimization needs.

Data-driven digital twins use machine learning and statistical modelling to infer degradation and forecast battery behaviour based on historical data. Neural networks such as LSTM or transformer architectures can model complex patterns in SOH evolution or predict RUL from short observation windows. These models are adaptive and can learn directly from field data, making them suitable for heterogeneous SLB fleets. Their main limitation is reduced interpretability and potential instability when extrapolating beyond the training domain, particularly in cases where SLBs exhibit nonlinear or irregular degradation due to prior usage.

Hybrid digital twins combine the strengths of both approaches. In a hybrid architecture, the physics-based model ensures structural consistency and enforces physical constraints, while the data-driven component corrects residual errors, captures unmodelled effects, or adapts parameters online. This combination is particularly powerful for SLBs, where initial conditions vary widely across cells and modules due to different usage patterns in EVs. Hybrid DTs have demonstrated superior accuracy in SOH tracking, adaptive RUL prediction, and anomaly identification under uncertain operating conditions.

An equally important dimension is the deployment architecture of the digital twin. Cloud-based DTs operate on centralized computational infrastructure and can handle large datasets, complex optimization problems, and long-horizon degradation forecasting. They are well suited for aggregators or virtual power plant (VPP) operators who manage large fleets of SLBs. Their downside is communication latency and dependence on stable connectivity, which limits their use in fast-response applications. Edge-based DTs execute locally on BMS units or microgrid controllers, enabling real-time decision-making with minimal latency. This makes edge DTs ideal for frequency regulation, peak-shaving, and hybrid PV-battery systems, where immediate reaction is crucial. Hybrid cloud-edge

architectures combine these advantages: computationally intensive tasks are executed in the cloud, while real-time functions are maintained on the edge device, ensuring both responsiveness and analytical depth.

Digital twins enhance SLB operation by enabling predictive, health-aware control. Based on up-to-date SOH estimates, the DT can determine safe operating boundaries, adjust allowable discharge depths, optimize power setpoints, and recommend economically efficient strategies that balance revenue with degradation cost. For markets such as frequency containment reserve (FCR), the DT can dynamically adjust participation parameters to prevent accelerated ageing. In distributed systems, it supports coordination of multiple SLB units by estimating their relative degradation states and allocating tasks according to remaining useful life.

A critical advantage of DTs in SLB applications is their ability to manage uncertainty. Due to the stochastic nature of ageing processes and heterogeneous prior use, SLBs often deviate from nominal degradation patterns. Digital twins address this challenge by incorporating uncertainty modelling-using ensemble methods, probabilistic RUL forecasting, or scenario-based analysis. This capability is essential for operators who need to maintain reliability while managing fleets of non-uniform batteries with different levels of residual performance and degradation risk. Hybrid DTs, in particular, can continuously re-estimate degradation parameters and detect deviations from expected SOH or IDI trajectories, enabling early-warning diagnostics and proactive replacement scheduling.

In summary, digital twins represent a transformative tool for ensuring efficient, safe, and predictable integration of second-life batteries into modern energy systems. By combining real-time data, predictive models, and adaptive optimization algorithms, DTs create a unified framework for lifecycle-aware operation. Hybrid DTs deployed in cloud-edge configurations offer the most promising balance between accuracy, responsiveness, and scalability for SLB applications. Future research should focus on developing standardized DT platforms, harmonizing data formats in line with EU Battery Passport requirements, and integrating digital twins into PDCA-based operational loops to enable fully autonomous, self-optimizing battery systems.

Conclusions:

The analysis demonstrates that digital twin technology holds significant potential for enhancing the reliability and efficiency of second-life battery operation in modern energy systems. One of the key advantages of DT-based frameworks is their ability to substantially reduce operational uncertainty, which is particularly relevant for SLB assets characterized by heterogeneous and often unknown usage histories. By integrating real-time telemetry with model-based estimation, digital twins provide continuous insight into the actual degradation state of each module, enabling more accurate forecasting of performance and remaining useful life.

Another important outcome is the role of DTs in enabling predictive maintenance. SLBs batteries are inherently more prone to micro-defects, uneven ageing, and unexpected failure modes compared to new batteries. The diagnostic capabilities of DTs-supported by anomaly detection, trend deviation analysis, and hybrid model correction - allow operators to identify emerging risks at an early stage, minimizing downtime and preventing safety-critical events.

The study also highlights the architectural shift toward cloud-edge hybrid deployments. While cloud platforms provide the computational capacity required for long-horizon forecasting and fleet-level analytics, edge computing ensures low-latency decision-making essential for real-time operation in microgrids and distributed environments. This division of computational responsibilities has become a dominant design pattern for practical SLB implementations.

Finally, hybrid modelling approaches that combine physics-based representations with data-driven components demonstrate the highest predictive accuracy and robustness. Such models compensate for missing historical information, capture unmodelled degradation dynamics, and adapt to the stochastic behaviour typical of second-life batteries. As a result, hybrid digital twins offer a balanced framework capable of supporting both operational optimization and reliable forecasting under uncertainty.

Overall, digital twins provide a coherent and powerful foundation for managing second-life batteries across their entire operational trajectory. By addressing uncertainty, enabling predictive maintenance, leveraging cloud–edge architectures, and applying hybrid modelling, DTs significantly advance the reliability, safety, and economic viability of SLB deployment in future energy systems.

1. Kostenko, G., Zaporozhets, A., Babak, V., Uruskyi, O., Titko, V., Denisov, V. (2025). Second-Life EV Batteries Application in Energy Storage Systems for Sustainable and Resilient Power Sector. In: Systems, Decision and Control in Energy VII, vol 595. Springer, Cham. https://doi.org/10.1007/978-3-031-90466-0_1.
2. Al-Shetwi, Ali Q., Ibrahem E. Atawi, Mohamed A. El-Hameed, and Ahmad Abuelrub. 2025. “Digital Twin Technology for Renewable Energy, Smart Grids, Energy Storage and Vehicle-to-Grid Integration: Advancements, Applications, Key Players, Challenges and Future Perspectives in Modernising Sustainable Grids.” *IET Smart Grid*: e70026. https://doi.org/10.1049/stg2.70026.
3. Kostenko, G., & Zaporozhets, A. (2025). Trigger-Based PDCA Framework for Sustainable Grid Integration of Second-Life EV Batteries. World Electric Vehicle Journal, 16(10), 584. https://doi.org/10.3390/wevj16100584.

А.А. Герасименко

ЗАДАЧИ ЗАСТОСУВАННЯ ШТУЧНОГО ІНТЕЛЕКТУ В КІБЕРБЕЗПЕЦІ

Основна роль ШІ при вирішенні завдань кібербезпеки полягає у підвищенні швидкості та масштабованості захисних механізмів, що є критично важливим для протидії сучасним, високоавтоматизованим кіберзагрозам. При застосуванні ШІ вирішується низка прикладних задач кібербезпеки, наприклад – системи захисту отримають властивості автоматичного аналізу величезні обсяги даних (Big Data), виявлення складних шаблонів та ідентифікації аномалії. Ці властивості забезпечують побудову ефективних систем виявлення вторгнень (IDS), класифікації шкідливого програмного забезпечення та проактівного аналізу вразливостей до того, як ними зможуть скористатися. Системи на основі ШІ оптимізують реагування на інциденти через платформи SOAR (Security Orchestration, Automation, and Response), значно скорочуючи час від виявлення до нейтралізації загрози. Більше того, ШІ забезпечує покращений захист особистості шляхом впровадження поведінкових біометричних технологій, які постійно перевіряють легітимність користувача на основі унікальних моделей взаємодії. ШІ також використовує обробку природної мови (NLP) для виявлення фішингових та шахрайських повідомлень, які стають усе більш складними через генеративний ШІ. Незважаючи на значні переваги, інтеграція ШІ у сферу безпеки створює нові виклики, які є предметом сучасних досліджень. Основною проблемою є вразливість самих моделей ШІ до атак супротивника, коли зловмисники маніпулюють даними, щоб обійти захист. У відповідь зростає інтерес до розробки стійких систем та впровадження пояснюваного ШІ (XAI), який дозволяє людям розуміти логіку прийняття рішень системою. Крім того, федеративне навчання активно досліджується як метод тренування моделей ШІ із збереженням конфіденційності та приватності даних [1,2].

Сфера цього дослідження включає: Окреслення можливостей використання алгоритмів машинного навчання у завданнях кібербезпеки шляхом аналізу публікацій. Визначення домінуючого спрямування використання штучного інтелекту у завданнях кібербезпеки. Визначення викликів та загроз використання штучного інтелекту у завданнях кібербезпеки.

Головна роль штучного інтелекту полягає в тому, щоб доповнювати та покращувати кібербезпеку шляхом автоматизації процесів і підвищення точності виявлення [3]. Розглянемо конкретні завдання.

Завдання виявлення та прогнозування загроз. ШІ використовує машинне навчання для встановлення нормальної базової поведінки мережі та користувачів. Будь-яке відхилення (аномалія) від цієї базової поведінки, яке може вказувати на нову, раніше невідому атаку (атака нульового дня), виявляється миттєво.

Завдання реагування в реальному часі. Коли виявляється загроза, ШІ може автоматично вживати заходів, таких як блокування шкідливого трафіку, ізоляція зараженої точки доступу або застосування виправлень без участі людини. Це значно скорочує час реагування.

Завдання аналізу даних та SIEM. ШІ (особливо генеративний ШІ) допомагає Центрам операцій безпеки (SOC), скорочуючи мільйони логів подій із систем SIEM (Системи управління інформацією та подіями безпеки) у зрозумілі звіти та рекомендує дії для реагування.

Завдання захисту кінцевих пристроїв та ідентифікації. Посилена автентифікація та швидше виявлення шкідливого ПЗ та програм-вимагачів на пристроях.

Завдання фільтрації фішингу та спаму. Обробка природної мови (NLP) дозволяє ШІ аналізувати контекст, тон та патерни електронних листів для виявлення складних фішингових атак.

Вищезазначене формує фокус пошуку наукових публікацій.

Крім переваг, нові технології приносять і нові ризики. Щонайменше слід враховувати наступне:

Зловмисне використання ШІ (адверсарний ШІ). Зловмисники використовують ШІ для створення високотехнологічного шкідливого програмного забезпечення, автоматизованих атак та переконливих дипфейків для соціальної інженерії.

Вартість та інтеграція. Впровадження рішень ШІ потребує значних інвестицій у технології, інфраструктуру та висококваліфікований персонал. Інтеграція з успадкованими системами може бути складною.

Проблема «чорного ящика». Моделі глибокого навчання ШІ часто є непрозорими; важко пояснити, чому система прийняла конкретне рішення. Це ускладнює довіру до системи та аудит її рішень (пояснюваність).

Вразливість до атак на системи ШІ. Хакер може маніпулювати вхідними даними для моделі ШІ (адверсарні атаки), змушуючи її неправильно класифікувати шкідливий файл як безпечний (помилково негативний результат).

Якість даних та упередженість. Якщо модель навчена на низькоякісних або упереджених даних, вона може генерувати хибнопозитивні результати або, що гірше, ігнорувати реальні загрози.

Також доцільно визначити, що впровадження штучного інтелекту створює нові, складні виклики. А саме:

• Швидкість і масштаб. ШІ може обробляти та аналізувати петабайти даних 24/7, що фізично неможливо для людських команд, забезпечуючи безперервний моніторинг.

• Проактивний захист. Здатність передбачати загрози — аналіз тенденцій дозволяє ШІ прогнозувати майбутні вектори атак та пріоритизувати усунення вразливостей до того, як їх буде використано.

• Зменшення людських помилок. ШІ автоматизує рутинні, повторювані завдання (наприклад, управління патчами або аналіз журналів), звільняючи людей для вирішення більш складних стратегічних проблем.

• Боротьба з новими типами атак. Завдяки машинному навчанню ШІ може вивчати та адаптуватися до нових, невідомих схем атак (поліморфне шкідливе програмне забезпечення).

Для отримання статистичних даних про публікації ми будемо шукати «бібліометричний аналіз штучного інтелекту в кібербезпеці» у наукових пошукових системах (наприклад, Google Scholar або ResearchGate). Приклади знайдених досліджень [4] показують загальні тенденції.

До 2015 року початкова стадія. Базові дослідження, в основному використовуючи класичне машинне навчання (Random Forest, SVM) для простих завдань (фільтри спаму). Кількість публікацій – низька. Повільне лінійне зростання.

2015–2018 Фаза прискорення. Поява перших значущих робіт із застосуванням глибокого навчання, особливо для виявлення вторгнень. Кількість публікацій – середня. Початок експоненціального зростання.

2019–2021 Вибуховий ріст. Масове застосування AI/ML/DL у всіх підгалузях кібербезпеки. Зростання досліджень з безпеки IoT та поведінкової аналітики. Кількість публікацій – висока. Дуже швидке експоненціальне зростання.

2022–2025 Активне зростання та складність. Прориви в генеративному штучному інтелекті (GenAI), акцент на Adversarial ML та стійкості систем. Кількість публікацій досягає піку. Пік експоненціального зростання

Ця робота пропонує аналіз публікацій з основних тем використання ШІ в кібербезпеці.

Основний висновок аналізу такий - домінуючий фокус: Найбільша частка дослідницьких зусиль присвячена системам виявлення вторгнень (IDS) та класифікації шкідливого ПЗ. Це підкреслює пріоритет наукової спільноти у використанні ШІ для виявлення аномалій у реальному часі та боротьби з найпоширенішими й найбільш руйнівними загрозами.

1. Yazici, I. Shayea, and J. Din, "A survey of applications of artificial intelligence and machine learning in future mobile networks-enabled systems," Engineering Science and Technology, an International Journal, vol. 44, pp. 0-0, 2023.
2. Lande D., Novikov O., Alekseichuk L. Application of Large Language Models for Assessing Parameters and Possible Scenarios of Cyberattacks on Information and Communication Systems // Theoretical and Applied Cyber Security. Vol. 6 No. 1 (2024). DOI: 10.20535/tacs.2664 29132024.1.315242.
3. F. Hang, L. Xie, Z. Zhang, W. Guo, and H. Li, "Artificial intelligence enabled fuzzy multimode decision support system for cyber threat security defense automation," Journal of Computer Virology and Hacking Techniques, vol. 19, no. 2, pp. 0-0, 2023.
4. O. S. Albahri and A. H. AlAmoodi, “Cybersecurity and Artificial Intelligence Applications: A Bibliometric Analysis Based on Scopus Database,” Mesopotamian Journal of CyberSecurity, vol. 2023, pp. 158–169, Sep. 2023, doi: 10.58496/MJCSC/2023/018.

К.Ю. Бровко, О.В. Великогорський

ЗАСТОСУВАННЯ ЦИФРОВИХ ДВІЙНИКІВ ДЛЯ ОЦІНКИ ТЕХНІЧНОГО СТАНУ ЕНЕРГОБЛОКУ АЕС

Атомна енергетика, будучи одним із найважливіших джерел стабільної та низьковуглецевої енергії, пред'являє найвищі вимоги до безпеки та надійності. Кожна година простою енергоблоку атомної електростанції (АЕС) означає значні фінансові втрати, а будь-яка технічна несправність несе потенційні ризики. У таких умовах пріоритетним завданням стає перехід від планово-запобіжного ремонту до технічного обслуговування за фактичним станом. Ключовим інструментом цього переходу виступає технологія цифрового двійника. Цифровий двійник – це не просто 3D-модель або комп'ютерна програма. Це динамічна, віртуальна копія фізичного об'єкта або системи, що імітує його поведінку в режимі реального часу. Для енергоблоку АЕС цифровий двійник – це комплексна математична модель, що об'єднує:

– фізичні параметри: характеристики матеріалів, геометрія обладнання;

– інженерні моделі: закони термодинаміки, гідравліки, механіки та ядерної фізики;

– оперативні дані в режимі реального часу: показники тисяч датчиків (тиск, температура, вібрація, потік нейтронів тощо);

– історичні дані: архіви роботи, журнали ремонтів, результати інспекцій.

Таким чином, цифровий двійник синхронно взаємодіє зі своїм фізичним прототипом, постійно оновлюючись на основі потоків даних [1]. Оцінка технічного стану перетворюється з періодичної процедури на безперервний моніторинг. До ключових напрямків застосування можна віднести:

1. Прогностичне обслуговування критичного обладнання. Найважливіша функція цифрового двійника – передбачення відмов. Моделюючи роботу такого обладнання, як головні циркуляційні насоси, турбіни, парогенератори, двійник аналізує тенденції в даних (наприклад, зростання вібрації або температури підшипників). На основі машинного навчання та предиктивних алгоритмів система може точно спрогнозувати залишковий ресурс деталі та запланувати її заміну до моменту критичного зносу, уникаючи аварійного простою.

2. Моніторинг старіння та деградації матеріалів. Корпус реактора, трубопроводи першого контуру зазнають впливу високих температур, тисків та радіації. Цифровий двійник інтегрує дані про режими експлуатації та результати фізичних обстежень (наприклад, вимірювання твердості металу). Це дозволяє моделювати процеси старіння та накопичення втоми матеріалів, оцінюючи залишковий ресурс ключових компонентів на десятиліття вперед.

3. Оптимізація технологічних режимів. Цифровий двійник дозволяє проводити безпечні «експерименти» у віртуальному просторі. Інженери можуть змоделювати, як змінити параметри роботи ядерного реактора

(потужність, температура теплоносія) для підвищення ефективності, не втручаючись у реальний процес. Це допомагає знайти оптимальні режими, що мінімізують знос обладнання та максимізують виробіток енергії.

4. Підтримка прийняття рішень і тренування персоналу. У разі нештатної аварійної ситуації цифровий двійник може стати потужним тренажером та ситуаційним центром. Оперативний персонал може відпрацювати сценарії ліквідації аварії на віртуальній копії, а керівництво – отримати прогноз розвитку подій та оцінку наслідків для прийняття обґрунтованих рішень.

5. Планування модернізації та капітального ремонту. Перед проведенням будь-якої модернізації її можна віртуально «впровадити» в цифровий двійник. Система проаналізує, як нове обладнання вплине на загальну динаміку блоку, чи не виникнуть непереносимі навантаження на суміжні системи. Це дозволяє уникнути дорогих помилок на етапі проектування.

На рис. 1 зображено інтерфейс користувача системи керування та моніторингу, який імітує роботу цифрового двійника енергоблоку атомної електростанції з реактором типу ВВЕР-1000.

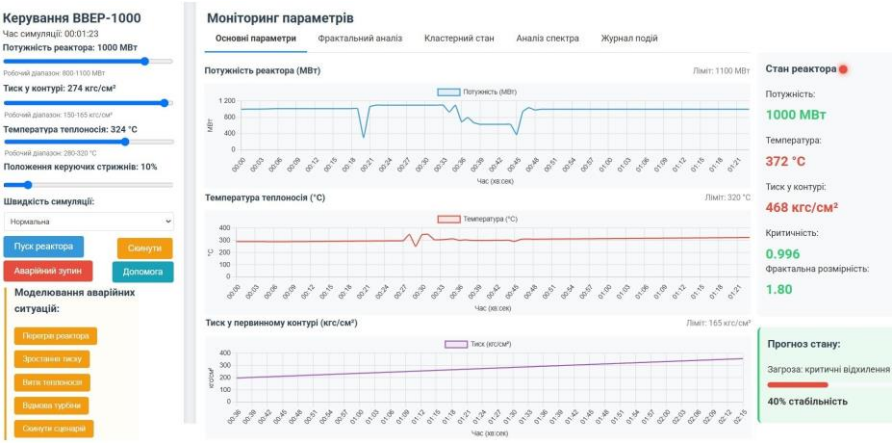

Рисунок 1 – Інтерфейс прототипу цифрового двійника енергоблоку АЕС

Цей інтерфейс наочно демонструє, як технологія цифрового двійника інтегрується в операційну діяльність для контролю технічного стану та моделювання різних сценаріїв. Він поєднує в собі функції панелі керування, системи візуалізації даних у режимі реального часу та тренажера для відпрацювання нештатних аварійних ситуацій.

У верхній лівій частині екрану відображаються ключові експлуатаційні параметри енергоблоку на поточний момент симуляції. Потужність реактора

становить 1000 МВт, що відповідає номінальному рівню, але привертає увагу той факт, що температура теплоносія складає 324 °C, що перевищує верхню межу рекомендованого робочого діапазону, яка вказана як 280-320 °C. Це може сигналізувати про потенційну аварію, яку система цифрового двійника виявляє та відображає для оператора. Одночасно тиск у контурі знаходиться в нормі, а положення керуючих стрижнів становить 10%, що вказує на стан роботи на повну потужність.

Центральну частину інтерфейсу займають інтерактивні елементи керування. Користувач може керувати швидкістю симуляції, ініціювати пуск реактора або його аварійну зупинку. Найбільш важливою з точки зору оцінки технічного стану є секція «Моделювання аварійних ситуацій». Вона дозволяє операторам та інженерам у віртуальному безпечному середовищі активувати різні сценарії відмов, такі як «Перегрів реактора», «Зростання тиску», «Витік теплоносія» та «Відмова турбіни». Ця функція є практичною реалізацією прогностичного та тренувального потенціалу цифрового двійника, дозволяючи вивчити поведінку системи та перевірити алгоритми дій у критичних ситуаціях без будь-якого ризику для реального обладнання.

Нижня частина екрану присвячена глибокому аналізу та моніторингу параметрів. Крім стандартних графіків зміни потужності та температури в режимі реального часу, система пропонує перехід до розширених інструментів аналітики. Це включає «Фрактальний аналіз» для виявлення складних паттернів у поведінці системи, «Кластерний стан» для групування станів обладнання, «Аналіз спектра», для обробки вібраційних сигналів, що критично важливо для діагностики механічного обладнання, та «Журнал подій» для фіксації всіх операцій та відхилень. На графіку температури можна побачити різкий сплеск, який візуалізує те саме відхилення технологічного параметра, на яку вказувало числове значення, що підтверджує здатність цифрового двійника надавати дані в різних форматах для кращого сприйняття.

Таким чином, даний інтерфейс є універсальним інструментом, який об'єднує в собі оперативне керування, безперервний моніторинг технічного стану, прогностичну аналітику та тренувальні функції. Він служить ідеальним прикладом того, як цифровий двійник забезпечує глибоке розуміння роботи складного об'єкта, дозволяючи не лише фіксувати поточний стан, але й передбачати його зміни та готувати оперативний та інженерно-технічний персонал до дій у будь-яких, навіть найкритичніших, умовах.

1. Budanov, P., Brovko, K., Melnikov, V., Yakymchuk, M., Kononov, V., Kyrysov, I., Nosyk, A., Karpenko, O., Kalnoy, S., & Khomiak, E. (2025). Construction of an information model of the digital twin of the technological process in a power unit at a nuclear power plant. *Eastern-European Journal of Enterprise Technologies, 4(8*(136), 39–49. https://doi.org/10.15587/1729-4061.2025.335712.

О.О. Висоцька, А.М. Давиденко

ДОСЛІДЖЕННЯ ЕТАПІВ ПОПЕРЕДНЬОЇ ОБРОБКИ ЕЛЕКТРОННИХ ЛИСТІВ ПРИ ВИЯВЛЕННІ ФІШИНГОВОГО СПАМУ

На сьогоднішній день вже майже не залишилось людей, які не отримували б спам. Такі небажані електронні повідомлення зазвичай розсилаються масово через різноманітні засоби розповсюдження інформації, такі як: електронна пошта, месенджери, соціальні мережі, тощо. Хоча, треба зазначити, що, якщо останнім часом спам, як правило має вигляд електронних повідомлень, то раніше такі повідомлення розповсюджувались у вигляді паперових листівок. Тобто, можна сказати, що задача виявлення спаму є актуальною досить давно. Але в даній роботі розглядається варіант саме електронних спам-повідомлень. Існує різні види спаму, але одним з найнебезпечніших є фішинговий спам [1], тому що саме він може призвести до крадіжки особистих даних, фінансових втрат та інших серйозних наслідків. Зазвичай він має форму електронного листа або текстового повідомлення, яке нібито надійшло від законного джерела, максимально імітуючи справжність джерела. У повідомленні може бути запропоновано одержувачу перейти за посиланням або надати особисту інформацію, щоб нібито оновити обліковий запис, підтвердити свою особу або отримати приз.

Розвиток штучного інтелекту, в даному випадку, має як позитивні так негативні наслідки. З одного боку, використання різноманітних інструментів штучного інтелекту спрощує зловмисникам задачу створення та масового розповсюдження фішингового спаму. З іншого боку, на даний момент значна частина технологій виявлення фішингового спаму заснована на використанні штучного інтелекту.

Для роботи переважної більшості технологій виявлення фішингового спаму необхідна наявність репрезентативної бази даних навчальних зразків електронних листів (для виявлення спаму у електронній пошті) або повідомлень у іншій формі (для виявлення спаму у месенджерах, соціальних мережах, тощо). Також, для ефективного застосування відповідних технологій принципово важливим є виконання попередньої обробки зразків електронних листів, що аналізуються. Тому, саме цим двом етапам процесу виявлення фішингового спаму присвячена основна частина даної роботи.

Збір та попередня обробки електронних листів для створення бази даних навчальних зразків включає наступні етапи:

1. ***Збір зразків фішингових електронних листів.*** Цей етап є важливим кроком у процесі розробки системи виявлення фішингового спаму. Є два основних способи накопичення вибірки фішингових електронних листів: налаштувавши обліковий запис електронної пошти honeypot або використовуючи загальнодоступний набір даних фішингових електронних листів.

Honeypot – це спеціальна створена приманка для зловмисника, яка

спонукає його зробити вторгнення в систему або іншу незаконну дію. Існують різні види honeypot, в даному випадку доцільне використання email-приманки. Для цього в першу чергу необхідно налаштувати обліковий запис електронної пошти honeypot. Обліковий запис електронної пошти honeypot – це обліковий запис електронної пошти-приманки, налаштований на залучення та перехоплення фішингових електронних листів. Це можна зробити, створивши обліковий запис електронної пошти з унікальною адресою, яку легко вгадати, наприклад «phishing@example.com», а потім опублікувавши адресу на веб-сайтах або форумах, які, як відомо, досить часто відвідують фішери. Це залучить фішерів, які потім надсилатимуть фішингові листи на цей обліковий запис електронної пошти. Відповідно, накопичені фішингови листи на цій пошті-приманці, доцільно використовувати при створенні бази даних навчальних зразків електронних листів для виявлення фішингового спаму.

Інший спосіб накопичення відповідної бази даних електронних листів – це використання загальнодоступного набору даних фішингових електронних листів. У відкритому доступі в мережі Інтернет існують різні загальнодоступні набори даних фішингових електронних листів, які доцільно використовувати для відповідних дослідницьких цілей. Ці набори даних, як правило, збираються організаціями, які займаються кібербезпекою, або дослідниками, які опублікували свої дані для використання іншими. В якості прикладів таких наборів можна навести: набір даних PhishTank [2], набір даних міжнародної робочої групи по боротьбі з фішингом APWG [3] і набір даних SpamHaus [4].

Зібравши вибірку фішингових електронних листів, важливо мати на увазі, що вибірка має бути репрезентативною для типів фішингових електронних листів, з якими система стикатиметься в реальному світі, і має бути збалансованою, тобто мати однакову кількість фішингових і легітимних електронних листів. Це допоможе навчити класифікатор мати високу точність у виявленні фішингових електронних листів.

2. Вибір дані: необхідно витягнути (виділити) відповідні дані з електронних листів, такі як адресу електронної пошти відправника, вміст електронного листа та будь-які URL-адреси чи посилання, які містяться в електронному листі.

Цей крок передбачає ідентифікацію та вилучення відповідних даних із електронних листів, які можна використовувати для навчання алгоритму виявлення. Розглянемо більш детально, які саме дані доцільно аналізувати для виявлення фішингового спаму.

Електронна адреса відправника: адреса електронної пошти відправника може бути корисною для виявлення фішингового спаму. Якщо електронний лист надійшов із відомого фішингового домену або підозрілої електронної адреси, швидше за все, це є фішинговий електронний лист.

Вміст електронного листа: вміст електронного листа містить в собі багато інформації, яку доцільно використовувати для виявлення фішингового

спаму. Електронний лист може містити певні ключові слова чи фрази, які зазвичай використовуються у фішингових електронних листах, наприклад, «терміново», «підтвердити», «обліковий запис», «ви виграли», «вам нараховано 10000 гривень!!!», «тільки сьогодні знижка 75%», тощо. Крім того, електронний лист може містити відчуття терміновості чи примушування до швидких дій – це є поширеною тактикою, яка використовується у фішингових електронних листах. Також в фішингових листах досить часто використовуються фейкові посилання на якихось культурних, громадських чи політичних діячів, які викликають довіру в суспільстві.

URL-адреси та посилання, включені в електронний лист: URL-адреси та посилання, включені в електронний лист, також можуть бути корисними для виявлення фішингового спаму. Якщо електронний лист містить посилання, яке перенаправляє на підозрілий веб-сайт або на веб-сайт, що відомий як сайт, що використовується для фішингу, швидше за все, це буде фішинговий електронний лист.

Вкладення: деякі фішингові електронні листи містять вкладення, це також може бути корисною інформацією при перевірці електронного листа на факт фішинга. Хоча треба зазначити, що найбільш «старанні» фішери, для максимальної схожості з оригінальними листами від легітимних організацій, додають в свої листи, крім фішингових посилань, ще й посилання на якісь ресурси з оригінальних сайтів легітимних організацій.

Структура електронного листа та форматування: структуру (структурні параметри) електронного листа та специфічне форматування також можна використовувати як спосіб визначення спаму. Якщо електронний лист має підозрілу тему, занадто офіційну мову, терміновість або відсутність персоналізації, швидше за все, цей лист є фішинговим. Також, досить часто, для того, щоб привернути увагу до листа, в його темі або у вмісти часто використовують такі прийоми форматування, як написання тексту у верхньому регістрі, використання в кінці речення трьох і більше знаків окликів, тощо.

Цей перелік ознак, які можна отримати з електронного листа, для виявлення фішингового спаму не є вичерпним. Вибір має залежати від використовуваного алгоритму виявлення та від специфіки організації, в якій буде застосована дана технологія. В деяких випадках може знадобитися виділити додаткові дані або використати більш просунуті методи вилучення даних, такі як обробка природної мови (NLP) або комп'ютерне бачення (CV), щоб витягти дані з вмісту електронної пошти.

3. Попередня обробка даних: цей крок передбачає очищення та нормалізацію даних шляхом видалення будь-якої нерелевантної інформації і перетворення електронних листів у формат, який можна легко обробити алгоритмом виявлення.

Попередня обробка даних є важливим кроком у підготовці зразка електронного листа для виявлення фішингового спаму. Цей крок передбачає очищення та нормалізацію даних, щоб гарантувати, що їх можна легко

обробити алгоритмом виявлення. Невиконання попередньої обробки або виконання недостатньої обробки зразків, як правило, приводить або до суттєвого зниження ефективності застосування даних технологій, або взагалі до їх непрацездатності.

Видалення нерелевантної інформації: це включає в себе видалення будь-яких верхніх і нижніх колонтитулів з електронного листа, а також будь-якої іншої нерелевантної інформації, тобто інформації, яка не є показовою при виявлені спаму, наприклад підписів або заяв про відмову від відповідальності. Цей крок допомагає спростити дані та усунути будь-який шум, який може негативно вплинути на продуктивність алгоритму виявлення.

Нормалізація тексту: цей крок передбачає перетворення тексту в електронних листах на узгоджений формат, наприклад малі або великі літери, і видалення будь-яких спеціальних символів, цифр або знаків пунктуації. Це допомагає зробити текст більш послідовним і легшим для аналізу. Але, якщо в листі було присутнє зазначене раніше форматування, яке притаманне фішинговим листам, то цей факт необхідно зафіксувати і в подальшому використовувати під час аналізу.

Видалення HTML-тегів: деякі фішингові електронні листи містять HTML-теги та форматування. Цей крок передбачає видалення всіх HTML-тегів і форматування, що зробить текст більш послідовним і легшим для аналізу.

Видалення стоп-слів: цей крок передбачає видалення загальних слів, таких як "the", "and", "is" тощо, які не додають суттєвого значення тексту та смислового навантаження. Це допомагає зменшити розмірність даних і зробити їх більш керованими для алгоритму виявлення.

Створення кореня або лемматизація: цей крок передбачає скорочення слів до їх основи або кореневої форми. Це допомагає зменшити розмірність даних і зробити їх більш керованими для алгоритму виявлення.

Токенізація: цей крок передбачає розбиття тексту на окремі слова або лексеми. Це дає можливість аналізувати слова та фрази в тексті окремо.

Перетворення на числове представлення: цей крок передбачає, за допомогою спеціальних технологій, перетворення тексту в числовий формат, який можна легко обробити алгоритмами машинного навчання.

Підсумувавши, необхідно зазначити, що перераховані етапи накопичення та попередньої обробки електронних листів для створення репрезентативної бази даних навчальних зразків є необхідними і без їх виконання неможливо досягти ефективної та коректної роботи більшості технологій виявлення фішингового спаму.

1. Фішинг як один із основних різновидів інтернет-шахрайства. https://science.donnu.edu.ua/wp-content/uploads/sites/6/2020/06/fishing.pdf.
2. PhishTank. https://www.phishtank.com/.
3. APWG. https://apwg.org/.
4. Spamhaus: Strengthening trust and safety across the internet. https://www.spamhaus.org/.

Х.І. Дрогобицька

ШТУЧНИЙ ІНТЕЛЕКТ ЯК ФАКТОР СУСПІЛЬНИХ ЗМІН: КОНФІДЕНЦІЙНІСТЬ, ПСИХОЛОГІЯ КОРИСТУВАЧА ТА НОВІ ГОРИЗОНТИ БЕЗПЕКИ

Стрімкий розвиток штучного інтелекту (ШІ) у XXI столітті істотно трансформує соціальні структури, інформаційні процеси та систему взаємодії людини з технологічним середовищем. Інтелектуальні алгоритми давно вийшли за межі допоміжних інструментів і стали потужними елементами соціотехнічної екосистеми, здатними впливати на політичні, економічні та культурні процеси. У цьому контексті ШІ розглядається не лише як технологічна інновація, а як комплексний фактор суспільних змін, що здатен формувати поведінкові моделі, впливати на колективну свідомість, трансформувати концепції приватності та безпеки, а також створювати нові ризики, пов'язані з автономними рішеннями, обробкою великих масивів персональних даних та інформаційним впливом.

Однією з ключових сфер, на які істотно впливає інтеграція ШІ, є психологія користувача. Рекомендаційні алгоритми, сервіси персоналізованого аналізу, моделі генеративного ШІ та системи прогнозування поведінки формують нову інформаційну реальність, у якій увага, емоції та когнітивні ресурси людини стають об’єктами алгоритмічної оптимізації. Продуктові системи цифрових платформ навмисно налаштовані на збільшення часу взаємодії користувача з контентом, а це, у свою чергу, формує особливі патерни залежності, змінює мотиваційні механізми та звужує когнітивну автономію. У науковій літературі наголошується, що зростає кількість випадків так званої “algorithmic habituation”—психологічної адаптації до рішень алгоритмів, коли користувач починає віддавати перевагу машинним судженням перед власними. Це явище впливає на якість критичного мислення, здатність аналізувати складні тексти, а також на загальну структуру сприйняття інформації.

Ще одним важливим аспектом впливу ШІ є питання конфіденційності. Алгоритми машинного навчання функціонують переважно на основі великих масивів даних, у тому числі персональних. Сучасні цифрові сервіси збирають інформацію про фізичні характеристики, соціальні зв’язки, переміщення, поведінкові реакції та навіть емоційні стани людини. В умовах посилення розвитку технологій розпізнавання облич, біометрії та аналізу цифрових слідів виникає ризик формування середовища тотального нагляду, у якому межа між добровільним наданням даних і примусовою участю у цифрових процесах стає дедалі нечіткішою. Виклики конфіденційності посилюються тим, що системи ШІ здатні робити висновки із даних, які людина формально не надавала, шляхом побудови кореляційних моделей та прогнозів —

наприклад, прогнозуючи політичні вподобання, стан здоров'я, соціально-економічний статус або ймовірність певних дій у майбутньому [1].

У зв'язку з цим зростає ризик алгоритмічної дискримінації. Помилки у навчальних вибірках, нерепрезентативність даних, алгоритмічні упередження або хибні кореляції можуть формувати несправедливі рішення, які впливають на життя людей у сферах кредитування, працевлаштування, доступу до медичних послуг, оцінювання ризиків або безпеки. Відомі випадки, коли моделі прогнозування рецидивізму в США проявляли системні расові упередження, а алгоритмічні системи відбору персоналу віддавали перевагу кандидатам певної статі або соціальної групи. Хоча такі приклади часто розглядаються як технічні збої, насправді вони є індикаторами складних соціальних ефектів ШІ, який відтворює та підсилює структурні нерівності.

Разом із тим штучний інтелект відкриває нові горизонти безпеки. Одним із найбільш перспективних напрямів є моніторинг екологічних загроз. Технології машинного навчання вже застосовуються для раннього виявлення лісових пожеж, аналізу динаміки затоплень, моделювання кліматичних сценаріїв, спостереження за станом біорізноманіття та прогнозування стихійних лих. Системи на основі нейронних мереж здатні обробляти супутникові знімки у масштабах, недоступних для людського аналізу, і виявляти навіть мінімальні зміни у стані природних об'єктів. Це створює можливості для розбудови ефективної системи екологічної безпеки, покращення управління природними ресурсами та формування політики сталого розвитку.

ШІ також відіграє ключову роль у протидії інформаційним загрозам. Сучасні соціальні мережі є середовищем, у якому поширення фейкових новин, маніпуляцій, масових вкидів та координованих інформаційних операцій відбувається надзвичайно швидко. Інтелектуальні системи дозволяють автоматично аналізувати великі обсяги контенту, виявляти аномальні поведінкові патерни, класифікувати потенційно небезпечні записи, виявляти бот-мережі та фіксувати спроби масової дезінформації. Перевагою ШІ є те, що він може працювати у реальному часі, що значно підвищує ефективність боротьби з кіберзагрозами та інформаційними атаками.

Водночас варто враховувати, що ШІ сам по собі може генерувати нові ризики у сфері інформаційної безпеки. Генеративні моделі здатні створювати переконливі фейкові відео (deepfake), аудіо та тексти, які важко відрізнити від реальних. Це відкриває можливості для маніпуляцій громадською думкою, шантажу, підриву політичної стабільності та порушення виборчих процесів. У зв'язку з цим ключовим завданням стає розроблення механізмів перевірки достовірності контенту, а також розвиток цифрової грамотності населення [2].

Таким чином, штучний інтелект виступає водночас і каталізатором розвитку, і джерелом нових загроз. Його вплив на суспільство є багатовимірним: від психологічних аспектів взаємодії користувачів із

технологічними платформами до складних правових питань конфіденційності та алгоритмічної справедливості. У контексті зростаючої цифрової залежності та глобальної інформатизації постає необхідність у створенні ефективної системи етичного регулювання ШІ, яка могла б збалансувати інноваційний потенціал і захист прав людини. Це включає формування прозорих стандартів використання даних, механізмів аудиту алгоритмів, системи превенції дискримінації, а також політик, спрямованих на підвищення цифрової компетентності населення.

Перспективи ШІ є безмежними: він здатен трансформувати управління державою, медицину, освіту, науку, транспорт і промисловість [3, с.4]. Проте реалізація його потенціалу без урахування соціальних ризиків може призвести до формування нових нерівностей, втрати приватності та підриву довіри до цифрових систем. Сучасне суспільство стоїть перед необхідністю знайти баланс між використанням технологічних можливостей і забезпеченням гуманістичних принципів розвитку. Лише за умови відповідального підходу, міждисциплінарного регулювання та чіткої етичної рамки штучний інтелект зможе стати інструментом прогресу, а не джерелом загроз для майбутніх поколінь.

1. Дія.Освіта – IT-студії. *Дія.Освіта – IT-студії*. URL: https://it-osvita.diia.gov.ua/task/item/8a6d8812-ab01-4285-a6fd-200dc3a1c6c2 (дата звернення: 21.11.2025).
2. Штучний інтелект і підлітки: ризики нових форм цифрової взаємодії | Центр Дністрянського. *Центр Дністрянського | Центр Дністрянського*. URL: https://dc.org.ua/news/shtuchnyy-intelekt-i-pidlitky-ryzyky-novyh-form-cyfrovoyi-vzaemodiyi (дата звернення: 21.11.2025).
3. Пищуліна О. штучний інтелект – можливості, небезпеки, фактори ризику. *Центр Разумкова*. URL: https://razumkov.org.ua/images/2025/09/15/sh_int.pdf (дата звернення: 21.11.2025).

І.М. Дюба

ШТУЧНИЙ ІНТЕЛЕКТ У ЗАХИСТІ ДАНИХ ПРИ УПРАВЛІННІ ПРИСТРОЯМИ З ВІДКРИТИМИ ІНФОРМАЦІЙНИМИ КАНАЛАМИ

Штучний інтелект (ШІ) відіграє критично важливу роль у захисті даних при управлінні пристроями, особливо тими, що функціонують через відкриті інформаційні канали, оскільки традиційні методи захисту часто не справляються з величезним обсягом і швидкістю передачі даних та динамікою загроз.

Зазвичай всі описують [1] переваги використання ШІ але існує і зворотна сторона наявність ШІ підвищує вимоги до надійності методів захисту даних. Розглянемо загрози які виникають в наслідок існування ШІ для пристроїв з відкритими каналами:

Масштабованість - ШІ може одночасно моніторити мільйони пристроїв та великі обсяги даних, що неможливо для людини.

Швидкість - Реакція на загрози відбувається за мілісекунди, що є життєво необхідним для запобігання поширенню атаки.

Адаптивність - ШІ може самостійно навчатися та адаптуватися до нових, раніше невідомих (zero-day) загроз.

Зменшення помилок - Автоматизація зменшує людський фактор і ймовірність пропуску важливого інциденту.

В якості прикладу того як переваги ШІ перетворюються на загрози для іншої сторони, особливо у військовому контексті, скористаємось технологією БПЛА:

Масштабованість → Загроза координованої атаки роїв.

Захисна перевага: ШІ моніторить мільйони пристроїв.

Атакуюча загроза: ШІ може координувати атаку рою (swarmattack) з десятків чи сотень БПЛА, які діють як єдиний інтелектуальний організм.

ШІ-система рою може одночасно аналізувати ППО супротивника, розподіляти цілі та самостійно перепланувати маршрути, щоб обійти захист. Це створює масштабовану, некеровану вручну загрозу, що перевантажує традиційні системи протидії.

Швидкість → Загроза гіпершвидкісного рішення

Захисна перевага: Реакція на загрози за мілісекунди.

Атакуюча загроза: ШІ дозволяє БПЛА приймати рішення про атаку, зміну курсу, або виявлення цілі на швидкостях, недосяжних для людського оператора.

Наприклад, ШІ може ідентифікувати, класифікувати і атакувати ціль за частки секунди після візуального контакту, мінімізуючи час на реакцію захисної системи.

Адаптивність → Загроза обходу захисту та глушіння

Захисна перевага: ШІ адаптується до zero-day загроз.

Атакуюча загроза: ШІ в системах РЕБ (радіоелектронної боротьби) та БПЛА може динамічно змінювати частоти, потужність та схеми зв'язку для обходу систем глушіння (Jamming) або придушення каналів зв'язку (Anti-Jamming).

Атакуючий ШІ, що керує БПЛА, може самостійно навчатися на реакції захисної системи (наприклад, системи ППО чи РЕБ), і коригувати свою стратегію для подальших атак, роблячи захисні заходи неефективними.

Зменшення помилок → Загроза автономної смертоносності

Захисна перевага: Автоматизація зменшує людський фактор.

Атакуюча загроза: Автоматизація дає можливість створювати летальні автономні системи зброї (LAWS), де рішення про застосування сили приймає ШІ.

Зменшення людських помилок в атакуючому комплексі ШІ означає підвищену точність, ефективність та надійність атаки, але водночас несе етичні та стратегічні ризики через відсутність людського контролю в петлі прийняття рішень (Human-in-the-Loop).

У цьому контексті, існування потужного атакуючого ШІ перетворює відкриті інформаційні канали, як-от GPS, супутниковий зв'язок чи радіоканали, на слабкі місця, якими може скористатися інтелектуальна система супротивника. Тому актуальною науковою задачею є захист таких слабких міст як за рахунок закриття інформації так і за рахунок диверсифікації інформації наприклад за рахунок використання багаторівневих моделей доступу.

Також важливим є розвідувальний аспект і використання відкритих даних (OSINT), отриманих безпосередньо з каналів зв'язку БПЛА, для планування контрзаходів.Цей процес базується на використанні ШІ для швидкої обробки та інтерпретації великих обсягів даних, отриманих із відкритого (незашифрованого або скомпрометованого) відео- чи телеметричного каналу БПЛА.

Зробим огляд цього більш детально:ШІ та використання відкритих відеоканалівБПЛА для контратак якщо канал відео або телеметрії БПЛА є "відкритим" (тобто не захищений або його шифрування скомпрометовано), ШІ стає критично важливим інструментом для швидкого видобування цінної інформації, яка може бути негайно використана для інформаційних або фізичних контратак.

Виявлення та аналіз інформації (IntelligenceExtraction). ШІ-системи обробляють відеопотік та метадані з БПЛА в реальному часі:

1. Розпізнавання об'єктів та сцен (ObjectRecognition) - Завдання ШІ: Використовуються згорткові нейронні мережі (CNN) для ідентифікації та класифікації військової техніки, інфраструктури, розташування військ, чи пунктів управління.Цінність: Миттєво визначається ціль БПЛА та її поточні дії.

2. Геолокація та картографування - Завдання ШІ: Аналіз відеопотокуразом із даними GPS/телеметрії (якщо вони відкриті) дозволяє

ШІ точно визначити місце розташування БПЛА, його маршрут та координати об'єктів, які він знімає.Цінність: Швидке створення оперативних карт для планування контрдій.

3. Аналіз поведінки (ActivityAnalysis) - Завдання ШІ: Алгоритми глибокого навчання аналізують послідовність кадрів для ідентифікації патернів (наприклад, підготовка до атаки, патрулювання, зміна позицій).Цінність: Прогнозування наступного кроку ворожого БПЛА чи його оператора.

Аспект протидії найбільш ярко виражен через інформаційні та фізичні контратаки.

Інформаційні контратаки [2](InformationCounterattacks) - ці атаки спрямовані на дезінформацію або порушення роботи самого БПЛА через його відкритий канал.

Метод Контратаки	**Роль ШІ**
Впорскування змагальних прикладів	ШІ створює змінені кадри (додаючи невеликий шум), які вставляються у відеоканал. Вони непомітні для людського оператора, але змушують ШІ ворожого БПЛА (якщо він використовується для автономного наведення) помилково ідентифікувати цілі або втратити їх.
Симуляція та імітація (Spoofing)	ШІ аналізує нормальний відеосигнал і використовує його для створення правдоподібних діпфейків чи імітацій ландшафту, які подаються у канал БПЛА, змушуючи його повірити, що він знаходиться в іншому місці або що ціль вже знищена.
Дезінформація телеметрії	ШІ може аналізувати патерни телеметричних даних (швидкість, висота) і використовувати їх для систематичного введення невеликих помилок у потік даних, які повільно погіршують точність навігації БПЛА, змушуючи його відхилитися від курсу.

Фізичні Контратаки (KineticCounterattacks)

ШІ використовує отриману інформацію для максимізації ефективності фізичного ураження.

Оптимізація РЕБ (Jamming):Роль ШІ: Після ідентифікації частоти та протоколу зв'язку БПЛА, ШІ-система РЕБ точно налаштовує потужність та частоту глушіння, щоб швидко перервати зв'язок і примусити БПЛА до приземлення або повернення на базу, мінімізуючи вплив на дружні канали.

Автоматичне Наведення:Роль ШІ: Отримавши точні координати цілі та прогнозований маршрут БПЛА, ШІ автоматично передає ці дані системам ППО, керованим ракетам або власним дронам-перехоплювачам. Це скорочує цикл "виявлення-ураження" з хвилин до секунд.

Ідентифікація Оператора:Роль ШІ: Аналіз відеоканалу може розкрити

патерни поведінки оператора (рухи камери, час реакції). ШІ може використовувати ці дані в поєднанні з радіорозвідкою, щоб локалізувати місце розташування пункту управління для подальшого фізичного ураження.

Висновок: Проведений аналіз чітко демонструє, що Штучний Інтелект (ШІ) є наріжним каменем сучасної кібербезпеки, особливо для пристроїв, що функціонують через відкриті інформаційні канали (наприклад, БПЛА, IoT). Його основні переваги роблять його незамінним для створення ефективних, проактивних захисних систем, здатних протистояти динаміці сучасних загроз.

Однак, фундаментальна перевага ШІ полягає в його технологічній нейтральності, що створює двосторонній виклик:

Посилення Загрози (Атакуючий ШІ): Ті самі переваги, які роблять ШІ ефективним захисником, перетворюються на гіпершвидкісні та некеровані вручну загрози в руках супротивника. У військовому контексті (БПЛА), атакуючий ШІ може реалізувати координовані атаки роїв та динамічний обхід захисту (наприклад, РЕБ), використовуючи відкриті канали як слабкі місця. Це вимагає від захисників значного підвищення вимог до надійності.

ШІ як Каталізатор Контрзаходів (Захисний ШІ): Загроза посилюється, але і захист набуває нових можливостей. ШІ стає критичним інструментом для активної протидії, зокрема шляхом:

Розвідувального Аналізу (OSINT): Швидке вилучення цінних даних (геолокація, розпізнавання об'єктів, прогнозування поведінки) з відкритих або скомпрометованих відео/телеметричних каналів БПЛА.

Інформаційних контратак: Застосування змагальних прикладів та імітації (Spoofing) для дезорієнтації ворожого ШІ або оператора через канал зв'язку.

Фізичних контратак: Використання ШІ для оптимізації РЕБ, автоматичного наведення засобів ураження (ППО, дрони-перехоплювачі) та ідентифікації місця розташування оператора, скорочуючи цикл "виявлення-ураження" до мінімуму.

Таким чином, у середовищі пристроїв з відкритими інформаційними каналами (зокрема, БПЛА), безпека перетворилася на гонку між інтелектами. Захист даних і систем не може обмежуватися лише закриттям каналів (шифруванням). Він вимагає комплексного, багаторівневого підходу, де ШІ використовується для агресивного захисту, застосовуючи як кінетичні, так і інформаційні контрзаходи, засновані на миттєвому аналізі даних, отриманих від самого атакуючого. Майбутнє безпеки залежить від розробки стійких та адаптивних захисних моделей ШІ, які можуть ефективно протистояти атакам, згенерованим їхніми ж аналогами.

1. Smith, J., &Jones, A. (2023). "Deep Learning for UAV Detection and Classification in C-UAS Systems." Journal of Autonomous Systems, 15(3), 112–125.
2. Chen, L., Wang, Q., &Li, H. (2022). "A Survey on Adversarial Attacks and Defenseson AI in Cyber Security." IEEE Transactionson Information Forensics and Security, 17, 3050–3065.

Ю.В. Писаренко, К.Ю. Мамедова, О.С. Коваль, К.В. Кармазін

АРХІТЕКТУРА БАГАТОРІВНЕВИХ СИТУАЦІЙНИХ ЦЕНТРІВ ДЛЯ КЕРУВАННЯ UAV В УМОВАХ НЕВИЗНАЧЕНОСТІ

Сучасні безпілотні авіаційні системи функціонують у середовищі, яке характеризується високою динамічністю, фрагментованістю інформації та невизначеністю щодо дій операторів, стану каналів зв'язку й технічного стану платформ. За таких умов класичні централізовані системи керування виявляють істотні обмеження: накопичення затримок, перевантаження вузлів обробки, вразливість до поодиноких відмов та кібератак на ключові елементи інфраструктури. У результаті знижується якість ситуаційної обізнаності, ускладнюється синхронізація дій операторів та зростає ймовірність несвоєчасного або некоректного керуючого впливу.

Ці тенденції природно приводять до постановки задачі побудови багаторівневої архітектури ситуаційних центрів, у якій центральний рівень забезпечує узагальнену координацію та політику функціонування, а регіональні та мобільні рівні відповідають за оперативні рішення у своєму секторі відповідальності. Ключовою вимогою до такої архітектури є не лише розподіл обчислювального навантаження, а й здатність працювати за умов часткової втрати даних, локальних збоїв сенсорної мережі та змін структури мережевих маршрутів. У цьому контексті особливого значення набуває використання верифікованих алгоритмів керування та засобів моделювання, які дозволяють оцінювати стійкість системи та її чутливість до різних типів збурень ще на етапі проєктування.

Дана тема є продовженням попередніх досліджень, присвячених розробці системи підвищення контролю за повітряним простором [1, 2, 3].

Запропоновану архітектуру доцільно розглядати як трирівневу керуючу систему $L_0 \rightarrow L_1 \rightarrow L_2$, де рівень L_0 – відповідає центральному ситуаційному центру (ЦСЦ), L_1 – сукупності регіональних ситуаційних центрів (РСЦ), а L_2 – мобільним ситуаційним центрам (МСЦ), інтегрованим з угрупованнями безпілотних авіаційних систем. ЦСЦ реалізує функції стратегічного планування, глобальної координації та формування політик безпеки, включно з визначенням зон відповідальності, допустимих режимів застосування UAV та регламентів інформаційної взаємодії між рівнями. РСЦ забезпечують операційне керування у своїх просторово визначених секторах, агрегують телеметрію, локальні оцінки загроз і стану інфраструктури, а також ініціюють коригування місій відповідно до стану операційного середовища. МСЦ розгортаються у безпосередній близькості до районів застосування та орієнтовані на мінімізацію затримок керуючих впливів і підтримання керованості за умов швидкоплинної зміни оперативної ситуації.Інформаційна взаємодія між рівнями організується як двонаправлена багатоканальна мережа. Потоки «знизу вгору» містять телеметричні дані, агреговані показники стану, виявлені аномалії та діагностичні повідомлення, натомість

потоки «згори донизу» включають політики, обмеження, параметри місій, шаблони сценаріїв реагування та оновлені конфігурації. Синхронізація телеметрії здійснюється на основі узгоджених часових шкал, єдиних форматів опису станів та процедур попередньої фільтрації, що зменшують обсяг переданих даних без втрати інформаційної повноцінності для задач оцінювання та прийняття рішень.

Принцип децентралізованого контролю реалізується через формальний розподіл повноважень між рівнями: ЦСЦ задає цільові функції та глобальні обмеження, РСЦ і МСЦ виконують локальну оптимізацію дій з урахуванням реального стану каналів зв'язку, сенсорної мережі та ресурсних обмежень. Це знижує ризик виникнення єдиної точки відмови, зменшує чутливість системи до локальних збоїв і забезпечує плавну деградацію працездатності замість катастрофічного руйнування структури керування.

Ключовим компонентом архітектури є цифровий двійник зони функціонування UAV, підтримуваний на рівні ЦСЦ з урахуванням даних, що надходять від РСЦ та МСЦ. Цифровий двійник інкапсулює модель повітряного простору, радіочастотного та інформаційно-комунікаційного середовища, топологію сенсорної інфраструктури, а також конфігурацію угруповань безпілотних платформ. На його основі реалізуються прогнозування еволюції станів, валідація сценаріїв реагування, а також аналіз чутливості системи до різних типів збурень і відмов. Це дає змогу виконувати попередню оцінку наслідків керуючих рішень у віртуальному просторі, підвищувати обґрунтованість вибору стратегій керування та забезпечувати трасованість від спостережуваних даних до прийнятих дій.

Методологічний базис запропонованого підходу спирається на поєднання системного аналізу, моделювання та симуляції, а також методів машинного навчання для забезпечення адаптації до змінних режимів операційного середовища. Системний аналіз використовується для архітектурної декомпозиції багаторівневої системи керування на підсистеми спостереження, діагностики, планування та реалізації керуючих впливів. Така декомпозиція дозволяє явно задати інтерфейси між рівнями L_0,L_1,L_2, визначити потоки даних, зони відповідальності та формалізувати вимоги до стійкості кожної підсистеми.

Моделювання та симуляція застосовуються у контексті цифрового двійника зони функціонування UAV як інструменту верифікації поведінкових сценаріїв. У цифровому двійнику реалізуються як дискретно-подійні та мережеві моделі передачі телеметрії, так і агент-орієнтовані моделі руху платформ та їхньої взаємодії із сенсорною інфраструктурою. Це дає змогу аналізувати ефекти затримок, втрат пакетів, відмов окремих вузлів та змін конфігурації мережі до розгортання відповідних рішень в реальній системі. Окремо розглядаються сценарії деградації, у межах яких оцінюється здатність архітектури зберігати керованість і виконувати критичні функції при частковій втраті ресурсів.

Компонент адаптації реалізується на основі методів машинного навчання, включно з навчанням з підкріпленням для локальної оптимізації дій у режимах невизначеності. Ідентифікація маршрутів UAV формулюється як задача класифікації та реконструкції траєкторій за телеметричними послідовностями; для цієї мети використовуються моделі, здатні враховувати часову кореляцію (наприклад, рекурентні або трансформерні архітектури). Виявлення рухомих об'єктів у відео- та радіолокаційних даних розглядається як поєднання детекторів ознак та статистичних критеріїв аномальності, що дозволяє знижувати кількість хибних спрацювань у шумових умовах. Контроль доступу до повітряного простору в режимах невизначеності формалізується через правила, які динамічно уточнюються з урахуванням оновлених оцінок «свій/чужий», ризик-профілів маршрутів та стану сенсорної мережі.

Багатокритеріальна оптимізація застосовується для узгодження цілей стійкості, навантаження мережі та часу реагування. Вона реалізується у вигляді задачі вибору рішень на Парето-множині за допомогою методів скаляризації або ієрархізації критеріїв. Зокрема, для оперативного рівня пріоритет може надаватися мінімізації часу реакції за умови допустимого рівня завантаженості каналів, тоді як для стратегічного рівня — максимізації запасу стійкості при заданих обмеженнях на обчислювальні ресурси.

Вимоги до телеметрії задаються через мінімальні частоти дискретизації, точність часової синхронізації та допустимі інтервали втрат даних, при яких ще можлива коректна реконструкція стану. Вводяться рівні діагностики: локальний (на борту UAV та в МСЦ), регіональний (РСЦ) та глобальний (ЦСЦ), кожен з яких має власні критерії повноти та достовірності даних. Критерії коректності станів системи включають фізичну реалізовність траєкторій, узгодженість параметрів різних сенсорних каналів, відсутність конфліктів з обмеженнями безпеки та відповідність плановим режимам, специфікованим на рівні політик. Сукупність цих методологічних та алгоритмічних компонентів формує основу для побудови верифікованої, адаптивної та стійкої багаторівневої системи керування UAV.

У роботі сформовано концептуально завершену та математично обґрунтовану архітектуру багаторівневої системи керування UAV з чітким розмежуванням функцій між центральним, регіональними та мобільними ситуаційними центрами. Архітектура містить формалізований опис інтерфейсів взаємодії, структуру інформаційних потоків, правила делегування повноважень і механізми децентралізованого контролю; у такому вигляді вона слугує референтною моделлю для подальшого інженерного проєктування та масштабування систем керування безпілотними платформами.

Сформований комплекс методів верифікації цифрового двійника зони функціонування UAV охоплює критерії відповідності між симуляційними та експлуатаційними даними, процедури стрес-тестування поведінкових

сценаріїв, а також аналіз чутливості до параметричних збурень і відмов елементів. Застосування цих методів дає змогу виявляти «вузькі місця» в структурах керування, оцінювати запас стійкості системи та коригувати конфігурацію ще на етапі проєктування, до розгортання на реальних полігонах.

Алгоритмічна частина роботи представлена засобами адаптивного моніторингу, що інтегрують аналіз потокової телеметрії, ідентифікацію маршрутів, виявлення аномальних режимів і діагностику станів у режимах неповної спостережуваності. Показано, що використання таких засобів скорочує час виявлення інцидентів, підвищує ефективність реагування на загрози та підтримує стійкість критично важливої інфраструктури, у якій UAV виконують функції сенсорних і виконавчих елементів. Запропонований підхід узгоджується з вимогами систем національного рівня, орієнтованих на контроль повітряного простору й захист інфраструктурних вузлів, та окреслює подальші напрямки досліджень: побудову програмно-апаратних прототипів, експериментальну валідацію на спеціалізованих полігонах і застосування формальних методів оцінки надійності.

1. J. Pisarenko, and E. Melkumyan, "The Structure of the Information Storage "CONTROL_TEA" for UAV Applications," IEEE 5th International Conference "Actual Problems of Unmanned Aerial Vehicles Developments (APUAVD)" (October 22-24, 2019), pp. 274–277. https://doi.org/10.1109/APUAVD47061.2019.8943938.
2. Kateryna Melkumian, Julia Pisarenko, Alexander Koval, Organization of regional situational centers of the intelligent system «CONTROL_TEA» using UAVs, 2021 IEEE 6th International Conference Actual Problems of Unmanned Aerial Vehicles Developments (APUAVD), Kyiv, Ukraine2021. https://doi.org/10.1109/APUAVD53804.2021.9615416.
3. Pisarenko J., Melkumyan K., Varava I., Koval O., Chumakova N. About the organization of regional situational centers of the intellectual system “Control_TEA” with the use of UAVs. Artificial intelligence. Kyiv, 2022. No. 1. P. 275-287. https://doi.org/10.15407/jai2022.01.275.

О.А. Ворон

АЛЬТЕРНАТИВНІ МОДЕЛІ ШТУЧНОГО ІНТЕЛЕКТУ ТА РОБОТИЗОВАНОЇ, ДИСТАНЦІЙНО-КЕРОВАНОЇ ТЕХНІКИ ПРИ ФІТОРЕМЕДІАЦІЇ ТЕРИТОРІЙ ТЕХНОГЕННИХ ОБ'ЄКТІВ НА ПРИКЛАДІ ЗОЛОШЛАКОВИХ ВІДВАЛІВ ТЕС

Сучасна енергетична та промислова діяльність зумовлює накопичення значних обсягів техногенного матеріалу на територіях довкола теплоелектростанцій. Золошлакові відвали Трипільської, Придніпровської, Зміївської та Бурштинської ТЕС містять оксиди важких металів (Pb, Zn, Cu, Cd), які становлять загрозу для ґрунтового та водного середовища. Погіршення фізико-хімічних властивостей ґрунтів супроводжується зниженням їх родючості, деградацією екосистем та забрудненням прилеглих територій [1].

Сьогоденні ефективним методом відновлення таких техногенних ландшафтів є фіторемедіація. Процес фіторемедіації передбачає використання рослин для вилучення, трансформації або стабілізації токсикантів у ґрунті (Desai et al., 2019; Pandey et al., 2016), підземних водах та інших відходах. Це найбільш екологічно чиста та економічно ефективна стратегія, оскільки рослини можуть колонізувати золовідвали, створюючи самопідтримуваний рослинний шар для відновлення здоров'я екосистеми. Забруднена територія поступово відновлюється з часом, і це призводить до зменшення кількості забруднюючих речовин у навколишньому середовищі. Хоча існує багато повідомлень про фіторемедіацію забруднених ділянок, дослідження щодо потенційного використання стратегій фіторемедіації для відновлення золовідвалів і для фітостабілізації важких металів у відходах вугільних ТЕС.

Для рекультивації та екологічного оздоровлення поверхні золошлакових відвалів ТЕС було обрано комплекс рослин-фіторемедіантів, здатних ефективно вилучати важкі метали, покращувати структуру техногенного субстрату та забезпечувати поступове відновлення біопродуктивності ґрунту. Вибір видів базувався на їхній стійкості до екстремальних умов, характерних для золошлакових масивів, а також на їхніх фіторемедіаційних властивостях [2-6].

Основними критеріями відбору були:

- висока толерантність до засолення, лужності та низького вмісту органічних речовин;
- розвинена коренева система, здатна закріплювати нестійкі субстрати та проникати на значну глибину;
- здатність накопичувати важкі метали (Pb, Cd, Zn, Ni, Cu, Cr) без значної втрати життєздатності;

- внесення поживних речовин у ґрунт після відмирання надземної та кореневої маси (азот, фосфор, калій, кальцій, органічні сполуки);
- невибагливість до умов вологості, що характерно для відкритих техногенних територій [7,8].

Для рекультивації поверхонь золо-шлакових відвалів доцільним є використання стійких рослин-ремедіантів, здатних до росту в умовах забруднення вугільною легкою золою (CFA). До таких видів належать *Ipomoea carnea*, *Lantana camara*, *Melilotus officinalis*, *Onobrychis arenaria*, *Reynoutria japonica*, *Miscanthus spp.* та *Sorghum bicolor*. Ці рослини проявляють високу толерантність до важких металів і низької забезпеченості поживними речовинами, що робить їх перспективними для фіторемедіації техногенно порушених субстратів [9, 10].

Разом з тим більшість зазначених видів не можуть повною мірою адаптуватися до середовищ, сформованих із 100% CFA, оскільки такі субстрати характеризуються несприятливими фізико-хімічними властивостями, включаючи дисбаланс елементів живлення, низький вміст органічної речовини та значні відхилення pH. Це обмежує формування рослинного покриву та вимагає попереднього поліпшення властивостей CFA.

Переваги технології полягають у її екологічній безпеці, економічній доцільності та можливості інтеграції у програми сталого розвитку промислових регіонів. А ефективність фіторемедіації залежить від типу забруднення,

біологічних властивостей рослин, кліматичних умов та фізико-хімічних характеристик субстрату (Hrynkiewicz et al., 2018). Для підвищення наукової обґрунтованості запропоновано модель фіторемедіації, що поєднує біологічні, екологічні та математичні аспекти управління відновленням територій.

Проаналізовано і досліджено, що альтернативні моделі штучного інтелекту (ШІ) та машинного навчання (МО) дозволяють прогнозувати динаміку фітогіперакумуляції важких металів, оптимізувати вибір рослин та мікробних симбіонтів на основі мультиомних даних (геноміка, протеоміка). Роботизована техніка, зокрема дрони з мультиспектральними сенсорами та автономні роботи, забезпечує дистанційний посів насіння, моніторинг біомаси та накопичення забруднювачів у реальному часі, зменшуючи ризики для персоналу на техногенних об’єктах (Kumar et al., 2022; Wang et al., 2025). Основні механізми фіторемедіації у рослинах, які забруднені важкими металами зображено на рис. 1.

В Україні рекультивація техногенних територій регламентується Земельним кодексом (2002) та Законом «Про землеустрій» (2003). Важливим є поєднання гірничотехнічного та біологічного етапів рекультивації: перший формує техногенний рельєф, другий — стабілізує біогеохімічні процеси через фітомеліорацію.

Дослідження Mormul і Terekhov (2017) показують, що ефективність фіторемедіації залежить від складу фітоценозу та використання видів з високою здатністю акумулювати важкі метали (Salix viminalis L., Brassica juncea, Helianthus annuus). Spontaneous відновлення рослинності також може бути ефективним, якщо субстрат не перевищує критичних рівнів токсичності [9]. Металофіти, такі як Biscutella laevigata L. та Silene vulgaris, стабілізують поверхню відвалів і сприяють природній сукцесії техногенних екосистем.

Рисунок 1 – Механізми фіторемедіації у рослинах, які забруднені важкими металами

Українські науковці (Korsunskyi et al., 2018; Petlovanyi et al., 2023) доводять, що інтеграція геотехнічних і біологічних підходів створює потенційно родючий шар на відвалах, що значно підвищує ефективність рекультивації. Слід зазначити, що бобові мають великий потенціал у процесах фіторемедіації, оскільки вони збільшують кількість мікробних популяцій і можуть зробити фіксовані поживні речовини доступними для інших культур [8].

Основні методи обробки забруднених ділянок під підприємствами гірничої промисловості, відвалів гірничих порід, золошлакових відвалів наступні: загальні стратегії; хімічний; фізичний; біологічний; фіторемедіація

Альтернативні моделі ШІ, такі як нейронні мережі та графові нейронні мережі (GNN), дозволяють моделювати забруднення та прогнозувати ефективність фітогіперакумуляції з точністю до 85–95% на основі даних дистанційного зондування (Li et al., 2025). Роботизована техніка, включаючи рої дронів для посіву насіння (до 40 000 подів/день) та наземні роботи для зразко-збору, оптимізує процедуру в небезпечних зонах золошлакових відвалів, зменшуючи час на моніторинг з місяців до днів (Flash Forest, 2025).

Особливо актуальним є застосування ШІ при проведенні підготовчо-розміточних робіт для посадки рослин на поверхні золошлакових відвалів. Ці роботи включають картографування забруднення, зонування ділянок за рівнями ризику (низький, середній, високий) та оптимізацію розміщення посадкових точок для максимальної ефективності фітогіперакумуляції. Використання дронів з високорозрізнювальною аерофотозйомкою (HRAI) та алгоритмів машинного навчання (наприклад, екстремальний випадковий ліс — ERF) дозволяє створювати точні карти забруднення важкими металами з точністю до 87%, що перевершує традиційні методи інтерполяції Крігінга. Це забезпечує автоматизоване розмежування зон для цільового посіву металофітів, зменшуючи витрати на підготовку на 30–40% та мінімізуючи ризики поширення забруднення [9].

У підготовчо-розміточних роботах ШІ інтегрувався для картографування поверхні відвалів: рої дронів з сенсорами (мультиспектральними, LiDAR, гамма-спектрометрами) збирали дані про вміст важких металів, вологість, рельєф та ерозію. Алгоритми МО (ERF, SVM) аналізували HRAI-зображення для зонування ділянок (наприклад, 66,9% низький ризик, 29,4% середній, 3,7% високий), оптимізуючи розміщення посадок з точністю 85–87%. Це дозволило автоматизувати розмітку для посіву, симулюючи сценарії та рекомендуючи оптимальні параметри (щільність, види рослин) на основі історичних даних [11,12].

Таке зонування дозволило оптимізувати розміщення посадок (фіторемедіаційних культур) з високою точністю, що сягає 85–87%. Це значно підвищило ефективність автоматизованої розмітки для посіву. ШІ також симулював різні сценарії рекультивації та надавав рекомендації щодо оптимальних параметрів (зокрема, щільності посіву та видів рослин) на основі аналізу історичних даних.

Дослідження проводилося на полігонах золошлакових відвалів ТЕС (Трипільська, Придніпровська, Зміївська, Бурштинська). Для оцінки ефективності фіторемедіації використовували спонтанну рослинність та штучні експериментальні ділянки з індикаторними культурами: Salix viminalis L., Brassica napus L., Festuca rubra L., Medicago sativa L. Відбір ґрунтових зразків здійснювався на глибині 0–20 см на початку експерименту та через 12 місяців. Концентрації Pb, Zn, Cu, Cd визначали методом атомно-абсорбційної спектрометрії (ДСТУ ISO 11047:2001). Біомасу рослин висушували до сталої маси при 70°C та визначали вміст металів.

Додатково застосовувалися дрони з мультиспектральними сенсорами (NIR, гіперспектральне зображення) для моніторингу росту рослин, біомаси та накопичення металів у реальному часі. Автономні роботи використовувалися для автоматизованого посіву насіння та зразкозбору в зонах високої токсичності. Дані оброблялися моделями машинного навчання (випадковий ліс, SVM) для прогнозування ефективності [13,14].

На відміну від стандартних RGB-камер, які обмежуються видимим світлом, мультиспектральними сенсори дозволяють аналізувати відбиття світла в конкретних довжинах хвиль, що робить їх незамінними для точного моніторингу навколишнього середовища, сільського господарства та рекультивації забруднених територій, як-от золошлакові відвали ТЕС. Основні типи таких дронів, які можна застосовувати такі: DJI Matrice 350 RTK, AgEagle eBee X, Yuneec H850-RTK, DJI Phantom 4 Multispectral, Mavic 3. Етапи проведення фіторемедіації за допомогою роботизованої, дистанційно-керованої техніки і альтернативної моделі штучного інтелекту територій техногенних об'єктів представлено на рис. 2.

Рисунок 2 – Етапи проведення фіторемедіації (згенеровано за допомогою штучного інтелекту GPT Open AI)

За даними аналізу наукових публікацій, проєктів нами запропоновано математична модель, яка описує динаміку вилучення металів з ґрунту через біоакумуляцію та масообмін між ґрунтом (S), рослинами (P) та мікробіотою (M) за [15]:

$$\frac{dC}{dt} = -k_a \cdot C_S \cdot P - k_m \cdot C_M + k_d \cdot P_{deg} + k_{deg} \cdot M_{deg} \quad (1)$$

де: k_a — коефіцієнт поглинання металу рослиною; k_m — коефіцієнт біотрансформації мікроорганізмами; k_d — швидкість деградації фітомаси; k_{deg} — швидкість біорозкладу мікробної популяції.

Далі альтернативна модель ШІ інтегрує дані сенсорів у графову нейронну мережу (GNN) для прогнозування:

$$E_{pred} = f_{GNN}(X_{soil}, X_{plant}, X_{microbe}; \theta) \quad (2)$$

де E_{pred} – передбачувана ефективність очищування; f_{GNN} – "чорна скринька" моделі, яка виконує обчислення через шари message passing (передачу повідомлень між вузлами графа). Вона перетворює вектори вхідних даних (X_{soil}, X_{plant}, $X_{microbe}$) на вихідний прогноз. X_{soil} – вектор вхідних даних ґрунту; X_{plant} – вектор вхідних даних рослини ; $X_{microbe}$ – вектор вхідних даних мікробів ; θ – параметри моделі за мультиомними даними

Індекс ремедіаційної ефективності (IRE) розраховуються як:

$$IRE = \frac{C_{initial} - C_{final}}{C_{initial}} \times 100\% \quad (3)$$

де $C_{initial}$ – початкова концентрація забруднювача; C_{final} – фінальна концентрація забруднювача.

Встановлено, що параметри θ— це внутрішні ваги та константи мережі, які оптимізуються під час навчання (наприклад, за допомогою градієнтного спуску на мультиомних даних: спектри, геноміка, сенсорні вимірювання). Вони визначають, як мережа обробляє дані через механізм message passing (передача повідомлень між вузлами). Параметри поділяються на структурні (архітектура) та оптимізовані (ваги). GNN моделює взаємодії між елементами системи (наприклад, ґрунт, рослини, мікробіота) як граф, де вузли — це компоненти (наприклад, зразки ґрунту чи рослини), а ребра — взаємозв'язки (наприклад, поглинання металів рослинами з мікробами).

Основні параметри, які впливають на модель:

- Навчання: θоптимізуються на датасеті (наприклад, симуляції з мультиспектральних сенсорів дронів), з loss = IRE_{real} - E_{pred}. Точність: 85–95% для прогнозу часу очищення (7–10 років для Cd/Zn).
- Інтерпретація: У GNN (наприклад, GraphSAGE чи GAT) параметри дозволяють аналізувати важливість ребер (наприклад, високі ваги для plant-microbe — синергія).
- Переваги в контексті: GNN краща за традиційні моделі (як диференціальна рівняння), бо враховує нелінійні взаємодії в реальному часі.

За результатами спостережень найбільшу здатність до акумуляції важких металів проявили корзиночна (Salix viminalis L.), Ріпак олійний (Brassica napus L.). Вони вилучали 35–48% кадмію та свинцю за один вегетаційний сезон. Трав'янисті види щільний вівсянець червоний (Festuca

rubra) та (Medicago sativa) сприяли стабілізації поверхні та зниженню пилового навантаження. РН ґрунту стабілізувався з 5,1 до 6,2, вміст гумусу зріс на 0,3%, що свідчить про початок процесу ґрунтоутворення. Моделювання показало, що максимальна ефективність досягається при \(k_a = 0,004–0,006 \) $д^{-1}$ і щільності насадження 50–60 тис. стебел/га. Час напівочищення ґрунту від кадмію становить близько 7 років, а від цинку — 10 років. Модель ШІ (GNN) передбачила ефективність ремедіації з похибкою <5%, оптимізувавши точкові посіви дронами в зонах із високим вмістом Pb, а дистанційний моніторинг у реальному часі дозволив враховувати сезонні коливання накопичення металів, підвищивши загальну ефективність до 55–65%. Інтеграція ШІ-дронів для зонування та поєднання фіторемедіаційних і мікробіологічних технологій збільшила точність картування до 87%, сконцентрувавши посадки на 29,4% середньоризикових ділянок та підвищивши IRE на 15%, одночасно скоротивши витрати на моніторинг на 40% порівняно з традиційними методами [15].

Висновки

1. Використання стратегій фіторемедіації через вирощування рослин на звалищах CFA відкриває значний потенціал для відновлення деградованих земель та генерації економічних доходів (Трипільська, Придніпровська, Зміївська, Бурштинська), здатним зменшувати вміст важких металів у ґрунті на 35–48% за один вегетаційний сезон.

2. створювання штучної стійкої екосистеми за допомогою толерантних рослин, які формуватимуть надійний рослинний покрив забезпечує обмежувальні несприятливі умови для розвитку рослин можна усунути, комбінуючи бобові культури з іншими видами, оскільки вони оптимізують властивості CFA для кращого росту.

4. Математична модель фіторемедіації дозволяє прогнозувати динаміку очищення ґрунту, оптимізувати вибір рослин та оцінювати ефективність технології без тривалих польових експериментів. Комбінування фіторемедіаційних та мікробіологічних методів підвищує ефективність відновлення до 60%, підтверджуючи важливість комплексного підходу.

5. Альтернативні моделі ШІ (GNN, МО) та роботизована техніка (дрони, автономні роботи) оптимізують процедуру посіву, моніторингу та прогнозування, підвищуючи точність на 20–30% та зменшуючи ризики в техногенних зонах. В підготовчо-розміточних роботах (картографування з HRAI та зонування МО) дозволяє точно розмежовувати ділянки для посадки, скорочуючи підготовчий етап на 30–40% та підвищуючи загальну ефективність фітогіперакумуляції.

6. Запропонована послідовність (геотехнічна стабілізація → дистанційний посів дронами → ШІ-моніторинг → біоакумуляція → оцінка IRE) забезпечує стале відновлення з акцентом на золошлакові відвали ТЕС.

1. Desai, S., et al. (2019). Phytoremediation of heavy metals: Mechanisms and approaches. *Environmental Science and Pollution Research, 26*, 12345–12360.
2. Pandey, V., et al. (2016). Advances in phytoremediation technologies for metal-contaminated soils. *Journal of Environmental Management, 182*, 89–101.
3. Hrynkiewicz, K., et al. (2018). Role of plant-microbe interactions in soil remediation. *Applied Microbiology and Biotechnology, 102*, 399–412.
4. Chetverik, M.S., & Bubnova, E.A. (2010). Formation of the technogenic geological environment and its relationship with natural. *Visnyk Kryvorizkogo tekhnichnogo universytetu, (25)*, 83–87.
5. Ворон, Е.А. (2010). Свойства создаваемой почвы при послойной горнотехнической и биологической рекультивации. *Науковий вісник НГУ, (5)*, 23–28.
6. Ковров, О., Федотов, В., & Зворигін, К. (2019). Обґрунтування технології фіторемедіації деградованих ландшафтів на основі екосистемного підходу. *Технологічний аудит і резерви виробництва, 6*(3(50)), 4–9. https://doi.org/10.15587/2312-8372.2019.185204.
7. Неджімі, Б. (2021). Фіторемедіація: стійка екологічна технологія для деконтамінації важких металів. *SN Appl. Sci., 3*, 286. https://doi.org/10.1007/s42452-021-04301-4.
8. Kumari, B., & Singh, S.N. (2011). Phytoremediation of metals from fly ash through bacterial augmentation. *Ecotoxicology, 20*, 166–176. https://doi.org/10.1007/s10646-010-0568-y.
9. Li, J., et al. (2025). Unified artificial intelligence framework for modeling pollution dispersion in phytoremediation sites. *Scientific Reports, 15*, 20083.
10. Babii K., Voron O., Chetveryk M., Lewicka E., Naworyta W. Phytoremediation of technogenic objects: Procedure, sequence, mathematical model // Geo-Technical Mechanics. – 2025. – № 172. – С. 172. – DOI: https://doi.org/10.15407/geotm2025.172.172.
11. Banda, M. F., et al. (2024). A phytoremediation approach for the restoration of coal fly ash polluted sites: A review. *Heliyon, 10*(23), e40741. https://doi.org/10.1016/j.heliyon.2024.e40741.
12. Flash Forest. (2025). Drone swarms for precision soil remediation in contaminated areas. *Sustainability Directory Reports*.
13. ABJ Academy Global. (2025). Top 6 Drones with Multispectral Cameras for Agriculture and Environmental Monitoring. Retrieved November 17, 2025, from https://abjacademy.global/drone-blog/top-6-drones-with-multispectral-cameras-for-agriculture-and-environmental-monitoring/.
14. Unmanned Systems Technology. (2025). New Multispectral Drone Camera Enhances Precision Agriculture & Smart Farming. Retrieved November 17, 2025, from https://www.unmannedsystemstechnology.com/2025/08/new-multispectral-drone-camera-enhances-precision-agriculture-smart-farming/.
15. Zhang, H., et al. (2023). Modeling phytoremediation of heavy metal contaminated soils using machine learning. *Journal of Hazardous Materials, 443*, 130168.

E.V. Stadnik

THE IMPACT OF ARTIFICIAL INTELLIGENCE ON SOCIETY AND USER PSYCHOLOGY

The rapid integration of Artificial Intelligence (AI) into the fabric of modern society represents one of the most significant technological shifts in human history. Its transformative power is reshaping economies, redefining communication, and revolutionizing education and healthcare. While AI promises unprecedented efficiency, innovation, and solutions to complex global challenges, its pervasive influence also introduces profound social, ethical, and psychological dilemmas [1]. The purpose of this study is to conduct a comprehensive analysis of the multifaceted impact of AI technologies on social institutions and the mental state of users. This paper will systematically explore both the beneficial opportunities and the potential adverse consequences, aiming to provide a balanced perspective on how AI is reconfiguring the human experience. The urgency of this research is underscored by the need to develop proactive strategies that can harness AI's potential while mitigating its risks, ensuring a future where technology enhances, rather than undermines, societal well-being and individual psychological health.

Social Impact and Ethical Challenges

Economic Transformation and the Labor Market. The economic impact of AI is dual-edged, offering both remarkable opportunities for growth and serious challenges to social stability. On one hand, AI-driven automation significantly enhances productivity across various sectors, from manufacturing and logistics to personalized medicine and education. Intelligent systems can optimize supply chains, accelerate drug discovery, and create adaptive learning platforms tailored to individual students. This leads to cost reduction, the creation of novel products and services, and the emergence of entirely new industries and high-skilled job categories, such as AI ethicists, data curators, and machine learning engineers [2].

On the other hand, this progress comes at the cost of *structural unemployment*. Unlike previous industrial revolutions that primarily automated manual labor, AI is capable of automating cognitive and analytical tasks. Professions in data analysis, accounting, translation, and even elements of creative work like graphic design and content writing are increasingly susceptible to automation. This creates a critical paradox: while high-tech jobs remain unfilled due to a skills gap, millions of workers in traditional sectors face obsolescence. This dynamic threatens to *deepen socio-economic inequality*, creating a "digital divide" between a technologically affluent elite and a large segment of the population left behind, potentially leading to social unrest and increased polarization [3].

Algorithmic Bias and Systemic Discrimination. One of the most insidious social threats posed by AI is the perpetuation and amplification of *algorithmic bias*. AI systems are not inherently objective; they learn patterns from historical data. If this training data reflects existing societal prejudices related to race,

gender, age, or socioeconomic status, the AI will inevitably inherit and often exacerbate these biases. This can lead to systemic discrimination in critical areas of life. For instance, in hiring, AI-powered resume screening tools may systematically disadvantage applicants from certain demographics. In finance, credit-scoring algorithms might assign lower ratings to residents of specific neighborhoods, a practice known as "redlining." Within the criminal justice system, predictive policing models have shown racial biases, leading to over-policing in minority communities [4]. Consequently, a technology touted for its objectivity can become a powerful engine for institutionalizing injustice, making discrimination more efficient and harder to detect.

The Erosion of Privacy and the Rise of Surveillance Capitalism. The business models of many leading tech companies are predicated on the mass collection and analysis of user data to train AI systems. This has given rise to what is termed "surveillance capitalism," where human experience is treated as a free raw material to be translated into behavioral data for prediction and monetization [5]. AI algorithms power the targeted advertising, content recommendation, and user engagement strategies that underpin this model. The psychological impact is a pervasive sense of being watched, leading to self-censorship and a chilling effect on free expression. The erosion of privacy is not merely a personal inconvenience but a fundamental threat to individual autonomy and democratic norms, as it concentrates unprecedented knowledge and power in the hands of a few corporations. (Figure 1).

Figure 1 – Visualization of algorithmic collection and analysis of user behavioral data [11]

Psychological Aspects of Human-AI Interaction

Cognitive and Behavioral Changes. The daily interaction with AI systems is fundamentally altering human cognition and behavior. While AI assistants and chatbots provide instant access to information and support, reducing cognitive load and stress in certain situations, they also foster significant negative trends.

1. **Erosion of Critical Thinking and Cognitive Offloading:** the convenience of algorithmic recommendations in search engines, social media feeds, and navigation apps encourages a state of *informational passivity*. Users increasingly delegate the tasks of information retrieval, verification, and synthesis to AI. This leads to the "cognitive offloading" effect, where the brain, to conserve

energy, avoids engaging in deep, analytical thought. The result is an atrophy of critical thinking skills, making individuals more susceptible to misinformation and less capable of independent problem-solving [6].

2. **Social Isolation and the Illusion of Companionship:** AI-powered virtual agents and social robots are designed to simulate empathy and conversation. For some users, particularly the lonely or socially anxious, these interactions can provide a semblance of connection. However, this is a dangerous substitute. These relationships are inherently one-sided and lack the reciprocity, shared vulnerability, and depth of human connection. Over-reliance on AI for social interaction can therefore *exacerbate feelings of loneliness and social isolation* in the long term, as it displaces the motivation and opportunity to cultivate genuine human bonds [7].

3. **Impact on Self-Perception and Mental Health:** the AI algorithms that curate social media feeds and other content platforms are optimized for "engagement," often prioritizing sensational, polarizing, or idealized content. This creates a "filter bubble" or a "distorted reality" where users are constantly exposed to curated highlights of others' lives and algorithmically amplified trends. This environment fosters constant *social comparison*, which is strongly linked to decreased self-esteem, body image issues, and increased levels of anxiety and depression, especially among adolescents and young adults whose identities are still forming [8]. (Figure 2)

Figure 2 – Visualization of the “filter bubble” phenomenon
and information overload in social networks [12]

Anthropomorphism and Emotional Dependence. A key psychological phenomenon in human-AI interaction is *anthropomorphism*—the human tendency to attribute human-like qualities, emotions, and intentions to non-human entities. AI systems are deliberately designed to trigger this response through the use of natural language, personalized greetings, and empathetic-sounding responses. This process can be described by a formula (1) that reflects the dependency of trust levels on personalization and perceived social presence:

$$`T = k * P * S` \quad (1)$$

where `T` is the level of trust in the AI; `P` is the degree of personalization in the interaction; `S` is the perceived social presence; and `k` is a coefficient

accounting for individual psychological characteristics of the user (e.g., predisposition to loneliness, level of social anxiety).[9]

The more a system mimics human behavior, the higher the emotional dependence it can foster. This raises ethical concerns about the potential for manipulative design and the formation of dysfunctional "one-sided relationships" that can be particularly harmful to vulnerable populations [9].

The Crisis of Agency and Identity. In a world where algorithms increasingly predict our preferences, make decisions for us, and shape our informational environment, a fundamental question arises about the *erosion of human agency*—the capacity for individuals to act independently and make their own free choices. When AI dictates what news we see, what music we listen to, and who we connect with, it subtly shapes our preferences, beliefs, and values. This is especially dangerous for adolescents, whose identity formation is a core developmental task. The constant behavioral modification by algorithmic recommendations can impede the development of authentic tastes, critical beliefs, and a stable sense of self, leading to a state of "algorithmic identity" where personal preferences are outsourced to machines [10].

Conclusions. In conclusion, the integration of Artificial Intelligence into society is characterized by a profound duality. The undeniable benefits of increased productivity, scientific advancement, and personalized services are counterbalanced by serious threats to social cohesion, economic equality, democratic integrity, and individual psychological well-being. The risks of structural unemployment, algorithmic bias, mass surveillance, cognitive degradation, and emotional manipulation are not hypothetical; they are already manifesting.

To navigate this complex landscape and minimize these risks, a multi-faceted and proactive approach is essential. This must include:

- ✓ The urgent development of a robust legal and regulatory framework that mandates algorithmic transparency, accountability, and audits to detect and mitigate bias [4].
- ✓ Strong data protection laws that curb the excesses of surveillance capitalism and restore a measure of individual privacy and autonomy [5].
- ✓ A massive investment in education and digital literacy, focusing not just on technical skills but, crucially, on fostering critical thinking, media literacy, and ethical reasoning from an early age [6].
- ✓ Further interdisciplinary research must be prioritized to deepen our understanding of the long-term psychological effects of human-AI interaction, to develop ethical design principles for AI, and to create frameworks for human-centered AI that augments rather than replaces human capabilities [7].

The goal is not to halt technological progress, but to steer it deliberately towards a future that is equitable, psychologically healthy, and fundamentally human.

1. Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.
2. World Economic Forum. (2023). Future of Jobs Report 2023.
3. Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation?
4. O'Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Publishing Group.
5. Zuboff, S. (2019). The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. PublicAffairs.
6. Carr, N. (2014). The Glass Cage: Automation and Us. W. W. Norton & Company.
7. Turkle, S. (2017). Alone Together: Why We Expect More from Technology and Less from Each Other. Basic Books.
8. Twenge, J. M. (2017). iGen: Why Today's Super-Connected Kids Are Growing Up Less Rebellious, More Tolerant, Less Happy. Atria Books.
9. Nass, C., & Yen, C. (2010). The Man Who Lied to His Laptop: What Machines Teach Us About Human Relationships. Current.
10. Pariser, E. (2011). The Filter Bubble: What the Internet Is Hiding from You. Penguin Press.
11. Unsplash. Algorithmic bias visualization [Електронний ресурс]. URL: https://images.unsplash.com/photo-1560472354-b33ff0c44a43?ixlib=rb-4.0.3&ixid=M3wxMjA3fDB8MHxwaG90by1wYWdlfHx8fGVufDB8fHx8fA%3D%3D&auto=format&fit=crop&w=800&q=80/
12. Stockcake. Social media generation [Електронний ресурс]. URL: https://stockcake.com/i/social-media-generation_295430_60132/

V. Lylo

THE IMPACT OF AI USE ON SOCIETY AND USER PSYCHOLOGY

Artificial intelligence as a scientific field emerged in the mid-20th century, and in recent decades it has undergone rapid development. Research has focused on creating automated intelligent systems, forming conceptual models, and introducing unique methods and approaches. Many ideas of artificial intelligence have already been implemented in technologies that have become part of everyday life [3, p. 59].

Artificial intelligence can be defined as a system or service toolkit capable of collecting, processing, and adapting user data (or publicly available datasets) and generating new outputs in response to user queries. It is one of the most dynamic and influential technologies of the modern era, transforming various spheres of human activity—from healthcare and education to business, logistics, and security. The rapid development of AI creates numerous opportunities for innovation and efficiency but also generates complex ethical and legal challenges. Modern society must address issues related to privacy, security, algorithmic transparency, and risks of discrimination. Although AI is now an integral part of life, it requires responsible regulation to prevent misuse and ensure safe integration.

As part of a study on AI implementation, a survey was conducted among respondents of different age groups to examine perceived ethical risks. According to the results, the majority (51.2%) expressed concern about the potential replacement of human labor by AI, particularly in logistics and customer service. Meanwhile, 30% believe that AI can support employees by automating routine tasks, and 24% expect that professions such as analysts, marketers, and lawyers may become partially automated.

Regarding the main ethical risks of AI in daily life, 30.4% indicated threats to privacy and unauthorized data collection. Another 32% emphasized the lack of algorithmic transparency and the need for oversight of AI-based decisions. Concerns about discrimination and algorithmic bias were shared by 17.6%, while 20% were worried about technological dependence and the decline of critical thinking.

Opinions on responsibility for addressing ethical aspects of AI were divided: 34.4% placed responsibility on developers, 23.2% on governmental institutions, and 14.4% on international organizations. At the same time, 28% argued for shared responsibility among all the above groups.

A total of 28.2% of respondents stated that AI development can benefit society if proper control systems are implemented, while 29% considered AI a potential threat. Concerning consequences for misuse of AI, 25.8% supported criminal liability, 26.6% preferred fines and sanctions, 12.9% advocated for public disclosure of incidents, and 34.7% supported complete prohibition of AI use in cases of abuse.

Positive prospects of AI development were also highlighted: 33.9% noted improvements in medical care, particularly diagnostics and treatment; 27.4% pointed to automation of routine tasks and expanded creative opportunities; and 23.4% emphasized the reduction of human error in critical fields [2, pp. 19–21].

Artificial intelligence offers considerable benefits, such as personalized learning, mental-health support, and enhanced communication. However, it also brings psychological and social challenges, including digital fatigue, loneliness, technostress, and reduced interpersonal interaction. Excessive reliance on AI may negatively affect social skills and emotional intelligence, contributing to isolation and increased anxiety. Furthermore, the growing use of AI intensifies concerns over data privacy, job displacement, and long-term socio-emotional well-being [1, pp. 1–5].

AI technologies have also begun to permeate the creative domain. Neural networks capable of processing massive datasets and generating coherent, sophisticated responses raise questions about the human role in an era increasingly influenced by artificial reasoning. Systems such as ChatGPT can produce knowledge that in some cases surpasses human performance, prompting debates about authorship, originality, and creative identity.

However, there is also a critical dimension to AI expansion. Some applications are not directed toward constructive purposes. In the military sphere, for example, AI-driven systems pose risks of catastrophic consequences. Although AI enhances human capabilities, this enhancement can simultaneously become a source of self-destruction, challenging the belief that society is progressing toward greater rationality. Ray Kurzweil predicts that the era of technological singularity may arrive imperceptibly - a stage at which technological development accelerates so rapidly that distinguishing human from machine control becomes nearly impossible. Kozlovets (2024) similarly reflects on the growing influence of AI on human existence and the increasing difficulty of maintaining autonomy.

Artificial intelligence fundamentally transforms modern society and user perception. Digitalization and automation reshape professional activity, reducing the role of traditional labor and altering approaches to self-realization. Unequal access to technologies intensifies digital inequality and complicates adaptation for certain social groups. For users, AI affects thinking, emotional states, and behavior: information bubbles emerge, critical reflection decreases, and technological dependence grows. Nevertheless, future generations raised in technology-saturated environments may be more psychologically prepared for these challenges.

AI has become a defining force of social transformation. While it enhances efficiency, expands opportunities, and supports human activity across various domains, it simultaneously produces new ethical, psychological, and socio-economic risks. It influences how people perceive information, communicate with one another, and construct their identities within an increasingly digitized world. The formation of information bubbles, the decline of critical thinking, and the growing dependence on algorithmic systems highlight the need for conscious and responsible AI integration.

To ensure that AI becomes a tool of progress rather than a source of societal fragmentation, balanced regulation, algorithmic transparency, and the development of digital literacy are essential. Equally important is the preservation of human autonomy, emotional resilience, and psychological well-being. The future of AI will depend on whether society can align technological advancement with ethical responsibility, ensuring that innovation strengthens rather than undermines human values and social cohesion.

1. Klimova, B., & Pikhart, M. (2025). *Research on the impact of artificial intelligence on student and academic well-being in higher education: A mini review*. Frontiers in Psychology, 16. https://doi.org/10.3389/fpsyg.2025.1498132.
2. Nechytailo, L. Ya., Bloshchytsiak, I. N., & Vorobiova, S. S. (2024). *Ethical challenges and regulation of artificial intelligence in the future: Perspectives and threats.* In *Science, education, technology and society: problems and prospects* (pp. 19–21). Tampere, Finland: Scholarly Publisher ICSSH.
3. Pchelianskyi, D., & Voinova, S. (2019). *Artificial intelligence: Development perspectives and trends.* Automation of Technological and Business Processes, 11(3), 59–64. https://doi.org/10.15673/atbp.v11i3.1500.

Н.В. Попик, Д.З. Магеррамов

ПРАКТИЧНІ АСПЕКТИ ВПРОВАДЖЕННЯ ШТУЧНОГО ІНТЕЛЕКТУ В СИСТЕМУ СТРАТЕГІЧНОГО УПРАВЛІННЯ ПІДПРИЄМСТВОМ

Сучасне бізнес-середовище характеризується високою динамічністю, періодичною економічною нестабільністю, посиленням конкурентного тиску, глобалізаційними викликами та технологічними зрушеннями. У таких умовах значно зростає рівень стратегічних ризиків, що ускладнюють процеси прогнозування, планування та прийняття управлінських рішень. Традиційні аналітичні методи часто виявляються недостатньо ефективними через повільність, обмеженість у врахуванні великої кількості факторів і суб'єктивність оцінок менеджерів.

На цьому тлі особливого значення набувають інструменти штучного інтелекту (ШІ), які здатні працювати з великими масивами даних, виявляти складні закономірності, забезпечувати автоматизоване прогнозування та моделювання. У цьому контексті особливого значення набуває дослідження ролі та потенціалу ШІ у трансформації стратегічних підходів, а також визначення практичних шляхів його впровадження в діяльність сучасних компаній.

Стратегічне управління охоплює формування місії, бачення, цілей підприємства, аналіз зовнішнього та внутрішнього середовища, вибір стратегічних альтернатив і розробку напрямів розвитку. В умовах нестабільності цей процес ускладнюється через низку чинників: швидкі зміни ринкової кон'юнктури, зростання варіативності поведінки споживачів, високий рівень технологічної мінливості, ризики глобальних криз та форс-мажорів.

Ризики, що впливають на стратегічні рішення, класифікують на економічні, фінансові, виробничі, інноваційні, соціальні та екологічні. Їх комплексність та взаємозалежність потребують застосування цифрових інструментів, здатних забезпечити оперативну обробку й аналіз даних. Обмеження традиційних методів (SWOT, PESTEL, матриця Ансоффа, експертні оцінки) проявляються у високій суб'єктивності, неможливості моделювання складних сценаріїв і недостатній швидкості оновлення інформації. Саме тому інтеграція ШІ стає важливою складовою сучасного стратегічного інструментарію.

Використання ШІ у стратегічному менеджменті охоплює широкий спектр технологій, таких як машинне навчання (ML), глибинні нейронні мережі (Deep Learning), інтелектуальний аналіз даних (Data Mining), системи підтримки прийняття рішень (DSS), цифрові двійники (Digital Twins), обробка природної мови (NLP).

Використання цих інструментів дозволяє формувати якісно нову систему стратегічного аналізу, засновану на об'єктивних даних (табл.1).

Таблиця 1 – Інструменти ШІ та їхнє призначення у стратегічному управлінні підприємством

Інструмент ШІ	Основні можливості та застосування інструменту ШІ
Машинне навчання (ML)	Створення моделей прогнозування попиту, цінових коливань, поведінки клієнтів, змін у постачальних ланцюгах, макроекономічних трендів.
Глибинні нейронні мережі (Deep Learning)	Аналіз неструктурованих даних: текстів, зображень, аудіо та сигналів; виявлення ринкових трендів, аналіз споживчих відгуків і контенту соціальних мереж.
Інтелектуальний аналіз даних (Data Mining)	Виявлення прихованих закономірностей, взаємозв'язків і трендів, які неможливо визначити традиційними методами аналізу.
Системи підтримки прийняття рішень (DSS) з компонентами ШІ	Формування оптимальних стратегічних альтернатив на основі багатофакторного аналізу, моделювання та оцінювання варіантів рішень.
Цифрові двійники (Digital Twins)	Симуляція стратегічних сценаріїв розвитку, оцінка потенційних наслідків управлінських рішень без ризику для реальних активів.
Обробка природної мови (NLP)	Аналіз новин, документів, ринкових тенденцій, соціальних медіа; відслідковування інформаційних сигналів, що впливають на стратегічні рішення.

Інтеграція інтелектуальних технологій суттєво розширює можливості стратегічного управління підприємством. Використання ШІ підвищує точність прогнозування завдяки аналізу великої кількості змінних і значних масивів даних, що створює передумови для більш обґрунтованого планування. Це також забезпечує динамічність стратегічного управлінського процесу, дозволяючи оперативно реагувати на швидкі зміни зовнішнього середовища. Інтелектуальні системи сприяють ранньому виявленню ризиків і слабких сигналів, завдяки чому зменшується вплив потенційних загроз. Крім того, вони дають змогу формувати багатоваріантні сценарії розвитку, що підвищує стійкість підприємства та його здатність протистояти невизначеності. Застосування ШІ мінімізує суб'єктивність управлінських рішень і водночас забезпечує оптимізацію використання ресурсів, сприяючи зростанню продуктивності та зміцненню конкурентних позицій компанії.

Таким чином, ШІ стає потужним каталізатором підвищення ефективності стратегічного управління.

Процес упровадження технологій штучного інтелекту в діяльність підприємства вимагає комплексного та системного підходу. Насамперед необхідно забезпечити розбудову сучасної цифрової інфраструктури та ефективних механізмів збору даних, що є базою для роботи інтелектуальних систем. Важливо також формувати професійні команди аналітиків і фахівців із data science, які володіють компетенціями для розробки, налаштування та використання інструментів ШІ. Окремої уваги потребує розроблення політики безпеки даних, що охоплює питання кіберзахисту та регламентації доступу до інформації. Паралельно з цим підприємство має вибудовувати гнучкі бізнес-процеси, здатні адаптуватися до автоматизованого аналізу та нових цифрових рішень. Важливою умовою успішної інтеграції ШІ є інвестиції у цифрову трансформацію, які створюють технічні та організаційні передумови для використання інтелектуальних технологій. Крім того, необхідно враховувати етичні й правові аспекти, пов'язані з алгоритмічною відповідальністю, захистом персональних даних та автоматизацією прийняття рішень, що забезпечує безпечне та відповідальне застосування ШІ в управлінні.

Приклади успішного застосування ШІ вже спостерігаються у виробництві (передиктивне планування), логістиці (оптимізація маршрутів), маркетингу (персоналізація), фінансовій сфері (моделювання ризиків) і сервісних компаніях (автоматизація клієнтської взаємодії).

Впровадження ШІ у стратегічне управління дає підприємству низку вагомих переваг:

- ✓ зниження рівня стратегічної та операційної невизначеності;
- ✓ пришвидшення процесів аналізу та прийняття рішень;
- ✓ підвищення конкурентоспроможності завдяки точнішому прогнозуванню;
- ✓ покращення організаційної гнучкості та стійкості;
- ✓ оптимізацію витрат і ресурсного забезпечення;
- ✓ підвищення якості стратегічного планування й контролю;
- ✓ формування довгострокових переваг за рахунок цифрової зрілості.

Таким чином, використання ШІ є не лише інструментом підвищення ефективності, а й важливим елементом стратегії успішного розвитку підприємства.

Штучний інтелект виступає ключовою технологією, що трансформує систему стратегічного управління підприємством. Він забезпечує можливість аналізувати складні та швидкозмінні процеси, прогнозувати ризики, адаптувати стратегії та приймати обґрунтовані рішення на основі даних. У сучасних умовах нестабільності впровадження ШІ стає необхідною передумовою стійкого розвитку підприємств, підвищення їх конкурентоспроможності та здатності швидко реагувати на зовнішні виклики.

Таким чином, інтеграція інструментів ШІ у стратегічне управління є стратегічною інвестицією, що формує нові можливості для зростання, інноваційності та довгострокової ефективності підприємств.

1. Мельник О. Г., Руда М. В. Стратегічні аспекти цифрової трансформації бізнесу. Менеджмент та підприємництво в Україні: етапи становлення та проблеми розвитку. 2024. № 2 (12). С. 196–209. DOI: 10.23939/smeu2024.02.196.
2. Вербівська Л. В., Дзюба Т. В. — „Цифрова трансформація підприємництва: стратегічні виклики та управлінські рішення“. Інвестиції: практика та досвід, № 12 (2025), С. 60–66. DOI: 10.32702/2306-6814.2025.12.60.
3. Вербицька Г. Л. Використання штучного інтелекту та машинного навчання при плануванні інноваційної діяльності підприємств з метою забезпечення їх економічної безпеки // Нові інформаційні технології управління бізнесом : зб. тез VII Всеукр. наук.-практ. конф. Київ : Спілка автоматизаторів бізнесу, 2024.
4. Храпач В.О. Різновиди штучного інтелекту та можливості і проблеми його використання при стратегічному плануванні в економіці. Економіка та суспільство.
   URL: https://economyandsociety.in.ua/index.php/journal/article/view/2715.

А.П. Завражна

ШТУЧНИЙ ІНТЕЛЕКТ ТА ЙОГО ВЗАЄМОЗВ'ЯЗОК З ПРАВАМИ ЛЮДИНИ

Штучний інтелект може принести значну користь суспільству, але також викликає занепокоєння щодо потенційного негативного впливу на права людини. Проявлений інтерес вчених до перспектив розвитку штучного інтелекту та його взаємозв'язку з правами людини обумовлений інтенсивним впровадженням у життєдіяльність суспільства та держави цифрових технологій.

Використання технологій штучного інтелекту стало повсюдним у сучасному суспільстві, воно полегшує повсякденне життя мільйонів людей у всьому світі – інновації, що відбуваються в медицині, освіті, економіці, транспорті тощо завдяки використанню штучного інтелекту, вражають і замінюють життя.

Технології штучного інтелекту можуть допомогти у прийнятті обґрунтованих та ефективних рішень, а також знизити витрати на виробництво. Однак в той же час, це може також призвести до необґрунтованих, дискримінаційних або расистських висновків, або до рішення, помилково прийнятого за точне, оскільки вони, як правило, автоматичні та обґрунтовуються на кількісних показниках.

У зв'язку з цим, зростаюче використання технологій штучного інтелекту порушує важливі питання щодо конфіденційності, дискримінації та відповідальності у разі шкоди, заподіяної такими інтелектуальними системами, які так чи інакше контролюють життя мільйонів людей у всьому світі.

Перетин штучного інтелекту та прав людини є однією з найгостріших правових та етичних проблем нашого часу. Для того, щоб повною мірою оцінити складність цієї проблеми, необхідно спочатку чітко зрозуміти основні принципи прав людини, закріплені в міжнародному праві.

У Сполучених Штатах Америки, Європейському Союзі та інших країнах бентежить занепокоєння з приводу використання урядів і великі технологічні компанії технологій штучного інтелекту. Для того, щоб вторгатися в життя громадян і збирати інформацію про їхню поведінку, що є явним порушенням прав людини.

Безперечно, застосування та технології штучного інтелекту надають як позитивний, так і негативний вплив на життя людей. Полегшуючи повсякденні завдання, і в той же час вони створюють загрозу їхній конфіденційності та безпеці. Особливо коли великі технологічні компанії зловживають технологією штучного інтелекту для збирання даних про своїх клієнтів. В інших випадках та системи штучного інтелекту продемонстрували расистську або дискримінаційну поведінку, що викликає все більшу занепокоєність в нашому суспільстві.

Штучний інтелект може бути корисним джерелом інформації, оскільки він спрощує процес доступу до інформації для людей. Однак на додаток до корисного аспекту використання технології штучного інтелекту існує і незаперечний негативний аспект, який полягає у негативному впливі штучного інтелекту на повсякденне життя людей. Особливо це проявляється у сферах з недоторканністю приватного життя, дискримінацією та створює загрозу їх основним правам у деяких випадках.

Дискримінація та упередженість є одними з найбільших проблем, пов'язаних із штучним інтелектом, що може негативно позначитися на правах людини, особливо на праві на дискримінацію.

У сфері конфіденційності стає більш очевидною суперечність між перевагами технології штучного інтелекту та ризиками для прав людини. Причина в тому, що штучний інтелект може мати серйозні наслідки для конфіденційності, особливо якщо він використовується для збирання та обробки особистих даних.

Для вирішення проблем, пов'язаних із штучним інтелектом та конфіденційністю даних, деякі країни почали приймати закони про захист, що регулюють використання персональних даних системами штучного інтелекту. Наприклад, Європейський Союз, ухваливши свою стратегію штучного інтелекту, зробив кілька значних кроків до організації експертної групи високого рівня, яка опублікувала свій перший план дій у лютому 2019 року, спрямований на просування надійного штучного інтелекту за допомогою етичних засад, політики та інвестиційні рекомендації [1].

Виходячи з того, що «всі переваги інформаційних технологій повинні реалізовуватися, тільки якщо вони відповідатимуть певним цінностям та етичним принципам» [2], автори описують конкретні дії держав у цьому напрямку та зазначають, що у багатьох країнах здійснюється діяльність із розробки національних стратегій у сфері штучного інтелекту. Наприклад, у Канаді діє Торонтська декларація про захист прав на рівність та дискримінацію у системах машинного навчання 2018 р. Вона закликає уряд дотримуватись керівних принципів та міжнародних стандартів прав людини, приватні компанії – враховувати ризики неминучої появи системних помилок через неповні дані з машинного навчання, вчених – консолідувати зусилля з вивчення негативних наслідків, пов'язаних з інформаційними технологіями. Дослідники звертають увагу на розроблений Лабораторією цифрової інтеграції Канади з глобальних питань у 2018 році проєкт стратегії «Штучний інтелект: наслідки для прав людини та зовнішньої політики». В даному документі звертається увага на вплив штучного інтелекту на такі права, як право на рівність, недоторканність приватного життя, вільне вираження думки та інше, а також способи, якими ці дії можна попередити або усунути.

У статті оцінюються наслідки використання штучного інтелекту у життєдіяльності країни та його впливу на права людини. Зокрема, автори акцентують увагу на використанні штучного інтелекту у сфері кримінального правосуддя, фінансів, охорони здоров'я, інформаційних та трудових відносин та освіти.

Усунення несприятливого впливу штучного інтелекту на права людини автори бачать у реалізації принципу належної обачності. Належна обачність є важливим першим кроком до визначення, пом'якшення та усунення несприятливих наслідків впливу на права людини. Тому як мінімум публічні політичні зусилля мають бути спрямовані на забезпечення того, щоб усі, хто бере участь у побудові цих систем, виявляли належну обачність, яка забезпечить повагу до прав людини.

Слід зазначити ті небезпеки, які підстерігають людину і сучасне суспільство у сфері розвитку цифровізації. Застосування заснованих на штучному інтелекті рейтингів повністю відкидає принцип рівності громадян, повертає суспільство в кастове минуле та призводить до управління на основі страху та насильства держави над безправною оцифрованою особистістю.

На сьогоднішній день використовуються кілька основних підходів до моделювання соціальних процесів з використанням штучного інтелекту. Одним із найпоширеніших методів є агентно-орієнтоване моделювання, яке дозволяє моделювати взаємодію окремих агентів у соціальних системах. Також використовуються методи машинного навчання, такі як нейронні мережі, які дозволяють навчати моделі на великих масивах даних [3].

Метою нової цифрової Стратегії Європейського Союзу є досягнення Європейським Союзом мирного лідерства у сфері створення штучного інтелекту. Ця Стратегія стала першим юридичним документом, який прописує принцип наявності контролю людини над штучним інтелектом у питаннях регулювання цих розробок.

Впровадження штучного інтелекту в різні сфери життєдіяльності породжує не лише низку актуальних питань, але може викликати конфлікт низки універсальних прав людини. Наприклад, зіткнення права на інформацію та права на недоторканність приватного життя в інформаційному суспільстві [4]. Межі між зазначеними правами були встановлені законом, проте реалізація цифрових технологій змінила обставини та породила конфлікт норм та принципів. На прикладі закріплених у Конституції Іспанії в ст. 18.1 права на недоторканність приватного життя та у ст. 20.1 права на свободу інформації автори аналізують сучасну практику та проблеми їх реалізації.

Впровадження технологій штучного інтелекту відрізняється від технологічних проривів минулого століття швидкістю змін, що відбуваються. Це означає, що працівникам доводиться постійно адаптуватися до нових технологічних умов і створювати нові навички, а іноді й професії в межах одного покоління. Однак попередні індустріальні революції торкалися хоча б двох поколінь працівників.

Оскільки представники бізнесу в першу чергу зацікавлені в прибутку та ефективності виробництва, вони готові замінити працівників робототехнікою. Коли це стає вигідно, тобто коли витрати на робототехніку стають значно нижчими від витрат на працівників-людей. З іншого боку роботодавці

наголошують на нестачі робочих спеціальностей. Серед представників молодих поколінь ці спеціальності не менш популярні, а отже з виходом на пенсію нинішніх робітників їх не буде кому замінити. Це робить використання робототехніки ще більш актуальним.

Таким чином, штучний інтелект може як змінювати, так і порушувати права людини, і для того, щоб зробити штучний інтелект кориснішим, не створюючи при цьому загрози правам людини, повинен бути набір етичних та юридичних принципів, що охоплюють теми, пов'язані з прозорістю, конфіденційністю, безпекою та дискримінацією.

Нові закордонні та українські ініціативи та рішення підтверджують важливість дотримання при використанні штучного інтелекту принципів етики.

1. E. K. Mpinga, N. K. Z. Bukonda, S. Qailouli, P. Chastonay, Dove press (February 2022). *Artificial Intelligence and Human Rights: Are There Signs of an Emerging Discipline?* A Systematic Review. https://www.dovepress.com/artificial-intelligence-and-human-rights-are-there-signs-of-an-emerging-peer-reviewed-fulltext-article-JMDH.
2. Raso F., Hilligoss H., Krishnamurthy V., Bavitz C., Kim L. (2018). *Artificial Intelligence & Human Rights: Opportunities & Risks*. Berkman Klein Center for Internet & Society research publication. P. 3-62. Mode of access: https://dash.harvard.edu/bitstream/handle/1/38021439/2018-09_AIHumanRights.pdf?sequence=1&isAllowed=y (дата звернення: 10.09.2020).
3. Симонов Д.І., Симонов Є.Д. (2024). Моделювання динаміки соціальних процесів за допомогою штучного інтелекту. Збірник тез доповідей XXIV міжнародної науково-технічної конференції «Штучний інтелект та інтелектуальні системи – AIIS'2024». Київ, Інститут проблем штучного інтелекту, 18–19.10.2024. 274 с. С. 162-165.
4. Torres F.T., Berdun M.I. (2019). El dret a la informacio versus la proteccio del dret a la intimitat en la societat de la informacio. La nova LO de proteccio de dades i el dret a l`oblit // Journal of research and analysis [online]. Vol. 36, N 1. P. 117-131. Mode of access: https://www.raco.cat/index.php/Comunicacio/article/view/103391 (дата звернення: 10.09.2020).

Д.П. Луцюк

ВПЛИВ ВИКОРИСТАННЯ ШІ НА СУСПІЛЬСТВО, ПСИХОЛОГІЯ КОРИСТУВАЧІВ

Сьогодні ми майже щодня стикаємося зі штучним інтелектом у різних сферах життя. Він присутній у медицині, навчанні, бізнесі, транспорті та навіть у розважальних сервісах. ШІ дозволяє швидше виконувати рутинні завдання, легше знаходити потрібну інформацію і загалом підвищує ефективність роботи. Проте разом із цими зручностями виникають і певні проблеми — соціальні, моральні та психологічні, які потребують особливої уваги та аналізу.

Соціально-економічний вплив ШІ проявляється в трансформації ринку праці та бізнес-моделей. Автоматизація рутинних процесів зменшує потребу в деяких професіях, водночас створюючи нові робочі місця у сферах, пов'язаних із розробкою, обслуговуванням та впровадженням технологій [1]. ШІ використовується для прогнозування поведінки клієнтів у комерції, персоналізації послуг і оптимізації маркетингових стратегій, що підвищує ефективність підприємств. Крім того, його застосування у фінансовому секторі дозволяє швидко аналізувати великі масиви даних, підвищуючи точність прийняття рішень, але водночас створює нові ризики щодо безпеки даних і стабільності систем.

Вплив ШІ на суспільство має й соціально-психологічний аспект. Системи рекомендацій у соціальних мережах формують своєрідні «інформаційні бульбашки», що обмежують контакт користувачів із різними точками зору та можуть підсилювати соціальну поляризацію. Крім того, постійна взаємодія з алгоритмами стимулює пасивне споживання контенту та знижує активне критичне мислення. Соціальна довіра до ШІ формується на основі очікувань і страхів: з одного боку, люди вірять у його потенціал вирішувати складні проблеми, з іншого — побоюються втрати контролю або упередженості алгоритмів [2].

Психологічний вплив ШІ на користувачів є не менш значним. Люди часто приписують системам «людські» риси, очікуючи емоційної підтримки від чат-ботів або голосових помічників. Це явище, відоме як антропоморфізація, формує довіру, але водночас може викликати ілюзію, що машина «розуміє» чи «відчуває», хоча насправді вона лише обробляє дані за алгоритмами [3]. Надмірна залежність від віртуальних асистентів здатна знижувати самостійність прийняття рішень, призводити до когнітивної лінощі та зменшувати активність критичного мислення. Крім того, надмірне використання ШІ для емоційної підтримки може замінювати живе спілкування, що підвищує ризик самотності та емоційної залежності, особливо серед підлітків.

З іншого боку, ШІ відкриває значні можливості, наприклад:

— В освіті він дозволяє персоналізувати процес навчання, адаптувати його під індивідуальні потреби учня та надавати доступ до великої кількості інформації

— У медицині ШІ використовується для діагностики, прогнозування захворювань і планування терапії

— У бізнесі та сфері послуг він оптимізує процеси, підвищує ефективність та полегшує комунікацію з клієнтами.

— У транспорті та логістиці технології ШІ допомагають планувати маршрути, контролювати безпеку та ефективніше керувати ресурсами.

— Крім того, ШІ може сприяти розвитку емоційного інтелекту, наприклад через системи, що «слухають» та реагують на емоційний стан користувачів.

Штучний інтелект значно полегшує життя людей з обмеженими можливостями. У таблиці 1 показано, як різні технології ШІ можуть підтримувати людей з вадами зору, слуху та з інклюзивними потребами, роблячи їхнє повсякденне життя більш доступним і зручним.

Таблиця 1

Категорія користувачів	Допомога від ШІ
Люди з вадами зору	Озвучує текст, описує навколишнє середовище, розпізнає об'єкти, читає дорожні знаки, допомагає орієнтуватися та виконувати повсякденні завдання, підвищує самостійність.
Люди з вадами слуху	Перетворює мову у текст у реальному часі, допомагає спілкуватися у навчанні, на роботі та в соціумі, перетворює аудіосигнали на візуальні підказки, полегшує сприйняття інформації.
Люди з інклюзивними потребами	Адаптує навчальний матеріал під індивідуальні потреби, підбирає темп і формат подачі інформації, пропонує інтерактивні вправи, забезпечує рівний доступ до освіти та розвиток навичок.

Джерело сформовано автором на основі [1].

Проте, разом із перевагами існують і серйозні ризики. Психологічні включають зниження критичного мислення, формування емоційної залежності, ізоляцію та заміну живих соціальних контактів взаємодією з машинами. Соціальні та політичні ризики пов'язані з можливістю маніпуляцій через алгоритми, регуляторними прогалинами та глобальною нерівністю у доступі до технологій. Технічні загрози включають непередбачувану поведінку систем, потенційні зловживання та питання відповідальності за рішення, ухвалені ШІ.

Хоча ШІ може створювати певні ризики, він також має багато позитивного впливу на наше повсякденне життя та психіку. Різні сервіси на базі ШІ здатні допомагати людям справлятися зі стресом, організовувати свій день і відстежувати власний прогрес у навчанні або роботі. Віртуальні помічники та мобільні додатки, які використовують штучний інтелект, можуть мотивувати, нагадувати про важливі справи і допомагати відчувати себе більш організованим і впевненим.

Майбутнє взаємодії людини і ШІ, на думку експертів, повинно базуватися на балансі: поєднання автоматизації та людського контролю. Важливим є розвиток етичних стандартів, «алгоритмічної грамотності» та правил використання ШІ, щоб його впровадження не загрожувало правам людини та соціальному благополуччю. При цьому ШІ має підтримувати, а не замінювати людську здатність до мислення, моральних рішень та емпатії.

Отже, вплив ШІ на суспільство та психологію користувачів є комплексним і багатогранним. Він відкриває нові можливості для освіти, медицини, бізнесу та соціальної взаємодії, але водночас створює ризики для психічного здоров'я, критичного мислення та соціальної згуртованості. Ключовим завданням є усвідомлення цих ризиків і відповідальна взаємодія з технологіями, яка дозволяє зберегти людські цінності та психічне благополуччя.

1. Касянчук А. В., Гоц В. В., Попович Н. Л., Хроленко В. М. *Вплив штучного інтелекту, комп'ютерного зору та машинного навчання на життя людини*. Управління розвитком складних систем. 2023.
   URL: https://urss.knuba.edu.ua/ua/zbirnyk-55/article-1716 (дата звернення: 21.11.2025).
2. Клозе Р. *Штучний інтелект: суспільства, розділені страхом і вірою*. Грані: науково-теоретичний альманах. 2024.
   URL: https://grani.org.ua/index.php/journal/article/view/2128 (дата звернення: 22.11.2025).
3. Федотенко Я., Шакуров Є. *Вплив штучного інтелекту на розвиток емоційного інтелекту*. Юний дослідник. 2025. № 3.
   URL: https://ojs.iod.gov.ua/index.php/ejd/article/view/64?utm_source (дата звернення: 21.11.2025).

Д.Р. Лопата

ШТУЧНИЙ ІНТЕЛЕКТ ТА ПРАВО НА ПРИВАТНІСТЬ: РИЗИКИ МАСОВОГО НАГЛЯДУ

Штучний інтелект сьогодні все активніше інтегрується в сфери безпеки, комунікації та управління, що має значний вплив на якість і масштаби контролю над громадянами. Технології розпізнавання облич, аналізу поведінки та автоматичної класифікації людей набувають поширення як у державних, так і в приватних структурах. Це створює новий формат цифрового середовища, в якому приватність стає обмеженим ресурсом [1]. Основна загроза полягає в тому, що штучний інтелект дозволяє здійснювати постійне спостереження в режимі реального часу. Таке спостереження дозволяє ідентифікувати людину навіть у великому натовпі. Дослідження українських фахівців показують, що масове відеоспостереження з використанням штучного інтелекту може порушувати основні стандарти прав людини [1].

Біометричні дані — зображення обличчя, відбитки пальців, особливості голосу — стають основним інструментом тотального контролю. На відміну від паролів або ідентифікаційних даних, біометричні дані не можна змінити в разі витоку. Тому масовий збір цих даних становить серйозну небезпеку. Системи штучного інтелекту не тільки розпізнають особу, але й можуть передбачати її дії. Такі передбачення ґрунтуються на аналізі моделей поведінки. Однак кожне передбачення містить ризик помилки. Помилки алгоритмів, як підтверджують українські правозахисники, неодноразово призводили до дискримінаційних результатів [2]. Якщо система навчається на упереджених або нерепрезентативних даних, вона відтворює ті самі упередження. Це може проявлятися у вигляді неправильних «співпадінь» під час ідентифікації осіб. Такі помилки зазвичай стосуються людей з темнішою шкірою, жінок або молоді. Іншим питанням є профілювання. Це процес, під час якого ШІ створює прогнозні профілі осіб. Профіль містить потенційні ризики, політичні уподобання, схильності та соціальні зв'язки. Все це може втручатися в приватне життя громадянина. Проблема полягає в тому, що людина навіть не знає, які дані про неї збираються. Більше того, невідомо, як і з якою метою ці дані використовуються. Українські юридичні журнали неодноразово наголошували на недостатній прозорості алгоритмічних рішень [3]. Масовий нагляд може здійснюватися таємно.

Іншою важливою проблемою є зберігання даних. Великі обсяги відео, фотографій та метаданих накопичуються на серверах. Вони можуть зберігатися роками. У разі хакерської атаки ці дані можуть стати відкритими для доступу. Дослідження з кібербезпеки, проведені в Україні, попереджають про недостатній захист таких систем [4]. Це означає, що навіть добросовісне використання ШІ не може гарантувати конфіденційність. Відсутність комплексного регулювання створює ризик довільного застосування

контролю. Крім того, немає чітких критеріїв для оцінки ризиків алгоритмів. Європейський Союз прийняв закон про штучний інтелект, що включає категорію «високого ризику». Україна також потребує подібної класифікації [2, 3]. Експерти наголошують, що ШІ може використовуватися лише в тих випадках, коли це є розумним і пропорційним. Масовий контроль у більшості випадків не відповідає цим вимогам. Крім того, люди мають право знати, як система приймає рішення і це вимагає прозорості алгоритмів. Важливе значення має також освіта, де громадяни повинні розуміти, як працюють цифрові технології. Це дозволить їм свідомо оцінювати ризики бо без цього навіть найкраще законодавство буде неефективним. ШІ може бути корисною в галузі громадської безпеки. Вона допомагає розслідувати злочини та зменшує навантаження на поліцію. Однак ризик зловживання є дуже високим. Україна повинна рухатися в напрямку європейських стандартів, що передбачає створення незалежних регуляторних органів. Також необхідно застосовувати оцінку впливу на права людини. Жодна система на основі ШІ не повинна запускатися без такої оцінки. Необхідно розробити єдині стандарти перевірки алгоритмів і вони повинні включати аудит, тестування та експертну оцінку. Дані повинні збиратися в мінімальному обсязі і тільки в разі необхідності. Зайві дані повинні бути видалені.

Отже, Україна повинна взяти до уваги ці практики і розуміти, що штучний інтелект є корисним інструментом, але може легко перетворитися на засіб масового спостереження. Тільки комплексне законодавче регулювання забезпечить безпечне використання технології. Необхідно запровадити механізми правового контролю, прозорості та підзвітності і тільки тоді штучний інтелект перестане бути загрозою і стане інструментом розвитку.

1. Уповноважений Верховної Ради України з прав людини. «Штучний інтелект: як захистити права людини у цифрову добу». https://ombudsman.gov.ua/uk/news_details/shtuchnij-intelekt-yak-zahistiti-prava-lyudini-u-cifrovu-dobu?utm_source.
2. Омбудсмен + EU4DigitalUA + Мінцифра. «Рекомендації з відповідального використання ШІ» https://ombudsman.gov.ua/news_details/eu4digitalua-ombudsman-office-and-mintsyfra-presents-artificial-intelligence-guidelines?utm_source.
3. Президент України. Указ № 119/2021 «Стратегічні цілі забезпечення права на приватність». https://www.president.gov.ua/documents/1192021-37537?utm_source.
4. Національна школа суддів України. «Штучний інтелект та право на приватність: баланс між інноваціями та захистом персональних даних» https://court.gov.ua/storage/portal/supreme/prezentacii_2025/119_AI_personal_data_protection_bernaziuk.pdf?utm_source.

С.Ю. Демянко, А.І. Вавіленкова

ШТУЧНИЙ ІНТЕЛЕКТ ЯК ІНСТРУМЕНТ ПІДВИЩЕННЯ ОПЕРАТИВНОЇ ОБІЗНАНОСТІ ДЕРЖАВНИХ СТРУКТУР: OSINT, ЦИФРОВІ ДВІЙНИКИ ТА МОДЕЛЮВАННЯ ЗАГРОЗ

Повномасштабне вторгнення російської федерації змусило Україну фактично перезібрати підходи до оцінювання ситуації в кіберпросторі та довкола об'єктів критичної інфраструктури. Систематичні удари по енергетичних об'єктах, поєднання кібератак із фізичними ураженнями та потужний інформаційний тиск показали, що класичні підходи до звітності й ручного аналізу не дають достатньої швидкості. Сучасна система безпеки має спиратися не лише на фіксацію подій, а й на оперативне розуміння контексту та завчасне бачення можливих наслідків.

За таких умов критичною стає взаємодія «OSINT – штучний інтелект – цифрові двійники». Відкриті джерела формують широку та динамічну картину подій, ШІ забезпечує опрацювання великих, різнотипних масивів даних, а цифрові двійники відтворюють поведінку конкретних об'єктів і середовищ у вигляді моделей. У сукупності це дає якісно інший рівень оперативної обізнаності: з'являється можливість бачити не лише вже здійснені дії противника, а й тренди, що до них ведуть.

Оперативна обізнаність і роль OSINT у державних системах кібербезпеки: можливості підсилення за допомогою штучного інтелекту

Оперативна обізнаність — це вміння швидко скласти цілісну картину подій на основі даних від технічних систем, операторів інфраструктури, органів влади та відкритого інформаційного простору. У воєнний час саме OSINT часто стає першим джерелом сигналів: фото та відео наслідків ударів, супутникові зображення, повідомлення про перебої в роботі систем, публічні заяви й реакція суспільства.

Водночас OSINT за своєю природою фрагментований: інформація з'являється нерівномірно, різними мовами, з різним рівнем довіри до джерел. Під час масштабних атак обсяги таких даних різко зростають, а маніпулятивний контент поширюється особливо швидко. Аналітичні підрозділи фізично не здатні опрацювати весь потік у часових рамках, що вимірюються хвилинами чи годинами. Тому відкриті джерела мають інтегруватися з технічними телеметричними даними як рівнозначний компонент картини ситуації. Ключовим завданням стає відфільтрувати інформаційний шум, пов'язати окремі сигнали між собою й співвіднести їх зі станом конкретних об'єктів, а це вже зона відповідальності інструментів ШІ.

Штучний інтелект як механізм підсилення OSINT

Штучний інтелект фактично розширює можливості аналітика, виступаючи «множником» його продуктивності. Він бере на себе рутинні

операції зі збирання, попереднього структурування та класифікації даних, а також допомагає виявляти приховані зв'язки у великому потоці повідомлень. Мовні моделі групують матеріали за тематиками, фіксують повторювані наративи й синхронні викиди інформації на різних майданчиках, що може вказувати на скоординовані інформаційні операції.

У роботі з візуальним контентом алгоритми комп'ютерного зору дають змогу оцінити характер пошкоджень, виконати геолокацію зображень і порівняти супутникові знімки «до» та «після» події. Це скорочує час первинного аналізу й дозволяє зосередити людські ресурси на об'єктах із найвищим рівнем ризику.

Алгоритми пошуку аномалій допомагають виявляти нетипові відхилення в роботі технологічних систем. Коли такі спостереження поєднуються з OSINT-даними, стає простіше ідентифікувати комбіновані операції, у яких кіберудари підсилюються інформаційним впливом. У результаті функція аналітика поступово зміщується від ручного перегляду потоків повідомлень до інтерпретації зведених продуктів, сформованих ШІ.

Цифрові двійники критичної інфраструктури

Цифровий двійник — це віртуальна модель реального об'єкта, яка постійно актуалізується за рахунок даних від датчиків, телеметрії, диспетчерських систем та, за потреби, відкритих джерел. Для енергетики та інших критичних галузей це означає створення єдиного інформаційного середовища, у якому аналізується поточний стан, прогнозується поведінка систем і оцінюються можливі наслідки атак.

Завдяки такому підходу можна швидко визначати, які елементи інфраструктури зазнали ураження, де існує ризик каскадних відмов і як зміниться робота мереж за умов повторних ударів. Штучний інтелект коригує параметри моделей, підраховує наслідки різних сценаріїв, тоді як OSINT дає додаткову інформацію безпосередньо з місць подій. Для України цифрові двійники стають базовим елементом нового режиму планування, відновлення та підвищення стійкості енергетичного сектору.

Моделювання загроз та сценарний аналіз

Поєднання технічної телеметрії, відкритих джерел і цифрових двійників відкриває шлях до системного сценарного аналізу й моделювання загроз. Алгоритми машинного навчання працюють із часовими рядами, статистикою інцидентів та графами взаємодії між елементами, виокремлюючи повторювані патерни та нетипові послідовності подій. Це дозволяє бачити не лише окремі епізоди, а й характер дій противника загалом.

Сценарні моделі застосовуються й для оцінювання ефектів управлінських рішень: перенаправлення резервів, зміни налаштувань мережі, зміни пріоритетів відновлювальних робіт. Через неминучу неповноту вихідних даних аналіз будується на діапазонах можливих значень, що дає змогу вибудовувати більш гнучкі підходи до реагування.

Виклики, ризики та обмеження

Ефективність застосування ШІ безпосередньо визначається якістю вхідних даних. Відкритий інформаційний простір насичений дезінформацією, емоційними повідомленнями та спеціально створеними фейковими сигналами, які противник може запускати під час операцій. Якщо такі дані потрапляють до моделей без відбору й контролю, результати аналізу спотворюються.

Додатково діють правові обмеження: державні органи мають дотримуватися вимог щодо захисту персональних даних, навіть якщо інформація формально отримана з відкритих джерел. Організаційний вимір включає підготовку фахівців, уніфікацію процедур, налагодження механізмів перевірки результатів та запровадження правил використання рекомендацій ШІ, щоб уникнути сліпої довіри до алгоритмів.

Перспективи для України

Інтегровані системи оперативної обізнаності, які поєднують OSINT, інструменти ШІ та цифрові двійники, стають одним із ключових чинників стійкості держави. У короткостроковій перспективі вони скорочують час реагування на масовані удари. У середньостроковій — забезпечують спільний аналітичний простір для взаємодії різних відомств. У довгостроковій — наближають практики України до стандартів НАТО та ЄС, водночас формуючи власні компетенції роботи з даними та застосування штучного інтелекту в системах національної безпеки.

1. ENISA. (2020). Artificial intelligence cybersecurity challenges. European Union Agency for Cybersecurity. https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges.
2. ENISA. (2021). Securing energy infrastructure in the EU. European Union Agency for Cybersecurity. https://www.enisa.europa.eu/publications/securing-energy-infrastructure.
3. NATO Cooperative Cyber Defence Centre of Excellence. (2021). Artificial intelligence and the future of cyber defence. https://ccdcoe.org.
4. Griffor, E. R. (Ed.). (2018). Handbook of system safety and security: Cyber-physical systems, digital twins and resilience. Springer.
5. Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. Science, 361(6404), 751–752. https://doi.org/10.1126/science.aat5991.

Д.О. Грицкевич

ПРОСТОРОВО-ОРІЄНТОВАНІ МОДЕЛІ ШТУЧНОГО ІНТЕЛЕКТУ ЯК ІНСТРУМЕНТ ОПТИМІЗАЦІЇ ДЕРЖАВНОГО УПРАВЛІННЯ ПРИРОДНИМИ РЕСУРСАМИ

Вступ.

Швидка цифровізація державного управління та зростання обсягів просторових даних формують потребу у створенні моделей штучного інтелекту (ШІ), здатних комплексно аналізувати природні ресурси, прогнозувати зміни й оптимізувати територіальне планування. У сфері управління природними ресурсами важливими є не лише швидкість оброблення даних, а й можливість виявлення нових закономірностей, що дозволяють ухвалювати стратегічно виважені рішення. Просторово-орієнтовані моделі ШІ інтегрують алгоритмічну аналітику та геоінформаційні дані, забезпечуючи новий рівень підтримки державних рішень (Longley et al., 2021).

Виклад основного матеріалу.

Завдяки розвитку глибинного навчання, моделі на основі U-Net та ResNet демонструють високу точність у семантичній сегментації супутникових знімків, виділенні структур лісового покриву, зон деградації, вирубок та інших природних об’єктів. Ці алгоритми дозволяють автоматизовано опрацьовувати дані, які раніше потребували тривалої ручної інтерпретації (Zhang et al., 2021). Поєднання CNN-моделей з LIDAR-даними дає змогу отримувати значно точнішу інформацію про рельєф, стан біомаси та структуру лісових насаджень. Такі дані дозволяють бачити просторову динаміку територій у більш деталізованому вигляді, що важливо для прийняття рішень у сфері природокористування (Ma et al., 2019).

Методи машинного навчання, наприклад Random Forest, Gradient Boosting або SVM, використовують для розподілу територій за різними ознаками: від оцінки рекреаційного потенціалу до визначення екологічних ризиків. У поєднанні з алгоритмами кластеризації – DBSCAN чи HDBSCAN – такі моделі допомагають виділяти ділянки, подібні за характеристиками. Це корисно під час планування туристичної інфраструктури чи аналізу навантаження на природні зони (Malczewski & Rinner, 2022).

Для прогнозування довгострокових змін у природних системах застосовують нейронні мережі, здатні працювати з часовими рядами, зокрема LSTM та GRU. Такі моделі ефективно відстежують динаміку деградації земель, поширення лісових пожеж або коливання рекреаційного навантаження протягом року (European Environment Agency, 2023). Це дозволяє перейти від реагування «постфактум» до заздалегідь прорахованих управлінських рішень.

Одним із сучасних напрямів є використання Graph Neural Networks (GNN). Ці моделі розглядають територію як систему пов’язаних між собою елементів, що дозволяє оцінювати просторові взаємозв’язки: наприклад, як зміни в одній частині регіону впливають на сусідні ділянки або які території є

ключовими для підтримання екологічної рівноваги. Це дає змогу оцінювати територіальні взаємозалежності: як транспортна доступність впливає на рекреаційну цінність, як зміни в одному районі транслюються на сусідні, які ділянки є критичними для відновлення екосистем (Wu et al., 2020). Такий підхід забезпечує системність аналізу, властиву сучасним цифровим платформам управління територіями.

Алгоритми підкріпленого навчання (Reinforcement Learning) можуть застосовуватися для моделювання різних сценаріїв державного управління – наприклад, оптимізації зон рекреації, розподілу ресурсів служб екоконтролю чи планування інфраструктурних втручань. За допомогою RL можливе тестування тисяч варіантів дій без втручання у реальні екосистеми, що істотно зменшує ризики (Silver et al., 2017).

В Україні також зростає інтерес до інтеграції просторових даних у державні сервіси. Зокрема, розвиток Національної інфраструктури геопросторових даних та відкритих ресурсів Держгеокадастру створює підґрунтя для впровадження інтелектуальних систем підтримки рішень, які можуть автоматично аналізувати рекреаційний потенціал, стан земель і екологічні ризики (Бердник, 2022).

Висновки.

Таким чином, просторово-орієнтовані моделі ШІ відкривають можливості не лише для автоматизації оброблення складних екологічних та територіальних даних, але й для радикального підвищення якості державного управління природними ресурсами. Інтелектуальні системи дозволяють скорочувати час аналізу у десятки разів, виявляти приховані патерни у просторі, прогнозувати ризики та формувати сценарії сталого розвитку, що робить їх стратегічним інструментом майбутнього.

1. Longley, P. A., Goodchild, M. F., Maguire, D. J., & Rhind, D. W. (2021). Geographic information science and systems (5th ed.). Wiley.
2. Zhang, L., Zhang, L., & Du, B. (2021). Deep learning for remote sensing data: A technical tutorial on the state of the art. IEEE Geoscience and Remote Sensing Magazine, 9(2), 63–89.
3. Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., & Johnson, B. (2019). Deep learning in remote sensing applications: A meta-analysis and review. ISPRS Journal of Photogrammetry and Remote Sensing, 152, 166–177.
4. Malczewski, J., & Rinner, C. (2022). Multicriteria decision analysis in geographic information science (2nd ed.). Springer.
5. European Environment Agency. (2023). Recreational and green infrastructure mapping in Europe. https://www.eea.europa.eu.
6. Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2020). A comprehensive survey on graph neural networks. IEEE Transactions on Neural Networks and Learning Systems, 32(1), 4–24.
7. Silver, D., Schrittwieser, J., et al. (2017). Mastering the game of Go without human knowledge. Nature, 550, 354–359.
8. Бердник, О. В. (2022). Геоінформаційні системи у плануванні територій: методичні основи та практика. Вісник геодезії та картографії, (4), 15–24.

Ю.В. Писаренко, К.В Кармазін

ІНТЕЛЕКТУАЛЬНА МУЛЬТИАГЕНТНА АРХІТЕКТУРА ДЛЯ АНАЛІЗУ ТЕЛЕМЕТРІЇ ТА ПОТОКІВ ДАНИХ У РЕАЛЬНОМУ ЧАСІ

Стрімке зростання обсягів даних, які генеруються сенсорними мережами, безпілотними системами та мультимедійними сервісами, створює нові виклики для інформаційних інфраструктур реального часу. Цифрові екосистеми – від регіональних ситуаційних центрів моніторингу дронів до глобальних сервісів потокового відео – потребують високорівневих інтелектуальних механізмів, здатних оперативно аналізувати, фільтрувати та інтерпретувати дані в умовах нестабільності, динаміки та обмежених ресурсів. Саме тому ключового значення набуває використання штучного інтелекту (ШІ), машинного навчання, генеративних моделей та мультиагентних систем, що дозволяють формувати стійкі та резильєнтні моделі обробки інформації в реальному часі. Дана тема є продовженням попередніх досліджень, присвячених розробці системи підвищення контролю за повітряним простором «Control_TEA» [1, 2, 3].

У контексті інформаційного середовища функціонування дронів застосування ШІ забезпечує автоматизацію збору, фільтрації та аналізу телеметрії, виявлення рухомих об'єктів, а також прогнозування маршрутів і загроз. Інтелектуальні алгоритми дозволяють оптимізувати розміщення сенсорів і радарів, розпізнавати ситуації «свій-чужий» та підвищувати ефективність регіональних і мобільних ситуаційних центрів. Аналогічні механізми відіграють ключову роль у сучасних потокових сервісах, де необхідно забезпечити стабільний Quality of Experience (QoE) за умов варіативного мережевого трафіку. Розробка нових метрик, зокрема Resilient Quality of Experience (RQE), дозволяє оцінювати не лише середню якість, але й стабільність системи при зовнішніх збуреннях.

Нехай $X(t)$ – потік даних (телеметрія, відео, сенсорні дані), а $C(t)$ – пропускна здатність каналу зв'язку. Задача ШІ-системи – знайти оптимальну політику π^*, що максимізує функціонал якості:

$$\pi^* = arg\max_{\pi} \mathbb{E}\left[R(X(t), C(t), a_t)\right],$$

де a_t – дія агента (вибір бітрейту, маршруту, режиму передачі, пріоритету), а функція винагороди може мати вигляд:

$$R = \alpha \cdot QoE - \beta \cdot D_{lag} - \gamma \cdot P_{loss},$$

де D_{lag} – delay lag, затримка; P_{loss} – packet loss, втрати пакетів; α – коефіцієнт важливості QoE; β – коефіцієнт штрафу за затримку; γ – коефіцієнт штрафу за втрати пакетів.

Це дозволяє балансувати між якістю, затримкою та втратами даних у динамічних умовах.

Ситуаційні центри, дрони, edge-вузли та клієнтські пристрої потокового відео можуть бути представлені як агенти:

$$A = \{a_1, a_2, \ldots, a_n\},$$

які взаємодіють через стан середовища:

$$s_{t+1} = f(s_t, a_t^1, a_t^2, \ldots, a_t^n).$$

Кооперативна ціль системи визначається як:

$$\max_{\{\pi_i\}} \sum_{i=1}^{n} R_i,$$

що дозволяє забезпечити узгодженість дій (наприклад, оптимальний розподіл частот, маршрутизація потоків, керування пріоритетами трафіку).

Схема мультиагентної архітектури наведено на рисунку 1.

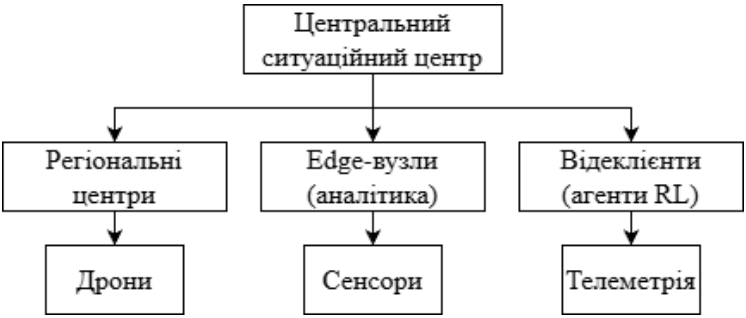

Рисунок 1 – Схема мультиагентної архітектури

Обидві системи – «Control_TEA» та адаптивне потокове відео – мають спільні риси:

1) необхідність обробки великих обсягів даних у реальному часі;
2) мінливість середовища (мережа, траєкторії, навантаження);
3) критичність рішень (моніторинг територій, безпека, онлайн-сервіси);
4) важливість прогнозування майбутнього стану системи.

У проєктах, присвячених техноекологічним подіям, ШІ забезпечує розпізнавання рухомих об’єктів, оптимізацію маршрутів, аналіз рельєфу для розміщення радарів та облік дронів. Водночас у відеосервісах LSTM/GRU-моделі прогнозують пропускну здатність, RL-алгоритми оптимізують вибір бітрейту, а генеративні моделі (GAN, diffusion networks) реконструюють спотворені кадри.

Узагальнений підхід передбачає побудову цифрового двійника середовища:

$$\hat{X}(t+1) = F_\theta(X(t), X(t-1), \ldots, C(t)),$$

де F_θ – нейромережева модель прогнозування.

Далі інтегрований агент приймає рішення:

$$a_t = \pi(F_\theta(X), s_t).$$

Таким чином формується петля інтелектуальної адаптації: Потік даних → Прогноз → Оптимізація → Дія → Збір нових даних → …

Практична значущість:

1) Дрони та техноекологічні події: моніторинг територій, контроль повітряного простору, розпізнавання загроз, оптимізація розміщення сенсорних вузлів.

2) Потокові сервіси: стабільність QoE/RQE, оптимізація трафіку 4G/5G, зменшення буферизації, підвищення ефективності edge-обробки.

3) Спільний ефект: формування резильєнтної цифрової інфраструктури, здатної до самоадаптації, кооперації агентів та стійкої роботи у динамічних умовах.

Поєднання методів машинного навчання, прогнозування, генеративних моделей та мультиагентної координації забезпечує ефективну обробку великих потоків даних у реальному часі як у системах дронів, так і у мультимедійних сервісах. Запропонований підхід формує нову концепцію цифрової резильєнтності, в якій різнорідні агенти взаємодіють для забезпечення стабільності, надійності й оптимального управління інформаційними потоками.

1. J. Pisarenko, and E. Melkumyan, "The Structure of the Information Storage "CONTROL_TEA" for UAV Applications," IEEE 5th International Conference "Actual Problems of Unmanned Aerial Vehicles Developments (APUAVD)" (October 22-24, 2019), pp. 274–277. https://doi.org/10.1109/APUAVD47061.2019.8943938.
2. Kateryna Melkumian, Julia Pisarenko, Alexander Koval, Organization of regional situational centers of the intelligent system «CONTROL TEA» using UAVs, 2021 IEEE 6th International Conference Actual Problems of Unmanned Aerial Vehicles Developments (APUAVD), Kyiv, Ukraine2021. https://doi.org/10.1109/APUAVD53804.2021.9615416.
3. Pisarenko J., Melkumyan K., Varava I., Koval O., Chumakova N. About the organization of regional situational centers of the intellectual system “Control_TEA” with the use of UAVs. Artificial intelligence. Kyiv, 2022. No. 1. P. 275-287. https://doi.org/10.15407/jai2022.01.275.

І.М. Коваль, С.А. Головня

КЛАСТЕРИЗАЦІЯ АНОМАЛЬНИХ БЛОКЧЕЙН-ТРАНЗАКЦІЙ ЗА ДОПОМОГОЮ АЛГОРИТМІВ ШТУЧНОГО ІНТЕЛЕКТУ

Швидке зростання блокчейн-мереж та децентралізованих фінансових інструментів супроводжується різким збільшенням шахрайських схем, відмиванням коштів та приховуванням першоджерела походження коштів. Псевдонімні та анонімні блокчейн-операції фактично унеможливлюють якісну роботу фінансового моніторингу. Відсутність централізованого підходу до роботи з будь-якими блокчейн-інструментами змушує шукати шляхи відстеження походження коштів за допомогою машинного навчання. У даній роботі розглядається один з методів ідентифікації підозрілих блокчейн-транзакцій за допомогою алгоритмів штучного інтелекту. У цьому контексті застосування кластеризації та виявлення аномалій відкриває нові можливості для автоматичного аналізу та виявлення потенційно небезпечних патернів поведінки.

Блокчейн-транзакції містять чітко виражену структуру [1], зокрема обсяг переказів, часові інтервали, пов’язані між собою адреси, взаємодію із смарт-контрактами та інші ознаки, що формують поведінковий профіль конкретного користувача. На основі цих даних алгоритми машинного навчання можуть здійснювати пошук прихованих закономірностей, які недоступні при перевірці стандартними інструментами аналізу даних.

Для кластеризації даних застосовуються класичні методи машинного навчання, зокрема k-means, hierarchical clustering, DBSCAN та HDBSCAN. Алгоритми на основі щільності [2] (DBSCAN/HDBSCAN) добре зарекомендували себе у задачах аналізу нерівномірно розподілених даних, характерних для блокчейн-мереж. Вони дозволяють автоматично виділити кластери транзакцій з подібними характеристиками, а також відфільтрувати «зашумлені» дані – тобто потенційні аномалії, серед яких зустрічаються шахрайські дії або транзакції, спрямовані на маскування першоджерела фінансових активів.

Методи виявлення аномалій, такі як Isolation Forest та автоенкодери, дозволяють будувати моделі «нормальної» поведінки та виявляти операції, що суттєво відхиляються від типового патерна. Використання графових нейронних мереж (GCN, GraphSAGE) дає змогу аналізувати структуру зв’язків між учасниками транзакцій та виявляти підозрілі ланцюги переказів, які утворюють складні графові структури.

Для демонстрації запропонованого підходу було використано набір даних [3], що включає типову інформацію про транзакцію в блокчейні Ethereum (Рис. 1), зокрема, хеш транзакції, номер блоку, часову мітку, адресу відправника, адресу отримувача, обсяг переданих активів та комісію мережі.

Transaction Hash	Method	Block	Date Time (UTC)	From	To	Amount
0x3c17d94c17...	Transfer	23875766	2025-11-25 12:19:11	0x18bb8969...c12f1b290	0xa6b4854f...3679AF17E	0.012948 ETH
0xb96333cd37...	Transfer	23875766	2025-11-25 12:19:11	0x6C46D6A9...Ce6EbBa16	Tether: USDT Stabl...	0 ETH
0xda8904edca...	Transfer	23875766	2025-11-25 12:19:11	0x41b30801...EA7c6EF05	FixedFloat 1	0.00898866 ETH

Рисунок 1 – Типовий набір даних про транзакцію у блокчейні Ethereum

У межах демонстраційного експерименту для аналізу транзакцій було обрано програмне середовище KNIME Analytics Platform [4]. Після виконання попередньої обробки даних (видалення номінальних полів, нормалізації числових ознак та формування ознакового простору) було застосовано алгоритми кластеризації k-means та DBSCAN (Рис.2).

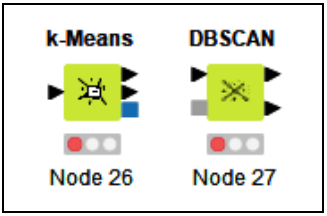

Рисунок 2 – Ноди алгоритмів кластеризації k-means та DBSCAN у середовищі KNIME Analytics Platform

Алгоритм DBSCAN успішно виділив три окремі кластери, які відповідають різним типам поведінки, та коректно ідентифікував підозрілі транзакції як окремий щільнісний кластер. Натомість k-means частково об’єднав аномальні операції з типовими транзакціями, що демонструє перевагу методів на основі щільності у виявленні нетривіальних та аномальних патернів у блокчейн-мережах.

Важливим етапом побудови моделей кластеризації є формування ознакового простору. У контексті аналізу блокчейн-транзакцій найінформативнішими виявилися такі характеристики, як кількість унікальних контрагентів, середній інтервал між операціями, варіативність обсягу переказів, частота взаємодії з певними типами смарт-контрактів та коефіцієнт подібності до відомих ризикових адрес. Використання таких поведінкових ознак дозволяє формувати чіткіші та відокремлені кластери, а також підвищує точність ідентифікації підозрілих транзакцій. Проведений аналіз підтверджує, що поведінкові метрики відіграють суттєвішу роль, ніж окремі числові показники (наприклад, сума або комісія), що підкреслює важливість контекстного підходу у системах моніторингу блокчейн-активності.

Результати кластеризації можуть застосовуватися для підвищення якості політики AML/KYC у різних фінансових інструментах, моніторингу потоків у DeFi-протоколах та відстеження шахрайських схем, які пов'язані з «міксерним» відмиванням коштів. Так звані «міксери» зараз є чи не найбільшою проблемою, яка блокує якісне відстеження походження коштів інструментами фінансового моніторингу, але за допомогою правильно побудованих кластерів підозрілі транзакції будуть відстежуватись. Такий підхід може бути інтегрований у комплексні системи аналізу ризиків та вбудований у процеси автоматичного маркування транзакцій у режимі реального часу, за рахунок підключення інструментів машинного навчання.

Водночас існують виклики, пов'язані з відсутністю маркованих даних, складністю інтерпретації результатів кластеризації, динамічністю поведінки учасників мережі та великою кількістю транзакцій, які потребують ефективних методів обчислення. Проте сучасні алгоритми штучного інтелекту демонструють значний потенціал у підвищенні прозорості та безпеки блокчейн-екосистем і можуть стати ключовим інструментом у протидії фінансовим зловживанням у децентралізованих середовищах.

1. Victor, F., & Weintraud, A. (2019). Detecting address clusters in Ethereum. In *Financial Cryptography and Data Security* (pp. 653–673). Springer. DOI:10.1007/978-3-030-51280-4_33.
2. Hahsler, M., Piekenbrock, M., & Doran, D. (2019). *dbscan: Fast density-based clustering with R*. Journal of Statistical Software, 91(1), 1–30. DOI:10.18637/jss.v091.i01.
3. Etherscan. (n.d.). *Ethereum transaction overview*. https://etherscan.io/.
4. KNIME AG. (n.d.). *KNIME Analytics Platform* [Computer software]. https://www.knime.com/.

UDC 004

A. Osypchuk, O. Chyzhmotria

IMPROVED OBJECTIVITY OF AUTOMATED LLM ASSESSMENT THROUGH GRADE-LIKE-A-HUMAN METHODOLOGY INTEGRATION

Automated analysis of open-ended student responses has become an essential component of modern digital learning platforms. Many systems, including our own platform, which is built around the Semantic Kernel AI module, already incorporate LLM-driven semantic analysis to assist instructors in evaluating short written answers. However, despite their linguistic competence, current LLM-based scoring pipelines remain limited in their capacity to maintain consistency, justify decisions, adapt rubrics, and detect subtle grading anomalies. These limitations prevent automated assessment from reaching the level of nuanced human reasoning that experienced instructors demonstrate when grading real student work [1].

Our previous work established a baseline system that processes student responses, extracts semantic patterns, and performs automated scoring using static rubrics. While effective for structured or fact-based questions, the system struggled with open-ended responses where student reasoning may legitimately diverge from expected formulations. These shortcomings highlighted a systemic issue: static rubrics and single-pass LLM scoring cannot fully capture the diversity of student understanding. This motivates our current effort to move beyond "LLM as a grader" toward a more reflective, multi-stage reasoning pipeline [2, 3].

The present study explores the integration of the Grade-Like-a-Human (GLAH) methodology into our platform to address this shortcoming. GLAH reframes grading as a human-like multi-agent process involving iterative refinement, contextual awareness, and post-hoc verification. Instead of relying on fixed rubrics and single-shot scoring, GLAH structures assessment as a loop involving rubric optimisation, batch-based comparison, self-reflection, and anomaly detection. Prior work has demonstrated that such reflective pipelines improve grading reliability on datasets such as OS and Mohler; however, existing studies do not address integration challenges within real educational platforms or the architectural changes required to operationalise GLAH at scale [4].

Our research contributes in three principal ways. First, we develop an extended rubric-optimisation module that adapts GLAH's score-distribution-aware sampling to the characteristics of our platform's student population. Our version dynamically chooses representative samples from current student submissions and repeatedly modifies rubric descriptors based on semantic patterns found through LLM reflection. Second, we implement consistency-oriented scoring using batching prompts and structured self-reflection steps, evaluating their effect on reducing randomness across multiple grading cycles within the same question set. Third, we integrate a post-grading review engine based on group comparison and re-grouping, designed to detect hallucinated or contradictory scores by examining cross-similarity among graded responses. This module is capable of routing ambiguous cases to manual experts or triggering an automated re-evaluation pipeline.

To support these new components, we outline a revised system architecture that introduces three new layers: (1) a rubric refinement layer, (2) a consistency-enhanced scoring layer, and (3) a reflective anomaly-detection layer. These layers operate in conjunction with the existing Semantic Kernel-based preprocessing engine and are coordinated through a centralised control module that manages iterative reasoning cycles.

The system will be evaluated using live datasets from our platform to examine robustness across diverse domains and answer styles. Key metrics include grading stability, inter-run consistency, agreement with expert graders, comprehensiveness of the rubric, and the anomaly-recovery rate. By integrating and adapting GLAH within a real educational platform, the study aims to produce actionable design principles for next-generation systems capable of human-level reflectiveness and reasoning.

1. Осипчук А. (2025) Система автоматизованого аналізу студентських робіт. https://ztu.edu.ua/en/site/graduation-works-view?id=9472717.
2. Yucheng C., Li H., Kaiqi Y. (2025) A LLM-Powered Automatic Grading Framework with Human-Level Guidelines Optimization. Proceedings of the 18th International Conference on Educational Data Mining, 31-41. https://educationaldatamining.org/EDM2025/proceedings/2025.EDM.long-papers.80/.
3. Owen H., Bill R., Libby H., Joshua McG. (2023) Can LLMs Grade Short-Answer Reading Comprehension Questions? An Empirical Study with a Novel Dataset. International Journal of Artificial Intelligence in Education. Vol. 35, 651–676 (2025). https://doi.org/10.1007/s40593-024-00431-z.
4. Wenjing X., Juxin N., Chun J. X. & Nan G. (2024). Grade Like a Human: Rethinking Automated Assessment with Large Language Models. https://arxiv.org/pdf/2405.19694.

Я.В. Сліпчук

ЗАСТОСУВАННЯ ШТУЧНОГО ІНТЕЛЕКТУ В АНАЛІЗІ ТА МОДЕЛЮВАННІ ТУРИСТИЧНИХ КЛАСТЕРІВ РЕГІОНУ

Вступ

Сучасні тенденції цифрової трансформації сприяють суттєвому розширенню інструментів аналізу територій, що, у свою чергу, впливає на ефективність управління регіональними туристичними системами. У міру зростання обсягів даних, що описують туристичні ресурси, поведінку відвідувачів, інфраструктуру та просторові зв'язки, постає потреба у використанні технологій, здатних обробляти великі інформаційні масиви та виявляти закономірності, які складно виявити традиційними методами.

Геоінформаційні системи є засобом об'єднання географічної, статистичної та предметної інформації, що формує основу для всебічного вивчення місцевостей. Одночасно з прогресом ГІС активно вдосконалюються методи машинного інтелекту, що дають змогу автоматизувати аналітичні процедури, створювати моделі групування, ідентифікувати відхилення, передбачати маршрути переміщення туристичних груп та визначати перспективи їх майбутнього розширення. На противагу звичайним статистичним методам, машинний інтелект може пристосовуватися до змін у даних, що робить його особливо дієвим у мінливих обставинах, таких як туризм, де зовнішні чинники мають важливе значення (Процик, Дорош, 2022).

Інформаційні системи та штучний інтелект створює потужний методологічний інструментарій, який дозволяє комплексно досліджувати туристичні території, аналізувати їхнє функціональне наповнення, враховуючи природні, культурні, соціальні та інфраструктурні компоненти. Таке поєднання сприяє більш точному визначенню туристичних кластерів – просторових угруповань об'єктів, що утворюють єдину систему функціональних взаємодій і мають стратегічне значення для розвитку регіону.

Виклад основного матеріалу

Проблематика моделювання туристичних кластерів набуває особливого значення у зв'язку з необхідністю оптимізації просторової організації туристичної діяльності та підвищення конкурентоспроможності туристичних регіонів. Туристичний кластер, як правило, включає сукупність об'єктів і територій, що доповнюють один одного та створюють цілісний туристичний продукт. Для його визначення необхідно здійснити детальний аналіз просторової взаємодії об'єктів, що стає можливим лише за умов використання спеціалізованих алгоритмів і програмних інструментів.

Одним із найбільш поширених методів, застосовуваних у кластеризації геопросторових даних, є DBSCAN – алгоритм, що формує кластери на основі

щільності точок та дозволяє виявляти як компактні скупчення, так і складні просторові структури (Dzikri et al., 2025). Його ключова перевага полягає у відсутності необхідності попереднього визначення кількості кластерів, що є критично важливим у туристичному аналізі, де просторові закономірності часто не мають чітких меж. Крім того, DBSCAN здатний відфільтровувати «шум» – точки, що не належать до жодного кластеру, але можуть бути використані для подальшого аналізу.

Надзвичайно важливою є можливість інтегрування в ГІС різноманітних джерел даних, зокрема топографічних карт, даних дистанційного зондування Землі, GPS-треків користувачів, інформації з туристичних порталів, мобільних застосунків, статистичних звітів, даних місцевих адміністрацій та сервісів купівельної активності. Завдяки цьому створюються інформаційні моделі, що охоплюють різні рівні деталізації й дозволяють здійснювати багатовимірний аналіз туристичних територій (Šoltésová et al., 2025).

Методи ШІ суттєво розширюють можливості просторової аналітики, включаючи прогнозування, виявлення прихованих залежностей та визначення ключових чинників розвитку туристичного середовища. Сучасні алгоритми машинного навчання дають змогу визначати ступінь взаємозалежності туристичних об’єктів, аналізувати часову динаміку відвідуваності, сегментувати структури туристичних потоків відповідно до поведінкових, сезонних або тематичних характеристик. У сфері формування туристичних об'єднань, ці підходи допомагають встановити головні цілі, перспективні області для розширення та передбачити майбутній прогрес регіону.

Застосування інтелектуальних систем виглядає багатообіцяючим для ідентифікації відхилень у географічних даних, що можуть вказувати на виникнення невідомих раніше цікавих місць, погіршення стану певних областей, зміни у звичках відвідувачів або створення нових об'єктів. Використання методів ідентифікації відхилень дає можливість контролювати туристичні території в режимі реального часу, що є важливим для оперативного реагування на зміни в туристичній сфері (Yin et al., 2022).

Практична реалізація процесу кластеризації туристичних об’єктів може бути виконана за допомогою інструментів програмування, серед яких ключову роль відіграє мова Python. У поєднанні з бібліотеками GeoPandas, scikit-learn і Folium стає можливим здійснення імпорту даних, їхнього очищення, кластеризації, просторового аналізу та інтерактивної візуалізації. Зокрема, Folium дозволяє створювати інтерактивні веб-карти, які можуть бути інтегровані у системи підтримки прийняття рішень, використовуватися органами влади та аналітичними центрами (Мацюк, 2021).

Отримані результати моделювання є важливими для практичного використання у сфері стратегічного планування, просторового розвитку та інвестиційної діяльності. На їх основі можна формувати туристичні стратегії, оптимізувати навантаження на інфраструктуру, визначати перспективні

території для модернізації, розробляти тематичні маршрути, підвищувати якість управлінських рішень та створювати конкурентоспроможні туристичні продукти. Аналітичні моделі також можуть бути використані для оцінювання рівня функціональної цілісності туристичних кластерів, визначення їхніх сильних і слабких сторін та формування рекомендацій щодо розвитку (Інститут географії НАН України, 2022).

Висновки

Застосування штучного інтелекту у сполученні з геоінформаційними системами (ГІС) утворює значний потенціал для моделювання туристичних кластерів та аналізу просторової організації туристичних зон. Завдяки використанню кластеризаційних алгоритмів, зокрема DBSCAN, можливо з високою точністю визначати особливості розміщення туристичних об'єктів, враховуючи як їхні просторові параметри, так і описові ознаки. Моделі штучного інтелекту допомагають прогнозуванню, виявленню відхилень та визначенню сприятливих напрямків розвитку галузі. Такий інтегрований підхід, що об'єднує можливості ГІС і ШІ, закладає основу для формування розумних систем підтримки прийняття рішень, сприяє стратегічному плануванню та забезпечує умови для стійкого розвитку туристичних територій.

1. Бердник О.В. Геоінформаційні системи у плануванні територій: методичні основи та практика. *Вісник геодезії та картографії*, 2023, №4, С. 15–24.
2. Процик М.Р., Дорош О.О. Застосування ГІС-технологій у дослідженнях просторової структури туристичних ресурсів. *Науковий вісник ЧНУ ім. Ю. Федьковича*, 2022.
3. Мацюк В.С. Картографічне моделювання туристичних об’єктів засобами ГІС. *Географія та туризм*, 2021, №58.
4. Інститут географії НАН України. Аналітичні підходи до просторового моделювання територій України. Київ: НАН України, 2022.
5. Šoltésová M., Iannaccone B., Štrba Ľ., Sidor C. Application of GIS Technologies in Tourism Planning and Sustainable Development: Case Study of Gelnica. *ISPRS International Journal of Geo-Information*, 2025, 14(3), 120. DOI: 10.3390/ijgi14030120.
6. Dzikri A., Aknuranda I., Indradjad A. Hierarchical Clustering-Based Geospatial Analysis for Tourism Systems. *Engineering Technology & Applied Science Research*, 2025, 15(4), 24794–24799.
7. Zeng H., Tang C., Zhou C., Zhou P. Machine Learning Approaches to Spatial Analysis of Tourism Resources. *Sustainability*, 2025, 17(2), 744.

А.О. Бочарова, О.М. Супрун

ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ ТЕХНІК ПРОМПТИНГУ ДЛЯ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ

Сучасні великі мовні моделі (LLM) демонструють значні можливості у розв'язанні широкого спектру завдань. Однак якість їхніх відповідей значною мірою залежить від способу формулювання вхідного запиту (prompt) [1]. Це робить актуальним питання систематичного дослідження та порівняння різних технік промптингу для оптимізації взаємодії з мовними моделями [2].

Мета дослідження полягає в експериментальному порівнянні ефективності основних технік промптингу (baseline, zero-shot, few-shot та chain-of-thought) на завданнях різних класів, а також розробленні практичного інструменту для автоматизації створення оптимізованих промптів.

Методологія експерименту. Для проведення дослідження була використана модель GPT-3.5-turbo [3]. Експеримент охоплював чотири класи завдань: математичні задачі, класифікація тексту, генерація текстового контенту та логічні задачі. Для кожного класу було підготовлено набір тестових прикладів, які обробляли за допомогою чотирьох типів промптів:

- Baseline – базовий запит без додаткових інструкцій чи контексту;
- Zero-shot – запит із короткою інструкцією, що описує завдання без прикладів;
- Few-shot – промпт з кількома прикладами правильних відповідей [4];
- Chain-of-Thought (CoT) – запит, що налаштовує модель пояснювати хід міркувань покроково [2].

Оцінювання результатів здійснювалося за допомогою метрик точності та релевантності. У завданнях із детермінованими відповідями (математика, логіка, класифікація) використовувалося пряме порівняння з еталонними результатами. Для генерації тексту застосовувалась перевірка наявності ключових слів у згенерованому контенті, які відповідали меті завдання.

Як показано на рис. 1, для *математичних завдань* найвищу точність (100%) продемонструвала техніка Chain-of-Thought. Покрокове міркування дозволило моделі правильно розв'язувати складні багатоетапні задачі, включно комбінаторні обчислення та задачі на рух. Інші техніки показали значно нижчі результати.

У завданнях на *класифікацію* найефективнішими виявилися few-shot (100%), zero-shot та baseline (89%) підходи. Короткі інструкції або приклади надавали моделі достатньо контексту для правильного визначення категорій. Слід зауважимти, що CoT у цій категорії показав дещо нижчий результат (67%), оскільки надмірні міркування іноді призводили до відхилення від основного завдання.

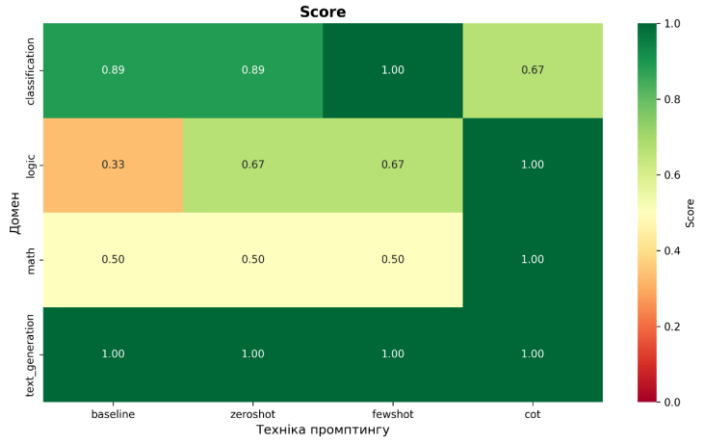

Рисунок 1 – Результати оцінювання точності відповідей для різних технік промптингу

Для *генерації тексту* усі техніки продемонстрували однаково високі результати (100%). Це пояснюється відсутністю однозначно правильної відповіді та високою здатністю моделі GPT-3.5-turbo генерувати зв'язний текст.

У *логічних завданнях* Chain-of-Thought знову виявився найефективнішим (100% точності). Покрокове обґрунтування допомогло моделі уникнути інтуїтивних помилок, характерних для інших підходів. Zero-shot та few-shot надали лише 67% правильних відповідей через поспішні висновки без логічного аналізу.

Практична реалізація. На основі результатів дослідження було розроблено Telegram-бот Prompt Helper Bot, який автоматизує процес створення оптимізованих промптів. Бот реалізований на Python із використанням Telegram API та OpenAI API [3]. Користувач надсилає початковий запит і обирає бажану техніку промптингу: zero-shot, few-shot, chain-of-thought, role prompting або auto. Бот генерує покращений промпт і за бажанням може відразу надіслати його до GPT-5 для отримання відповіді. Режим «auto» автоматично визначає оптимальну стратегію на основі типу завдання.

Функціональні можливості бота включають:

- аналіз вхідного запиту та визначення оптимального типа покращення;
- генерацію оптимізованого промпту відповідно до обраної техніки;
- можливість регенерації результату;
- пряму інтеграцію з GPT для отримання відповіді в реальному часі.

Приклад використання розробленого Telegram-бота зображено на рисунку 2.

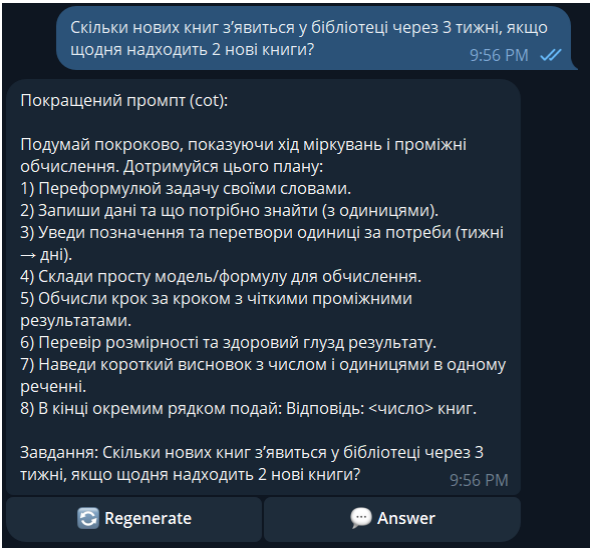

Рисунок 2 – Приклад використання Telegram-бота Prompt Helper Bot для створення покращеного промпту методом Chain-of-Thought

Висновки. Виконане дослідження підтверджує суттєвий вплив техніки промптингу на якість відповідей великих мовних моделей. Chain-of-Thought виявився найефективнішим для завдань, що потребують багатоетапного міркування (математика, логіка), тоді як few-shot та zero-shot краще підходять для класифікації та задач із чіткою структурою. Розроблений Telegram-бот демонструє можливість практичного застосування результатів дослідження для допомоги користувачам у формуванні ефективних запитів до LLM.

Подальші дослідження можуть включати тестування на більш потужних моделях (GPT-4, Claude), розширення набору класів завдань та розробку гібридних технік промптингу, що автоматично адаптуються до специфіки завдання.

1. Sahoo, P. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. arXiv. https://doi.org/10.48550/arXiv.2402.07927.
2. Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2025). Unleashing the potential of prompt engineering for large language models. Patterns, 6(6), Article 101260. https://doi.org/10.1016/j.patter.2025.101260.
3. OpenAI. (2023). GPT-3.5 Turbo fine-tuning and API updates. OpenAI Blog. https://openai.com/index/gpt-3-5-turbo-fine-tuning-and-api-updates/.
4. Liu, C. (2025, February 20). A review of LLM prompting techniques. Medium. https://medium.com/@celine-liu/a-review-of-llm-prompting-techniques-bc153834bb50.

В.С. Новік

МЕТОДИЧНІ ПІДХОДИ ДО АВТОМАТИЗОВАНОЇ ОЦІНКИ ВГОДОВАНОСТІ ВЕЛИКОЇ РОГАТОЇ ХУДОБИ НА ОСНОВІ ГЛИБОКИХ МОДЕЛЕЙ КОМП'ЮТЕРНОГО ЗОРУ

Автоматизована оцінка вгодованості великої рогатої худоби (Body Condition Score, BCS) є важливим напрямом розвитку сучасного точного тваринництва, оскільки рівень вгодованості безпосередньо впливає на продуктивність тварин, їх репродуктивну здатність та економічну ефективність фермерських господарств. Традиційні способи оцінювання базуються на візуально-пальпаторних шкалах і суттєво залежать від досвіду спеціаліста, що робить їх суб'єктивними та трудомісткими.

Розвиток глибоких моделей комп'ютерного зору дозволяє побудувати автоматизовані системи оцінювання BCS, які забезпечують відтворюваність, масштабованість і можливість обробки великих масивів зображень. У сучасних дослідженнях широко застосовуються CNN- та Transformer-архітектури, проте підвищення точності сегментації та здатності моделі розрізняти тонкі морфологічні відмінності між рівнями BCS залишається актуальним завданням.

Для дослідження використовувався великий набір зображень великої рогатої худоби у форматі JPEG стандартної роздільної здатності 1024×576. Датасет охоплює п'ять рівнів BCS: 3.25, 3.5, 3.75, 4.0 та 4.25, що структуровано у табл. 1.

Таблиця 1 – Розподіл зображень за категоріями BCS

Категорія BCS	Кількість зображень	Відсоток від вибірки
3.25	7,536	14.07%
3.5	13,256	24.74%
3.75	14,255	26.61%
4.0	12,556	23.44%
4.25	5,963	11.13%
Загалом	53,566	100%

Зображення змінюються за умовами зйомки, ракурсу, рівня освітлення та фону, що забезпечує варіативність навчальних даних. Розподіл на трейн/валідацію/тест здійснювався стратифіковано для збереження пропорцій класів.

Оскільки вихідні анотації містили лише bounding boxes, застосовано ітеративний підхід pseudo labeling, який об’єднує детекційні та сегментаційні моделі сучасного покоління, наведено на рис. 1.

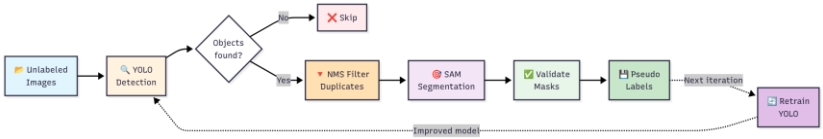

Рисунок 1 – Процес Iterative pseudo labeling

Метод включає такі етапи:

- YOLOv11x генерує первинні bounding boxes з високим порогом упевненості.
- Bounding boxes передаються як prompts до SAM 2.1 Large (Segment Anything Model).
- SAM створює детальні полігональні маски тіла тварини.
- Виконується фільтрація дублікатів, перевірка валідності полігонів, нормалізація координат.
- Сформовані маски конвертуються у формат COCO JSON.
- Маски використовуються для донавчання YOLOv11x-seg, після чого цикл повторюється.

На рис. 2 подано приклади одержаних масок після кількох ітерацій pseudo labeling.

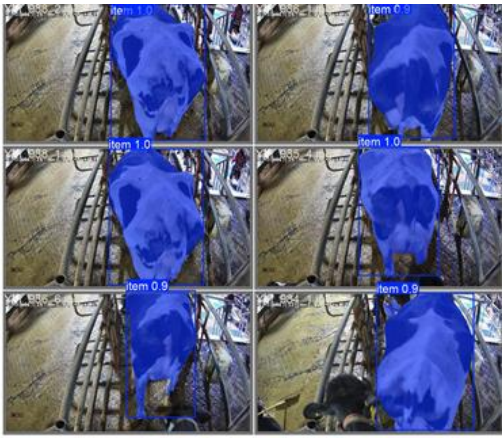

Рисунок 2 – Результат отриманих масок методом Iterative pseudo labeling

Після формування масок використовувалася архітектура Mask R-CNN, яка виконує:

- сегментацію об’єктів,
- класифікацію BCS,
- генерацію високоточних полігональних масок.

- Mask R-CNN складається з таких компонентів:
- Backbone — моделі для екстракції ознак: ResNet-50, ResNeXt-101, Swin Transformer Tiny;
- Feature Pyramid Network (FPN) — поєднання ознак різних масштабів;
- Region Proposal Network (RPN) — генерація областей інтересу;
- RoI Align — точне вирівнювання ознак об’єкта;
- Mask Head — формування бінарної маски.

Swin Transformer є ключовою інновацією методів, застосованих у роботі. На рис. 3 наведено схему його роботи.

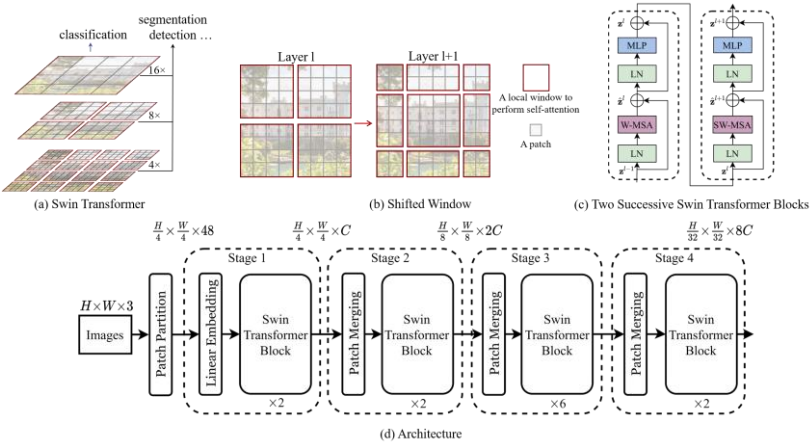

Рисунок 3 – Схема зміщених вікон та архітектура swin transformer [1]

1. Shifted Window Attention — увага в локальних вікнах із періодичним зсувом, що забезпечує обмін інформацією між сусідніми областями.
2. Patch Merging — формування ієрархічних рівнів ознак.
3. Обробка з високою просторовою роздільністю на ранніх шарах.
4. Сумісність із FPN, що полегшує інтеграцію з Mask R-CNN.

Завдяки цьому Swin Transformer здатний описувати тонкі морфологічні відмінності, що є важливим для розмежування близьких категорій BCS.

Під час тренування застосовувалися такі техніки:

- Automatic Mixed Precision (AMP) для оптимізації використання GPU-пам’яті;
- аугментації: Resize (1024×576), RandomFlip;
- оптимізатори: AdamW (для Swin-T) та SGD (для CNN-бекбонів);
- планувальники learning rate: cosine annealing та MultiStepLR;
- навчання у фреймворку MMDetection 3.3.0.

Після сегментації Mask R-CNN використовує семантичні та морфологічні ознаки виділеного тіла тварини для визначення категорії BCS.

Класифікація здійснюється у межах основної архітектури без ручного опису окремих анатомічних зон — модель самостійно формує внутрішні просторові уявлення завдяки механізмам уваги.

Висновки. Методичний підхід до автоматизованої оцінки вгодованості великої рогатої худоби базується на поєднанні сучасних компетенцій комп’ютерного зору: формування полігональних масок методом iterative pseudo labeling, ієрархічних моделей типу Swin Transformer та архітектури Mask R-CNN для сегментації екземплярів.

Застосовані методи забезпечують комплексне відтворення контурів тіла, точну локалізацію об’єкта та можливість аналізу його морфологічних особливостей. Такий підхід створює основу для побудови масштабованих систем моніторингу стану тварин і автоматизації процесів прийняття рішень у точному тваринництві.

1. Liu Z., Lin Y., Cao Y., Hu H., Wei Y., Zhang Z., Lin S., Guo B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2021. P. 10012-10022. DOI: 10.1109/ICCV48922.2021.00986.
2. Dandıl E., Çevik K.K., Boğa M. Automated Classification System Based on YOLO Architecture for Body Condition Score in Dairy Cows. Veterinary Sciences, 2024.
3. Shi W. et al. Automatic estimation of dairy cow body condition score based on attention-guided 3D point cloud feature extraction. Computers and Electronics in Agriculture, 2023.
4. Siachos N. et al. Development and validation of a fully automated 2D surveillance system for BCS monitoring of dairy cows. Journal of Dairy Science, 2024.
5. Angel T.R. et al. Comparison of manual and automated body condition scoring in dairy cows. Veterinary Record, 2024.
6. Lewis R. et al. Automated Body Condition Scoring in Dairy Cows Using 2D Imaging and Deep Learning. AgriEngineering, 2025.

С.О. Євдокимов

МОДЕЛЮВАННЯ ПОВЕДІНКИ КОРИСТУВАЧА В УМОВАХ КІБЕРАТАКИ ЧЕРЕЗ МОБІЛЬНІ ДОДАТКИ

В сучасних умовах цифровізації фейкові мобільні додатки стали одним із основних векторів кібератак у 2024-2025 роках. За даними Google [1] та Zimperium [2], понад 13 млн шкідливих застосунків було виявлено лише за один рік, а 95% випадків зараження спричинені завантаженням зі сторонніх джерел. Дослідження [3-5] показують, що успішність атаки залежить не лише від технічної досконалості шкідливого програмного забезпечення (ШПЗ), а й від поведінкової моделі користувача, який виступає найслабшою ланкою системи безпеки. Моделювання поведінки користувача під час таких атак дозволяє зрозуміти, як саме відбувається компрометація пристрою, які психологічні фактори активують помилки, та як штучний інтелект (ШІ) може зупинити атаку на ранніх етапах.

У типовому сценарії атаки користувач переходить за фішинговим посиланням, отриманим через SMS або месенджер, та ініціює завантаження інсталяційного пакета програми (наприклад, з розширенням .apk). В процесі встановлення власник смартфону власноруч надає застосунку критичні права доступу: до читання повідомлень (табл. 1), служби спеціальних можливостей (Accessibility) та функцій накладання вікон (Overlay). Отримавши ці права, шкідливе ПЗ (наприклад, трояни Xenomorph чи Triada) здатне перехоплювати паролі та відображати підроблені вікна поверх справжніх банківських застосунків.

Таблиця 1 – Результати аналізу впливу фейкових додатків

Показник	Поточний стан (без захисту)	Динаміка (Ефект)
Частка троянів у структурі атак	20-25%	62.5%
Середній час до втрати коштів	7–11 хв	97%
Середні фінансові втрати	50 000 грн	95%
Критичні дозволи	Accessibility, доступ до сповіщень, Overlay, доступ до SMS	75–100%

В табл. 1 показано ключові показники впливу фейкових мобільних додатків на безпеку користувачів та продемонстровано ефективність запропонованих методів захисту (динаміка).

Аналіз структури мобільного шкідливого ПЗ свідчить, що частка банківських троянів становить 20–25%. Це узгоджується зі світовими тенденціями зростання фінансово мотивованих атак, описаними у роботах [6, 7–9]. Такий рівень поширення вказує на те, що зловмисники систематично використовують фейкові застосунки як основний вектор компрометації.

Впровадження захисних механізмів дозволяє знизити успішність таких атак на 62,5%.

Критичним фактором залишається швидкість компрометації. Згідно з даними [6, 10, 11], середній час від моменту встановлення шкідливого ПЗ до втрати коштів становить лише 7–11 хвилин. Це зумовлено автоматизованим перехопленням облікових даних. Застосування алгоритмів виявлення загроз здатне покращити час реакції та нівелювати загрозу з ефективністю до 97%.

Суттєвий економічний ризик підтверджується показником середніх фінансових втрат, які сягають 50 000 грн на одну жертву, що корелює з оцінками, наведеними у джерелі [12]. Прогнозована динаміка показує можливість зниження фінансових збитків на 95% завдяки своєчасному блокуванню транзакцій або шкідливої активності.

Окрему увагу в таблиці приділено технічним аспектам атаки. На основі джерел [6, 7, 10, 13] визначено перелік критичних дозволів, якими найчастіше зловживають фейкові додатки: Accessibility Service, доступ до сповіщень, Overlay (накладання вікон) та доступ до SMS. Саме ці можливості Android забезпечують програмі повний контроль над пристроєм. Контроль та обмеження цих дозволів демонструє найвищу ефективність захисту (динаміка 75–100%), фактично унеможливлюючи роботу ШПЗ.

Поведінкова модель в умовах атаки базується на ірраціональній реакції на соціальну інженерію (рис. 1). Зловмисники використовують основні психологічні тригери: страх, терміновість, бажання швидкого збагачення та ілюзія вигоди. Аналіз наслідків підтверджує, що ігнорування базових правил кібергігієни призводить до значних фінансових втрат, які варіюються від списання коштів за підписки до втрати криптоактивів.

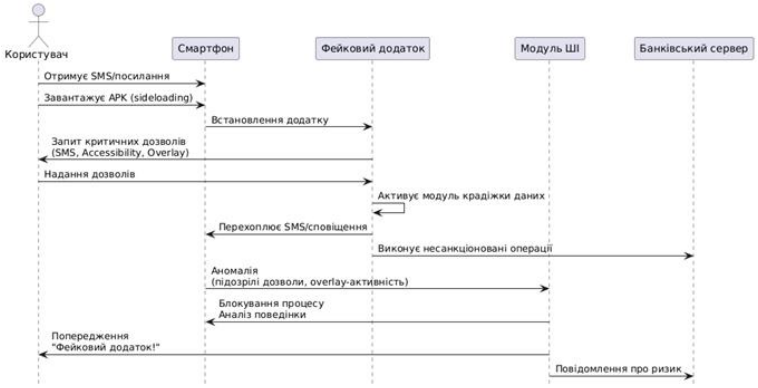

Рисунок 1 – Модель поведінки користувача

На рис. 1 представлено модель взаємодії користувача, шкідливого додатку та системи виявлення загроз на основі ШІ. Модель враховує соціально-психологічні стимули (терміновість, страх, вигода), технічні дії користувача (встановлення додатку, надання доступів) та реакцію системи ШІ.

Застосування систем на основі ШІ, що аналізують поведінку додатків (overlay-активність, аномальні дозволи, підозрілі мережеві запити), дозволяє:

1. знизити ймовірність інфікування на 60–70% [14-15],
2. скоротити час виявлення троянів у 30–40 разів [16],
3. попередити більшість несанкціонованих банківських операцій [17],
4. мінімізувати вплив людського фактору [18].

Таким чином, моделі ШІ є ключовим інструментом для запобігання та блокування атак через фейкові додатки та підвищення загального рівня кіберзахисту мобільних систем.

Висновок. Ефективна протидія фейковим додаткам вимагає переходу від суто технічного захисту до комплексного підходу, що включає поведінковий аналіз. Системи безпеки повинні враховувати патерни поведінки користувача, автоматично, блокуючи спроби надання критичних дозволів підозрілим додаткам, встановленим через сайдлоадінг. Перспективним напрямком подальших досліджень є розробка адаптивних алгоритмів на базі штучного інтелекту, здатних у реальному часі виявляти ознаки соціальної інженерії та попереджати користувача про маніпулятивний вплив ще до моменту завантаження шкідливого контенту.

1. How we kept Google Play & Android app ecosystem safe in 2024 [Електронний ресурс] / Google Security Blog. – 2025. – Режим доступу: https://security.googleblog.com/2025/01/how-we-kept-google-play-android-app-ecosystem-safe-2024.html.
2. Global Mobile Threat Report 2024 [Електронний ресурс] / Zimperium. – 2024. – Режим доступу: https://www.zimperium.com/global-mobile-threat-report/.
3. Doffman Z. Google Play Store Warning—95% Of 'Malicious Apps' Come From Sideloading [Електронний ресурс] / Zak Doffman // Forbes. – 2024. – Режим доступу: https://www.forbes.com/sites/zakdoffman/2024/10/11/google-play-store-new-app-warning-for-samsung-galaxy-s24-pixel-9-pro-android/.
4. 2024 Data Breach Investigations Report (DBIR) [Електронний ресурс] / Verizon Business. – New York, 2024. – 102 p. – Режим доступу: https://www.verizon.com/business/resources/reports/dbir/.
5. Check Point Cyber Security Report 2024 [Електронний ресурс] / Check Point Software Technologies. – 2024. – Режим доступу: https://www.checkpoint.com/cyber-security-report-2024/.
6. Mobile Banking Heists Report 2023 [Електронний ресурс] / Zimperium, Inc. – Dallas : Zimperium Mobile Security, 2023. – 26 p. – Режим доступу: https://lp.zimperium.com/hubfs/MAPS_MTD/WP/FIN/2023_Mobile_Banking_Heists_Report.pdf (дата звернення: 27.11.2025).
7. Mobile Security Report 2024 [Електронний ресурс] / Check Point Research. – Tel Aviv : Check Point Software Technologies, 2024. – 34 p. – Режим доступу: https://research.checkpoint.com/2024/mobile-security-report (дата звернення: 27.11.2025).
8. Trellix Cyber Threat Report 2024 [Електронний ресурс] / Trellix (McAfee Enterprise + FireEye). – Santa Clara : Trellix Labs, 2024. – 48 p. – Режим доступу:

https://www.trellix.com/en-us/security-intel/reports.html (дата звернення: 27.11.2025).
9. ESET Threat Report Q4 2024 [Електронний ресурс] / ESET Research. – Bratislava : ESET Security Research, 2024. – 36 p. – Режим доступу: https://www.eset.com/int/security-report (дата звернення: 27.11.2025).
10. Internet Crime Report 2024 [Електронний ресурс] / Federal Bureau of Investigation (FBI). Internet Crime Complaint Center. – Washington, D.C. : U.S. Department of Justice, 2024. – 38 p. – Режим доступу: https://www.ic3.gov/Media/PDF/AnnualReport/2024_IC3Report.pdf (дата звернення: 27.11.2025).
11. Mobile Threat Report 2024 [Електронний ресурс] / Lookout, Inc. – San Francisco : Lookout Mobile Endpoint Security, 2024. – 32 p. – Режим доступу: https://www.lookout.com/resources/reports/mobile-threat-report (дата звернення: 27.11.2025).
12. Android Malware & Permissions Abuse Analysis Report 2024 [Електронний ресурс] / Zimperium, Inc. – Dallas : Zimperium Labs, 2024. – 18 p. – Режим доступу: https://www.zimperium.com/resources (дата звернення: 27.11.2025).
13. Annual Threat Assessment 2024 [Електронний ресурс] / Google Threat Analysis Group (TAG). – Mountain View : Google LLC, 2024. – 18 p. – Режим доступу: https://blog.google/threat-analysis-group (дата звернення: 27.11.2025).
14. Google blocks 2.36 million malicious apps from Play Store in 2024 [Електронний ресурс] / Bitdefender. – 2024. – Режим доступу: https://www.bitdefender.com/blog/hotforsecurity/google-blocks-2-3-million-malicious-apps-from-play-store-in-2024-what-you-need-to-know.
15. AI in Cyberdefense: Deception at a Glance [Електронний ресурс] / VirusTotal, Google Cloud Security. – 2024. – Режим доступу: https://www.virustotal.com/gui/intelligence-overview.
16. Cost of a Data Breach Report 2024 [Електронний ресурс] / Ponemon Institute, IBM Security. – New York : IBM Corporation, 2024. – 60 p. – Режим доступу: https://www.ibm.com/reports/data-breach.
17. Hi-Tech Crime Trends 2023/2024 [Електронний ресурс] : Annual Report / Group-IB. – Singapore, 2024. – Режим доступу: https://www.group-ib.com/resources/research-hub/hi-tech-crime-trends/.
18. Yevdokymov, S. Overview of modern cyber security solutions for cyber-physical systems [Text] / S. Yevdokymov // Інформаційні управляючі системи і технології (ІУСТ-ОДЕСА-2024): матеріали XII Міжнародної науково-практичної конференції (23-25 вересня 2024 р., Одеса) / вип. ред. В.В. Вичужанін. – Одеса: Видавничий дім "Гельветика", 2024. – С. 55–59.

В.Д. Передер

ВИКОРИСТАННЯ МЕТОДІВ ШТУЧНОГО ІНТЕЛЕКТУ ДЛЯ ВИЯВЛЕННЯ АНОМАЛЬНОЇ АКТИВНОСТІ У ДЕЦЕНТРАЛІЗОВАНИХ СИСТЕМАХ ЕЛЕКТРОННОГО ГОЛОСУВАННЯ

Вступ.

Децентралізовані системи електронного голосування на основі блокчейну розглядаються як інструмент забезпечення прозорості та незмінності виборчих даних, оскільки транзакції голосування захищені криптографічно та доступні для верифікації. Архітектура таких систем передбачає використання смарт-контрактів, IPFS та криптографічних механізмів для реєстрації виборців і підрахунку голосів, що підтверджено в роботі Ghorpade-Aher [1]. Одночасно сучасні дослідження показують, що зростання загроз у виборчих процесах — від бот-активності до динамічних графових атак у мережі — потребує інтеграції методів штучного інтелекту, здатних виявляти аномалії в режимі, близькому до реального часу. На це вказують як міжнародні аналітичні звіти International IDEA, так і наукові публікації у сфері кібербезпеки виборів [2].

Результати досліджень.

Першим ключовим напрямом виявлення аномалій є аналіз транзакцій блокчейну на основі безвчительських моделей. Як показано у великому огляді Cholevas C. та ін., автоенкодери, ізоляційні ліси та алгоритми типу one-class SVM здатні формувати профілі "нормальної" активності у смарт-контрактах та виявляти відхилення, які можуть відповідати масовим підробним голосам або несанкціонованим викликам функцій [4]. Ці методи дозволяють відстежувати нетипові серії транзакцій, різкі зміни у розподілі gas-витрат, чи синхронні операції з новостворених адрес.

Другий напрям — графовий аналіз, що враховує структуру зв’язків між гаманцями, смарт-контрактами та адміністративними вузлами. Моделі на основі графових нейронних мереж (GNN) продемонстрували високу ефективність у задачах виявлення підозрілих кластерів адрес, пов’язаних зі спробами маніпуляції або централізованої бот-активності. Дослідження Sharma S. та ін. показує, що застосування GNN (Inspection-L) значно підвищує точність класифікації потенційно шахрайських вузлів у блокчейнах на прикладі транзакційного графа Elliptic [5]. У контексті e-voting такі моделі можуть ідентифікувати групи адрес, що нехарактерно тісно взаємодіють та голосують у стислі часові вікна.

Третім напрямом є поведінковий аналіз користувачів — часових патернів голосування, взаємодії з гаманцем, пристроїв та спроб автентифікації. У роботі Starovoitenko M. продемонстровано, що саме поведінкові аномалії найчастіше супроводжують спроби маніпуляцій і можуть бути ефективно виявлені за допомогою моделей ШІ, які аналізують

послідовності дій виборців та ризикові патерни у реєстрах [3]. Для децентралізованих голосувань це означає можливість виявлення «нічних хвиль» голосування, активності з однієї підмережі або повторних підозрілих авторизацій.

Висновки.

Інтеграція методів ШІ у децентралізовані системи голосування створює ефективний інструмент для виявлення фальсифікацій, Sybil-атак, аномальних транзакцій та підозрілих поведінкових патернів. Unsupervised-моделі, графові нейронні мережі та поведінковий аналіз доповнюють властивості блокчейну — прозорість і незмінність — гнучкими інструментами адаптивного захисту. Водночас застосування таких систем має відповідати вимогам прозорості алгоритмів, збереження анонімності виборців та регуляторним нормам (наприклад, AI Act), що визначають високий рівень ризиків для виборчих технологій.

1. Ghorpade-Aher J., Dhodapkar N. Decentralized Voting Application Using Blockchain Technology. 2024. с. 336-345. URL: https://www.atlantis-press.com/article/126012684.pdf.
2. International IDEA. Artificial Intelligence in Electoral Management. 2024. с. 5–14 URL: https://www.idea.int/publications/catalogue/artificial-intelligence-electoral-management.
3. Starovoitenko M. Artificial intelligence in the electoral process: new dimensions of cyber threats and cybersecurity. 2025. с. 233–235, 234, 293–303, 315–317, 351–360. URL: https://politpsy.org/index.php/popp/article/download/194/177.
4. Cholevas C. та ін. Anomaly Detection in Blockchain Networks Using Unsupervised Learning: A Survey. Algorithms, 2024. URL: https://doi.org/10.3390/a17050201.
5. Sharma S. та ін. Graph Neural Network-Based Anomaly Detection in Blockchain Network. с. 909–910 URL:2023.https://www.researchgate.net/publication/372058639_Graph_Neural_Network-Based_Anomaly_Detection_in_Blockchain_Network.

А.Р. Коновалова

ЕТИЧНІ СТАНДАРТИ СТВОРЕННЯ КОНТЕНТУ ЗА ДОПОМОГОЮ ШІ: ВПЛИВ ГЕНЕРАТИВНИХ МОДЕЛЕЙ НА МАРКЕТИНГ І СПОЖИВАЧА

Стрімкий розвиток генеративних технологій штучного інтелекту у 2023–2025 роках докорінно трансформував підходи до створення маркетингових комунікацій. Моделі нового покоління - текстові, візуальні, аудіо- та відеогенератори - дали можливість брендам миттєво масштабувати виробництво контенту, персоналізувати рекламні повідомлення та підвищувати ефективність взаємодії зі споживачем. Проте паралельно зі зростанням можливостей посилюються й виклики, пов'язані з етикою, безпекою, достовірністю та відповідальністю при використанні ШІ.

У 2025 році питання прозорості та етичності AI-контенту стає особливо актуальним. З одного боку, маркетингова індустрія активно впроваджує автоматизовані системи для генерації візуалів, сценаріїв, рекламних відео та навіть deepfake-матеріалів, що дозволяє скорочувати витрати та пришвидшувати виробничі процеси. З іншого боку, саме такі інструменти несуть ризики маніпуляції, підміни реальності, викривлення сприйняття та порушення довіри споживача - ключового активу сучасного бренду.

Окрему загрозу становить використання deepfake-технологій у рекламі, коли зображення або голоси людей можуть бути змінені без їхньої згоди, а сам контент - навмисно або ненавмисно вводити аудиторію в оману. У поєднанні з алгоритмічною персоналізацією такі практики можуть впливати на купівельні рішення значно сильніше, ніж традиційні маркетингові інструменти, що ставить під сумнів межу між креативом та маніпуляцією.

Генеративні моделі вже інтегровані в робочі процеси багатьох брендів: від персоналізованих рекламних оголошень до автоматизованих сценаріїв SMM. Водночас споживачі дедалі більше вимагають прозорості: у дослідженнях 2025 року понад 3 з 4 респондентів заявили, що хочуть знати, коли контент створено ШІ. Це створює подвійне завдання для маркетингу - використовувати можливості ШІ, не втрачаючи довіри [1].

Регулювання також рухається - ЄС уже впроваджує AI Act і в 2025 році працює над кодексом практики щодо маркування AI-контенту; це означає, що питання прозорості та маркування можуть незабаром стати обов'язковими. Брендам потрібно готуватися сьогодні, аби завтра не опинитися поза законом чи репутаційно вразливими [2].

Станом на 2025 рік інструменти штучного інтелекту активно інтегрувалися в роботу українських медійників і стали ключовим елементом сучасних маркетингових комунікацій. Найчастіше ШІ використовується для автоматизації рутинних процесів: 62% медійників застосовують його для розшифровки інтерв'ю, 21% - для виправлення текстових помилок, а 19% - для генерації заголовків та лідів. Зростає й роль генеративних інструментів:

31% фахівців створюють за допомогою ШІ візуальний контент, що поступово змінює підхід до креативності, персоналізації та швидкості виробництва контенту [3].

У контексті стрімкого впровадження ШІ постає питання: де саме проходить межа між допустимою креативністю та прихованою маніпуляцією? Генеративні моделі здатні створювати контент, який викликає сильні емоції, підсилює увагу та формує більш персоналізований досвід взаємодії зі споживачем. Однак саме ця емоційність і персоналізація можуть перетворитися на інструмент впливу, що виходить за межі інформування або художнього прийому. Маніпуляція починається там, де штучний інтелект використовується для навмисної підміни реальності, приховування авторства, створення неправдивих або спотворених образів чи повідомлень, що спонукають споживача до рішень, які він би не прийняв за умов повної поінформованості. Тому креативність у роботі з ШІ має залишатися прозорою, а цінність контенту - ґрунтуватися на етичності та достовірності, а не на технічній можливості впливати на свідомість аудиторії.

Водночас ШІ починає формувати не лише технічну, а й змістову частину медіапродукту: 21% журналістів використовують його для написання статей, а 17% - для адаптації матеріалів під різні платформи та аудиторії. Таке розширення функціоналу підсилює ефективність маркетингових комунікацій, однак актуалізує питання етики, прозорості та відповідальності. Окрему роль відіграє аналітика: 14% застосовують ШІ для роботи з великими даними, а 7% - для моніторингу репутації, що свідчить про поступовий перехід до алгоритмічного ухвалення стратегічних рішень. Це підкреслює необхідність формування стандартів безпеки та маркування AI-контенту, щоб запобігти маніпуляціям і зберегти довіру аудиторії [3].

Алгоритми аналізують поведінкові дані, визначають тригери та емоційні патерни й можуть автоматично пропонувати варіанти, спрямовані на підсилення реакцій - від підвищення зацікавленості до стимуляції покупки. Проблема виникає тоді, коли креатив перестає виконувати інформаційну або художню функцію і стає способом прихованого переконання, що не усвідомлюється споживачем. У таких випадках маркетингове повідомлення переходить межу етичності, навіть якщо його формально створено в рамках допустимих інструментів. Тому важливим завданням сучасних фахівців є формування етичних рамок, які б дозволили застосовувати AI-креатив відповідально, без втручання в автономію вибору користувача.

Останні дослідження, зокрема аналітичний звіт Texty, засвідчують стрімке поширення deepfake-маніпуляцій у TikTok, де ШІ використовується для створення фейкових відео з “участю” українських журналісток. Такі ролики, що нерідко набирають десятки тисяч переглядів, підміняють реальність - ведучі нібито озвучують вигадані новини або репліки, які відповідають проросійським наративам, зокрема про “втому Заходу” чи зневіру в українській обороні. Автори дослідження фіксують організоване

масове виробництво deepfake-контенту, який алгоритми TikTok ще й активно просувають, майже не реагуючи на скарги користувачів [4]. Використання впізнаваних облич журналісток легітимізує дезінформацію для широкої аудиторії, водночас створюючи значні репутаційні та психологічні ризики для самих медійниць. У цих умовах експерти наголошують на необхідності вироблення чітких правил і механізмів реагування з боку медіаспільноти та платформ, оскільки відсутність системної політики дозволяє deepfake-технологіям залишатися потужним інструментом інформаційного впливу.

Поширення deepfake-відео демонструє, що технології ШІ вже не просто інструмент для створення креативного контенту - вони стають чинником інформаційної вразливості, здатним масштабувати маніпуляції та підривати довіру до медіа. У ситуації, коли підроблені "новини" легко маскуються під професійний контент, а алгоритми платформи підсилюють їхнє охоплення, суспільство фактично позбавлене можливості відрізнити достовірну інформацію від штучно згенерованої. Це актуалізує нагальну потребу у впровадженні чітких маркерів та індикаторів безпеки для AI-контенту, які б дозволяли користувачу ідентифікувати штучно створені матеріали, оцінювати їхню надійність та уникати прихованого впливу. Формування таких стандартів - один із ключових етапів підвищення прозорості та відповідальності використання ШІ в медіа та маркетингових комунікаціях.

В Україні почали формуватися конкретні індикатори безпеки для AI-контенту на основі добровільних етичних стандартів і саморегуляції. Так, 14 провідних IT-компаній підписали Добровільний кодекс поведінки щодо етичного використання ШІ, який включає принципи прозорості, контролю з боку людини, приватності та безпеки [5]. Міністерство юстиції додатково розробило практичні рекомендації з безпечного використання ШІ, що передбачають оцінку ризиків, підзвітність і прозорість алгоритмів [6].

Індикатори безпеки для AI-контенту формуються навколо трьох ключових принципів: прозорості, підзвітності та перевірюваності. Перш за все, безпечний контент має бути чітко маркований як створений або модифікований штучним інтелектом, щоб аудиторія не плутала його з оригінальними матеріалами. Другий індикатор - можливість відстежити джерело та процес створення матеріалу: хто ініціював запит, хто редагував результат, які моделі використовувалися. Третій - наявність внутрішніх політик редакцій чи компаній, які вимагають перевірки фактів, контролю якості та регулярного аудиту ризиків, пов'язаних із використанням ШІ. Усе це поступово формує культуру відповідального й безпечного використання AI-інструментів у медіа та креативних індустріях.

Узагальнюючи, сучасні креативні інструменти штучного інтелекту відкривають величезні можливості для медіа та рекламної індустрії, проте разом із цим загострюють питання відповідальності й етичних меж. Вибір між творчістю та маніпуляцією сьогодні визначається не лише намірами автора, а й тим, наскільки прозоро і чесно він комунікує з аудиторією.

Приклади масового поширення deepfake-контенту в українському інформаційному просторі демонструють, наскільки легко інновації можуть бути використані для дезінформації та підриву довіри. Саме тому розвиток індикаторів безпеки для AI-контенту - від маркування до перевірки достовірності - стає ключовим елементом інформаційної гігієни. Лише поєднання технологічного прогресу з етичними стандартами дозволить забезпечити захист суспільства та зберегти цінність правдивої комунікації у цифрову епоху.

1. Trust: transparency earns trust, and right now there isn't enough of either. *Baringa*. URL: https://www.baringa.com/en/insights/balancing-human-tech-ai/trust/ (дата звернення: 25.02.2025).
2. High-level summary of the AI Act. *EU Artificial Intelligence Act*. URL: https://artificialintelligenceact.eu/high-level-summary/ (дата звернення: 27.02.2024).
3. Машкова Я. Українські медіа та штучний інтелект. Як редакції залучають ШІ для створення контенту?. *Інститут масової інформації*. URL: https://imi.org.ua/monitorings/ukrayinski-media-ta-shtuchnyj-intelekt-yak-redaktsiyi-zaluchayut-shi-dlya-stvorennya-kontentu-i62217 (дата звернення: 28.06.2024).
4. Троян В. ШІ у TikTok масово генерує фейки з українськими медійницями. *Інститут масової інформації*. URL: https://imi.org.ua/news/doslidzhennya-texty-shi-u-tiktok-masovo-generuye-feyky-z-ukrayinskymy-zhurnalistkamy (дата звернення: 19.11.2025).
5. Ковальова А. Низка великих українських IT-компаній підписала кодекс використання ШІ у своїх продуктах – хто долучився. *AIN*. URL: https://ain.ua/2024/12/17/kodeks-vikoristannia-si-u-produktax/ (дата звернення: 17.12.2024).
6. Мін'юст підготував рекомендації щодо безпечного використання штучного інтелекту. *Центр Дослідження Законодавства України*. URL: https://ualaw.org/minyust-pidgotuvav-rekomendacziyi-shhodo-bezpechnogo-vykorystannya-shtuchnogo-intelektu/ (дата звернення: 04.08.2025).

В.С. Третяк

АРХІТЕКТУРНІ АСПЕКТИ ЗАБЕЗПЕЧЕННЯ ВІДМОВОСТІЙКОСТІ ТА КОНФІДЕНЦІЙНОСТІ ДІАЛОГОВИХ СИСТЕМ НА БАЗІ LLM

Масштабна інтеграція великих мовних моделей (LLM) у системи автоматизованої технічної підтримки та корпоративні комунікаційні платформи створює нові виклики для архітекторів інформаційних систем. Якщо на етапі зародження технології пріоритетом була якість генерації тексту, то в умовах промислової експлуатації на перший план виходять питання інформаційної безпеки. Забезпечення тріади CIA (конфіденційність, цілісність, доступність) для LLM-систем ускладнюється їхньою вразливістю до специфічних векторів атак, таких як ін'єкції промптів (Prompt Injection) та атаки на виснаження ресурсів (Resource Exhaustion). Нездатність системи ефективно розподіляти навантаження під час пікових запитів може призвести до відмови в обслуговуванні (DoS), що є критичним для сервісів реального часу [1].

Першочерговою проблемою при проєктуванні є вибір моделі розгортання: використання хмарних API (наприклад, OpenAI, Anthropic) або імплементація локальних моделей (LLaMA, Mistral) у власному контурі (On-premise).

У рамках дослідження проведено порівняльний аналіз цих підходів через призму безпеки даних та стійкості до зовнішніх впливів. Аналіз показав, що хмарні рішення, хоч і забезпечують високу якість генерації без необхідності підтримки власної інфраструктури, несуть значні ризики конфіденційності. Передача даних третій стороні унеможливлює повний контроль над їх життєвим циклом, що є критичним для державних та фінансових установ [2].

Результати порівняння ключових характеристик наведено у табл. 1.

Таблиця 1 – Порівняльний аналіз безпекових характеристик архітектур

Характеристика безпеки	Хмарна архітектура (API)	Локальна архітектура (On-premise)
Конфіденційність (Privacy)	Низька. Ризик перехоплення даних та їх використання для донавчання моделей провайдера.	Висока. Дані циркулюють виключно в захищеному периметрі організації.
Стійкість до DoS-атак	Забезпечується провайдером. Проте існує ризик блокування акаунта через підозрілу активність.	Залежить від локальних ресурсів. Висока вразливість до перевантаження GPU/TPU.

Характеристика безпеки	Хмарна архітектура (API)	Локальна архітектура (On-premise)
Доступність (Availability)	Змінна. Залежить від стабільності інтернет-каналу та навантаження на сервери API.	Контрольована. Залежить від внутрішньої інфраструктури та черг запитів.
Контроль доступу (IAM)	Обмежений. Базується на API-ключах.	Повний. Інтеграція з LDAP, Active Directory та рольовими моделями.

Як видно з таблиці, локальне розгортання є кращим для забезпечення конфіденційності, але воно перекладає ризики забезпечення доступності на власну інфраструктуру. Саме обмеженість локальних ресурсів робить систему вразливою до сплесків навантаження.

Моделювання відмовостійкості. Для оцінки стійкості локальної системи до перевантажень було застосовано математичне моделювання на основі теорії масового обслуговування. Система розглядається як модель M/M/n, де вхідний потік запитів є пуассонівським, а час генерації відповіді розподілено експоненційно. Це дозволяє формалізувати залежність між інтенсивністю запитів та часом перебування в системі.

Чисельні експерименти, проведені в середовищі Python, показали, що система має чітко виражену «точку відмови». При наближенні коефіцієнта завантаження до одиниці, час очікування зростає експоненційно [3].

Експериментальні дані демонструють таку динаміку:

1. При використанні 2 серверів: критичне зростання затримок (понад 10 секунд) спостерігається вже при інтенсивності 6 запитів/с. Така система є нестійкою до навіть незначних DDoS-атак.

2. При масштабуванні до 8 серверів: система здатна стабільно обробляти потік до 15 запитів/с із затримкою менше 1 секунди.

Це підтверджує, що статична конфігурація ресурсів є небезпечною. Без механізмів автоскейлінгу локальна LLM може бути паралізована навіть легітимним піковим трафіком, що порушує принцип доступності.

Запропоновані архітектурні рішення. Для підвищення рівня безпеки та надійності запропоновано впровадження гібридної архітектури захисту:

1. Семантичний фаєрвол: використання легковагових моделей-класифікаторів (наприклад, BERT) на вході системи. Це дозволяє відфільтрувати до 70% сміттєвого трафіку та шкідливих промптів ще до того, як вони потраплять на обробку до ресурсоємної LLM.

2. Пріоритезація черг: впровадження алгоритмів планування, які надають пріоритет коротким транзакційним запитам, запобігаючи блокуванню черги «важкими» генеративними задачами [4].

3. Динамічне керування потужностями: моніторинг довжини черги в реальному часі та автоматичне розгортання квантованих версій моделей при досягненні критичних показників.

Висновки. Побудова безпечної діалогової системи вимагає балансу між конфіденційністю даних та відмовостійкістю. Локальне розгортання моделей нівелює ризики витоку інформації, проте вимагає впровадження інтелектуальних механізмів розподілу навантаження. Результати моделювання доводять, що без адаптивного керування чергами та масштабування ресурсів система залишається вразливою до атак на доступність.

1. Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., & Zhang, Y. (2024). A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly. https://www.semanticscholar.org/paper/A-Survey-on-Large-Language-Model-(LLM)-Security-and-Yao-Duan/383c598625110e0a4c60da4db10a838ef822fbcf.
2. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). Llama: Open and efficient foundation language models. https://www.researchgate.net/publication/368842729_LLaMA_Open_and_Efficient_Foundation_Language_Models.
3. Miao, X., Oliaro, G., Zhang, Z., Cheng, X., Wang, Z., Zhang, Z., Wong, R. Y. Y., Zhu, A., Yang, L., Shi, X., Shi, C., Chen, Z., Arfeen, D., Abhyankar, R., & Jia, Z. (2024). SpecInfer: Accelerating Generative LLM Serving with Speculative Inference and Token Tree Verification. https://dl.acm.org/doi/10.1145/3620666.3651335.
4. Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., & Stoica, I. (2023). Efficient memory management for large language model serving with pagedattention. https://dl.acm.org/doi/10.1145/3600006.3613165.

Я.В. Денісова, П.О. Самусь

ШТУЧНИЙ ІНТЛЕКТ І БЕЗПЕКА

В останній час, завдяки стрімкому науковому розвитку, у повсякденне життя активно впроваджуються технології штучного інтелекту (ШІ). Це стосується не лише окремих громадян, а й підприємств різних секторів економіки. Сучасні тенденції формують нові стандарти цифрової трансформації у більшості галузей, включно з харчовою промисловістю. Підприємства, що прагнуть інноваційного розвитку, стикаються з необхідністю інтеграції систем ШІ для підвищення ефективності, покращення взаємодії з користувачами та забезпечення конкурентоспроможності.

У контексті швидкого розвитку цифрових технологій питання безпеки використання ШІ постає особливо гостро. Підприємства, впроваджуючи інтелектуальні системи, стикаються з низкою викликів: інтеграцією ШІ за обмежених ресурсів, ризиками кіберзагроз, збільшенням витрат на імплементацію, технічною складністю інтеграційних процесів і необхідністю дотримання сучасних нормативно-правових вимог.

Враховуючи міжнародні тенденції та глобальний характер цифровізації, питання безпечності технологій ШІ стає важливим також у контексті захисту персональних даних та запобігання дискримінаційним або упередженим алгоритмічним рішенням. Водночас світова практика демонструє значні переваги впровадження штучного інтелекту, зокрема чат-ботів: зменшення кількості помилок через людський фактор, відсутність упередженого ставлення, зниження ймовірності фальсифікації інформації, оптимізація операційних витрат та підвищення рівня обслуговування клієнтів.

Матеріалами дослідження є сучасні наукові публікації та міжнародні стандарти, що розкривають питання безпеки, надійності та етичності застосування технологій штучного інтелекту.

Методологічною основою дослідження є:

1. Системний підхід, що дозволяє розглядати впровадження ШІ в діяльність підприємства як комплексну взаємодію технічних, організаційних і правових елементів.

2. Метод верифікації та валідації моделей ШІ, застосований для визначення ключових технічних і функціональних загроз.

3. Метод експертної оцінки, що дав змогу узагальнити практичні рекомендації щодо впровадження чат-ботів у бізнес-процеси на основі думок фахівців у сфері кібербезпеки та цифрової трансформації.

У роботі враховано рекомендації Європейського агентства з кібербезпеки ENISA щодо безпечного використання ШІ в цифрових сервісах [1], а також сучасні методики оцінювання прозорості та валідації моделей машинного навчання [2].

Проведений аналіз дав змогу визначити ключові технічні, організаційні, правові та економічні аспекти впровадження ШІ у діяльність інноваційних підприємств харчового сектору.

1. Технічні ризики. Встановлено, що моделі ШІ мають високу чутливість до атак типу *adversarial*, які можуть спричинити спотворення результатів обробки даних [1]. Чат-боти вразливі до маніпулятивних або шкідливих запитів (prompt injection), що може призвести до неправильних відповідей чи небажаної поведінки системи. Додатковим ризиком є недостатня якість навчальних даних, що впливає на коректність роботи моделі.

2. Організаційні виклики. Інтеграція ШІ потребує значних початкових витрат на розробку, тестування та технічну підтримку систем. Нестача кваліфікованих фахівців з кібербезпеки та даних ускладнює впровадження. Важливим бар'єром залишається також внутрішній психологічний та корпоративний спротив — недовіра працівників, побоювання змін у кадровій структурі та небажання адаптуватися до нових технологій [2].

3. Ефективність інтеграції ШІ. Використання чат-ботів сприяє зменшенню помилок, пов'язаних із людським фактором, та забезпечує швидшу реакцію на запити клієнтів. Завдяки алгоритмам обробки природної мови (NLP) чат-боти видають більш точні та змістовні відповіді й здатні формувати персоналізовані рекомендації на основі мінімального набору вхідних даних [3]. Стабільність і передбачуваність роботи таких систем підвищують якість обслуговування.

4. Позитивний економічний ефект. Виявлено, що впровадження ШІ дозволяє суттєво скоротити витрати на ручну обробку замовлень, збільшити пропускну здатність систем, автоматизувати рутинні операції та прискорити обслуговування на 40–60% залежно від типу запитів [4]. Це підтверджує економічну доцільність застосування ШІ навіть у малих інноваційних підприємствах.

5. Правові та нормативні обмеження. Використання ШІ потребує дотримання вимог щодо захисту персональних даних (GDPR, Закон України «Про захист персональних даних») та прозорості алгоритмічних рішень. Нині відсутні уніфіковані стандарти відповідальності за дії автономних систем, що ускладнює регулювання їх застосування на практиці [5].

6. Ключові фактори безпеки. Для забезпечення надійності роботи ШІ необхідні регулярна валідація моделей, фільтрація заборонених і шкідливих запитів, мінімізація даних, що збираються, а також постійний моніторинг аномальної поведінки системи [6]. Ці заходи дозволяють суттєво знизити ризики некоректної роботи ШІ та зміцнити кіберзахист підприємства.

Отримані результати свідчать, що інтеграція ШІ, зокрема чат-ботів, здатна значно підвищити ефективність бізнес-процесів підприємства та скоротити операційні витрати. Водночас успішне впровадження вимагає забезпечення високого рівня кіберзахисту, регулярної валідації моделей і дотримання нормативних вимог. Це робить безпеку та надійність систем ШІ ключовими елементами цифрової трансформації харчових підприємств.

1. Abomakhelb A., Jalil K. A., Buja A. G., Alhammadi A., Alenezi A. M. A Comprehensive Review of Adversarial Attacks and Defense Strategies in Deep Neural Networks [електронний ресурс]. – Technologies. – 2025. – Vol. 13, No. 5. – Article 202. – Режим доступу: https://doi.org/10.3390/technologies13050202.
2. ENISA. *Artificial intelligence cybersecurity challenges* [електронний ресурс]. – Режим доступу: https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges.
3. Hariguna T., Ruangkanjanases A. Assessing the impact of artificial intelligence on customer performance: a quantitative study using partial least squares methodology [електронний ресурс]. – Data Science and Management. – 2024. – Vol. 7, No. 3. – P. 155–163. – DOI: 10.1016/j.dsm.2024.01.001. – Режим доступу: https://www.sciencedirect.com/science/article/pii/S2666764924000018.
4. McKinsey Global Institute. *The economic potential of generative AI: the next productivity* frontier [електронний ресурс]. – Режим доступу: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai.
5. European Commission. *Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (AI Act)* [електронний ресурс]. – Режим доступу: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.
6. OECD. *AI principles and risk management guidelines* [електронний ресурс]. – Режим доступу: https://oecd.ai/en/ai-principles.

А.Ю. Волинець

ШІ ОПТИМІЗАЦІЯ РЕАКТИВНОГО ЯДРА НА ОСНОВІ DAG-МОДЕЛІ З ВИКОРИСТАННЯМ ТЕОРІЇ ПУЧКІВ

У сучасних реактивних і розподілених системах штучного інтелекту особливо важливими є узгодженість станів, коректність поширення змін та контроль за залежностями між обчислювальними вузлами. Порушення причинно-наслідкових зв'язків або некоректна мутація станів можуть призвести до помилок або небезпечної поведінки системи, що є критичним для високонадійних застосувань у режимі реального часу.

У роботі запропоновано підхід до оптимізації реактивного ядра шляхом моделювання його у вигляді орієнтованого ациклічного графа (DAG) із накладеною пучковою структурою. Така модель забезпечує формальний опис локальних станів, правил їх трансформації та умов глобальної узгодженості. Використання теорії пучків дозволяє поєднати топологічну структуру графа з алгебраїчним апаратом аналізу станів, що забезпечує чітке формальне трактування залежностей у реактивних обчисленнях.

Реактивне ядро подається як частково впорядкована множина, індукована DAG. Кожному вузлу відповідає локальний простір станів (stalk), у якому зберігаються дані, метадані та інваріанти безпеки. Кожне ребро визначає відображення (restriction map), що задає правило переходу або передачі інформації між вузлами. Глобальна конфігурація системи визначається як глобальна секція пучка — набір локальних станів, що не суперечать обмеженням. Відсутність глобальної секції інтерпретується як сигнал некоректності або небезпеки [1].

1. Структурна оптимізація DAG.

Пучково-посетна модель дає змогу виявляти топологічно еквівалентні фрагменти графа, мінімізувати надлишкові зв'язки та скорочувати кількість вузлів без втрати семантики. Це зменшує розмір внутрішнього представлення, обсяг пам'яті та кількість необхідних оновлень при зміні вхідних даних.

2. Аналіз поширення оновлень через sheaf-дифузію.

Процес пропагації змін у реактивному ядрі інтерпретується як дискретна дифузія на графі. На основі відповідного коцепного комплексу будується sheaf-лапласіан, спектральні властивості якого дають змогу оцінити швидкість повернення системи до стабільного стану та вплив локальних збурень на глобальну конфігурацію. Власні значення слугують мірою інерційності системи, а власні вектори — "режимами колективної динаміки".

3. Інтеграція безпекових інваріантів у пучкову модель.

Безпекові обмеження формалізуються як частина структури пучка та простору допустимих секцій. Якщо конфігурація локальних станів не допускає жодної глобальної секції, така конфігурація означає порушення

політик безпеки. Це дає можливість формально виявляти небезпечні сценарії ще до реалізації в програмному коді, що є критичним для автономних систем, інтелектуального керування та енергетичних об'єктів.

Узагальнена модель реактивного ядра подається у вигляді кортежу (1):

$$R = (G, D, \Phi), \quad (1)$$

де G — орієнтований ациклічний граф обчислень, що задає причинно-наслідкову структуру системи;

D — пучкова структура на G, яка визначає локальні простори станів у вузлах та відображення між ними;

Φ — система обмежень (логічних, алгебраїчних, безпекових), яка визначає множину допустимих глобальних секцій графа та відсікає небажані конфігурації [2].

Такий формальний опис дає змогу поєднати топологічні властивості графа (наявність циклів, компонент зв'язності, «вузьких місць») з алгебраїчними операціями над станами і дає підґрунтя для подальшого використання методів спектрального аналізу, гомології та когомології пучків для оцінювання складності, стійкості та надійності реактивних ядер [3].

Запропонований підхід поєднує теоретичну формалізацію з практичним застосуванням та має перспективи використання у високонадійних і критичних системах, зокрема в енергетиці, системах оборонного призначення, автономних агентах, інтелектуальних середовищах керування та складних інформаційно-аналітичних платформах.

Результати дослідження можуть бути використані для проєктування безпечних реактивних систем, підвищення стабільності складних обчислювальних середовищ, формування нових архітектур на основі sheaf-моделей, а також для побудови спеціалізованих нейромережевих архітектур, де механізми дифузії та узгодженості явно задані на рівні моделі [4].

Ключові слова: штучний інтелект, DAG, реактивні системи, теорія пучків, безпечні обчислення, узгодженість, оптимізація.

1. Ayzenberg, A., Magai, G., Gebhart, T., Solomadin, G. Sheaf theory: From deep geometry to deep learning (2025) // arXiv preprint. – arXiv:2502.15476.
2. Kipf, T. N., Welling, M. Semi-Supervised Classification with Graph Convolutional Networks (2017) // Proceedings of the International Conference on Learning Representations (ICLR 2017).
3. Robinson, M. Topological Signal Processing (2014). – Berlin: Springer.
4. Leray, J. Sur les espaces fibrés et les homologies des espaces (1950). – Paris: Secrétariat mathématique.

В.В. Мохор, О.В. Цуркан, Р.П. Герасимов, В.П. Яшенков,
Т.М. Клименко, В.І. Павловська

АНАЛІЗ ВИКОРИСТАННЯ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ В АТАКАХ СОЦІАЛЬНОЇ ІНЖЕНЕРІЇ

Відповідно до [1, 2], великі мовні моделі (англ. Large Language Models, LLM) є окремою категорією моделей глибокого навчання. Вони навчаються на великих обсяг даних. Цим обумовлюється їхня здатність генерувати контент для виконання різноманітних завдань [2]. Практичне реалізування такої можливості досягається з огляду на особливості побудови великих мовних моделей. Передумовою і водночас запорукою успішності їх застосовності є навчання на основі великих обсягів даних. Після цього здійснюється їх удосконалювання з огляду на специфіку завдань. Наприклад [2], генерування, узагальнювання тексту; створювання чат-ботів, генерування коду та аналізування настроїв. Серед них виокремлюється і діяльність стосовно порушення властивостей інформації [1]. При цьому одним з найпростіших способів виконання дій у її межах є надання з боку «користувача» відповідних підказок (англ. Prompt Engineering) [2]. Тому аналізування використання великих мовних моделей в атаках соціальної інженерії є актуальним завданням.

Використання великих мовних моделей умовно можна поділити на два напрями. Першим з них об'єднуються можливості традиційних реалізацій, наприклад [1]: GPT, DeepSeek, Gemini, LLaMA. Другим – «спеціалізовані» або «зловмисні» великі мовні моделі, наприклад [1]: WormGPT, FraudGPT, VirusGPT, PoisonGPT. Перший спосіб характеризується орієнтованістю на спроби практичного використання типових можливостей [1, 3], наприклад, генерування тексту. Насамперед це стосується моделювання сценаріїв взаємодії з потенційною жертвою – працівником організації. Отриманими при цьому правдоподібними сценаріями приховується реалізування загрози соціальної інженерії. Особливо такі прояви спостерігаються при фішингу як окремого її різновиду. Так, згенерованим шаблонам електронних листів притаманна стилістична грамотність, відсутність помилок, терміновість дій і обмеженість часу. Результативність даних дій обумовлюється умінням зловмисника формулювати промти. Обмеженість використання першого способу для атак соціальної інженерії зводиться до дотримання етичних норм, тобто відмовляння від промтів зловмисного характеру. Хоча залежно від контексту іноді спостерігаються і приклади двозначності окремих тлумачень.

Виокремленого обмеження позбавлене використання «спеціалізованих» великих мовних моделей у межах другого способу [1, 3–5]. Оскільки створення кожної з них орієнтоване на виконання персоналізованих завдань. Зокрема реалістичні діалоги генеруються без втручання людини за допомогою WormGPT [1, 3]. Це доповнюється існуванням можливості отримання нового контенту незалежно від наявних наборів даних. Етичних обмежень позбавлене і використання FraudGPT [1, 3]. Його застосовність зводиться до генерування контенту оманливого характеру та, як наслідок, реалізування

життєвого циклу атаки соціальної інженерії. Створення вкладень до фішингових електронних листів можливе завдяки Virus-GPT [4]. Оскільки дана персоналізована велика мовна модель призначена для генерування шкідливого програмного забезпечення. Автоматизування атак соціальної інженерії до того ж досягається і завдяки використанню PoisonGPT [4, 5]. Її практичне застосування орієнтоване перш за все на отримання і поширення спотвореної інформації. Це дозволяє робити хибні судження, спонукати до хибних дій і, як наслідок, вводити в оману працівника організації.

Як окремий напрям використання великих мовних моделей варто виокремити їхню орієнтованість на протидіяння атакам соціальної інженерії [1]. Наприклад, створювання агентів-нападників і агентів-жертв, виявляння та ізолювання спроб реалізування загроз соціальної інженерії, підвищування обізнаності. Завдяки цьому можливе моделювання потенційних сценаріїв атак соціальної інженерії [1, 6]. При цьому формуються психологічні профілі з огляду на уразливості до маніпулювань і впроваджуються відповідні заходи протидії.

Отже, використання великих мовних моделей в атаках соціальної інженерії характеризується багатоаспектністю – і нападання, і протидіяння. Особливість типових рішень пов'язується з необхідністю дотримання перш за все етичних норм. Так унеможливлюється застосовність великих мовних моделей зловмисниками для реалізування загроз соціальної інженерії. Подолання виокремленого обмеження досягається розробленням і використанням персоналізованих рішень. Це дозволяє зловмисникам обходити етичні норми і, як наслідок, автоматизовувати реалізування загроз соціальної інженерії.

1. Мохор В. В., Цуркан О. В., Герасимов Р. П., Клименко Т. М., Яшенков В. П. Соціоінженерний аспект використання генеративного штучного інтелекту. *Кібербезпека енергетики* : матеріали науково-практичної конференції Інституту проблем моделювання в енергетиці ім. Г.Є. Пухова Національної академії наук України (Київ, 29 травня 2024 р.). Київ : ІПМЕ ім. Г. Є. Пухова НАН України, 2024. С. 114, 115.
2. What are large language models (LLMs)? URL: https://www.ibm.com/think/topics/large-language-models (accessed on: 27.11.2025).
3. Falade P.V. Decoding the Threat Landscape : ChatGPT, FraudGPT, and WormGPT in Social Engineering Attacks, *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*. 2023. Vol. 9, Iss. 5. P. 185–198. DOI: https://doi.org/10.32628/CSEIT2390533.
4. Hill J. The Rise of Malicious AI : 5 Key Insights from an Ethical Hacker. URL: https://abnormal.ai/blog/key-insights-ethical-hacker (accessed on: 27.11.2025).
5. Huynh D., Hardouin J. PoisonGPT : How We Hid a Lobotomized LLM on Hugging Face to Spread Fake News. URL: https://blog.mithrilsecurity.io/poisongpt-how-we-hid-a-lobotomized-llm-on-hugging-face-to-spread-fake-news/ (accessed on: 27.11.2025).
6. Kumarage T. et al. Personalized Attacks of Social Engineering in Multi-turn Conversations: LLM Agents for Simulation and Detection. Cryptography and Security. 2025. arXiv:2503.15552. DOI: https://doi.org/10.48550/arXiv.2503.15552.

К.Е. Вітвіцька, В.В. Кочина

ІЛЮСТРАЦІЯ РЕАЛЬНИХ ЗАСТОСУВАНЬ ШТУЧНОГО ІНТЕЛЕКТУ В РІЗНИХ СФЕРАХ

Ілюстрація реальних застосувань (ШІ) в різних сферах життя стає дедалі актуальнішою в сучасному світі. Штучний інтелект вже активно впроваджується в медицину, фінанси, освіту, промисловість та багато інших галузей, змінюючи підходи до вирішення традиційних завдань. Штучний інтелект – одна з важливих обставин, що дозволяє нам оцифрувати наш світ. Його забезпечують різні галузі: державне управління, медицина, наука, оборона, освітній процес, бізнес-сфера. Використання штучного інтелекту передбачає автоматизацію найбільш трудомістких процесів, підвищення точності аналізу, швидшу обробку, нові форми послуг.

Впровадження технологій також створює нові проблеми етики, безпеки та відповідальності. По-перше, однією з найактивніших сфер штучного інтелекту – державне управління. Міністерства Європейського союзу та США використовують технологію для проектування рішень, аналізу великих обсягів інформації, прогнозування ризиків, оптимізації адміністративних послуг. Це GPT@EC, моніторинг фінансів, US digital permit approval system та багато іншого. В Україні відбувається WisX center of excellence запуск, «Інтелектуальні асистенти» і розробка штучного інтелекту для державного язика «Дія». По-друге, важливу роль штучний інтелект відіграє в медицині [1]. Він допомагає аналізувати медичні знімки, прогнозувати той чи інший патологічний процес, прискорювати дослідження нових лікарських препаратів, вести мініторінг обладнання. Наприклад, алгоритми машинного навчання можуть аналізувати рентгенівські знімки, виявляючи ознаки пневмонії або раку на ранніх стадіях. Це дозволяє лікарям швидше та точніше ставити діагнози, що в свою чергу підвищує шанси на успішне лікування. Етичні виклики, пов’язані з використанням медичних даних для навчання моделей штучного інтелекту, потребують комплексного підходу, що враховує сучасні технологічні можливості, етичні принципи та інтереси пацієнтів [2]

У фінансовій сфері штучний інтелект допомагає в управлінні ризиками, виявленні шахрайства та автоматизації торгівлі. Багато банків і фінансових установ використовують алгоритми для аналізу транзакцій у реальному часі, що дозволяє виявляти підозрілі операції та запобігати фінансовим втратам.

Освіта також зазнає змін завдяки ШІ. Системи адаптивного навчання, які використовують штучний інтелект, можуть аналізувати прогрес студентів і пропонувати індивідуальні навчальні плани. Це дозволяє кожному учневі отримувати знання в зручному для нього темпі, що підвищує ефективність навчального процесу.

Штучний інтелект допомагає органам публічного управління швидше аналізувати великі обсяги даних і приймати обґрунтовані рішення. Завдяки

алгоритмам машинного навчання можливо передбачати потреби громадян та оптимізувати розподіл ресурсів. Використання чат-ботів та віртуальних помічників покращує комунікацію між державними установами та громадянами. ШІ сприяє автоматизації рутинних процесів, знижуючи бюрократію та підвищуючи ефективність роботи державних служб [3]. Впровадження штучного інтелекту в публічне управління вимагає дотримання етичних стандартів і захисту персональних даних громадян.

У промисловості штучний інтелект застосовується для оптимізації виробничих процесів, прогнозування технічних збоїв та управління ланцюгами постачання. Наприклад, системи, що базуються на ШІ, можуть аналізувати дані з датчиків на виробничих лініях та виявляти аномалії, що дозволяє зменшити витрати на обслуговування та підвищити продуктивність.

Ефективна інтеграція ШІ вимагає цифрової грамотності: від формулювання чіткого запиту до перевірки даних та критичного аналізу логіки відповідей, запобігання автоматичному оприлюдненню згенерованого тексту як опублікованого; ШІ є інструментом для підвищеного підсилювання, але не замінює його.

Таким чином, ілюстрація реальних застосувань в різних сферах підтверджує його потенціал у трансформації традиційних процесів та покращенні якості життя. З розвитком технологій можна очікувати ще більшої інтеграції ШІ в повсякденне життя, що відкриває нові можливості для інновацій та вдосконалення.

1. Штучний інтелект у медицині: переваги та недоліки // Go Online Law School. – URL: https://www.goonlinelawschool.com/post/ai-in-medicine-pros-cons (дата звернення: 23.11.2025).
2. Інформаційні системи та технології в медицині (ISM-2018): зб. наук. праць І Міжнар. наук. - практ. конф., 28-30 листоп. 2018 р. / М-во освіти і науки України, Харків. нац. ун-т радіоелектроніки, Укр. Асоціація "Комп'ютерна Медицина"; редкол.: В. М. Левикін та ін. – Харків: Друкарня Мадрид, 2018. – 300 с.: іл. – ISBN 978-617-7683-32-1.
3. Штучний інтелект у публічному управлінні: поради, факти і практичні рішення // Публікації Порталу EU4PAR. – 08.04.2025. – URL: https://par.in.ua/information/publications/420-shtuchnyi-intelekt-u-publichnomu-upravlinni-porady-fakty-i-praktychni-rishennia (дата звернення: 23.11.2025).

І.М. Чвир, М.Є. Сидоренко

РОЛЬ ШТУЧНОГО ІНТЕЛЕКТУ У ВДОСКОНАЛЕННІ НАВЧАННЯ АНГЛІЙСЬКОЇ МОВИ У ВИЩІЙ ОСВІТІ

Сучасна освіта дедалі активніше впроваджує новітні технології, серед яких ключову роль нині виконує штучний інтелект, який поєднує технологічні інновації з педагогічним підходом. Викладачі та студенти стикаються з новими можливостями - адаптивними платформами для англійської, чат-ботами, автоматичними системами зворотного зв'язку. Метою цієї роботи є: проаналізувати, яким чином ШІ може впливати на навчання англійської мови у вищій освіті; з'ясувати його переваги та виклики; і запропонувати практичні рекомендації для викладачів та закладів. «Інтеграція штучного інтелекту та авторських методик викладачів дозволяє автоматизувати рутинні завдання в навчанні, вдосконалити мовну компетенцію, збільшити лексичний запас, покращити вимову та навчитись вільно реалізовувати здобуті мовленнєві вміння в різних ситуаціях» [1, с. 133].

Користь застосування ШІ у навчанні англійської мови по-перше, ШІ дозволяє персоналізувати процес вивчення англійської: системи аналізують відповіді студента, визначають слабкі місця (наприклад, вимова, лексику, граматику) і пропонують вправи, підлаштовані під потреби. Наприклад, дослідження показують, що у вищій освіті використання ШІ-інструментів для англійської мови сприяє покращенню навичок письма й вимови.

По-друге, ШІ може забезпечити миттєвий зворотний зв'язок: коли студент виконує вправу з англійської, система одразу показує помилки, дає підказки – це прискорює навчання й підвищує ефективність.

По-третє, ШІ-інструменти розширюють можливості онлайн-навчання англійської, особливо коли студенти перебувають у різних локаціях або працюють за гібридною/дистанційною формою: чат-боти, віртуальні репетитори, інтерактивні вправи - все це робить навчання більш гнучким.

Ці переваги дають змогу значно підвищити мотивацію студентів до вивчення англійської мови, створити середовище, орієнтоване на їхній темп і стиль навчання, а не універсально для всіх.

Питання та виклики впровадження ШІ у навчанні англійської мови Хоча переваг багато, слід враховувати кілька важливих проблем. Студент може звикнути до системи, що «підказує», і не розвивати навички автономного мислення або самостійного формулювання думок англійською мовою – особливо важливо для розвитку мовної компетенції. «Зазначено, що впровадження ШІ є важливою складовою цифрової трансформації освіти, яка передбачає персоналізацію освітнього процесу, автоматизацію рутинних завдань викладачів, розширення доступу до якісних освітніх послуг» [2, с. 27]. Інструменти штучного інтелекту можуть генерувати відповіді або

тексти, але не завжди враховують культурний контекст, різні варіанти англійської чи помилки, характерні для конкретного студента – це може призводити до поверхневого навчання.

Приватність та безпека даних: для персоналізації навчання англійської мови системи збирають дані про студентів – їх відповіді, час роботи, помилки – заклад мусить забезпечити захист цих даних. Зменшення живої взаємодії між викладачем та студентом: коли навчання сильно автоматизовано, ризикує бути втраченим живий обмін, обговорення, розвиток комунікативних навичок, що критично для вивчення англійської.

Крім того, надмірне використання інструментів ШІ може призвести до зниження креативності студентів. Якщо вони звикають до шаблонних формулювань або автоматичних підказок, процес вивчення мови стає механічним і менш цікавим. Ще однією проблемою є різниця у рівні цифрової грамотності: не всі викладачі та студенти здатні ефективно використовувати інструменти ШІ, що створює нерівність у можливостях навчання.

«Такі інструменти можна адаптувати до освітніх та навчальних цілей, проте основною умовою є те, що, використовуючи такий інструмент за традиційного підходу до навчання, викладачеві потрібно пояснювати ті чи ті нюанси, які пропонує сервіс у якості варіанту виправлення, або ж здобувачеві потрібно навчитися самостійно аналізувати та розуміти такі особливості, що, зі свого боку, потребує формування та розвитку додаткових компетенцій у здобувача» [3, с. 209].

Наприклад: спілкування лише з чат-ботами не замінить живої розмови з викладачем або іншими студентами. «Можна стверджувати, що вивчення ролі штучного інтелекту в процесах навчання іноземної мови буде позитивним стимулом до розвитку» [4, с. 6].

Узагальнення: щоб ШІ був ефективним у навчанні англійської мови, він має бути інтегрований продумано, з урахуванням педагогіки та потреб студентів.

Штучний інтелект має великий потенціал для вдосконалення навчання англійської мови у вищій освіті: він може персоналізувати процес, підвищити ефективність, розширити можливості доступу. «Побудова викладання іноземної мови з урахуванням сучасних тенденцій може надати учням відповідні для них стратегії навчання та сприяти досягненню кращих результатів у навчанні» [5, с. 185]. Однак його успішність залежить не лише від технології - важливо, як саме він впроваджується: чи враховано культурний контекст, чи забезпечено баланс між автоматизацією і людською взаємодією, чи підтримується розвиток самостійності студента.

Рекомендовано: викладачам та закладам впроваджувати ШІ інструменти поступово, навчати студентів і викладачів їх грамотному використанню, розробити політику захисту даних, забезпечити живу комунікацію і критичне

мислення. Лише за таких умов ШІ стане справжнім партнером у навчанні англійської мови, а не заміною.

Автори з різних куточків світу, маючи досвід і розуміння як технологічних, так і педагогічних питань, розглядають проблеми штучного інтелекту в освіті. Разом вони ставлять під сумнів ідею невпинного просування освіти, що все більше базується на технологіях, і натомість виступають за виважений підхід, який використовує сильні сторони як людських педагогів, так і технологічних досягнень.

1. Бондар, А. В., Самар, О. М., & Фурманчук, Н. М. (2024). Особливості використання штучного інтелекту як ефективного інструменту викладання іноземних мов. https://pednauk.cusu.edu.ua/index.php/pednauk/article/view/1851/1827.
2. Куцак, Л. В. (2025). Штучний інтелект у сучасній освіті: перспективи застосування та виклики. Методологічні проблеми впровадження цифрових технологій та інноваційних методик навчання. https://vspu.net/sit/index.php/sit/article/view/5679/5095.
3. Четверик, В. (2024). Ресурси зі штучним інтелектом у навчанні іноземним мовам: огляд можливостей та перспектив використання. https://vspu.net/sit/index.php/sit/article/view/5650/5078.
4. Давидюк, А. Р. (2024). Штучний інтелект як фактор змін у навчанні іноземних мов. Педагогічна академія. https://pedagogical-academy.com/index.php/journal/article/view/409/287.
5. Пономаренко, Н., Тимченко, Г., & Неустроєва, Г. (2024). Вплив штучного інтелекту на вивчення іноземної мови. Витоки педагогічної майстерності. https://sources.pnpu.edu.ua/article/view/318106.

М.О. Пісоцький

КОНТЕКСТНО-ЗАЛЕЖНЕ ВАГУВАННЯ ОЗНАК У ПРОГНОЗНИХ МОДЕЛЯХ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ ДЛЯ СИСТЕМ БЕЗПЕКИ

Штучний інтелект стає ключовим елементом сучасних систем безпеки [1], що працюють у реальному часі – від кіберзахисту до моніторингу технічного стану об'єктів критичної інфраструктури. Основою таких систем є методи інтелектуального аналізу даних, які виявляють закономірності у великих потоках інформації [2]: машинне навчання, кластеризація, а також нейронні мережі. Висока динамічність середовища, постійна зміна навантаження та поява нових типів атак призводять до зниження ефективності моделей, побудованих на статичних параметрах.

Ключовим елементом будь-якої моделі є набір ознак, що описує об'єкт аналізу: у класичних моделях ваги ознак визначаються на етапі навчання та залишаються фіксованими, проте у реальних умовах експлуатації систем безпеки контекст постійно змінюється, наприклад, залежно від часу доби, активності користувачів, типу трафіку чи рівня навантаження. Через це фіксовані ваги втрачають актуальність, а моделі перестають коректно реагувати на нетипові ситуації.

Проблема може бути вирішена шляхом застосування принципів контекстно-залежних обчислень [3-5], які дозволяють адаптувати алгоритм до умов середовища і створювати адаптивні моделі що враховують зовнішні умови. У системах безпеки це означає, що модель не лише аналізує дані, а й розуміє умови їх виникнення, динамічно змінюючи власні вагові коефіцієнти.

Контекст розглядається як множина характеристик $C = \{ c_1, c_2 , \ldots , c_k \}$, що впливають на інформативність ознак. Відповідно, ваги ознак w_i стають функцією від поточного контексту C_t, а модель набуває здатності до самоналаштування в реальному часі. Підхід контекстно-залежного вагування передбачає модифікацію функції оцінки інформативності ознак. У базовому випадку для кожної ознаки x_i визначається ваговий коефіцієнт:

$$w_i(C_t) = \lambda_i + \alpha_i \cdot \Phi_i(C_t) \qquad (1)$$

де λ_i – початкова (статична) вага ознаки, α_i – коефіцієнт контекстної чутливості, $\Phi_i(C_t)$ – міра релевантності ознаки в поточному контексті C_t.

Ця залежність описує зміну впливу ознаки в межах конкретного контексту. Контекст C_t формується з набору змінних: часу доби, типу події, джерела трафіку, рівня навантаження. Для кожного стану системи формується контекстний кластер, у межах якого переоцінюється інформативність ознак за допомогою інформаційно-ентропійного критерію, який широко застосовується у сучасних моделях виявлення аномалій у кіберфізичних системах [6].

Оновлення ваг у реальному часі виконується за правилом:

$$w_i^{(t+1)} = (1 - \eta)\ w_i(t) + \eta \cdot r_i(C_t) \qquad (2)$$

де $r_i(C_t)$ – оцінка релевантності ознаки, η – швидкість адаптації (0,1–0,3).

Таким чином модель оновлює ваги поступово, не потребуючи повторного навчання. Прототип системи реалізовано у середовищі Python із використанням бібліотек scikit-learn і NumPy. Архітектура складається з трьох компонентів:

1) контекстний моніторинг – виявляє зміни стану середовища та класифікує поточний контекст.

2) адаптивне вагування – оновлює коефіцієнти ознак за допомогою функції релевантності.

3) прогнозний модуль – використовує метод наближеного пошуку подібних контекстів (Approximate Nearest Neighbors, ANN), щоб швидко знаходити схожі стани та уточнювати ваги.

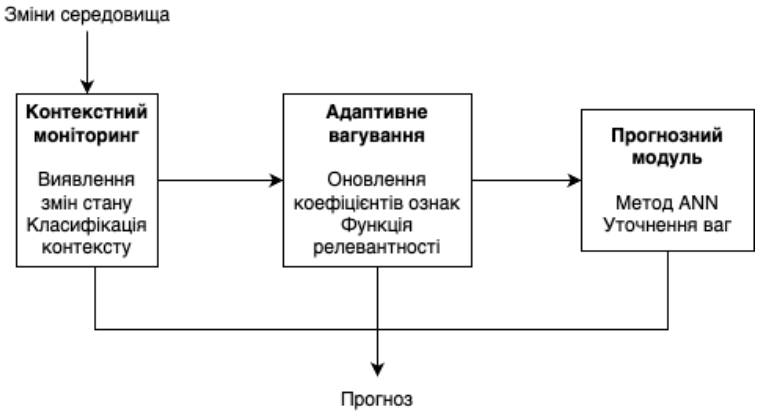

Рисунок 1 – Архітектура контекстно-залежного вагування ознак

Перевірку запропонованого підходу здійснено на відкритому наборі CICIDS2017 [7], який містить реальні мережеві трафіки нормальної активності й атак типу DoS, PortScan, Infiltration, DDoS. Для аналізу було відібрано 50 ознак, що характеризують мережеві сеанси, та сформовано 12 контекстних сценаріїв (денна/нічна активність, високі/низькі навантаження, тип користувача, інтенсивність з'єднань). Для отримання результатів було розглянуто базову і контекстно-залежну модель, а також використано наступні метрики - F1-score, MAE та False Positive Rate.

Таблиця 1 – Результати дослідження

Модель	F1-score	MAE	FP Rate
Базова	0,87	0,156	0,15
Контекстна	0,93	0,124	0,10

Таким чином, урахування контексту дозволило зменшити похибку прогнозу на 20%, а також знизити кількість хибнопозитивних спрацьовувань на 33%. Найбільше покращення спостерігалося під час зміни контексту “нічна активність/високі навантаження”, коли поведінкові патерни користувачів істотно відрізняються від звичайних.

Метод може бути інтегрований у системи SIEM або у модулі технічної діагностики промислових об’єктів. Контекстна адаптація ваг дозволяє пріоритезувати критичні параметри, наприклад, частоту запитів, тип протоколу чи температуру вузлів, залежно від реального стану середовища. Це створює основу для побудови самоналаштовуваних моделей безпеки, здатних працювати в умовах непередбачуваної динаміки.

Висновки

Робота присвячена підвищенню адаптивності прогнозних моделей інтелектуального аналізу даних у системах безпеки. Запропоновано метод контекстно-залежного вагування ознак, який дозволяє моделі автоматично змінювати вагові коефіцієнти залежно від поточного стану середовища. Використання контекстного модуля забезпечує стабільну роботу алгоритму навіть у ситуаціях різких змін поведінки користувачів або навантаження мережі. Результати досліджень засвідчили, що врахування контексту покращує точність класифікації інцидентів безпеки до 93%, знижує похибку прогнозу на 20% та істотно зменшує кількість хибнопозитивних спрацьовувань. Метод може бути використаний у практичних рішеннях для систем кіберзахисту, моніторингу інфраструктури та управління ризиками. Подальші дослідження спрямовані на поєднання контекстного вагування з глибинними нейронними архітектурами типу Transformer для обробки потокових подій у реальному часі.

1. Russell, S., & Norvig, P. (2021). Artificial intelligence: A modern approach (4th ed.). Pearson. ISBN 9780134610993.
2. Sarker, I. H., Colman, A., Han, J., & Watters, P. (2024). Context-Aware Machine Learning and Mobile Data Analytics: Automated Rule-based Services with Intelligent Decision-Making. Springer. https://doi.org/10.1007/978-3-030-88530-4.
3. Dey, A. K. (2001). Understanding and using context. Personal and Ubiquitous Computing, 5(1), 4–7. https://doi.org/10.1007/s007790170019.
4. Pisotskyi, M. and Botchkaryov, A. (2021) Online Video Platform with Context-aware Content-based Recommender System, Advances in Cyber-Physical Systems, 6(1), pp. 46–53. https://doi.org/10.23939/acps2021.01.046.
5. Pisotskyi, M. (2022) Approaches to Implementing Video Data Management Service with the Context-Aware Recommendation, Advances in Cyber-Physical Systems, 7(2), pp. 140–146. https://doi.org/10.23939/acps2022.02.140.
6. Moriano, P., Hespeler, S. C., Li, M., & Mahbub, M. (2025). Adaptive anomaly detection for identifying attacks in cyber-physical systems: A systematic literature review. Artificial Intelligence Review, 58, Article 283. https://doi.org/10.1007/s10462-025-11292-w.
7. Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2017). CICIDS2017 dataset. Canadian Institute for Cybersecurity. https://www.unb.ca/cic/datasets/ids-2017.html.

І.О. Деменко

ФОРМАЛЬНА МОДЕЛЬ ТЕСТУВАННЯ АЛГОРИТМІВ ШІ У КРИТИЧНИХ СИСТЕМАХ З ВИКОРИСТАННЯМ ЛОГІК LTL/CTL

Штучний інтелект активно впроваджується у критично важливі сфери, де помилки алгоритмів можуть спричиняти значні ризики для безперервності роботи та безпеки систем. Традиційні методи тестування не забезпечують достатнього охоплення можливих сценаріїв поведінки, особливо коли алгоритми ШІ демонструють складну або нечітку логіку прийняття рішень.

У таких умовах важливого значення набувають формальні методи, які дозволяють математично строго описувати та перевіряти властивості систем. Логіки LTL та CTL є фундаментальними інструментами для верифікації часових і розгалужених поведінкових моделей, а сучасні дослідження підтверджують ефективність їх застосування для аналізу алгоритмів ШІ [1-2].

Метою роботи є розробка формальної моделі тестування алгоритмів ШІ у критичних системах на основі логік LTL/CTL та демонстрація її застосування для перевірки властивостей безпеки та коректності.

У межах запропонованого підходу алгоритм штучного інтелекту, що використовується у критичній системі, представляється у вигляді формальної моделі переходів станів. Така модель дозволяє описати всі можливі конфігурації системи та логіку її зміни у часі, що є необхідною передумовою для застосування формальних методів верифікації.

Модель визначається як перехідна система

$$M = (S, S_0, R, L), \tag{1}$$

де S – множина станів, що відображають внутрішній стан алгоритму та ключові параметри середовища; $S_0 \subseteq S$ – множина початкових станів; $R \subseteq S \times S$ – відношення переходів, яке задає допустимі зміни стану; $L : S \to 2^{AP}$ – функція маркування, що визначає множину атомарних висловлювань AP, істинних у відповідному стані.

Вимоги до безпеки та коректності формулюються у вигляді логічних специфікацій, що описують очікувану поведінку системи. Для цього використовуються логіки LTL (Linear Temporal Logic), придатна для лінійних послідовностей станів, та CTL (Computation Tree Logic), яка дозволяє описувати розгалужені сценарії розвитку подій. Типовими властивостями є вимоги безпеки (safety), що гарантують уникнення небажаних станів, та вимоги живучості (liveness), які забезпечують досягнення необхідних реакцій у майбутньому.

Узагальнена постановка задачі тестування полягає у визначенні того, чи виконує модель M множину формул $\phi = \{\varphi_1, \varphi_2, \dots\}$, що описують бажані властивості. Формально завдання зводиться до перевірки виконання відношення:

$$M \vDash \varphi_i, \quad \forall \varphi_i \in \phi, \tag{2}$$

або до пошуку контрприкладів – послідовностей переходів, які демонструють порушення властивості.

Для автоматизованої перевірки використовується процедура *model checking*, яка виконує обхід простору станів моделі та визначає, чи існує траєкторія, що суперечить вимогам безпеки. Запропонована формальна модель дозволяє не лише строго визначити вимоги до алгоритмів ШІ, але й автоматизувати процес перевірки їх відповідності цим вимогам, що є критично важливим у застосуваннях із підвищеними вимогами до безпеки.

Для демонстрації застосування запропонованої формальної моделі було проведено експеримент із тестування алгоритму штучного інтелекту, інтегрованого у модуль моніторингу критичних подій. Алгоритм виконує класифікацію рівня загрози (low, medium, high) за вхідними параметрами, що характеризують стан середовища. Від коректності його роботи залежить своєчасність активації механізмів реагування системи.

На основі поведінки алгоритму було сформовано модель переходів, що містить такі стани:

- S_0: нормальний режим роботи;
- S_1: виявлено аномалію у вхідних даних;
- S_2: класифікація загрози як high;
- S_3: перехід системи у режим alarm;
- S_{err}: некоректна реакція алгоритму або неправильне рішення.

Відношення переходів R моделює можливі сценарії поведінки: від виникнення аномалії до активації системи попередження або, у разі помилки алгоритму, до переходу в аварійний стан.

Для перевірки відповідності алгоритму вимогам безпеки та коректності було визначено три ключові специфікації у логіці CTL:

1. Властивість відсутності небезпечних станів (safety):

$$\varphi_1 = AG(\neg error) \tag{3}$$

2. Властивість обов’язкової реакції на загрозу (реактивність):

$$\varphi_2 = AG(threat \rightarrow AF\ alarm) \tag{4}$$

3. Властивість коректності класифікації високого рівня ризику:

$$\varphi_3 = AG(HighRisk \rightarrow (classification = high)) \tag{5}$$

Ці властивості охоплюють критичні аспекти роботи системи: уникнення помилок, своєчасність реагування та правильність прийняття рішень алгоритмом.

Процедуру *model checking* застосовано до моделі переходів із використанням повного обходу простору станів. Для кожної властивості

визначено, чи існують контрприклади – траєкторії, що порушують специфікацію.

Перевірка показала, що алгоритм коректно реагує на загрози та не переходить у небезпечні стани, однак властивість φ_3 була порушена. Контрприклад відповідає ситуації, в якій алгоритм неправильно класифікує високий рівень ризику як medium, що може призвести до затримки реакції системи. Це вказує на необхідність уточнення моделі класифікації або навчальних даних.

Таблиця 1 – Результати перевірки властивостей алгоритму ШІ

№	Формула	Опис властивості	Результат	Коментар
1	$\varphi_1 = AG(\neg error)$	Відсутність аварійних станів	Виконується	Небезпечні стани недосяжні
2	$\varphi_2 = AG(threat \rightarrow AF\ alarm)$	Гарантована реакція на загрозу	Виконується	Система завжди переходить у alarm
3	$\varphi_3 = AG(HighRisk \rightarrow (classification = high))$	Коректна класифікація HighRisk	Порушується	Знайдено контрприклад: хибна класифікація

Результати експерименту підтвердили ефективність запропонованого формального підходу до тестування алгоритмів ШІ, оскільки він дозволяє не тільки перевірити відповідність властивостей, але й автоматично знаходити мінімальні контрприклади, що спрощує діагностику помилок.

У роботі запропоновано формальний підхід до тестування алгоритмів штучного інтелекту у критичних системах на основі логік LTL/CTL та модельної перевірки. Побудована модель переходів станів дає змогу строго визначати вимоги до безпеки, реактивності та коректності роботи алгоритму. Експеримент підтвердив, що метод забезпечує повне охоплення можливих сценаріїв і дозволяє автоматично виявляти логічні помилки у вигляді контрприкладів.

Попри відповідність більшості перевірених властивостей, виявлене порушення у класифікації високого рівня ризику демонструє здатність формальних методів виявляти критичні недоліки у поведінці алгоритму. Запропонований підхід може бути інтегрований у процеси автоматизованого тестування, а подальші дослідження передбачають розширення формальної моделі, автоматизовану генерацію тестових сценаріїв та адаптацію методу до більш складних моделей ШІ.

1. Bordais, B., Neider, D., & Roy, R. (2025). The complexity of learning LTL, CTL and ATL formulas. Symposium on Theoretical Aspects of Computer Science. https://doi.org/10.4230/LIPIcs.STACS.2025.19.
2. Bordais, B., Neider, D., & Roy, R. (2024). Learning branching-time properties in CTL and ATL via constraint solving. World Congress on Formal Methods. https://doi.org/10.48550/arXiv.2406.19890.
3. Kesavan, E. (2024). AI-driven and autonomous testing. International Scientific Journal of Engineering and Management. https://doi.org/10.55041/isjem01651.

А.А. Ублінських

ОГЛЯД ІСНУЮЧИХ МЕТОДІВ ВИЯВЛЕННЯ ВРАЗЛИВОСТЕЙ У СМАРТКОНТРАКТАХ ТА ПЕРСПЕКТИВИ ІНТЕГРАЦІЇ ШІ

Блокчейн-індустрія - нова галузь, котра динамічно розвивається і стикається з численними викликами, серед яких - питання безпеки. Неодноразово траплялися кібератаки, що призводили до втрати мільйонів доларів, а одним із векторів атак є смартконтракти. Смартконтракти - це програми, розгорнуті в блокчейн-мережах, які забезпечують автоматичне виконання зобов'язань між сторонами. Вони застосовуються у децентралізованих біржах (наприклад, Uniswap, Curve, Aerodrome), кредитних протоколах (наприклад, AAVE, Morpho) та інших децентралізованих додатках. Зловмисники досить часто знаходять у них вразливості й виводять значні кошти. Тому безпека смартконтрактів є надзвичайно актуальною. У цій роботі ми розглянемо існуючі інструменти, сервіси та платформи для пошуку вразливостей у смарт-контрактах, а також проаналізуємо можливі шляхи інтеграції штучного інтелекту (ШІ) у цю сферу.

Почнемо з середовищ розробки: на момент написання роботи найбільш популярними є Hardhat та Foundry. Обидва середовища надають достатній набір засобів для тестування смарт-контрактів, зокрема — за допомогою unit-тестів і fuzz-тестів.

Для автоматичного виявлення вразливостей застосовуються статичні аналізатори коду, такі як:

- Aderyn — статичний аналізатор, написаний на Rust. Дає змогу виявляти поширені помилки під час написання смартконтрактів шляхом парсингу коду. Його перевагою є модульна архітектура, що дозволяє додавати власні правила аналізу. Нині це один із найактуальніших статичних аналізаторів.
- Mythril — статичний аналізатор, який фокусується на технічних помилках в реалізації смарт-контрактів, однак не орієнтований на виявлення помилок у бізнес-логіці. Інструмент активно підтримується та регулярно оновлюється.
- Slither — аналізатор, створений на Python; забезпечує статичний аналіз, який допомагає виявляти потенційні вади в структурі коду.

Існують також платформи, які надають послуги аудиту смартконтрактів, наприклад: V12 і MythX. Такі сервіси автоматично виявляють як технічні, так і логічні вразливості, проте працюють за принципом «чорної скриньки», не розкриваючи деталі реалізації, та надаються на комерційній основі.

Крім того, існують платформи для публічних аудитів і програм «баг-баунті», такі як Code4rena, Immunefi, HackenProof. Проєкти, які планують розгортання власних контрактів, можуть оприлюднити їх там, щоб white-hat

хакери за винагороду виявили вразливості. Такий підхід справді допомагає - проте завжди залишається ризик, що деякі вразливості не будуть знайдені.

Є також компанії, які надають професійні послуги аудиту - наприклад, Hacken, Cyfrin, OpenZeppelin та інші. Вони надають найбільш глибокий аналіз як технічний так і логічний - проте такі аудити часто коштують дорого, і навіть у цьому випадку не можна гарантувати відсутність вразливостей.

Після аналізу існуючих методів виявлення вразливостей ми приходимо до висновку, що найбільший ризик становлять логічні вразливості - вони можуть призвести до непоправних наслідків, і сучасні методи не завжди здатні їх виявити.

У дослідженні [1] здійснено комплексний огляд підходів до застосування систем Intrusion Detection Systems (IDS) у контексті атак на блокчейн мережу Ethereum. У роботі проаналізовано широкий спектр алгоритмів і фреймворків глибокого машинного навчання (Deep learning models), які використовуються для виявлення аномалій та потенційних вразливостей у блокчейн-мережах.

Дослідження розглядає методи наступні методи навчання нейронних мереж:

- Supervised Learning,
- Semi-Supervised Learning,
- Unsupervised Learning,
- Reinforcement Learning.

Окремо розглянуто алгоритми, притаманні кожному підходу, а також їхню ефективність і обмеження під час аналізу транзакцій, поведінкових шаблонів та аномалій у мережі Ethereum.

У роботі подано огляд наявних наукових праць, у яких запропоновано фреймворки для виявлення аномалій та атак на основі глибокого машинного навчання, що дозволяє порівняти їх ефективність та сфери застосування.

Ключовий висновок дослідження полягає у тому, що більшість існуючих фреймворків базується саме на supervised learning, оскільки цей підхід забезпечує найвищу точність виявлення вразливостей. Проте таке навчання вимагає великих розмічених датасетів (labeled data), що значно збільшує витрати часу й ресурсів на підготовку моделей. Незважаючи на це, автори підкреслюють, що supervised-методи наразі залишаються найефективнішими для задач точного класифікаційного виявлення загроз та вразливостей.

Джерело [2] присвячене аналізу сучасних методів застосування штучного інтелекту для виявлення вразливостей у смартконтрактах Ethereum, зі спеціальним акцентом на інтеграцію підходів Explainable AI (XAI). У роботі розглянуто широкий спектр фреймворків глибокого машинного навчання, які використовуються для класифікації й пояснення потенційно небезпечних шаблонів у смартконтрактах.

У практичній частині дослідження побудовано Explainable DNN-модель та виконано її тренування на відкритому наборі даних SCsVul (2023/2024).

Враховуючи багатокласовий поділ на Non-Vulnerable, Arithmetic & Gas Issues, Control Flow & Logic Bugs, Access Control Issues, Runtime & DOS, отримано наступні результати:

- Score ≈ 97.74%
- Accuracy 97.73%
- F1-score 97.73%
- MCC 97.13%

Такі показники свідчать про те, що запропонована DNN-модель ефективно ідентифікує основні типи вразливостей у смартконтрактах.

Важливим внеском роботи є додавання пояснюваності до процесу прийняття рішення. Використовуючи LIME та SHAP, дослідження демонструє, які характеристики байт коду, AST, ABI чи opcode-послідовностей мають найбільший вплив на класифікацію. Це підвищує прозорість моделі і дозволяє розробникам та аудиторам перевірити коректність виявлених вразливостей, що є критично важливим у Metaverse та Web3-екосистемах.

Джерело [3] описує розроблений фреймворк CodeSpeak, який емулює процес мануального аудиту для пошуку вразливостей у смартконтрактах. У дослідженні фреймворк було протестовано на таких типах вразливостей, як Reentrancy, Timestamp Dependency, Overflow/Underflow та Delegatecall. Автори зазначають, що швидкість виявлення подібних вразливостей підвищилася до 98.7% порівняно з класичним процесом аудиту.

У роботі також детально розглядаються обмеження як нейронних мереж, так і моделей LLM. Основна проблема нейронних мереж полягає в тому, що після кожного великого оновлення блокчейн-мережі можуть з'являтися нові вектори атак, про які модель не знає, що робить їх виявлення неможливим без повторного навчання. Перенавчання ж потребує значних часових та обчислювальних ресурсів.

Ще одне обмеження LLM-моделей полягають у їх схильності до галюцинацій, що може призводити до появи хибнопозитивних результатів під час аналізу. У дослідженні це вирішили шляхом виконання додаткового етапу верифікації вразливості.

Дослідження демонструє високі результати для невеликих смартконтрактів, однак модель погано працює з великими контрактами (понад 1000 рядків коду).

Аналіз розглянутих джерел свідчить, що блокчейн-індустрія потребує швидкого та масштабованого підходу до виявлення вразливостей у смартконтрактах. Темп розвитку екосистеми, часті оновлення стандартів і поява нових моделей взаємодії між контрактами створюють потребу в інструментах, здатних оперативно адаптуватися та ефективно працювати в нових умовах.

У цьому контексті моделі великомасштабного машинного навчання (LLM) виглядають найбільш перспективними. Їхня ключова перевага -

відсутність необхідності у великих розмічених датасетах, які є критичним обмеженням для класичних нейронних мереж. LLM здатні швидко адаптуватися, демонструють високу універсальність, добре працюють з контекстними залежностями та природною мовою, що особливо важливо для аналізу логіки смартконтрактів. На відміну від глибинних нейронних мереж, які потребують тривалого навчання і швидко втрачають актуальність після появи нових векторів атак, LLM можуть бути оновлені та інтегровані значно швидше.

Водночас поточний стан досліджень показує, що наукових робіт, присвячених саме виявленню логічних вразливостей, усе ще недостатньо. Більшість існуючих підходів фокусуються переважно на технічних або структурних помилках, які простіше формалізувати та виявляти. Логічні ж вразливості потребують глибшого розуміння бізнес-логіки та поведінкових сценаріїв контракту, що робить їх значно складнішими для автоматизованого аналізу.

З огляду на це, перспективним напрямом подальших досліджень є комбінований підхід, який поєднує можливості LLM з уже перевіреними інструментами аналізу, такими як unit-тестування, fuzz-тестування та статичні аналізатори. Така інтегрована методологія може забезпечити більш повне покриття різних класів вразливостей та підвищити надійність процесу аудиту смартконтрактів.

1. Jiang, F., Chao, K., Xiao, J., Liu, Q., Gu, K., Wu, J., & Cao, Y. (2023). Enhancing Smart-Contract Security through Machine Learning: A Survey of Approaches and Techniques. Electronics, 12(9), 2046. https://doi.org/10.3390/electronics12092046.
2. Nkoro, E. C., Ahakonye, L. A. C., & Kim, D.-S. (2025). Explainable DNN for smart contract vulnerability detection in the Metaverse. High-Confidence Computing, 100374. https://doi.org/10.1016/j.hcc.2025.100374.
3. Chang, S., Geng, C., Huang, H., Wang, R., Li, Q., & Zhang, Y. (2026). CodeSpeak: Improving smart contract vulnerability detection via LLM-assisted code analysis. Journal of Systems and Software, 231, 112635. https://doi.org/10.1016/j.jss.2025.112635.

О.М. Леонтенко, А.О. Отамась

СОЦІАЛЬНО ВІДПОВІДАЛЬНЕ ВПРОВАДЖЕННЯ ШІ В СМАРТ-ЛОГІСТИКУ МІЖНАРОДНИХ ПЕРЕВЕЗЕНЬ

Сучасна логістика міжнародних перевезень переживає глибоку трансформацію, і одним із ключових чинників цих змін є впровадження штучного інтелекту. У глобальних ланцюгах постачання, де швидкість, точність і стійкість є вирішальними, ШІ стає необхідним елементом підвищення ефективності та конкурентоспроможності. Саме завдяки можливостям аналізу великих масивів даних і автоматизації складних процесів ШІ здатен формувати нову модель логістики — гнучку, прогнозовану й стійку до зовнішніх шоків [1].

У міжнародних перевезеннях штучний інтелект проявляє себе як технологія, що може впорядковувати транспортні потоки, оптимізувати маршрути, прогнозувати затори та коливання попиту, а також інтегрувати інформацію від тисяч сенсорів у єдину управлінську систему. Це дозволяє компаніям скорочувати витрати на пальне, зменшувати ризики незапланованих зупинок та підвищувати точність доставки, що вже доведено практиками логістичних операторів [2]. Крім того, іінтеграція IoT та систем машинного навчання підсилює надійність міжнародних перевезень, дозволяючи контролювати стан вантажу в реальному часі [3].

Однак активне впровадження ШІ несе не лише економічні переваги, але й суттєві соціальні виклики. Автоматизація логістичних процесів змінює структуру зайнятості, зменшуючи потребу в ручній праці, зокрема у водіях або працівниках складів, хоча водночас формує нові професійні ролі — аналітиків даних, операторів автономних систем, інженерів ШІ. Соціально відповідальний перехід до смарт-логістики має супроводжуватися програмами перекваліфікації та розвитку цифрових навичок серед працівників [4]. На мою думку, саме готовність компаній інвестувати в людей, а не лише в технології, визначатиме, наскільки гуманним і справедливим буде цей технічний прогрес.

Не менш важливою проблемою є захист даних, адже сучасні логістичні системи оперують сотнями тисяч чутливих записів: від GPS-маршрутів до комерційної інформації та персональних даних клієнтів. Вразливість штучного інтелекту до кібератак ставить під загрозу не лише окремі перевезення, але й функціонування глобальних ланцюгів постачання як елемента критичної інфраструктури. З огляду на зростання складності атак необхідними є багаторівнева аутентифікація, шифрування, регулярні аудити безпеки та впровадження міжнародних стандартів інформаційної безпеки [1]. Я вважаю, що проблема критичної інфраструктури виходить за межі технологічних рішень. Вона потребує політичної волі, координації держав і єдиних стандартів, щоб захистити міжнародні ланцюги постачання від системних ризиків.

Окрему увагу слід приділити етичним ризикам, які виникають під час прийняття рішень штучним інтелектом. Алгоритми можуть демонструвати упередженість, наприклад, у виборі маршруту, оцінюванні постачальників чи плануванні пріоритетів доставки, якщо навчаються на нерепрезентативних даних. Це може призводити до несправедливих бізнес-рішень і навіть дискримінації окремих груп клієнтів або регіонів [5]. Постає також питання відповідальності: хто має відповідати за збитки, якщо автономна система припускається помилки? Виробник, оператор чи компанія, яка ухвалила рішення застосувати таку технологію? На мою думку, правове регулювання у сфері ШІ досі серйозно відстає від технологічного розвитку, і без чітких норм компанії ризикують опинитися у правовій «сірій зоні».

Для мінімізації таких ризиків важливими є правові та методологічні підходи, серед яких — формування міжнародних і національних правил використання автономних систем, проведення незалежних аудитів ШІ, створення етичних кодексів для розробників і операторів. Крім того, логістичним компаніям необхідно розробляти внутрішні політики соціальної відповідальності, які забезпечують захист працівників та сприяють сталому розвитку технологій. Я переконана, що лише поєднання правових, технічних та соціальних інструментів здатне забезпечити справедливе, безпечне й етично виважене майбутнє міжнародної логістики.

Узагальнюючи, варто підкреслити, що соціально відповідальне впровадження штучного інтелекту в смарт-логістику має ґрунтуватися на балансі між економічними вигодами, етичними принципами та захистом критичної інфраструктури. В умовах зростаючої залежності світу від стійких ланцюгів постачання необхідно створити міжнародні стандарти безпеки та прозорості, інвестувати в підготовку кадрів і забезпечувати технологічну справедливість. Тільки так розвиток ШІ стане не загрозою, а ресурсом для формування нового, стійкого і відповідального глобального логістичного середовища.

1. Буров Є., Кулявець А. *Штучний інтелект у логістиці: можливості та виклики*. SISN, 2024. https://doi.org/10.23939/sisn2024.16.001.
2. Ekol Ukraine. *Штучний інтелект у логістиці: можливості та виклики*. https://ekol.com.ua/shtuchnyj-intelekt-u-logistyczi-mozhlyvosti-ta-vyklyky/.
3. Корсовська С.Р., Потапова Н.А. *Штучний інтелект у логістиці та вантажних перевезеннях*. https://jait.donnu.edu.ua/article/view/14040.
4. Судук Н., Герасимович І. (2025). *Застосування штучного інтелекту у виробничій логістиці: сучасні практики та перспективи розвитку.* Економіка та суспільство, (73). https://doi.org/10.32782/2524-0072/2025-73-40.
5. Rosemarie Santa Gonzalez, Ryan Piansky, Sue M Bae, Justin Biddle, Daniel Molzahn. *Beyond Algorithmic Fairness: A Guide to Develop and Deploy Ethical AI-Enabled Decision-Support Tools*. ArXiv, 2024. https://arxiv.org/abs/2409.11489.

Д.І. Бахтіяров, В.В. Тельних, Б.П. Котик

МЕТОДОЛОГІЧНІ АСПЕКТИ ОБРОБКИ ВЕЛИКИХ ДАНИХ У ТЕЛЕКОМУНІКАЦІЙНИХ СИСТЕМАХ З ВИКОРИСТАННЯМ ІНСТРУМЕНТІВ ШТУЧНОГО ІНТЕЛЕКТУ

Еволюція мережевої архітектури сучасних інформаційно-телекомунікаційних систем стандарту 5G та майбутнього 6G потребує суттєвих змін у системах управління мережевими ресурсами [1].

Коли трафік зростає експоненціально з різних джерел, таких як мобільні пристрої та датчики Інтернету речей, класичні детерміновані алгоритми обробки даних стають набагато менш ефективними. Основною проблемою є те, що для забезпечення відповідності стандартам QoS та QoE необхідна обробка інформаційних потоків у режимі реального часу. У цьому контексті науковий інтерес переходить до вдосконалення адаптивних методів аналізу даних, заснованих на штучному інтелекті (ШІ), особливо на глибокому навчанні (DL).

Проблеми з обробкою даних у телекомунікаційних системах безпосередньо пов'язані з випадковим характером навантаження на мережу. Аналіз часових рядів трафіку показує, що він має властивості самоподібності (фрактальності) та довгострокової залежності, що підтверджується значеннями експоненти Херста ($H > 0{,}5$). Традиційні статистичні моделі, включаючи авторегресійну інтегровану рухому середню ARIMA, залежать від припущень лінійності та стаціонарності, що заважає їм точно моделювати виражені сплески активності («burstiness»), типові для мультисервісного трафіку. Як результат, існує потреба в нелінійних апроксиматорах, які можуть знаходити приховані закономірності в багатовимірних наборах даних. Переважна методологія в сучасному інтелектуальному аналізі даних передбачає застосування штучних нейронних мереж (ANN). Рекурентні нейронні мережі (RNN) виділяються серед багатьох типів архітектур, оскільки мають механізми зворотного зв'язку, що дозволяють враховувати історію процесів.

Класичні RNN в процесі навчання на довгих послідовностях схильні до проблеми зникаючого градієнта, що в свою чергу компенсуються архітектурою Long Short-Term Memory (LSTM). Дана архітектура базується на математичному апараті, що описує процес проходження інформації через так звані вхідні вентилі i_t, забуття f_t та вихідні o_t. Математичне представлення оновлення стану комірки C_t наступне (1):

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t, \quad (1)$$

де $\odot$ означає «помноження елементів» (добуток Адамара), а $\tilde{C}_t$ – вектор нових значень-кандидатів.

Ця структура дозволяє алгоритму зберігати тільки важливу інформацію про довгострокові трафік-патерни та ігнорувати шум, який трапляється часто.

Останні дослідження [2-4] показують, що гібридизація методів має великий потенціал. Поєднання конволюційних нейронних мереж (CNN) та LSTM полегшує моделювання подій, що відбуваються в часі та просторі. У

цьому типі архітектури CNN використовуються для отримання просторових характеристик з матриць трафіку, що показують топологію мережі та навантаження комірки. LSTM, з іншого боку, використовуються для прогнозування того, як ці характеристики змінюватимуться з часом. Для роботи технології Network Slicing цей метод є дуже важливим, оскільки він повинен бути здатним динамічно розподіляти віртуалізовані ресурси між різними типами послуг (eMBB, URLLC, mMTC).

Методи федеративного навчання (FL) – це інший спосіб розробки, який вирішує проблеми конфіденційності даних. В архітектурі FL периферійні пристрої обробляють інформацію локально, і лише оновлені градієнти ваг локальної моделі надсилаються до центрального сервера. Це дає змогу будувати глобальні прогнозні моделі без об'єднання необроблених даних користувачів, що відповідає сучасним стандартам кібербезпеки [5].

Експериментальна валідація запропонованих методологій, як зазначено в роботах [3, 4], показує, що використання методів глибокого навчання зменшує середньоквадратичну похибку прогнозів на 15-22 % порівняно з традиційними статистичними методами. Ця особливість забезпечує проактивне управління мережею, мінімізацію затримки та оптимізацію енергоспоживання інфраструктури.

Підводячи підсумки, використання методів обробки даних на основі штучного інтелекту є необхідним кроком для розвитку систем управління інформаційно-телекомунікаційними системами. Подальші дослідження повинні зосередитися на зменшенні обчислювальної складності моделей нейронних мереж, щоб полегшити їх ефективне впровадження на апаратному забезпеченні з обмеженими ресурсами.

1. Kotyk, B., Bakhtiiarov D., et al. (2025). Neural network approach to 5G digital modulation recognition. CEUR Workshop Proceedings, 3925, 82-92.
2. Abdulrahman Yarali, "Big Data and Artificial Intelligence," in Intelligent Connectivity: AI, IoT, and 5G , IEEE, 2022, pp.299-326, doi: 10.1002/9781119685265.ch17.
3. A. S. Bale, H. G, S. Sandhya, B. M. Ningegowda, S. Ponnuru and K. Lal, "A Study on Optimization of the Intelligent Control System using AI," 2024 International Conference on Communication, Computing and Energy Efficient Technologies (I3CEET), Gautam Buddha Nagar, India, 2024, pp. 917-922, doi: 10.1109/I3CEET61722.2024.10993935.
4. Kotyk, B., Bakhtiiarov D., et al. (2025). Neural network approach to 5G digital modulation recognition under a priory uncertainty of parameters. CEUR Workshop Proceedings, 4024, 119–132.
5. G. P. Bhandari, A. Lyth, A. Shalaginov and T. -M. Grønli, "Artificial Intelligence Enabled Middleware for Distributed Cyberattacks Detection in IoT-based Smart Environments," 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan, 2022, pp. 3023-3032, doi: 10.1109/BigData55660.2022.10020531.

К.В. Діордій

ВПЛИВ ШТУЧНОГО ІНТЕЛЕКТУ НА ПСИХІКУ ЛЮДИНИ: МОЖЛИВОСТІ ТА ВИКЛИКИ

Нині цифрові технології є рушійною силою у всіх сферах суспільного життя. Нещодавно в кінці 2022 року компанія OpenAI презентувала світові цифрові сервіси на основі штучного інтелекту. До зазначених сервісів належать ChatGPT, DALL-E та інші. Вони дозволили компанії побити усі світові рекорди зі збільшення кількості відвідувачів. Згідно з даними The Guardian протягом двох місяців після запуску сервісів штучного інтелекту (далі ШІ) його відвідали сто мільйонів унікальних користувачів, водночас загальна кількість відвідувань на січень 2023-го року становила 590 млн (The Guardian, 2023), (Драч та ін., 2023).

Проте важливим питанням є вплив розвитку та загальнодоступності штучного інтелекту на психіку людини й суспільства. С. Рудницька у своєму дослідженні дійшла висновку, що чат-боти у період високої турбулентності є доступним інструментом психоемоційної стабілізації та продуктивним способом дискурсивної взаємодії. Вони здатні сприяти емоційній саморегуляції та зниженню напруги, поверненню відчуття контролю, допомогти структурувати особистісні смисли та у розвитку рівня саморефлексії (Рудницька, 2025).

Aakriti Varshney зазначає у висновках свого дослідження, що впровадження та використання штучного інтелекту є перспективним та ненадійним напрямом одночасно. ШІ може знизити стрес, створити відчуття дружніх стосунків та підтримати. З іншого боку, здатен викликати залежність, емоційну плутанину та сприяти розвитку когнітивних лінощів, адже люди можуть почати зловживати делегуванням йому завдань та прийняття рішень. Отже, варто брати до уваги, що ШІ це не просто функціональний інструмент, а психологічна присутність у житті користувача, тому розробникам слід потурбуватися, зокрема, й про емоційну безпеку людей під час його використання та зручність (Varshney, 2024).

Brooke N. Macnamara, Ibrahim Berber, M. Cenk Çavuşoğlu, Elizabeth A. Krupinski, Naren Nallapareddy, Noelle E. Nelson, Philip J. Smith, Amy L. Wilson-Delfosse, Soumya Ray у своєму дослідженні зазначають, що ШІ є корисним, наприклад, допомагає уникнути помилок у роботі, притаманних людині, здатен пришвидшити виконання завдань. Одночасно з тим люди здатні втратити або відмовитися розвивати навички, постійно делегуючи виконання тих чи інших завдань ШІ (Macnamara et al., 2024).

Jung-In Choi, Eunja Yang, Eun-Hee Goo описали у статті розроблену ними програму навчання з етики використання штучного інтелекту для учнів середньої школи, яка мала на меті підготувати їх етично правильного та ефективного використання ШІ. Основний акцент був зроблений на тому, щоб пояснити учням, що ШІ розроблений людиною та має запрограмовані

алгоритми. Водночас за прийняття рішення відповідальність все ще несе людина самостійно. Загальною метою навчальної програми було пояснити учням принципи роботи ШІ, як ефективно його використовувати та управляти ним, а також дослідити його характеристики. Для кращого розуміння програми, автори застосовували практичні вправи, відеоматеріали про ШІ, роботів та інше. Апробація програми показала, що вона є ефективною та дозволяє навчити здобувачів середньої освіти етично використовувати штучний інтелект, адже згідно з дослідженням ставлення до штучного інтелекту учнів зазнало змін після апробації програми та сприяло їх підготовці до життя у епоху цифрових технологій (Choi et al., 2024).

Таким чином, у підсумку зазначимо:

1. Впровадження штучного інтелекту має як позитивні, так і негативні наслідки. До позитивних можна віднести здатність стати інструментом для зниження рівня напруги та стресу, допомогти підвищити рівень рефлексії, повернути відчуття контролю, структурувати особистісні смисли, створити відчуття дружньої підтримки, пришвидшити виконання завдань та зменшити кількість помилок, допущених під час роботи. До негативних наслідків можна віднести ймовірність виникнення залежності, емоційної плутанини, зниження когнітивних здібностей, втрату/погіршення рівня навичок, внаслідок надмірного делегування завдань ШІ, ймовірність розвитку надмірної довіри та передача права приймати рішення за людину ШІ.

2. Щоби вирішити ці проблеми учені виокремили потребу у створенні навчальної програми та створили її для учнів середньої школи з безпечного етичного використання ШІ у повсякденному житті, апробація якої підтвердила свою ефективність.

Проте створення подібних комплексних навчальних програм необхідне не лише для школярів, а й для всіх, хто користується або буде користуватися штучним інтелектом. Окрім того, вони повинні бути доступними, апробованими, а отже якісними, щоби уникнути зазначених вище негативних наслідків впровадження штучного інтелекту.

Перспективи подальших досліджень: Подальші наукові розвідки будуть спрямовані на вивчення можливостей для розробки та апробації навчальних програм з використання ШІ для різних груп населення та програм з профілактики залежності від цифрових технологій.

1. The Guardian. (2023, February 2). ChatGPT reaches 100 million users two months after launch. URL: https:// www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app.
2. Драч, І. І., Петроє, О. М., Бородієнко, О. В., Регейло, І. Ю., Базелюк, О. В., Базелюк, Н. В., & Слободянюк, О. М. (2023). Використання штучного інтелекту у вищій освіті. Міжнародний науковий журнал «Університети і лідерство», (15), 66–82. URL: https://lib.iitta.gov.ua/id/eprint/741952/.

3. Рудницька, С. Ю. (2025). Вплив штучного інтелекту на трансформацію життєвої компетентності особистості в умовах соціальної турбулентності (с. 1–5). URL: https://lib.iitta.gov.ua/id/eprint/746734/1/Rudnitska_Svitlana_2025.pdf.
4. Varshney, A. (2024). Understanding human-AI interaction: Psychological effects and behavioral responses. Global Insights Journal Covers – Multidisciplinary Research Across Science, Technology, Humanities, and Social Sciences, 4(4), 26–37. URL: https://globalinsightsjournal.com/gij/index.php/journal/article/view/86/78.
5. Macnamara, B. N., Berber, I., Çavuşoğlu, M. C., Krupinski, E. A., Nallapareddy, N., Nelson, N. E., Smith, P. J., Wilson Delfosse, A. L., & Ray, S. (2024). Does using artificial intelligence assistance accelerate skill decay and hinder skill development without performers' awareness? Cognitive Research: Principles and Implications, 9(1), Article 46. URL://doi.org/10.1186/s41235-024-00572-8.
6. Choi, J.-I., Yang, E., & Goo, E.-H. (2024). The effects of an ethics education program on artificial intelligence among middle school students: Analysis of perception and attitude changes. Applied Sciences, 14(4), 1588. URL: https://www.mdpi.com/2076-3417/14/4/1588.

У.Р. Копілева-Ромащенко

АВТОРСЬКЕ ПРАВО В ЕПОХУ ШІ: ЧИ МОЖЕ АЛГОРИТМ БУТИ ТВОРЦЕМ?

Вступ

Дуже стрімки розвиток технологій штучного інтелекту очікувано створив проблеми та виклики для традиційної системи авторського права, яке раніше не було налаштоване на роботу із такими алгоритмами. Генеративні моделі, такі як ChatGPT, Midjourney, DALL-E та інші, які здатні створювати тексти, зображення, музику, коди та інший контент, який за якістю іноді вже не поступається творам людини, що ставить перед правовою системою нові питання: чи може алгоритм бути автором, і як захистити права усіх учасників процесу створення контенту за допомогою ШІ?

Правовий статус творів, створених штучним інтелектом

У більшості юрисдикцій світу наразі домінує принцип, згідно із яким авторське право може належати виключно людині. Прикладом може бути справа Thaler v. Permutter у США, де федеральний апеляційний суд у березні 2025 року підтвердив, що авторське право вимагає обов'язкової участі людини [1]. Стівен Талер намагався зареєструвати авторські права на картину "A Recent Entrance to Paradise", створену його системою ШІ DABUS, проте суд відхилив його позов, наголосивши на необхідності людського творчого внеску.

Аналогічну позицію займає також Бюро авторського права США, яке у липні 2024 року оприлюднило перші розділи комплексного звіт "Copyright and Artificial Intelligence", що базується на аналізі понад 10 000 коментарів від зацікавлених сторін [2]. Звіт підтверджує, що твори, які були створені виключно ШІ без значущої людської участі, не підлягають захисту авторським правом.

Україна обрала інноваційний шлях, запровадивши ще у 2023 році концепцію sui generis прав для об'єктів, створених без участі людини [3]. Це особливий правовий режим, який відрізняється від традиційного авторського права, проте надає певний рівень захисту творам, згенерованим ШІ.

Історичною подією стала перша в Україні реєстрація авторських прав на твори з використанням ШІ у жовтні 2024 року через Укрпатент [4], що свідчить про прагматичний підхід українського законодавства до врегулювання нових технологічних реалій. Реформа авторського права в Україні, розпочата ще у 2023 році, спрямована на гармонізацію із європейськими стандартами, зокрема із Дирктивою ЄС 2019/790 про авторське право на єдиному цифровому ринку [5].

У Європі ситуація лишається неоднозначною, адже, наприклад, Чеський Празький муніципальний суд у травні 2024 року виніс перше європейське рішення у справі про авторське право на контент, створений ШІ, залишивши можливість визнання авторства, якщо інструкції для ШІ були достатньо

деталізованими та персоналізованими [6]. Європейська Комісія у липні 2024 року оприлюднила звіти щодо ШІ та авторського права, намагючись при цьому знайти баланс між захистом інтелектуальної власності та стимулюванням інновацій [7].

Використання авторських творів для навчання ШІ

Одним із найгострішим питань є легальність використання захищених авторським правом творів для навчання моделей ШІ. Зокрема, у лютому 2025 року вийшло перше важливе рішення у справі Thomson Reuters v. ROSS Intelligence, яке стало прецедентом у питанні використання авторських баз даних для тренування ШІ [8].

Проте більш резонансною стала справа Bartz v. Anthropic у вересні 2025 року, де група авторів домоглася укладання мегаугоди на 1,5 млрд доларів з компанією Anthropic - це найбільша виплата за порушення авторських прав у історії США [9]. Позов будувався на тому, що компанія використала понад 7 мільйонів піратських копій книг для навчання своїх моделей.

Центральним у дискусіях є питання про те, чи є використання авторських матеріалів для навчання ШІ "справедливим використання" згідно із доктриною американського права. Congressional Research Service опублікував спеціальний звіт "Generative Artificial Intelligence and Copyright Law", який аналізує аргументи обох сторін [10]. Компанії-розробники ШІ стверджують, що навчання моделей є трансформативним використанням, яке не завдає шкоди ринку оригінальних твторів, а творці наполягають на тому, що їхніроботи використовуються для створення комерційних продуктів без їхньої згоди та компенсації.

Висновки

Питання "Чи може алгоритм бути творцем?" наразі отримало переважно негативну відповідь у більшості правових систем світу. Головним лишається принцип необхідності людського творчого внеску для виникнення авторських прав. Проте український досвід із запровадженням sui generis прав демонструє можливість альтернативних підходів до захисту інвестицій у створення ШІ-контенту.

Використання авторських творів для навчання моделей ШІ залишається предметом інтенсивних судових спорів, які поступово формують нову правову політику, а мегаугоди на мільярди доларів та десятки активних судових справ свідчать про те, що індустрія перебуває у стані трансформації, а баланс між захистом авторських прав та стимулюванням інновацій ще не знайдено.

1. Thaler v. Perlmutter: Federal Circuit Court of Appeals Decision on AI Authorship. United States Court of Appeals for the Federal Circuit. 2025. March. URL: https://cafc.uscourts.gov/.
2. Copyright and Artificial Intelligence: Report of the U.S. Copyright Office. Part 1. Washington, DC: U.S. Copyright Office, 2024. July. URL:

https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-1-Digital-Replicas-Report.pdf.
3. Mayidanyk L. Artificial Intelligence and Sui Generis Right: A Perspective for Copyright of Ukraine? AJEE Journal. 2023. Vol. 15. No. 2. P. 144–154. URL: https://ajee-journal.com/upload/attaches/att_1627907054.pdf.
4. Перша реєстрація авторських прав на твори з використанням ШІ в Україні. Державне підприємство «Українське агентство з авторських та суміжних прав» (Укрпатент). 2024. Жовтень. URL: https://ukrpatent.org/.
5. Ukraine's copyright reform: what's new in 2023? CMS Law-Now. 2023. January 23. URL: https://cms-lawnow.com/en/ealerts/2023/02/ukraine-reforms-copyright-regulation-system-in-line-with-eu-standards.
6. Czech Court denies copyright Protection of AI-Generated Work in First Ever Ruling. *Bird & Bird | International Law Firm*. URL: https://www.twobirds.com/en/insights/2024/czech-republic/czech-court-denies-copyright-protection-of-ai-generated-work-in-first-ever-ruling.
7. Guidance on Artificial Intelligence and Copyright. European Commission, Directorate-General for Communications Networks, Content and Technology. 2024. July. Brussels. URL: https://digital-strategy.ec.europa.eu/en/library/commission-publishes-guidelines-prohibited-artificial-intelligence-ai-practices-defined-ai-act.
8. Thomson Reuters v. ROSS Intelligence: Court Decision on Copyright and AI Training Data. United States District Court for the District of Delaware. 2025. February. Case No. 1:20-cv-00613. URL: https://www.ded.uscourts.gov/sites/ded/files/opinions/20-613_5.pdf.....
9. Press A. Judge approves $1.5 billion copyright settlement between AI company Anthropic and authors. *Santa Cruz Sentinel*. URL: https://www.santacruzsentinel.com/2025/09/25/anthropic-authors-copyright/?gad_source=1&gad_campaignid=23187953256&gbraid=0AAAAAqZCio_nYlLfkfEmvWNhmu-7lipBt&gclid=Cj0KCQiAoZDJBhC0ARIsAERP-F_3OynSCKruCf5vCbmxVu4rfUZq1QL77nDUUrPWYcVy5jvqPf4Il9IaAsmDEALw_wcB.
10. Generative Artificial Intelligence and Copyright Law / E. L. Silva. Congressional Research Service Report. Washington, DC: CRS, 2025. URL: https://www.congress.gov/crs-product/LSB10922.

С.В. Гаврилюк

РОЗРОБКА АЛГОРИТМУ ПОБУДОВИ ТА ОПТИМІЗАЦІЇ ТУРИСТИЧНИХ МАРШРУТІВ З ВИКОРИСТАННЯМ ГІС ТА ШТУЧНОГО ІНТЕЛЕКТУ

Вступ

Сучасна туристична галузь переживає активну цифрову трансформацію, що супроводжується зростанням вимог до точності навігації, персоналізації та безпеки пересування. Традиційні навігаційні системи не можуть забезпечити достатній рівень адаптивності, оскільки зазвичай оперують статичними даними та обмежено враховують індивідуальні особливості користувача. Туристи очікують на інтелектуальні рішення, здатні не лише будувати оптимальні маршрути, а й реагувати на зміни в режимі реального часу, аналізувати навколишнє середовище та підлаштовуватися під контекст ситуації [1].

Геоінформаційні системи (ГІС) тривалий час залишалися базовим інструментом для здійснення просторового аналізу в туристичній сфері, однак їх функціональність значною мірою обмежена застосуванням традиційних підходів до обробки даних. На відміну від цього, технології штучного інтелекту дають змогу працювати з багатовимірними залежностями, здійснювати прогнозування динаміки туристичних потоків, визначати потенційні ризики та знаходити оптимальні маршрути навіть у складних просторових умовах. Поєднання ГІС і ШІ відкриває можливість створення нових типів маршрутів, у яких враховуються не лише стандартні критерії, а й фактори середовища, поведінкові особливості користувачів, часові зміни та ситуативні впливи (2).

Використання інтелектуальних моделей забезпечує опрацювання значно ширшого спектра інформації — від традиційних картографічних джерел до динамічних потокових даних про погодні умови, рух транспорту та концентрацію людей. Такий підхід дозволяє формувати маршрути, що адаптуються до реальних умов і враховують рівень комфорту, можливі ризики, прогнозовані зміни в оточенні та загальну інтенсивність людських потоків. Створення алгоритму, який поєднує ГІС-аналітику з методами інтелектуальної оптимізації та прогнозування, є ключовим етапом у розвитку сучасних туристичних навігаційних систем (3).

Таким чином, актуальність проблеми полягає у формуванні такого алгоритмічного рішення, яке дозволить створювати маршрути, що є не лише оптимальними за формальними критеріями, а й адаптивними, безпечними, персоналізованими та стійкими до непередбачуваних факторів. Саме тому виникає потреба у комплексному підході, який охоплює багатошарову просторову обробку, прогнозні моделі та алгоритми інтелектуальної оптимізації [4].

Виклад основного матеріалу

Розроблений алгоритм оптимізації туристичних маршрутів базується на багаторівневій структурі обробки даних, де кожний інформаційний шар виконує окрему функцію, але водночас взаємодіє з іншими. На початковому етапі система збирає та структурує просторові дані, що охоплюють туристичні локації, транспортні мережі, інфраструктурні об'єкти, динаміку руху транспортних засобів, погодні характеристики, рівень завантаженості об'єктів та можливі ризикові параметри. Зібрані дані піддаються геокодуванню та нормалізації в межах ГІС середовища, що дозволяє забезпечити їх просторову сумісність та коректне накладання шарів [5].

Після цього система формує два ключових показника — індекс привабливості та індекс безпеки. Індекс привабливості відображає культурну, історичну, рекреаційну та емоційно-естетичну цінність об'єкта з урахуванням його популярності, рейтингу серед туристів, тривалості відвідування та типу активності, яку він пропонує. Індекс безпеки оцінює потенційні ризики, серед яких: щільність транспортних потоків, доступність медичної та допоміжної інфраструктури, георизики та історична частота інцидентів. Кожний з індексів отримує вагову оцінку, що дозволяє формувати багатофакторний профіль кожної локації [6].

Далі запускається оптимізаційний модуль, який поєднує модифікований алгоритм A*, що виконує локальний пошук оптимальних шляхів, з генетичним алгоритмом, який генерує різні конфігурації маршрутів, оцінює їх і поступово знаходить найкраще рішення. Такий механізм дозволяє уникати локальних мінімумів, що характерно для класичних алгоритмів маршрутизації, та досягати глобальної оптимальності. Додаткову цінність створює персоналізований модуль аналізу профілю користувача, який враховує рівень фізичної активності, темп пересування, втомлюваність, уподобання щодо типів локацій і бажаний рівень складності маршруту [7].

Окремою складовою системи виступає модуль прогнозування, завданням якого є аналіз можливих змін у навколишньому середовищі. У його основі — моделі машинного навчання, що дозволяють передбачати виникнення заторів, зростання черг, зміни погоди, поява транспортних обмежень та коливання інтенсивності туристичних потоків. Завдяки такій аналітиці під час формування маршруту враховується не лише актуальна ситуація, а й те, якою вона може стати найближчим часом, що робить планування точнішим і більш стійким до зовнішніх факторів. Наявність прогностичного модуля також забезпечує можливість гнучкого перепланування: система автоматично коригує маршрут при появі нових умов або загроз, що істотно підвищує практичність та надійність навігаційного рішення (8).

Результати моделювання свідчать про суттєві переваги алгоритму. Застосування інтелектуальних методів дозволило скоротити загальний час у дорозі, зменшити кількість затримок, уникати ризикових зон, покращити

послідовність переміщення та зробити маршрут комфортнішим і безпечнішим. Алгоритм продемонстрував здатність до навчання на основі історичних даних, що дає змогу підвищувати якість рішень при повторному використанні та розширювати можливість індивідуальної адаптації [9].

Висновки

У результаті проведеного дослідження сформовано комплексний алгоритм побудови та оптимізації туристичних маршрутів, який базується на інтеграції геоінформаційних технологій та методів штучного інтелекту. Запропонована методика дає змогу охопити значну кількість чинників, серед яких просторові властивості об'єктів, показники їх привабливості й безпеки, особисті характеристики користувача, прогнозні зміни у довкіллі та можливість використання альтернативних маршрутів. Такий алгоритм показує здатність працювати з багатьма критеріями одночасно, залишаючись гнучким, стійким до неочікуваних ситуацій та здатним вносити корективи у маршрут безпосередньо під час руху (10).

Подальший розвиток цієї системи може передбачати використання IoT-пристроїв для збору оперативних даних, аналіз потокової інформації про мобільність населення, застосування глибинних нейронних мереж для точнішого прогнозування поведінки туристів, а також розширення набору екологічних і безпекових параметрів, що впливають на формування маршруту.

Створення таких систем здатне забезпечити нову якість туристичної навігації та сприяти розвитку smart-tourism та інтелектуальних міських екосистем [11].

1. Борта В.М., Ляшенко Т.О. Геоінформаційні системи в туризмі: навчальний посібник. Київ: КНУКіМ, 2020.
2. Мельничук А.А., Журбас В.М. Основи штучного інтелекту: навчальний посібник. Київ: КНЕУ, 2021.
3. Клечановський А.В., Корнієнко І.В. ГІС-технології: теорія та практика. Харків: ХНУ ім. В.Н. Каразіна, 2019.
4. Шарапов О.І., Мартинюк О.В. Методи просторового аналізу у ГІС. Львів: ЛНУ ім. І. Франка, 2018.
5. Воробйов А.І. Геоінформаційні системи: навчальний посібник. Київ: НАУ, 2020.
6. Босик З.М. Основи штучного інтелекту. Київ: КНЕУ, 2021.
7. Герасименко В.В., Дроздов О.О. Інформаційні технології моделювання та оптимізації складних систем. Київ: КНУ ім. Т. Шевченка, 2017.
8. Литвин В.В., Рудень В.І., Ситник В.В. Інтелектуальні системи підтримки прийняття рішень. Львів: Видавництво Львівської політехніки, 2020.
9. Стасюк В.В. Моделювання та прогнозування у використанні просторових даних. Київ: НТУУ «КПІ ім. І. Сікорського», 2019.
10. Ковальчук В.П., Яровий І.Є. Інформаційні системи та технології в управлінні. Київ: КНЕУ, 2018.
11. Сенько В., Коваленко О. Технології інтелектуального аналізу даних: навчальний посібник. Харків: ХНЕУ ім. Кузнеця, 2020.

І.А. Полонка

ВПЛИВ ТЕХНОЛОГІЙ ШТУЧНОГО ІНТЕЛЕКТУ НА ТРАНСФОРМАЦІЮ ПРАВОВОЇ ПОВЕДІНКИ, ПРАВОСВІДОМОСТІ ТА ПРАВОВОЇ КУЛЬТУРИ

Сучасне правове життя розвивається в умовах стрімкої технологічної модернізації, і одним із центральних її драйверів є технології штучного інтелекту. Вони дедалі активніше проникають у правову сферу, змінюючи не лише технічні аспекти опрацювання інформації, а й механізми формування правосвідомості, характер правової культури та щоденні практики правової поведінки. На перший погляд може здаватися, що штучний інтелект є лише інструментом, покликаним спростити комунікацію між громадянином і правовою системою. Проте його вплив значно ширший: у певному сенсі він стає елементом середовища, у якому вибудовується структура правового мислення.

Однією з ключових змін є здатність алгоритмічних систем адаптувати правові положення до рівня розуміння конкретної людини. Це принципово відрізняє штучний інтелект від традиційних джерел правової інформації. У рекомендаціях ЮНЕСКО наголошується, що доступність і зрозумілість норм права є необхідною умовою формування якісної правової культури [1]. Алгоритми, які здатні пояснити зміст норми права, структурувати її, подати у наочному чи прикладному форматі, фактично виконують функцію посередника між правом і його адресатом. У такому підході проявляється новий тип правової соціалізації – коли особа не просто читає норму, а взаємодіє з нею у пояснювальному, контекстуалізованому форматі.

Разом із пояснювальними можливостями важливою є і аналітична функція штучного інтелекту. У документах ОЕСР підкреслюється, що алгоритмічні системи, які застосовуються в юридично значущих сферах, мають бути прозорими, підзвітними й керованими [2]. Це пов’язано з тим, що такі технології здатні прогнозувати можливі сценарії розвитку подій, а отже – формувати певне бачення того, яка модель поведінки є більш вірогідно «правильною». Якщо система підказує громадянинові, які строки подання документів є критичними або які наслідки може мати порушення певної процедури, то така взаємодія безпосередньо впливає на його поведінкові рішення. Проте водночас алгоритм не може виступати абсолютним орієнтиром: він функціонує на підставі певних наборів даних і методів оброблення, а тому потребує зовнішнього контролю.

Найбільш показовим прикладом ризику алгоритмічної упередженості є історія застосування системи COMPAS у кримінальному судочинстві США. Ця система використовувалася судами для оцінки ймовірності рецидиву та визначення ступеня небезпечності підсудних. На перший погляд COMPAS виглядала як нейтральний інструмент: вона аналізувала відповіді особи на

стандартизований набір запитань, порівнювала їх з історичними даними й формувала числовий прогноз ризику. Проте незалежне розслідування ProPublica виявило, що результати цієї системи були прямо залежними від того, які дані закладено в алгоритм і які соціальні упередження ці дані відтворювали [3]. Наприклад, було встановлено, що для афроамериканських підсудних система частіше формувала завищені прогнози рецидиву, тоді як для білих підсудних ризики нерідко занижувалися, навіть за наявності фактичних порушень у минулому.

Це не означає, що сама технологія була «упередженою» у технічному сенсі. Джерело проблеми лежало у структурі даних, на яких її навчали. Якщо вихідні масиви інформації містили соціально зумовлені деформації – наприклад, різну частоту затримань, умови проживання чи нерівний доступ до адвокатської допомоги, – то алгоритм фактично «успадковував» ці перекоси та відтворював їх у прогнозах. Таким чином, система, яка мала посилити об'єктивність правосуддя, на практиці могла створювати додаткові нерівності.

У міжнародному правовому дискурсі цей випадок набув статусу ключового етичного прецеденту. Він засвідчив, що навіть математично коректна модель може генерувати соціально несправедливі результати, якщо її робота залишається непрозорою або якщо формування алгоритму здійснюється без належного урахування структурних факторів нерівності. Саме тому COMPAS стала прикладом для численних наукових досліджень, у яких підкреслюється: упередженість не є властивістю штучного інтелекту як технології, але є наслідком того, які дані використовуються, як налаштовані методи оцінки та які ціннісні орієнтири закладені в основу системи.

Ситуація з COMPAS актуалізувала кілька ключових питань: необхідність аудиту алгоритмічних систем, забезпечення прозорості принципів їх роботи, відкритості даних, на яких вони навчаються, а також важливість багатодисциплінарного контролю – з боку юристів, соціологів, етиків і технічних фахівців. Адже саме на перетині цих сфер виявляється можливість вчасно розпізнати потенційні викривлення та запобігти тому, щоб автоматизовані рішення впливали на долі людей несправедливим чином. У підсумку кейс COMPAS перетворився на символ того, що навіть найсучасніші технології потребують етичного контролю, людського нагляду та необхідних гарантій захисту прав людини.

Крім ризиків, пов'язаних із дискримінаційними деформаціями, важливо враховувати і психологічний аспект. У суспільстві поступово формується звичка звертатися до автоматизованих порад при вирішенні правових питань. Якщо цей процес відбувається без паралельного розвитку правової грамотності, може виникнути тенденція до зниження критичного мислення. Людина, яка звикає покладатися на алгоритм, може почати сприймати рекомендацію як однозначну істину, а не як інструмент для аналізу. Саме тому міжнародні організації наголошують на потребі збереження автономії

особи у процесі прийняття рішень, навіть якщо штучний інтелект здатен надати розгорнуту консультаційну підтримку.

Україна сьогодні перебуває в ситуації, коли ці питання набувають особливої ваги. Держава активно розвиває електронні сервіси, а технології штучного інтелекту все частіше застосовуються в інформаційно-аналітичних процесах. Це створює можливості для підвищення відкритості та доступності адміністративних процедур, водночас ставлячи вимогу сформувати систему етичного та юридичного нагляду за впровадженням таких рішень. Українські реалії підтверджують: штучний інтелект може сприяти розвитку правової культури, але лише тоді, коли його використання ґрунтується на принципах відповідальності, прозорості та поваги до прав людини.

Підсумовуючи, можна сказати, що технології штучного інтелекту сьогодні формують нову якість правової поведінки. Вони підсилюють здатність громадян до осмисленого правового вибору, сприяють доступності правової інформації, розширюють інструментарій правової освіти та створюють умови для розвитку сучасної правової культури. Водночас вони демонструють і серйозні виклики, які потребують уваги – від ризиків упередженості до небезпеки втрати критичної автономії особи. Саме баланс між цими аспектами визначатиме майбутнє взаємодії суспільства з технологіями штучного інтелекту та майбутнє права як інструменту суспільного порядку.

1. UNESCO. (2021). Recommendation on the ethics of artificial intelligence. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000381137.
2. OECD. (2019). Recommendation of the Council on Artificial Intelligence. OECD Legal Instrument No. OECD/LEGAL/0449. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.
3. Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: Risk assessments in criminal sentencing. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Я.В. Анікієнко

ВПЛИВ ШІ НА СУСПІЛЬСТВО: ПРАВОВА ПЕРСПЕКТИВА ДЛЯ АДВОКАТІВ

Розвиток штучного інтелекту (далі - ШІ) сьогодні сильно впливає на життя суспільства, у тому числі й на роботу юристів. Використання ШІ в адвокатській діяльності допомагає автоматизувати звичні завдання, швидше опрацьовувати велику кількість інформації та полегшувати спілкуванн між адвокатом та клієнтом. Усе це відкриває нові можливості для більш ефективної роботи адвокатів.

У цьому контексті, Бєлов Д.М. та Бєлова М.В. стверджують, що «одним із основних напрямів автоматизації в юридичній сфері є опрацювання та аналіз великих обсягів правової інформації – судових рішень, нормативних актів, прецедентів та інших джерел. Технології ШІ здатні застосовувати алгоритми машинного навчання для опрацювання текстів і виявлення типових моделей у судових рішеннях. Це забезпечує швидкий доступ до потрібних даних і дає можливість ефективно досліджувати судову практику, що допомагає приймати більш виважені правові рішення» [2, с.317].

Загалом, автоматизація судових процесів за допомогою ШІ має великий потенціал: системи можуть аналізувати факти, аргументи та попередні рішення. Але тут варто наголошувати на тому, що остаточне рішення має залишатися саме за суддею, який враховує всі нюанси справи. Використання ШІ дозволяє працювати швидше, зменшує помилки та робить процес більш об’єктивним.

Однак, використання таких технологій супроводжується значними проблемами. На думку Матвєєва В.Ю., до них можна віднести: виникнення питань юридичної відповідальності за помилки, допущені алгоритмами, забезпечення прозорості та недопущення упередженості у рішеннях, а також безпеку обробки конфіденційної інформації клієнтів [1, с. 63].

По-перше, виникає питання відповідальності: хто має відповідати за неправильне рішення - адвокат, який застосував інструмент, чи розробник програмного забезпечення? У Франції, наприклад, вже обговорювалося, що використання прогнозних систем може спровокувати помилкові рішення адвоката і таким чином підірвати довіру клієнтів. Крім того, адвокати адвокати також стикаються з низкою етичних проблем під час застосування штучного інтелекту, особливо тоді, коли автоматизовані рішення здатні безпосередньо впливати на права та подальшу долю людини. Так, у січні 2024 року адвокатам Флориди було надано нові етичні вказівки щодо використання генеративних технологій штучного інтелекту в їхній професійній діяльності [1, с. 66].

По друге, використання сторонніх ШІ-платформ, особливо хмарних, завжди пов’язане з передачею даних клієнта третім особам – розробникам програмного забезпечення та власникам сервісів. Адвокат повинен запитати

себе: наскільки надійно захищені ці дані? Чи відповідає політика конфіденційності розробника вимогам профільного законодавства? Завантажуючи матеріали справи в онлайн-систему для аналізу, адвокат ризикує ненавмисно порушити адвокатську таємницю [3, с.251]. Тобто адвокат зобов'язаний не лише користуватися інструментом, а й ретельно перевіряти його на безпеку та надійність.

По-третє, цікавим і водночас складним є питання так званої «чорної скриньки» та права на справедливий суд. Багато сучасних систем штучного інтелекту працюють на основі машинного навчання, і часто їхню роботу неможливо повністю пояснити. Вони дають результат, але навіть розробники не завжди можуть чітко описати, як саме система до нього дійшла. Виходячи з цього, виникає проблема: як адвокат може переконливо обґрунтувати свою позицію в суді, якщо вона спирається на висновок ШІ, логіку якого він сам не може детально пояснити? Це суперечить принципу змагальності та праву сторони на зрозуміле й мотивоване рішення [3, с. 252].

Слід звернути увагу на те, що вже існує практика де адвокати у своїх процесуальних документах посилались на неснуюючі норми, які відповідно придумані ШІ. Зокрема в рішенні № 420/26668/25 від 11 листопада 2025 року, суд зазначає «представник позивача спирається на буквально вигадані норми, які не існують в національному законодавстві України, що притаманно неперевіреній роботі штучного інтелекту. В цьому аспекті суд звертає увагу представника позивача на гайд Council of Bars and Law Societies of Europe (CCBE) Guide on the use of Artificial Intelligence-based tools by lawyers and law firms in the EU (2022), в якому зазначається, що якщо адвокат застосовує ШІ-технології, він не може пасивно покладатись на них і тим самим уникати свого професійного обов`язку. Навіть при використанні ШІ інструментів адвокат має залишатися професійно відповідальним за якість, законність та етичність своєї практики. Використання ШІ не знімає з нього зобов`язань щодо конфіденційності, компетентності, незалежності і юридичної відповідальності. Якщо адвокат використав генеративний ШІ-інструмент, але не перевірив його результати, або передав дані клієнта так, що порушено конфіденційність, то це може вважатись недбалістю або порушенням професійних обов`язків. Відповідальність адвоката тут має три ключові аспекти: він має знати і розуміти інструменти (компетентність), забезпечувати конфіденційність клієнта та перевіряти / контролювати те, що генерує система» [4].

Враховуючи викладене, на нашу думку, зазначена проблема сьогодні є доволі поширеною і здатна спричинити серйозні, а подекуди й незворотні наслідки під час судового розгляду. Зокрема, вона може вплинути на кінцевий результат справи у випадку, коли суддя, не перевіривши повністю наведені в позовній заяві норми, помилково враховує положення, які фактично не існують або не мають юридичної сили. Така ситуація створює ризик ухвалення неправомірного рішення та порушення одного з ключових

засадничих принципів правосуддя – принципу законності та обґрунтованості судового рішення, що вимагає, щоб будь-яке рішення ґрунтувалося виключно на чинних нормах права та достовірних фактах.

Враховуючи викладене, органам адвокатського самоврядування в Україні, зокрема Національній асоціації адвокатів України, необхідно розробити та впровадити спеціальні роз'яснення або доповнення до Правил адвокатської етики щодо використання штучного інтелекту. При цьому ми можемо орієнтуватися на міжнародний досвід – наприклад, на згадані рекомендації Ради адвокатських та правничих товариств Європи, які можуть стати якісним зразком для таких нововведень.

Отже, можна зробити висновок, що штучний інтелект стає потужним інструментом у руках адвоката, здатним покращити юридичний аналіз, проте його застосування межує з високими професійними ризиками. Як свідчить практика, неконтрольоване використання генеративних моделей здатне призвести до спотворення правової дійсності та ухвалення необґрунтованих рішень Це зумовлює необхідність запровадження нової моделі «цифрової відповідальності» адвоката, заснованої на положенні: кожен алгоритмічний висновок потребує перевірки людиною. Лише за умови впровадження жорстких етичних фільтрів та усвідомлення того, що ШІ є лише допоміжним інструменто, а не арбітром, можливо забезпечити баланс між технологічним прогресом та неухильним дотриманням прав людини.

1. Матвєєв В.Ю. Штучний інтелект в адвокатській діяльності: можливості, виклики та перспективи правового регулювання. *Право і суспільство.* 2024. № 6. С. 63–69.
2. Бєлов Д., Бєлова М. Штучний інтелект в судочинстві та судових рішеннях, потенціал та ризики. *Науковий вісник Ужгородського Національного Університету.* Серія: Право. 2023. Вип. 2, № 78. С. 315–320.
3. Храпенко О. О., Меденцев А. М., Сперанський В. О. Етичні дилеми адвокатської діяльності в умовах застосування штучного інтелекту у правосудді. Юридичний науковий електронний журнал. 2025. № 6. С. 250–254.
4. Рішення Одеського окружного адміністративного суду від 11.11.2025 у справі № 420/26668/25. URL: https://reyestr.court.gov.ua/Review/131701679 (дата звернення: 26.11.2025).

І.В. Левицька

ШТУЧНИЙ ІНТЕЛЕКТ У ТУРИЗМІ

Сучасний розвиток туристичної галузі відбувається в умовах стрімкого поширення цифрових технологій, серед яких особливе місце посідає штучний інтелект. ШІ імітує людський розум і має різні функції та можливості. Залежно від цього, він поділяється на різні типи, які продемонстровано у табл. 1. Наразі штучний інтелект спеціалізується тільки на конкретних завданнях, а тому є слабким (вузьким) – використовується тільки Narrow AI. General AI та Superintelligent AI ще не існують і є теоретичними.

Таблиця 1 – Види штучного інтелекту [1, с.217]

Класифікація ШІ	Види штучного інтелекту
1. На основі можливостей	Вузький/слабкий ШІ – Narrow AI (Weak AI; Artificial Narrow Intelligence; ANI)
	Загальний/сильний ШІ – General AI (Strong AI; Artificial General Intelligence; AGI)
	Суперінтелект – Superintelligent AI (Artificial Super intelligence; ASI)
2. На основі функціональності	Реактивні машини (Reactive Machines)
	Обмежена пам'ять (Limited Memory)
	Теорія розуму (Theory of Mind)
	Самосвідомий ШІ (Self-aware AI)

Застосування штучного інтелекту суттєво розширює можливості туристичної сфери – від обробки даних до вдосконалення сервісів і взаємодії з мандрівниками. У результаті саме ця технологія стає каталізатором головних змін, що сьогодні визначають розвиток туризму. Основні трансформації, зумовлені впровадженням штучного інтелекту, охоплюють такі основні напрями [2; 3]:

– *Автоматизація процесів, планування та бронювання подорожей*. Штучний інтелект може персоналізувати подорожі, пропонуючи рекомендації та маршрути на основі уподобань, бюджету й попереднього досвіду мандрівника. AI-алгоритми аналізують великі обсяги даних – від розкладів рейсів до цін на готелі й популярності локацій – щоб формувати індивідуальні поради у туристичних застосунках. Крім того, чат-боти покращують процес бронювання, забезпечуючи миттєві відповіді, допомогу в реальному часі та автоматичну обробку бронювань, що економить час як мандрівникам, так і турагентам.

– *Обслуговування клієнтів та віртуальні помічники*. ШІ покращує обслуговування клієнтів у туристичній сфері, надаючи цілодобову підтримку через інтелектуальних віртуальних помічників. Вони допомагають не лише з бронюванням, а й дають персоналізовані рекомендації: наприклад, підкажуть найкраще місце для вечері відповідно до вашого настрою та уподобань, роблячи подорож більш комфортною та приємною. Також вони можуть

відповідати на запити та скарги, допомагати з організацією поїздок і навіть забезпечувати переклад у режимі реального часу, роблячи подорожі зручнішими для людей з усього світу.

– *Розумний транспорт і логістика*. Алгоритми ШІ оптимізують транспорт і логістику в туристичній індустрії, допомагаючи авіакомпаніям та круїзним лініям планувати ефективніші маршрути, зменшувати витрати на пальне, скорочувати час у дорозі й покращувати загальний досвід подорожей.

– *Аналіз даних і персоналізація*. ШІ може аналізувати дані з соціальних мереж, відгуків користувачів і вебсайтів, щоб визначати ключові фактори, які впливають на вибір мандрівників. Завдяки таким класифікаційним алгоритмам туристичні компанії глибше аналізують вподобання клієнтів і можуть формувати більш персоналізовані пропозиції.

– *Аналіз поведінки користувачів*. ШІ навчається на даних повсякденного досвіду користувача. Система запам'ятовує ваші попередні бронювання та уподобання, щоб наступного разу пропонувати ще більш точні рекомендації для подорожей.

– *Прогнозування попиту*. Штучний інтелект здатен прогнозувати популярність курортів, аналізуючи сезонність, кліматичні зміни, економічні показники та інші важливі фактори. Це дає змогу компаніям точніше планувати свої послуги й пропонувати мандрівникам більш актуальні варіанти відпочинку.

Однак, крім значних переваг не слід забувати, що ШІ не досконалий і може створювати фейкову або неточну інформацію і цим самим вводити туристів в оману. Наприклад, літня пара у Малайзії шукала неіснуючу канатну дорогу «Kuak Skyride». Туристи в Перу шукали вигаданий «Священний каньйон Юмантай», що поєднував дві реальні, але окремі локації. Блогерка в Японії на підказку ChatGPT опинилася на вершині гори, коли підйомник був зачинений. [4] Тож важливо завжди перевіряти ключову інформацію на офіційних сайтах, актуальних онлайн-картах та у свіжих відгуках інших мандрівників.

Отже, штучний інтелект істотно змінює туристичну сферу, роблячи сервіси більш персоналізованими, ефективними та технологічно інтегрованими. Його інтеграція змінює способи планування подорожей, організації сервісів і взаємодії туристів із дестинаціями, формуючи нову модель розвитку галузі.

1. Levytska, I. (2024). Artificial intelligence in digital branding of territories. In O. Prokopenko, K. Välk, & A. Neroda (Eds.), Smart machines and technologies at the service of mankind (pp. 216–228). Teadmus OÜ. https://conference.euas.eu/2024/wp-content/uploads/2024/12/Monograph_2024.pdf.
2. Ялалов, Д. & Гащ, К. (2023). Вплив штучного інтелекту на туристичну індустрію. Mpost. https://mpost.io/uk/the-impact-of-artificial-intelligence-on-the-travel-industry/.
3. SalesBox. (б.д.). Як штучний інтелект змінює туризм: персоналізовані рекомендації для ідеальної подорожі. https://salesbox.ua/blog/yak-shtuchnyi-intelynkt-zminyuye-turizm-personalizovani-rekomendatsii-dlya-idealnoi-podorozhi/
4. Довгопол, В. (2025). Туристи їдуть у вигадані місця за порадами ChatGPT. Мультиканальне медіа про подорожі dovkola.media. https://dovkola.media/turysty-idut-u-vyhadani-mistsia-za-poradamy-chatgpt.

О.В. Новікова, І.П. Суїма

SECURE DIGITAL ASSESSMENT: AI-DRIVEN TESTING TOOLS FOR ENGLISH PROFICIENCY

Secure digital assessment has become an urgent priority in modern education as artificial intelligence rapidly reshapes the ways English proficiency is measured, interpreted, and certified. The integration of AI-driven testing tools offers unprecedented opportunities for personalized evaluation, real-time analysis of learner performance, and scalable testing environments. At the same time, it raises essential questions concerning data privacy, algorithmic transparency, academic integrity, and the need for robust security protocols that protect test-takers, institutions, and the credibility of results [1]. As digital learning expands globally and online English assessment becomes increasingly common, educational institutions must rethink how to ensure that AI-enhanced testing remains both effective and secure.

English proficiency assessment historically relied on standardized, paper-based, and proctored examinations that positioned human examiners as gatekeepers of reliability and fairness. Today, however, AI-based systems can automatically evaluate reading, listening, writing, and even speaking. These tools use machine learning algorithms to analyze language patterns, detect grammatical accuracy, measure vocabulary depth, assess speech fluency and pronunciation, and track improvement over time. Their promise lies in the ability to process thousands of responses within seconds, offer adaptive item difficulty tailored to each learner's performance, and provide detailed feedback at a scale impossible for human assessors. Yet the shift from human scoring to AI scoring raises new concerns about the integrity of digital testing ecosystems [2].

Security in AI-driven English assessment can be conceptualized at several interconnected levels: technological security, data security, academic security, and ethical security. Technological security involves protecting the assessment platform itself from external attacks, unauthorized access, and manipulation. AI-based testing systems often rely on cloud infrastructure, storing test items, student profiles, and response datasets on remote servers. This makes them vulnerable to cyberattacks such as data breaches, DDOS, item harvesting, and automated cheating through bots. Ensuring platform security requires encrypted data transmission, secure authentication protocols, intrusion detection systems, and continuous monitoring of system vulnerabilities. Without such measures, AI-enhanced examinations can become targets for hackers seeking to exploit institutional weaknesses.

Data security is even more critical, especially because AI-driven assessments require large datasets to function effectively. AI needs access to user profiles, linguistic samples, voice recordings, keystroke patterns, and performance histories to train models, generate scores, and refine assessment algorithms. This learning process inevitably involves collecting and processing sensitive personal information. As a result, institutions must comply with data protection regulations such as GDPR, CCPA, and national education privacy laws. Transparency about

what data is collected, how it is used, and who has access to it is essential. The ethical risk is twofold: on the one hand, unauthorized access could expose sensitive information; on the other, misuse of learner data by commercial providers could result in surveillance, profiling, or monetization without informed consent. Secure digital assessment therefore requires a privacy-by-design approach that minimizes unnecessary data collection and ensures all processed information is anonymized or pseudonymized whenever possible [3].

Another dimension involves academic integrity and safeguarding the fairness of AI-driven English proficiency tests. As AI becomes more embedded in assessment, it not only evaluates test-taker responses but also creates new avenues for cheating. Students can attempt to exploit system vulnerabilities by using AI tools themselves during tests, ranging from grammar checkers to fully automated writing generators. Voice verification systems can be tricked with synthesized speech, while webcam monitoring algorithms can sometimes be bypassed. Ensuring integrity requires integrated proctoring mechanisms such as biometric authentication, keystroke dynamics, eye-tracking tools, and plagiarism detection specifically adapted to AI-generated content. However, each of these raises new questions about privacy and user rights. The challenge lies in achieving a balance between creating a secure testing environment and respecting the dignity and autonomy of learners.

Adaptive AI-driven systems provide additional benefits while simultaneously introducing security concerns. These systems adjust the difficulty of test items in real time based on student performance. Although this approach increases precision in estimating proficiency levels, the algorithm must be guarded against item exposure, prediction attacks, and reverse engineering. If students or external actors gain access to item banks or understand the adaptation patterns, test validity can be compromised. Therefore, secure item rotation strategies, encrypted item repositories, and dynamic item generation techniques are essential. Some AI-based platforms already use neural networks to generate novel test items on the fly, reducing the risk of item theft. Still, the reliability and validity of AI-generated items must be scrutinized to ensure they accurately measure language proficiency rather than reflect algorithmic quirks or biases within the training data.

Bias in AI-driven assessment is another challenge that intersects with security concerns. AI systems learn from existing datasets, and if these datasets reflect biases related to dialect, gender, accent, or socioeconomic background, the resulting scoring will inherit and amplify those biases. For English proficiency tests, accent bias is especially concerning. Speech recognition models trained predominantly on native or standardized accents may inaccurately score test-takers with diverse linguistic backgrounds. Mechanical scoring of writing may penalize rhetorical structures common in non-Western cultures. Ensuring the fairness and security of AI-driven assessment requires auditing datasets, examining algorithmic outputs, and implementing continuous bias mitigation mechanisms. Security, therefore, is not merely technical—it also involves securing equity and fairness in the scoring process.

Despite these challenges, AI-driven tools offer transformative pedagogical benefits. They can provide immediate feedback on speaking and writing tasks,

track learner progress longitudinally, highlight individual strengths and weaknesses, and create personalized learning pathways. This is particularly valuable in large-scale educational environments where human assessment would require extensive resources. Moreover, AI enhances test security by automating irregularity detection: unusual response patterns, abnormal timing, improbable score increases, or answers mirroring known cheating databases can be flagged instantly. AI can analyze digital behavior patterns that human proctors would never detect, such as subtle inconsistencies in typing rhythm or mismatched acoustic signatures in voice recordings. In this sense, AI simultaneously creates new vulnerabilities and new protections.

One of the major debates concerns the transparency of AI scoring mechanisms. Many commercial testing platforms use proprietary models that do not disclose how specific responses are evaluated. This lack of transparency raises concerns about validity, fairness, and accountability. For secure digital assessment to function effectively, test-takers and institutions must have confidence in the scoring system. Increasingly, educational researchers call for explainable AI (XAI) approaches in English proficiency assessment, where algorithms provide justifications for their scores or highlight specific linguistic features influencing decisions. Explainable scoring enhances both pedagogical value and ethical trust. Without transparency, AI-based scores may be questioned by institutions, employers, and students, undermining the credibility of digital certification.

Integration of AI-driven testing tools also requires institutions to develop clear policies and guidelines. Teachers must be trained in the effective use of AI tools, understanding their strengths and limitations. Students must be educated about digital safety, responsible use of technology, and ethical behavior during online assessments. Administrators must adopt standardized procedures for responding to suspected cybersecurity incidents or academic misconduct identified by AI systems. Moreover, the shift to secure digital assessment demands cooperation between educators, IT specialists, cybersecurity professionals, and AI researchers to ensure that technology aligns with pedagogical objectives and ethical values.

In the broader educational landscape, secure AI-driven English proficiency testing contributes to global accessibility. Online assessments supported by AI make it possible for students in remote or underserved regions to obtain internationally recognized certifications without traveling to testing centers. This creates opportunities for inclusive education but simultaneously places responsibility on institutions to protect vulnerable learners from exploitation, data misuse, or technological manipulation. Ensuring that AI-based assessment remains equitable, secure, and ethical is therefore central to building trust in digital learning ecosystems.

The future of secure digital assessment in English proficiency lies in the synergy between AI advancements and responsible governance. Continuous monitoring of systems, regular security audits, transparent scoring practices, and adherence to privacy standards will define the credibility of AI-based testing. At the same time, interdisciplinary collaboration will shape robust frameworks that protect learners' rights while promoting innovation.

In conclusion, AI-driven testing tools represent a paradigm shift in evaluating English proficiency, offering highly scalable, adaptive, and data-rich methods for measuring language skills. However, with great technological capability comes a profound responsibility to protect system integrity, personal data, fairness, and academic honesty. Ensuring secure digital assessment requires more than software solutions; it demands a holistic academic and ethical commitment to building trust, safeguarding learners, and designing AI systems that enhance rather than compromise the educational experience. As the digital transformation of language assessment accelerates, institutions must prioritize responsible AI implementation, ensuring that technological progress is matched by secure, equitable, and transparent testing practices that uphold the values of modern education.

1. Emirtekin, E. Large Language Model-Powered Automated Assessment: A Systematic Review. Applied Sciences, 15(10), 2025. P. 1–24.
2. Zhang, T., Erlam, R., & de Magalhães, M. Exploring the dual impact of AI in post-entry language assessment: Potentials and pitfalls. Annual Review of Applied Linguistics, published online June 2025.
3. Zhyhadlo, O., & Zaiarna, I. Artificial Intelligence-Driven Testing in EFL/ESP Classrooms. Artificial Intelligence in Education and Scientific Research, Vol. 106, No. 2, 2025, P. 45–62.

А.І. Коваль, Б.В. Ліщук

ВИКОРИСТАННЯ ТЕХНОЛОГІЙ ШТУЧНОГО ІНТЕЛЕКТУ ЩОДО ОПТИМІЗАЦІЇ ПРИЙНЯТТЯ РІШЕНЬ ПІД ЧАС ВИНИКНЕННЯ НАДЗВИЧАЙНИХ СИТУАЦІЇ ТА НЕБЕЗПЕЧНИХ ПОДІЙ

Сучасний безпековий простір України визначається високою інтенсивністю воєнних загроз, зростанням техногенних аварій і природних катастроф, що зумовлює необхідність оновлення механізмів реагування на надзвичайні ситуації. На мою думку, ці процеси потребують модернізації систем управління та впровадження новітніх рішень. Оскільки умови безпеки постійно ускладнюються, технології ШІ стають невід'ємним інструментом для підвищення ефективності роботи відповідних служб. Закон України «Про Національну поліцію» прямо визначає, що поліція зобов'язана забезпечувати публічну безпеку, охороняти права і свободи громадян, реагувати на надзвичайні події та взаємодіяти з іншими суб'єктами сектору безпеки [1]. За даними Державної служби України з надзвичайних ситуацій за 2022–2024 роки, кількість надзвичайних ситуацій воєнного, природного та техногенного характеру істотно зросла, що спричинило підвищення навантаження на всі служби реагування [2].

Традиційні методи управління кризовими подіями вже не відповідають масштабу сучасних викликів, а ефективність реагування значною мірою залежить від швидкості обробки інформації та узгодженості взаємодії служб [3]. На мою думку, поєднання аналітичних систем і автоматизованого збору даних дозволяє значно підвищити оперативність реагування. Оскільки обсяг інформації в умовах кризових ситуацій стрімко зростає, інтелектуальні системи стають ключовими для своєчасного ухвалення рішень. Кодекс цивільного захисту України встановлює загальні вимоги до організації системи цивільного захисту, але він не враховує повною мірою можливості сучасних цифрових технологій, зокрема ШІ [4].

Згідно з Глобальним звітом UNDRR 2023 року, технології штучного інтелекту значно підвищують точність прогнозування загроз, дозволяють проводити раннє виявлення небезпек і оптимізувати процеси управління ризиками [5]. У Стратегії НАТО щодо використання ШІ для кризового реагування зазначено, що країни-члени успішно інтегрують інтелектуальні системи у військові, поліцейські та гуманітарні операції для скорочення часу ухвалення рішень і підвищення точності аналізу [6].

OECD у звіті «Artificial Intelligence in Society» підкреслює, що ШІ дозволяє проводити аналіз великих масивів інформації у реальному часі, моделювати сценарії розвитку подій та оцінювати ризики на основі комплексних даних [7]. Наукові дослідження доводять ефективність алгоритмів машинного навчання для пошуку постраждалих, аналізу супутникових даних, визначення зон ураження та прогнозування наслідків надзвичайних ситуацій [8].

Європейська Комісія у звіті «Artificial Intelligence for Crisis Management» детально аналізує можливості інтеграції ШІ у національні системи цивільного захисту, зокрема використання інтелектуальних систем ситуаційної обізнаності, аналітики реального часу та автоматизованої координації служб реагування [9]. У США FEMA впроваджує інтелектуальні системи прогнозування розвитку природних катастроф, що дозволяє точніше оцінювати масштаби небезпеки та підвищувати ефективність підготовки рятувальних служб [10]. Europol активно застосовує технології аналізу великих даних для виявлення загроз безпеці, аналізу відеодоказів та моніторингу кризових подій [11].

Європейський механізм цивільного захисту використовує супутникову систему Copernicus, яка забезпечує автоматичний аналіз руйнувань, картографування зон лиха та оцінку рівня ризику на основі даних штучного інтелекту [12]. Public Safety Canada у своїх звітах зазначає успішне застосування мультиагентних систем ШІ для пошуково-рятувальних операцій, що скорочує час локалізації постраждалих і підвищує точність оцінки ситуації [13].

Група вчених університету INN створили програмне забезпечення для ситуаційного центру. Завдання штучного інтелекту Binom передбачати ризики та наслідки для 25 кейсів із надзвичайних ситуацій та небезпечних подій. Технологія штучного інтелекту складається із тисяч інтеграцій з чутливими індикаторами, що допомагають контролювати зміни критичних параметрів і завчасно інформувати служби про можливі ризики. На мою думку, подібні системи дозволяють значно скоротити час реагування на потенційні загрози.

Інтеграція інтелектуальних технологій у діяльність Національної поліції України забезпечує можливість автоматизувати аналіз відеопотоків, швидко формувати оперативну картину подій, здійснювати оцінку ризиків у реальному часі, оптимізувати маршрути евакуації, підвищувати точність документування наслідків та мінімізувати ризик людських помилок.

Європейська Комісія та OECD наголошують, що ключовим елементом розвитку кризового менеджменту є створення інтегрованих інформаційних систем, у яких дані від різних служб об’єднуються в єдину аналітичну платформу [7]. Оскільки Україна перебуває у стані постійних безпекових загроз, впровадження таких платформ є критично важливим для ефективної координації між поліцією, ДСНС, НГУ, МОЗ та місцевими адміністраціями.

Таким чином, використання технологій штучного інтелекту є ключовою умовою модернізації системи управління надзвичайними ситуаціями. Оскільки швидкість реагування та точність прогнозів напряму впливають на порятунок життя людей, інтелектуальні системи стають стратегічно важливим інструментом державної безпеки.

1. Закон України «Про Національну поліцію».
2. Державна служба України з надзвичайних ситуацій. Звіт 2022–2024.
3. Коваль А. Наукова робота...

4. Кодекс цивільного захисту України.
5. UNDRR Global Assessment Report 2023.
6. NATO Artificial Intelligence Strategy 2022.
7. OECD. Artificial Intelligence in Society 2019.
8. Ahmad M. Applications of AI in Emergency Management 2020.
9. European Commission. AI for Crisis Management 2022.
10. FEMA. AI Tools in Emergency Response 2023.
11. Europol. Use of Emerging Technologies 2022.
12. Copernicus EMS 2022.
13. Public Safety Canada 2021.
14 Binom Adviser documentation.

N. Vasyliv

USE OF ARTIFICIAL INTELLIGENCE IN SAFETY: FORECASTING AND PREVENTING ACCIDENTS AT CRITICALLY IMPORTANT FACILITIES

The aim of this study is to identify the potential of AI for forecasting and preventing accidents at CIFs and to develop recommendations for implementing such technologies. The main research tasks are:

1.To analyze the causes of accidents at CIFs and modern methods of prevention.

2.To review approaches for applying AI in safety systems.

3.To assess the effectiveness of accident forecasting using machine learning and deep data analysis algorithms.

4. To develop recommendations for integrating AI into comprehensive CIF safety systems.

Keywords: artificial intelligence, critically important facilities, accident forecasting, accident prevention, machine learning, safety, monitoring.

Introduction

Ensuring the safety and reliability of critical infrastructure facilities (CIFs) is a matter of growing importance in today's interconnected and technology-driven world. CIFs, which include energy plants, transportation networks, water supply systems, and communication hubs, form the backbone of modern society. Any failure or accident at these facilities can lead to severe consequences, including economic losses, environmental damage, disruption of essential services, and threats to human life [1].

The complexity of modern infrastructure, combined with the increasing integration of automated and digital systems, has made traditional safety management methods insufficient. Technical failures, human errors, organizational deficiencies, and extreme natural events often interact, creating highly unpredictable scenarios. These factors underscore the urgent need for advanced approaches to accident prevention and risk mitigation.

Artificial intelligence (AI) offers significant potential in addressing these challenges. By enabling predictive analytics, real-time monitoring, intelligent decision support, and simulation modeling, AI can anticipate hazards, reduce human error, and optimize preventive measures. Implementing AI in CIFs is not only a technological advancement but also a strategic necessity to maintain operational continuity, enhance public safety, and ensure resilience against both technical and environmental threats.

The relevance of this problem is further heightened by the increasing frequency of extreme weather events, growing cyber-physical interdependencies, and the global demand for uninterrupted services. As such, developing and integrating AI-based solutions for forecasting and preventing accidents at CIFs is critical for sustainable infrastructure management and societal well-being.

Methodology

Artificial intelligence encompasses a wide range of technologies used to enhance reliability, operational monitoring, and timely response to potential threats at critical infrastructure facilities. The main methods include [2-3]:

1. Machine Learning and Deep Neural Networks

These approaches allow the analysis of large volumes of data from sensors, industrial equipment, and information systems. Algorithms detect hidden patterns, predict equipment failures, identify anomalies in technological processes, and improve early warning accuracy for potential accidents.

Recurrent and convolutional neural networks are particularly effective, being applied to time series analysis and image processing, respectively.

2. Expert Systems

Expert systems rely on knowledge bases containing rules, models, and instructions for crisis situations. They assist operators in making real-time decisions by proposing optimal response options according to safety standards. Expert systems are especially useful in areas where human experience– engineers, dispatchers, emergency services – needs to be formalized.

3. Computer Vision Systems

These solutions are used for visual monitoring of facility conditions: detecting cracks, deformations, smoke emissions, or unauthorized objects or personnel in hazardous zones.
Computer vision is applied in video surveillance systems, robotic inspection complexes, drone surveys, and automated access control systems. Image analysis occurs in real time with minimal human intervention.

4. Integrated Cyber-Physical System (CPS) Platforms

Cyber-physical systems combine data from sensors, AI algorithms, automatic control mechanisms, and cybersecurity tools.

These platforms provide:

- ✓ continuous monitoring of technological processes,
- ✓ automatic adjustment of equipment parameters,
- ✓ detection of cyber and physical threats,
- ✓ coordinated interaction between different security systems.

They create a unified digital environment, enabling centralized infrastructure management and rapid response to deviations from normal operations.

Main Body

Causes of Accidents at Critical Infrastructure Facilities (CIFs)

Accidents at critical infrastructure facilities are typically complex events resulting from the interaction of multiple factors. Understanding these causes is essential for implementing effective preventive and mitigative measures. They can be broadly categorized as follows [4-5]:

1. Technical Causes

Technical failures are among the most common contributors to accidents at CIFs. They include:

- ✓ Equipment wear and degradation: Mechanical components, pipelines, or electrical systems may fail due to fatigue, corrosion, or insufficient maintenance.

✓ Software defects: Bugs, glitches, or vulnerabilities in control and monitoring software can lead to malfunctions or incorrect system responses.

✓ Automation system failures: Errors in automated control systems, including sensors and actuators, may result in unintended operational deviations, potentially triggering accidents.

Addressing technical causes requires rigorous maintenance, real-time monitoring, and the integration of predictive analytics to anticipate failures before they escalate.

2. Organizational Causes

Human and organizational factors significantly influence accident occurrence. Key issues include:

✓ Human errors: Mistakes made by operators, engineers, or maintenance personnel can lead to unsafe operations or delayed responses.

✓ Non-compliance with safety standards: Deviations from established safety protocols, regulatory requirements, or operational guidelines increase the risk of incidents.

✓ Insufficient process control and supervision: Poorly designed procedures or lack of oversight may allow minor issues to escalate into major accidents.

Mitigating organizational causes involves comprehensive training, adherence to safety standards, implementation of standard operating procedures, and deployment of decision-support systems.

3. Natural and Extreme Factors

Environmental conditions and extreme natural events can precipitate accidents at CIFs. These include:

✓ Natural disasters: Earthquakes, floods, hurricanes, and landslides can damage infrastructure and disrupt operations.

✓ Extreme weather conditions: Severe temperatures, storms, or heavy precipitation can affect equipment performance or create hazardous working conditions.

✓ Other environmental hazards: Lightning strikes, fires, or chemical leaks originating from environmental sources can trigger chain reactions affecting facility safety.

Addressing natural and extreme causes involves resilient design, risk assessment, emergency preparedness, and real-time environmental monitoring.

The Role of Artificial Intelligence in Forecasting and Preventing Accidents at Critical Infrastructure Facilities

Artificial intelligence (AI) has become a key tool for enhancing safety and operational reliability at critical infrastructure facilities (CIFs). By leveraging advanced data processing, predictive modeling, and automated decision support, AI enables organizations to anticipate, prevent, and mitigate accidents. The main contributions of AI can be described as follows [6-7]:

1. Big Data Analysis

AI systems can process vast volumes of heterogeneous data generated by CIFs, including:

✓ Sensor outputs from machinery, pipelines, and environmental monitors;

✓ System logs from SCADA, PLCs, and industrial control networks;

✓ Historical records of past accidents, maintenance, and operational deviations.

Through advanced analytics, AI identifies hidden patterns, correlations, and early warning signals that may remain undetected by human operators. Techniques such as clustering, anomaly detection, and feature extraction allow for continuous evaluation of operational health and risk factors.

2. Accident Forecasting

Machine learning (ML) and deep learning algorithms are applied to estimate the probability of equipment failures and critical incidents. Key capabilities include:

✓ Predictive maintenance: Forecasting when a component is likely to fail based on sensor trends and operational history;

✓ Scenario anticipation: Evaluating sequences of events that could escalate into accidents;

✓ Risk ranking: Prioritizing equipment, processes, or areas with the highest likelihood of failure.

By providing quantitative risk assessments, AI supports proactive management, helping to prevent accidents before they occur.

3. Real-Time Monitoring

AI enables continuous monitoring of operational parameters and environmental conditions. Systems can:

✓ Automatically detect deviations from normal operating ranges;

✓ Identify early signs of malfunction, wear, or unsafe conditions;

✓ Trigger alerts or alarms to operators for immediate intervention.

Integration with IoT devices and edge computing allows these analyses to occur in real time, significantly reducing response times to emerging hazards.

4. Decision Support

Intelligent decision-support systems use AI to recommend optimal corrective actions in complex operational environments. Benefits include:

✓ Reducing reliance on human judgment under pressure, thereby minimizing errors;

✓ Suggesting preventive or corrective interventions based on real-time data;

✓ Incorporating safety protocols, regulatory requirements, and past incident knowledge into recommendations.

Such systems enhance human performance by providing data-driven guidance in critical situations.

5. Simulation and Scenario Modeling

AI can simulate potential accident scenarios using digital twins or virtual models of CIFs. Applications include:

✓ Impact assessment: Estimating consequences of equipment failures, environmental hazards, or human errors;

✓ Optimization of preventive measures: Testing safety strategies and emergency plans in a virtual environment;

✓ Training and preparedness: Creating realistic scenarios for operator training and emergency drills.

Simulation modeling allows organizations to evaluate “what-if” situations without exposing real facilities to risk, improving readiness and resilience.

The combination of these capabilities allows AI to act as a comprehensive safety enhancement tool: it detects problems early, forecasts potential incidents, supports informed decision-making, and evaluates mitigation strategies. By bridging the gap between human expertise and automated monitoring, AI contributes to reduced accident rates, improved operational efficiency, and enhanced safety culture at CIFs.

Accident Prevention Measures Using Artificial Intelligence

Artificial intelligence (AI) plays a critical role in preventing accidents at critical infrastructure facilities (CIFs) by enabling technical, organizational, and informational measures. These measures collectively enhance operational safety, reduce human error, and ensure timely responses to potential hazards [8-9].

1. Technical Measures

AI-driven technical solutions focus on monitoring, detection, and automated intervention to prevent accidents. Key approaches include:

✓ Intelligent sensors: AI-equipped sensors continuously monitor equipment condition, environmental parameters, and operational processes. They detect deviations from normal patterns, such as unusual vibrations, temperature fluctuations, or pressure anomalies.

✓ Automated monitoring systems: These systems analyze real-time data streams using machine learning algorithms, identifying potential failures before they escalate into accidents.

✓ Emergency shutdown mechanisms: AI systems can automatically trigger protective actions, such as shutting down critical equipment or isolating hazardous areas, based on predictive risk assessment.

By implementing these technical measures, CIFs can respond to developing threats instantly, minimizing damage and preventing catastrophic failures.

2. Organizational Measures

Organizational measures involve the integration of AI into operational procedures and the human workforce. Key aspects include:

✓ Training personnel: Operators and engineers are trained to interpret AI-generated alerts, interact with decision-support systems, and follow recommended preventive actions.

✓ Development of response protocols: Clear procedures are established for reacting to AI warnings, including escalation steps, coordination between departments, and verification protocols.

✓ Integration into safety culture: Embedding AI tools into daily operations promotes a proactive approach to risk management, encouraging personnel to anticipate problems rather than react to crises.

These measures ensure that AI enhances human decision-making rather than replacing it, creating a synergistic safety framework.

3. Informational Measures

AI facilitates centralized monitoring and early-warning capabilities by integrating data from multiple sources:

✓ Centralized monitoring platforms: AI systems consolidate sensor data, operational logs, and historical incident records into unified dashboards for real-time situational awareness.

✓ Integration with external data: AI can incorporate information from meteorology services, energy networks, and other infrastructure systems to anticipate environmental or systemic hazards.

✓ Predictive alerts and analytics: By combining internal and external data, AI provides early warnings of potential accidents, enabling preemptive interventions.

Informational measures ensure that decision-makers have a holistic understanding of facility status, environmental threats, and potential risks, supporting faster and more effective preventive actions.

By combining technical, organizational, and informational measures, AI enables a multi-layered accident prevention strategy:

✓ technical systems detect and respond to immediate hazards;

✓ organizational measures ensure human operators effectively interpret and act on AI guidance;

✓ informational systems provide a broad situational context for anticipating risks.

This integrated approach improves operational reliability, reduces downtime, and enhances safety culture at CIFs.

Conclusions

The use of AI in security systems at critically important facilities significantly enhances the effectiveness of forecasting and preventing accidents. Intelligent algorithms analyze large datasets, detect anomalies, and predict critical scenarios, allowing timely responses to potential threats. Integrating AI into comprehensive CIF security systems is a strategic step to increase infrastructure resilience and minimize the consequences of emergencies.

1. Tsallis, C., Papageorgas, P., Piromalis, D., & Munteanu, R. A. *Application-Wise Review of Machine Learning-Based Predictive Maintenance: Trends, Challenges, and Future Directions. Applied Sciences.* 2025, 15(9): 4898. DOI: https://doi.org/10.3390/app15094898.
2. Falsk, R. (2023). AI for Predictive Maintenance in Industrial Systems. Handbuch Ansehen. https://doi.org/10.13140/RG.2.2.27313.35688.
3. Xu, S., Wang, J., Shou, W., Ngo, T., Sadick, A.-M., Wang, X. (2020). Computer Vision Techniques in Construction: A Critical Review. Archives of Computational Methods in Engineering, 28 (5), 3383–3397. https://doi.org/10.1007/s11831-020-09504-3.
4. Fernández, J. (2024). Integrating AI into Functional Safety Management. Safe and Explainable. Critical Embedded Systems based on AI. Available at: https://safexplain.eu/integrating-ai-into-functional-safety-management/.
5. Mahdavinejad, M. S., Rezvan, M., Barekatain, M., Adibi, P., Barnaghi, P., Sheth, A. P. (2018). Machine learning for internet of things data analysis: a survey. Digital Communications and Networks, 4 (3), 161–175. https://doi.org/10.1016/j.dcan.2017.10.002.

6. AI in worker management: involving people to prevent risks (2025). ENSHPO. Available at: https://www.enshpo.eu/ai-in-worker-management-involving-people-to-prevent-risks/.
7. Implementing safer AI worker management through policy and prevention (2024). European Agency for Safety and Health at Work (EU-OSHA). Available at: https://osha.europa.eu/en/oshnews/implementing-safer-ai-worker-management-through-policy-and-prevention.
8. An Occupational Safety and Health Perspective on Human in Control and AI (2022). BauA. Available at: https://www.baua.de/EN/Service/Publications/Essays/article3454
9. Reports of Fatalities and Catastrophes – Archive. OSHA. Available at: https://www.osha.gov/fatalities/reports/archive.

А.В. Гаєвський

МЕХАНІЗМИ МАНІПУЛЯЦІЇ В СОЦІАЛЬНІЙ ІНЖЕНЕРІЇ

Соціальна інженерія залишається одним із найнебезпечніших видів кіберзагроз сучасності, оскільки зловмисники орієнтуються не на технічні вразливості, а на людський фактор. Статистика 2024-2025 років демонструє критичну гостроту проблеми: фішингові атаки залишаються ключовим вектором компрометації у понад 70% інцидентів, причому до 68% випадків витоку даних причетний саме людський фактор [1].

За даними FBI IC3 за 2024 рік [2], кількість звернень про фішинг досягла 193 407 випадків, а сукупні заявлені втрати від соціально-інженерних атак перевищили 16 млрд доларів. Найбільш вражаючим є те, що атаки типу BEC, попри відносно невелику кількість інцидентів (2,5% від загального числа), становлять непропорційно великий фінансовий ризик — середній збиток від однієї операції складає 137132 доларів при собівартості атаки всього 170 доларів, що забезпечує рентабельність понад 80000%.

Модель швейцарського сиру, система HFACS та теорія когнітивних упереджень пояснюють, як атаки СІ експлуатують універсальні механізми людської психіки. Встановлено вісім ключових маніпулятивних прийомів: залякування, авторитет, терміновість, вигода, соціальний доказ, взаємність, приязнь та FOMO, кожен із яких операціоналізовано через лінгвістичні маркери.

Атаки характеризуються диференційованою вразливістю цільових груп: молоді люди найбільше сприйнятливі до FOMO (65% кліків), люди середнього віку — до авторитету та терміновості (40-50%), старші користувачі — до довіри (50-60%), а фінансові працівники становлять групу найвищого ризику з 28-30% успішністю BEC-атак.

Встановлено перелік 12 IoA-індикаторів, що охоплюють поведінкові та технічні аномалії (нічна активність, аномальна частота помилок, складні командні рядки, zero-day-подібні ознаки).

Порівняння чотирьох методів виявлення демонструє: Statistical Baseline найкраще дискримінує при низьких FPR (AUC=0.924), Isolation Forest досягає повного recall (AUC=0.932, TTD=42ms), Rule-Based виявляється найшвидшим і найточнішим (AUC=0.985, TTD=8ms), а LOF+Z-Score показує найменшу ефективність (AUC=0.834, TTD=124ms). Гібридний підхід Rule-Based із багатовимірною статистикою забезпечує оптимальний баланс між швидкістю, точністю та часом виявлення.

Протидія СІ вимагає узгодженої роботи на чотирьох рівнях. Навчання й обізнаність передбачає регулярні симуляційні тренінги з персоналізованими сценаріями, гейміфікацію, створення культури повідомлення про підозрілі листи без санкцій та встановлення швидких каналів репорту. Технічні засоби включають передові email-фільтри з SPF/DKIM/DMARC, багатофакторну

автентифікацію з захистом від MFA bombing, EDR/XDR системи та аналітику лінгвістичних патернів. Організаційні процеси охоплюють процедури двоканальної верифікації для критичних операцій (особливо фінансові трансакції), принцип four-eyes для важливих рішень та чіткі OSINT-розвідування на початку атаки. Поведінкова аналітика забезпечує інтеграцію IoA-моніторингу з email-системою для кроссмодального виявлення та автоматизовані системи раннього попередження на базі розроблених індикаторів. Комбінування цих рівнів забезпечує значне зменшення успішності атак (із 30-40% до 2-5%) та скорочує час виявлення до критичного мінімуму.

Забезпечення стійкості організацій передбачає залучення зовнішніх ресурсів — участь у професійних мережах обміну загрозами та створення галузевих альянсів для анонімного обміну досвідом атак. Запровадження метрик кіберстійкості (час виявлення, частка заблокованих листів, звіти працівників) та регулярні бенчмарки серед підрозділів створюють прозорий механізм постійного поліпшення. Гібридні алгоритми, гейміфікація та міжорганізаційна кооперація забезпечують еволюцію методик і дозволяють організаціям поступово досягати найвищих рівнів стійкості.

1. Verizon. (2025). 2025 data breach investigations report. https://www.verizon.com/business/resources/reports/.
2. Federal Bureau of Investigation. (2024). *Internet crime report 2024*. https://www.ic3.gov/.

О.А. Кобилянська

МЕТОДИ ЗАБЕЗПЕЧЕННЯ ДИФЕРЕНЦІЙНОЇ КОНФІДЕНЦІЙНОСТІ У ВЕЛИКИХ НАБОРАХ ДАНИХ

Актуальність дослідження визначена радикальними змінами у контексті приватності в умовах Big Data. Традиційні методи анонімізації виявилися неефективними при поєднанні великих масивів даних, що дозволяє зв'язувати та реідентифікувати записи через квазіідентифікатори та зовнішні джерела.

Гучні інциденти – AOL (2006)[1], Netflix Prize[2], Cambridge Analytica[3] – продемонстрували, що навіть «знеособлені» дані можуть бути повторно пов'язані з конкретними людьми. Одночасно GDPR, NIST SP 800-226 та інші міжнародні стандарти встановлюють вимогу про неможливість ідентифікації «ні прямо, ні опосередковано», що стимулює пошук методів з формальними гарантіями.

Диференційна конфіденційність у цьому контексті розглядається як формалізований механізм захисту, що дозволяє виконувати статистичний аналіз і навчання моделей на конфіденційних наборах даних, водночас математично обмежуючи внесок кожного окремого запису в опубліковані результати.

Одним із ключових результатів є демонстрація чітко вираженого компромісу між рівнем приватності, що задається параметром ε, та корисністю моделі, виміряною основними класифікаційними метриками.

За базової моделі без шуму ($\varepsilon \to \infty$) досягається точність 85.37% при ROC-AUC 90.81%, тоді як для найсуворішого режиму ($\varepsilon = 0.1$) точність падає до 46.33% при ROC-AUC 51.29%, що відповідає втраті корисності близько 45.7% відносно бенчмарку. Проміжні значення ε демонструють плавне відновлення якості: у діапазоні 2.0–5.0 точність підвищується до 60.52–70.64%, а ROC-AUC – до 65.95–73.57% при залишковій втраті 17–29% від базового рівня.

Особливо показовим є режим $\varepsilon = 10.0$, за якого досягається точність 74.47% та F1-score 52.71% при ROC-AUC 71.47% і відносній втраті лише 12.8%; ще вищі значення ε (20.0–50.0) наближають модель до немодифікованого класифікатора, забезпечуючи точність 79.64–84.61% та ROC-AUC 82.72–89.04% при мінімальній деградації (0.89–6.7%).

Таким чином, експериментальна крива «приватність–корисність» підтверджує наявність області раціональних компромісів, у якій помірне послаблення формальних гарантій ($\varepsilon \approx 5–10$) дозволяє зберегти більшу частину прогностичної здатності моделі.

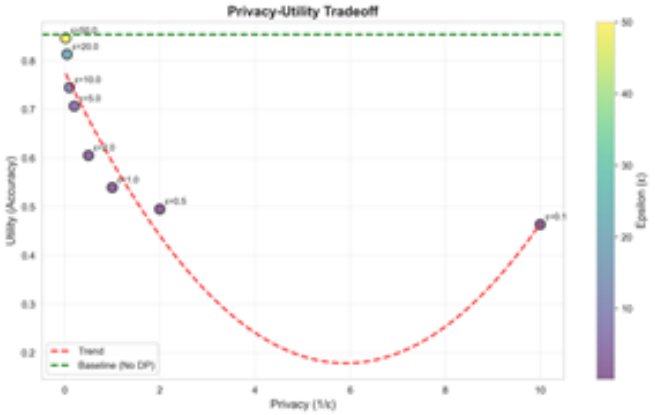

Рисунок 1 – Приватність – корисність за різних ε

На підставі отриманих результатів сформовано практичні рекомендації щодо вибору ε для різних класів даних.

Для медичних записів та генетичної інформації доцільно використовувати ε у межах 0.1–0.5, для фінансових транзакцій та кредитного скорингу рекомендовано 0.5–2.0 з високим рівнем захисту. Персональні профілі користувачів доцільно обробляти з ε = 1.0–5, тоді як поведінкові дані для рекомендаційних систем та A/B-тестування – у діапазоні 5.0–10.0 із помірним рівнем захисту. Для агрегованих статистик публічних звітів достатньо ε = 10.0–20.0, а для некритичних або демонстраційних наборів можуть використовуватись значення 20.0–50.0 за мінімальних формальних обмежень на витік.

1. AOL search data leak. (2006, August 9). The New York Times. https://www.nytimes.com/2006/08/09/us/09aol.html.
2. Narayanan, A., & Shmatikov, V. (2007). Robust de-anonymization of large sparse datasets (The Netflix Prize dataset). arXiv. https://arxiv.org/abs/cs/0610105.
3. The Cambridge Analytica files. (2018). The Guardian. https://www.theguardian.com/news/series/cambridge-analytica-files.

І.В. Блінов, П.В. Шиманюк

АСПЕКТИ ІНТЕЛЕКТУАЛІЗАЦІЇ МІСЬКИХ ЕНЕРГОМЕРЕЖ І МІКРОМЕРЕЖ НА ОСНОВІ ЦИФРОВИХ ДВІЙНИКІВ: ПРОБЛЕМАТИКА ТА ШЛЯХИ ВИРІШЕННЯ

Сучасний розвиток електроенергетики, зокрема і міських систем електропостачання характеризується зростанням складності інфраструктури, інтеграцією розподіленої генерації [1], зокрема відновлюваних джерел енергії, появою мікромереж (локальних електроенергетичних систем) [2, 3] та підвищенням вимог до надійності та адаптивності електропостачання [4, 5] в умовах роздрібного ринку електричної енергії України [6-8]. Міські енергетичні комплекси функціонують у динамічному середовищі, де поєднання систем розподілу з мікромережами [9-11] створює нові виклики для керування потоками енергії, балансування та забезпечення стійкості системи. У таких умовах традиційні підходи до моніторингу та аналізу роботи енергосистеми стають недостатніми, що вимагає впровадження інноваційних цифрових технологій для забезпечення надійної та безперебійної роботи як систем розподілу, так і мікромереж [12, 13], інтегрованих з нею, зокрема на за рахунок впровадження технологій «розумних мереж», систем Smart-моніторингу [14].

Важлива проблема в цих умовах, яка сьогодні потребує вирішення, полягає у відсутності інструментів, здатних повноцінно моделювати поведінку міської енергосистеми, включно з мікромережами, у реальному часі із врахуванням усіх змінних факторів та можливих сценаріїв розвитку подій. Енергетичні системи, що працюють у взаємодії з розосередженими джерелами генерації, установками зберігання енергії та мікромережами, набувають нового рівня складності, що робить їх вразливими до природних, техногенних та аварійних впливів. Наявні математичні та симуляційні підходи часто не здатні забезпечити своєчасний аналіз станів, не дозволяють програвати різноманітні сценарії відмов і не забезпечують накопичення досвіду та адаптації системи. Це особливо критично для міст, де поєднання основної мережі та мікромереж може спричиняти складні режими взаємодії, що потребують постійного контролю та прогнозування.

Актуальним стає створення цифрових інструментів, які дозволять забезпечувати стійку роботу всієї міської енергетичної інфраструктури, включно з мікромережами, в умовах ускладнення її структури, підвищення частоти аварійних ситуацій та збільшення ролі відновлюваних джерел енергії. У світовій практиці активно розвивається концепція цифрових двійників, що дозволяє створювати віртуальні аналоги складних технічних систем. Проте їх використання для міських енергосистем з інтегрованими мікромережами у режимі реального часу практично не реалізовується на практиці. Це обґрунтовує потребу створення комплексної системи цифрових двійників, яка буде синхронізована з реальною інфраструктурою, дозволить

моделювати як глобальні, так і локальні процеси, оцінювати стабільність взаємодії між основною мережею та мікромережами й прогнозувати функціонування системи за різних умов.

Розроблення методів і засобів побудови цифрових двійників міської енергетичної системи з урахуванням структури мікромереж, що забезпечить надійну експлуатацію інфраструктури, прогнозування аварійних і критичних станів та підвищить ефективність управління мережею в реальному часі потребує поєднання математичних моделей, симуляційних інструментів і методів машинного навчання для створення інформаційної системи, здатної відображати поточний стан всієї мережі, включно з локальними мікромережами, та забезпечувати підтримку прийняття рішень оператором.

Для досягнення цих цілей необхідно вирішити низку завдань. Першим завданням є розроблення числових моделей енергосистеми, що охоплюють сталий і перехідний режими роботи систем розподілу та мікромереж, включно з моделями потокорозподілу, балансування потужності та опису динамічних процесів. Наступним завданням є створення цифрових двійників на базі цих моделей у симуляційних середовищах, що дозволить відтворювати взаємодію між основною мережею і мікромережами, аналізувати стан елементів системи та моделювати можливі аварійні ситуації. Важливим завданням є розроблення моделі прогнозування станів енергосистеми на основі алгоритмів машинного навчання з урахуванням історичних та поточних даних про режими роботи як основної мережі, так і мікромереж. Для цього необхідним є формування бази даних показників роботи енергетичних об’єктів, створення спеціалізованих модулів аналізу та інтеграція прогнозної моделі в цифровий двійник для підтримки операторів систем розподілу та мікромереж у процесі прийняття рішень.

Очікуваним результатом розв’язання наведених завдань є створення комплексного інструменту цифрових двійників міської енергосистеми з інтегрованими мікромережами, здатного працювати в реальному часі та забезпечувати безперервний моніторинг, діагностику та прогнозування станів мережі. Такі системи дозволить своєчасно виявляти ризики, прогнозувати можливі пошкодження та відмови, підвищувати безвідмовність енергопостачання та оптимізувати взаємодію між центральною мережею та мікромережами. Впровадження цих рішень сприятиме підвищенню стійкості міської енергетичної інфраструктури, розвитку інтелектуальних енергетичних технологій та зміцненню енергетичної безпеки на місцевому та національному рівнях.

Фінансується за підтримки Національного фонду досліджень України в межах конкурсу «Передова наука в Україні» (Договір грантової підтримки № 241/0131) та держбюджетною темою «Розробка цифрового двійника мікромережі для оптимізації та керування децентралізованими мережами», (шифр «Двійник») №0125U002920.

1. Про схвалення Стратегії розвитку розподіленої генерації на період до 2035 року і затвердження операційного плану заходів з її реалізації у 2024 - 2026 роках. Розпорядження Кабінету Міністрів України.18.07.2024 №713-р URL: https://zakon.rada.gov.ua/laws/show/713-2024-%D1%80#Text.
2. Концепція впровадження “розумних мереж” в Україні до 2035 року. Розпорядження Кабінету Міністрів України 14.10.2022 № 908-р. URL: https://zakon.rada.gov.ua/laws/show/908-2022-%D1%80#Text.
3. Блінов І.В. Розвиток розподіленої енергетики в Україні з використанням технологій мікромереж (за матеріалами доповіді на засіданні Президії НАН України 5 березня 2025 р.). *Вісник НАН України*. 2025. № 5. С. 35—44. DOI: https://doi.org/10.15407/visn2025.05.035.
4. Денисюк С.П., Таргонський В.А., Артем’єв М.В. "Локальні електроенергетичні системи з активним споживачем: методи побудови та алгоритми їх функціонування." *Науковий журнал «Енергетика: економіка, технології, екологія»* 3 (2018): 7-22.
5. Blinov I, Radziukynas V, Shymaniuk P, Dyczko A, Stecuła K, Sychova V, Miroshnyk V, Dychkovskyi R. Smart Management of Energy Losses in Distribution Networks Using Deep Neural Networks. *Energies*. 2025; 18(12):3156. https://doi.org/10.3390/en18123156.
6. Правила роздрібного ринку електричної енергії. Постанова НКРЕКП 14.03.2018 № 312. URL: https://zakon.rada.gov.ua/laws/show/v0312874-18#Text.
7. Блінов І.В., Парус Є.В., Клименко О.Г., Клюзко О.І. Спосіб порівняльних оцінок комерційних пропозицій електропостачальників для споживачів без погодинного обліку електричної енергії. Енергетика: економіка, технології, екологія. 2023. № 2 (73). С. 36-42. DOI: https://doi.org/10.20535/1813-5420.3.2023.289654.
8. Кодекс систем розподілу. Постанова НКРЕКП 14.03.2018 № 310. URL: https://zakon.rada.gov.ua/laws/show/en/v0310874-18?lang=uk#Text.
9. Блінов І.В., Парус Є.В., Мірошник В.О., Шиманюк П.В., Сичова В.В. Модель оцінки доцільності переходу промислових споживачів до погодинного обліку електричної енергії на роздрібному ринку. Енергетика: економіка, технології, екологія. 2021. №1. С. 88-97. DOI: https://doi.org/10.20535/1813-5420.1.2021.242186.
10. Блінов І., Парус Є., Шиманюк П., Ворушило А. Модель оптимізації функціонування мікромережі з СЕС та установкою зберігання енергії. *Технічна електродинаміка*. 2024. № 5. С. 69-78 DOI: https://doi.org/10.15407/techned2024.05.069.
11. Парус Є.В., Блінов І.В. Оптимізація використання доступних енергоресурсів мікромережі за умов підтримки готовності до ізольованого режиму. *Технічна електродинаміка*. 2025. №5. С. 56-69. DOI: https://doi.org/10.15407/techned2025.05.056.
12. IEC TS 62898-1: 2017+AMD 1: 2023 Microgrids – Part 1: Guidelines for microgrid projects planning and specification. 2023. 78 p.
13. IEC TS 62898-2: 2018+AMD 1: 2023 Microgrids - Part 2: Guidelines for operation. 2018. 38 p.
14. Kyrylenko, O.; Denysiuk, S.; Bielokha, H.; Dyczko, A.; Stecuła, B.; Pazynich, Y. Smart Monitoring and Management of Local Electricity Systems with Renewable Energy Sources. Energies 2025, 18, 4434. https://doi.org/10.3390/en18164434.

П.В. Шиманюк, Сичова В.В.

ДОСЛІДЖЕННЯ КОНЦЕПЦІЙ ЦИФРОВОГО ДВІЙНИКА МІКРОМЕРЕЖІ

Децентралізація та зростання частки відновлюваних джерел енергії створюють нові виклики для управління енергосистемами [1–5]. Мікромережі стають ключовим елементом енергетичної трансформації, проте їхня складність потребує інноваційних цифрових рішень. Одним із таких рішень є цифрові двійники (DT- Digital Twin), що інтегрують фізичні об'єкти з їхніми віртуальними моделями [6–10].

Цифровий двійник – це динамічна кіберфізична система, яка безперервно синхронізується з даними сенсорів та систем моніторингу, створюючи динамічну модель мікромережі. DT не лише відображає стан системи, а й формує керуючі сигнали, що повертаються у фізичну мережу. В кіберфізичному середовищі поєднуються фізичні процеси, цифрового моделювання та аналітичні дані у у вигляді єдиної інтегрованої системи. Цифровий двійник здатний прогнозувати майбутні стани системи (генерацію, навантаження, відмови обладнання) та оптимізувати енергосистему. [9-14].

- Фізичний рівень – обладнання, сенсори, джерела даних.
- Цифровий рівень – моделі, алгоритми прогнозування, симуляції.
- Комунікаційний рівень – протоколи передачі даних, IoT, двостороння синхронізація.

Створення цифрового двійника складається з наступних етапів [11-15]:

- Збір даних (Використання IoT-сенсорів, SCADA, edge-пристроїв та протоколів MQTT, OPC UA, IEC 61850 для стандартизованого обміну даними).
- Моделювання компонентів (White-box: фізичні рівняння та енергетичні баланси. Grey-box: комбінація фізичних моделей та експериментальних даних. Black-box: ML/DL моделі для апроксимації нелінійних процесів).
- Інтеграція з IoT та хмарними платформами (Cloud computing: централізоване зберігання та масштабованість. Edge computing: локальна обробка даних для зменшення затримок. Багаторівнева архітектура: Device Layer → Edge Layer → Cloud Layer).
- Прогнозування та оптимізація (Прогнозування навантаження, генерації, стану ESS за допомогою ML/DL. Оптимізація роботи мікромережі: стохастичне програмування, MPC, генетичні алгоритми).
- Віртуальне тестування (Co-simulation та Hardware-in-the-Loop (HIL). Перевірка стратегій керування без ризику для фізичної інфраструктури. Використання ANSYS Twin Builder, Azure Digital Twins тощо).

У сучасних дослідженнях виділяють три основні архітектури DT [6,14]:

1) Three-Layer – проста та модульна структура, придатна для лабораторних прототипів.

2) SGAM – стандартизована багаторівнева інтеграція, орієнтована на промислові мікромережі [14].

3) CPS – замкнене керування у реальному часі, ефективне для автономних систем [9].

Таблиця 1 – Архітектури цифрових двійників

Архітектура	Переваги	Недоліки	Застосування
Three-Layer	Простота, масштабованість	Обмежена міждоменною інтеграцією	Академічні прототипи
SGAM	Стандартизація, багаторівнева інтеграція	Складність, великі ресурси	Smart campus, промислові мережі
CPS	Реальний час, автономність	Високі вимоги до кібербезпеки	Автономні мікромережі, IoT-системи

Основні проблеми впровадження DT у мікромережах:

- Точність моделей гібридних енергосистем. Гібридні енергосистеми мають складну нелінійну динаміку, що важко відтворити у цифровій моделі;
- Інтероперабельність між платформами, необхідність узгодження різних платформ, протоколів та стандартів;
- Моделювання у реальному часі потребує швидкої обробки даних та синхронізації між фізичним і цифровим рівнями;
- Захист даних від атак, забезпечення стійкості системи до кіберзагроз.
- Великі обсяги даних та складні алгоритми потребують потужних обчислювальних ресурсів;
- Перевірка адекватності цифрової моделі перед її застосуванням у реальних умовах.

Перспективи розвитку полягають у створенні розподілених цифрових двійників, інтегрованих із системами штучного інтелекту та edge computing, що забезпечить перехід від статичних моделей до самонавчальних енергетичних систем.

Цифрові двійники стають центральною технологією цифрової трансформації енергетики. Їх застосування у мікромережах відкриває нові можливості для оптимізації, прогнозування та підвищення надійності енергосистем. Подальші дослідження мають бути спрямовані на інтеграцію DT з AI, edge computing та системами розподіленого управління.

Дослідження виконано за держбюджетною темою «Розробка цифрового двійника мікромережі для оптимізації та керування децентралізованими мережами», (шифр «Двійник») №0125U002920.

1. Кириленко О.В., Денисюк С.П., Блінов І.В. Цифрова трансформація: сучасні тенденції тазавдання. *Праці Ін-ту електродинаміки НАН України*. 2023. Вип. 65. С. 5–14. DOI: https://doi.org/10.15407/publishing2023.65.005.
2. Блінов І.В., Палачов С.О., Парус Є.В. Функціонування сучасних систем енергомереджменту мікромережі в умовах підключення до систем розподілу електроенергії згідно із вимогами міжнародних стандартів. *Праці Ін-ту електродинаміки НАН України*. 2025. Вип. 71. С. 15-27.
3. Блінов, І.В., Парус, Є.В., Шиманюк, П.В. і Ворушило, А.О. Модель оптимізації функціонування мікромережі з СЕС та установкою зберігання енергії. *Технічна електродинаміка*. 2024. 5. С. 069.
4. Кириленко О.В., Блінов І.В., Зайцев Є.О., Палачов С.О.,Васильченко В.І. Впровадження міжнародних та європейських стандартів для розвитку ОЕС України згідно з концепцією Smart Grid. *Праці Ін-ту електродинаміки НАН України*. 2022. Вип. 63. С. 5-12.
5. Блінов І.В., Палачов С.О., Парус Є.В., Клименко О.Г. Сценарій використання мікромереж згідно з міжнародними стантартами. *Праці Ін-ту електродинаміки НАН України*. 2025. Вип. 70. С. 14-25.
6. Kumari, N., Sharma, A., Tran, B., Chilamkurti, N., & Alahakoon, D. A comprehensive review of digital twin technology for grid-connected microgrid systems: State of the art, potential and challenges faced. *Energies*. 2023. 16(14). pp. 5525.
7. Sahoo, B., Panda, S., Rout, P. K., Bajaj, M., & Blazek, V. Digital twin enabled smart microgrid system for complete automation: an overview. *Results in Engineering*. 2025. pp. 104010.
8. Bazmohammadi, N., Madary, A., Vasquez, J. C., Mohammadi, H. B., Khan, B., Wu, Y., & Guerrero, J. M. Microgrid digital twins: Concepts, applications, and future trends. IEEE Access. 2021. 10. pp. 2284-2302.
9. Wu, Y., Guerrero, J. M., Wu, Y., Bazmohammadi, N., Vasquez, J. C., Cabrera, A. J., & Lu, N. Digital twins for microgrids: Opening a new dimension in the power system. *IEEE Power and Energy Magazine*.2024. 22(1). pp. 35-42.
10. Jafari, M., Kavousi-Fard, A., Chen, T., & Karimi, M. A review on digital twin technology in smart grid, transportation system and smart city: Challenges and future. *IEEE Access*. 2023. 11. pp. 17471-17484.
11. Kabir, M. R., Halder, D., & Ray, S. Digital twins for iot-driven energy systems: A survey. IEEE Access. 2024.
12. Chen, H., Zhang, Z., Karamanakos, P., & Rodriguez, J. Digital twin techniques for power electronics-based energy conversion systems: A survey of concepts, application scenarios, future challenges, and trends. IEEE Industrial Electronics Magazine. 2022. 17(2). pp. 20-36.
13. Park, H. A., Byeon, G., Son, W., Jo, H. C., Kim, J., & Kim, S. Digital twin for operation of microgrid: Optimal scheduling in virtual space of digital twin. Energies. 2020. 13(20). pp. 5504.
14. Irsyad and Pradipta, Justin and Putra, Angga and Leksono, Edi, Microgrid digital twin: Implementation of Digital Twin Concept Based on Smart Grid Architectural Model (Sgam) and its Case Study. 2024.
15. Song, Z., Hackl, C. M., Anand, A., Thommessen, A., Petzschmann, J., Kamel, O., ... & Hauptmann, S. Digital twins for the future power system: An overview and a future perspective. Sustainability.2023. 15(6). pp. 5259.

Т.В. Коваленко

ДЕЯКІ РИЗИКИ ПРИ ВИКОРИСТАННІ ШТУЧНОГО ІНТЕЛЕКТУ

«Штучний інтелект - (ШІ, англ. artificial intelligence, AI) — розділ комп'ютерної лінгвістики та інформатики, який швидко розвивається, і зосереджений на розробці інтелектуальних машин, здатних виконувати завдання, які зазвичай потребують людського інтелекту. Ці завдання можуть варіюватися від простих дій, як-от розпізнавання мови чи зображень, до більш складних завдань, як-от ігри чи водіння автомобіля» [1].

Штучний інтелект – це здатність машини (комп'ютера) проявляти людські здібності, такі як розсуд, навчання, планування, творчість тощо. Штучний інтелект дозволяє технічним системам сприймати навколишнє середовище, справлятися з тим, що вони сприймають, вирішувати проблеми та діяти для досягнення певної мети. Комп'ютер отримує вже підготовлені або зібрані за допомогою власних датчиків (камер) дані, обробляє їх та виконує завдання.

Системи штучного інтелекту здатні у певній мірі адаптувати свою поведінку, аналізуючи попередні дії та працюючи автономно. Деякі технології штучного інтелекту існують вже більше 50 років, але досягнення у галузі обчислювальної потужності, доступності величезних об'ємів даних та нових алгоритмів в останні роки призвели до проривів у сфері штучного інтелекту.

Використання штучного інтелекту зростає дуже швидко. Штучний інтелект дозволяє обробляти та аналізувати великі обсяги даних та на їх основі приймати рішення. При цьому можуть використовуватися як персональні, так і конфіденційні дані, що викликає у світі великі суперечки.

Модель штучного інтелекту може включати мільйони та мільярди різноманітних параметрів. Кожен параметр впливає на результат таким чином, що його складно розглядати ізольовано.

Штучний інтелект використовує великий обсяг різноманітних даних, наприклад, текст та інформацію з датчиків. Складність розуміння, як системи штучного інтелекту приймають рішення або генерують результати, часто називають «проблемою чорної скриньки».

«З практичної точки зору, проблема «чорної скриньки» ШІ полягає у складності розшифрування міркувань, що лежать в основі прогнозів або рішень системи ШІ. Ця проблема особливо поширена в моделях глибокого навчання, таких як нейронні мережі, де кілька шарів взаємопов'язаних вузлів обробляють і перетворюють дані в ієрархічному порядку. Складність цих моделей і нелінійні перетворення, які вони виконують, роблять надзвичайно складним відстежити обґрунтування їхніх результатів. Микита Бруднов, генеральний директор BR Group – компанії, що займається маркетинговою аналітикою на основі штучного інтелекту, – розповів Cointelegraph, що

відсутність прозорості в тому, як ШІ-моделі приходять до певних рішень і прогнозів, може бути проблематичною в багатьох контекстах, таких як медична діагностика, прийняття фінансових рішень і судові розгляди, що значно впливає на подальше впровадження ШІ» [2].

Наприклад, фінансова установа використовує систему штучного інтелекту для визначення кредитоспроможності своїх клієнтів. Процес настільки складний, що фінансова установа не може зрозуміти, чому відмовлено у кредиті, а співробітники такої установи не в змозі пояснити клієнтам причину відмови. Відповідно клієнти не можуть оскаржити рішення фінансової установи.

Для ефективного вивчення систем штучного інтелекту необхідні великі обсяги даних. Чим більшим є обсяг даних, тим більше вірогідність, що всі варіанти, які повинні розпізнати модель, будуть представлені точно. Якщо даних замало, модель штучного інтелекту буде занадто сильно фокусуватися на окремих прикладах, моделі будуть неправдиво відтворювати дані.

Наприклад, відділ кадрів використовує систему штучного інтелекту для підбору персоналу. Система базується на даних, що відображають історичну дискримінацію жінок. Це призводить до неусвідомленої переваги кандидатів-чоловіків. Якщо раніше чоловіки отримували підвищення частіше, наприклад, у разі знаходження жінки у декретній відпустці, штучний інтелект помилково припускає, що стать є показником кар'єрного зросту.

Однією з юридичних проблем штучного інтелекту є питання захисту інтелектуальної власності. Чим більше вхідних даних отримує система штучного інтелекту, тим точніше та ефективніше вона стає. Для поповнення даних дуже часто використовується загально доступна інформація – наукові статті, газетні статті тощо. Однак ця інформація захищена авторським правом. Якщо штучний інтелект створює контент,схожий на матеріали, які захищені авторським правом, це є порушенням авторських прав.

Наприклад, «Регіональний суд Мюнхена постановив, що американська компанія OpenAI порушила німецьке законодавство про авторське право, використовуючи для навчання своїх мовних моделей тексти пісень, зокрема твори музиканта Герберта Гренемаєра.

Суд встановив, що під час тренування своїх моделей OpenAI використала матеріали з дев'яти німецьких пісень, серед яких Maenner та Bochum Гренемаєра. Позов проти компанії подало німецьке товариство захисту музичних прав GEMA, до складу якого входять композитори, автори текстів і музичні видавці. Організація звинуватила OpenAI у несанкціонованому копіюванні контенту, захищеного авторським правом, для навчання систем штучного інтелекту.

Суддя Ельке Шваґер зобов'язала OpenAI виплатити компенсацію за порушення, проте розмір відшкодування не розголошується. Представник GEMA Кай Велп висловив сподівання, що рішення суду стане підґрунтям для

подальших переговорів з OpenAI щодо справедливої винагороди власникам авторських прав» [3].

Завдяки штучному інтелекту можливо аналізувати та використовувати дані швидше. Системи штучного інтелекту значно пришвидшують автоматизоване прийняття рішень, економлять час та кошти.

Однак, наприклад, страхова компанія використовує штучний інтелект для більш ефективної обробки страхових випадків. Система штучного інтелекту рекомендує співробітникам приймати або відхиляти страхові випадки. У деяких випадках обґрунтовані страхові випадки можуть бути відхилені, оскільки алгоритм невірно інтерпретує певну закономірність. В результаті клієнти змушені подавати апеляції та довше чекати компенсації. додаткова робота збільшує витрати страхової компанії.

У багатьох випадках люди мають обмежену можливість заперечувати проти обробки своїх даних системами штучного інтелекту або ефективно керувати своїми уподобаннями. На сьогодні існує мало законодавчо закріплених принципів етичного використання систем штучного інтелекту.

Витік даних – це інцидент безпеки, при якому розкривається конфіденційна інформація. Системи штучного інтелекту вразливі до порушень безпеки. Витік або крадіжка даних можуть мати серйозні наслідки.

Наприклад, «Компанія Google опинилася в центрі судового позову, її звинувачують у незаконному зборі приватних даних користувачів через свій інструмент штучного інтелекту Gemini. Позов подали до федерального суду в Сан-Хосе, Каліфорнія, і він стосується роботи сервісів Gmail, Google Chat та Meet.

Раніше користувачі мали можливість самостійно активувати функцію Gemini. Але в жовтні 2025 року, як стверджує позивач, компанія Alphabet Inc., яка володіє Google, автоматично ввімкнула цей інструмент для всіх користувачів без їхнього відома чи згоди. Це, за словами позивача, дозволило Gemini отримати доступ до особистих повідомлень, електронних листів та вкладень, що зберігаються в акаунтах Gmail. У позові зазначили, що хоча користувачі можуть вимкнути Gemini, для цього потрібно глибоко зануритися в налаштування конфіденційності, що не є очевидним або простим кроком для більшості. До моменту деактивації, за твердженням позивача, Google має змогу переглядати повну історію приватного спілкування користувача, що викликає серйозні питання щодо порушення права на конфіденційність» [4].

Або, наприклад, лікарняний заклад використовує штучний інтелект для аналізування даних пацієнтів. З однієї сторони, відсутність даних може призвести до лікарняної помилки, з іншої сторони, якщо станеться витік або крадіжка конфіденційних медичних даних, що зберігаються в системі штучного інтелекту, конфіденційність даних багатьох пацієнтів опиниться під загрозою, і виникне ризик крадіжки особистих даних.

Відповідальне використання штучного інтелекту – це складне завдання. Регулярний контроль якості, високі стандарти безпеки та людський контроль мають вирішальне значення. Як компанії, так і приватні особи мають враховувати свої права та усвідомлювати ризики, пов'язані зі штучним інтелектом, щоб активно протистояти йому.

1. Штучний інтелект. URL: https://uk.wikipedia.org/wiki/%BD%D1%82%D0%B5%D0%BB%D0%B5%D0%BA%D1%82.
2. Проблема чорної скриньки ШІ: виклики та рішення для прозорого майбутнього. by Ezra Reguerra. 24 Листопада, 2023. URL: https://crypta.kyiv.ua/tehnologii-blokchejn/problema-chornoi-skrinki-shi-vikliki-ta-rishennja/.
3. Німецький суд визнав OpenAI винною у порушенні авторських прав, 11 листопада, 2025. URL: https://zn.ua/ukr/TECHNOLOGIES/nimetskij-sud-viznav-openai-vinnoju-u-porushenni-avtorskikh-prav.html.
4. Позов проти Google. Штучний інтелект Gemini нібито відстежує приватні листування без згоди користувачів. URL: https://techno.nv.ua/ukr/it-industry/google-zvinuvachuyut-u-nezakonnomu-zbori-danih-cherez-gemini-50560367.html.

М.С. Клименко

РОЗРОБКА БОРТОВОЇ СИСТЕМИ БПЛА ДЛЯ ПРОТИДІЇ ПЕРЕХОПЛЮВАЧУ

Висока інтенсивність застосування безпілотних літальних апаратів (БПЛА) широкої номенклатури є невід'ємною ознакою сучасних військових дій. Паралельно з цим активно розвиваються засоби протидії, серед яких особливе місце посідають антидронові літальні апарати-перехоплювачі, що дублюють та подеколи перевершують виконання завдань комплексів протиповітряної оборони. Внаслідок можливості запуску перехоплювачів практично збудь-якої оперативної ділянки суттєво знижується загальна ефективність застосування безпілотних комплексів та збільшується кількість втрат тактичних розвідувальних БПЛА.

Мета дослідження полягає у розробці програмної бортової системи, що на основі даних оптичного каналу обчислюватиме ймовірність уникнення ураження перехоплювачем та визначатиме оптимальний маневр ухилення. Отримані результати стануть основою для подальшого удосконалення бортової системи на основі експериментів у середовищі симуляції та дослідному зразку літального апарата. У роботі пропонується формалізувати задачу у вигляді дискретної моделі взаємодії двох апаратів у просторі, що дозволить здійснити оцінку можливості уникнення зіткнення відповідно до апріорних параметрів.

Виходячи із основних викликів уникнення зіткнення із дроном [1] для літакового тактичного БПЛА, важливим показником рішення є енергоефективність для збільшення тривалості та дальності виконання завдань. Моделювання маневру для уникнення зіткнення виконано виходячи із припущень стосовно конкретного сценарію поведінки перехоплювача (П) та розвідника (Р), а саме: Р маневрує від одного П одночасно, атака виконується із задньої напівсфери БПЛА, П спрямовується на центр проекції фронтальної камери Р (за вектором свого руху) із максимальною швидкістю, швидкості П та Р незмінні з моменту виявлення П до кінця маневру Р, наведення виконує оператор (що, спричинює затримку корекції напряму відносно візуального контролю від 1-3 с.) або бортовими алгоритмами коригування по захопленій цілі (затримка корекції напряму 0.05-0.2 с.) [2]. Потенційно розглядаються різні режими маневрування в залежності від типу ураження П: камікадзе (зіткнення), вибуховий пристрій (детонація при наближенні) або дистанційно бортовою стрілецькою зброєю [3]. Для формування первинної оцінки ситуації абсолютним успіхом вважається можливість уникнути зіткнення при атаці в крайню точку Р, виконуючи найгірший маневр виходу через протилежний бік проекції.

Відповідно до наведеної формалізації задачі мінімальна для успішного ухилення відстань S між П та Р може бути обчислена як:

$$S \geq (P_{Vmax} - R_{Vo})(t_{det} + t_m), \quad (1)$$

де P_{Vmax} – максимальна швидкість перехоплювача, R_{Vo} – швидкість Р на момент обчислення за вектором руху П, t_{det} – час виявлення П у полі зору Р (складається із 1/fps та обробки кадру моделлю машинного зору із обчисленням відносних кутів розташування та вектору руху), t_m – час маневру Р, необхідний для виходу із зони ураження П, у роботі обчислюється як мінімальний або максимальний (для потенційного та абсолютного варіантів ухилення) час за 4 напрямами відносно габаритів у проєкції Р.

За наведеним вище підходом розраховані теоретичні межі маневру для типових характеристик П та Р при атаці за різними напрямами. Враховано площу ураження, тобто проекцію Р за напрямом атаки П. Визначаючи цільовим типом Р БПЛА типу крило, було розраховано пропорційні зміни необхідного часу маневру відносно площі проекції. Час виявлення П напряму маневру ухилення Р, необхідний для початку коригування власної траєкторії, для успішного уникнення зіткнення не повинен бути меншим за t_m.

Розраховано мінімально необхідну для успішного маневру дистанцію між П та Р для різних кутів атаки (рис. 1). На основі розрахунків запропоновано відповідну програмну реалізацію прийняття рішення на основі штучної нейронної мережі, однак обчислювальна складність даного підходу істотно не переважає обчислювальну складність алгоритму розрахунку.

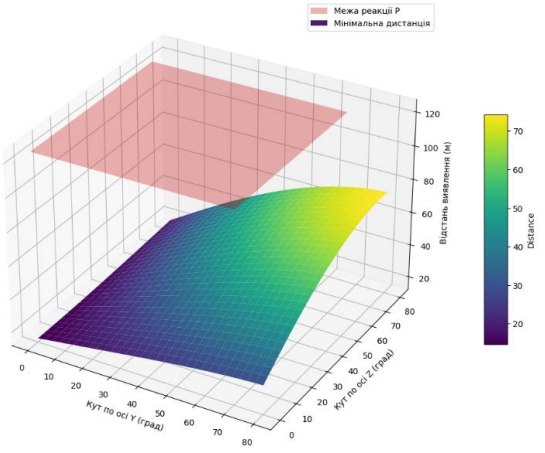

Рисунок 1 – Мінімальна дистанція між БПЛА для успішного маневру залежно від кутів атаки при швидкостях розвідника і перехоплювача 70 та 300 км/год відповідно

Експериментальне програмне забезпечення використовує модель машинного зору YOLOv12 для детекції цільового об'єкта й обчислення вектору його руху відносно камери. Використання стереокамери додатково надасть інформацію про орієнтовну відстань до об'єкта, що в дозволить враховувати динаміку зближення між БПЛА.

Висновки. Під час спроби уникнути перехоплення безпілотний літальний апарат опиняється у завідомо невигідних умовах, оскільки поступається перехоплювачу за динамічними характеристиками, зокрема швидкістю та доступними перевантаженнями. Така ситуація може бути формалізована як задача переслідування-ухилення, подібна до класичних моделей «хижак–жертва» або до задачі реагування воротаря під час пробиття пенальті. Оцінювання можливості здійснення маневрування виконано в межах інтервалу часу, який не перевищує період корекції траєкторії перехоплюючого апарата. Подальший розвиток дослідження передбачає аналіз ситуацій із варіативним зменшенням швидкості обох учасників протиборства, а також інтеграцію різних методів донаведення для уточнення поведінки перехоплювача та визначення меж маневреності БПЛА.

1. Harun, M. H., Abdullah, S. S., Aras, M. S. M., & Bahar, M. B. (2024). Recent Developments and Future Prospects in Collision Avoidance Control for Unmanned Aerial Vehicles (UAVS): A Review. International Journal of Robotics & Control Systems, 4(3). https://doi.org/10.31763/ijrcs.v4i3.1482.
2. Геращенко, М., Нестеренко, С., Ісаченко, О., Лось, А. і Сірик , О. (2021) «Моделі автоматичного наведення ударних безпілотних літальних апаратів на рухому ціль», Збірник наукових праць Державного науково-дослідного інституту випробувань і сертифікації озброєння та військової техніки, 2(8), с. 20-30. doi: 10.37701/DNDIVSOVT.8.2021.03.
3. Антишахедні та противертолітні: чеські розробники представили автономні дрони-перехоплювачі нового покоління [Електронний ресурс]. URL: https://armyinform.com.ua/2025/08/30/antyshahedni-ta-protyvertolitni-cheski-rozrobnyky-predstavyly-avtonomni-drony-perehoplyuvachi-novogo-pokolinnya.

Є.О. Вербова, В.В. Цуркан

СКЛАДНИКИ КІБЕРРИЗИКУ ОТРУЄННЯ ДАНИХ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ

Одним з десяти найбільш небезпечних кіберризиків безпеки великих мовних моделей є отруєння даних [1]. Під ним розуміється компрометування зловмисником їх навчальних наборів. Така діяльність призводить до отримання хибних результатів за умови нормальної поведінки великої мовної моделі. Прикладом такого зловмисного використання змагального штучного інтелекту є PoisonGPT [2]. Це прикладне застосування отруєння великої мовної моделі з відкритим кодом GPT-J-6B. Отриманий результат досягнуто використанням методу редагування фактичних зв'язків між її елементами (англ. Rank-One Model Editing) [3]. У такий спосіб досягається змінення ваг окремих шарів нейронів і, як наслідок, надання хибних відповідей відповідно до змінених навчальних наборів даних. Тому визначання складників кіберризику отруєння даних великих мовних моделей є актуальним завданням.

Ризик кібербезпеки отруєння даних великих мовних моделей визначається як вплив невизначеності на забезпечування непорушності цілісності навчальних наборів даних [4]. У даному формулюванні під впливом невизначеності розумітимемо «відсутність – наявність» інформації про відповідні вразливості, загрози, наслідки, заходи та засоби захищання. За основу її структурування взято підхід до виокремлювання і представляння взаємозв'язків між складниками ризику (рис. 1), який викладено в [5]. До них віднесено інформаційний актив, кіберзагрозу, агента кіберзагрози, кіберризик, захід та засіб захищання, розпорядник. Під інформаційним активом розуміються навчальні набори даних. Вони мають значення як для окремого користувача, так і великої мовної моделі. З одного боку, на основі цілісних навчальних наборів даних можливе отримання правдивих відповідей для сформульованих промтів. З іншого – унеможливлювання змінень ваг шарів нейронів і, як наслідок, алгоритмів функціонування великої мовної моделі. На це направлена діяльність агентів кіберзагроз – і внутрішнього, і зовнішнього зловмисників. Характерною особливістю першого різновиду на відміну від другого є обізнаність про велику мовну модель, заходи та засоби її захищання. Попри це діяльність кожного з них зводиться до реалізування кіберзагрози змінення навчальних наборів даних. Запобігання її реалізуванню можливе впроваджуванням відповідних заходів і засобів. Наприклад [2], відстежування походження і трансформування навчальних наборів даних за допомогою OWASP CycloneDX, ML-BOM. При цьому отруєння даних великих мовних моделей до того ж асоціюється з ризиком ланцюгів постачання [3, 8]. Його прояв пов'язується з упроваджуванням неправдивих навчальних наборів даних без змінення їх нормальної поведінки. У даному прикладі персоналізоване рішення PoisonGPT є «троянським конем» [3].

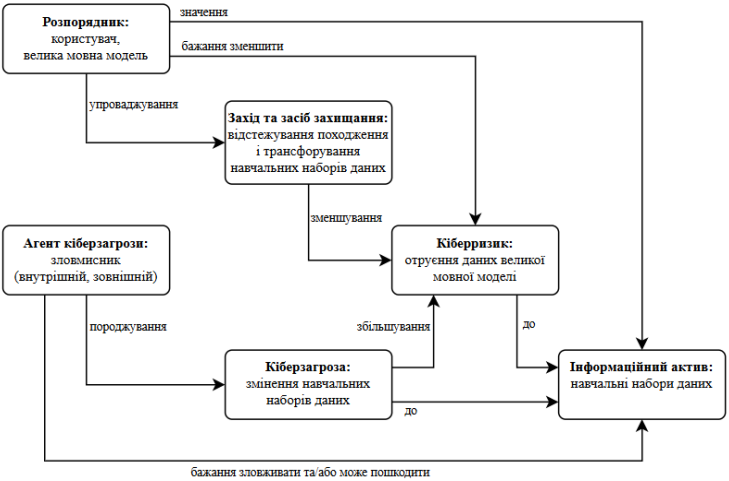

Рисунок 1 – Приклад представлення взаємозв'язків між складниками кіберризику отруєння даних великих мовних моделей [2, 6]

Отже, визначено складники кіберризику отруєння даних великих мовних моделей. Серед них виокремлено інформаційний актив, кіберзагрозу, агента кіберзагрози, кіберризик, захід та засіб захищання, розпорядника. За основу такого виокремлювання і встановлювання зв'язків взято міжнародний підхід з гармонізованого в Україні стандарту ISO/IEC 15408-1:2022. Окрему увагу приділено взаємозалежності кіберризиків отруєння даних великих мовних моделей, LLM04:2025, і ланцюгів постачання, LLM03:2025. Як підтвердження її існування використано персоналізоване рішення PoisonGPT.

7. 2025 Top 10 Risk & Mitigations for LLMs and Gen AI Apps. URL: https://genai.owasp.org/llm-top-10/ (accessed on: 25.11.2025).
8. LLM04:2025. Data and Model Poisoning. URL: https://genai.owasp.org/llmrisk/llm042025-data-and-model-poisoning/ (accessed on: 25.11.2025).
9. Khan A. PoisonGPT : Weaponizing AI for disinformation. URL: https://blog.barracuda.com/2025/09/11/poisongpt-weaponizing-ai-disinformation (accessed on: 25.11.2025).
10. Youssef P., Zhao Z., Seifert C., Schlötterer J. Tracing and Reversing Rank-One Model Edits. *Computation and Language*. 2025. arXiv: 2505.20819. DOI: https://doi.org/10.48550/arXiv.2505.20819.
11. ISO 31073:2022. Risk management. Vocabulary. [Valid from 2022-02-15]. URL: https://www.iso.org/standard/79637.html (accessed on: 25.11.2025).
12. ISO/IEC 15408-1:2022. Information security, cybersecurity and privacy protection. Evaluation criteria for IT security. Part 1 : Introduction and general model. [Valid from 2022-08-09]. URL: https://www.iso.org/standard/72891.html (accessed on: 25.11.2025).
13. LLM03:2025 Supply Chain. URL: https://genai.owasp.org/llmrisk/llm032025-supply-chain/ (accessed on: 25.11.2025).

А.І. Вавіленкова

ВИКОРИСТАННЯ ШТУЧНОГО ІНТЕЛЕКТУ ДЛЯ ВИЯВЛЕННЯ ТА НЕЙТРАЛІЗАЦІЇ КІБЕРЗАГРОЗ

Інтенсивний розвиток цифрових технологій спричинив різке зростання кількості та складності кібератак, що становлять загрозу для державних установ, бізнесу та критичної інфраструктури. Традиційні методи захисту, що базувалися на сигнатурному аналізі та вручну сформованих правилах, виявилися недостатньо ефективними в умовах постійного ускладнення технік зловмисників, зокрема використання експлойтів нульового дня, поліморфного шкідливого програмного забезпечення та соціальної інженерії. Саме тому використання технологій штучного інтелекту набуває особливої актуальності для процесу виявлення та нейтралізації кіберзагроз.

Використання штучного інтелекту в кібербезпеці передбачає автоматизацію задач, які раніше виконувалися аналітиками вручну, а також розширення аналітичних можливостей систем моніторингу за рахунок здатності алгоритмів виявляти складні патерни поведінки. Основною перевагою штучного інтелекту є його здатність працювати з великими масивами даних у режимі, наближеному до реального часу, що дозволяє швидко реагувати на нові загрози.

У даних матеріалах запропоновано використати алгоритм автоматичного аналізу логіко-лінгвістичних моделей електронних текстових документів [1] у компоненті пошуку та аналізу інформації SIEM-системи з метою подальшого порівняння системою математичних моделей вхідних повідомлень, що надходять в систему з математичними моделями сигнатур, закладених у базі знань системи.

На рис. 1 наведено архітектуру SIEM-системи з інтегрованим модулем автоматичного аналізу вхідних повідомлень.

Фактично представлена на рис. 1 архітектура SIEM-системи забезпечує виконання чотирьох основних процесів: збір даних, консолідація даних, сповіщення про події та політика подальших дій [2]. Усі вони працюють на повторній та постійній основі.

Збір даних – це частина, де система збирає всю можливу інформацію про події в мережі. Джерелами цієї інформації є комп'ютери в мережі, сервери та мережеві засоби. Крім того, SIEM може збирати інформацію з програмного забезпечення кібербезпеки, якщо воно наявне та налаштоване для обміну даними з інформацією про безпеку та системами керування подіями. На цьому етапі формується набір даних, необхідний для подальшого аналізу.

Завданням блоку «Джерела подій» є максимально повне, але кероване охоплення інформації, що може надходити в систему, визначення пріоритетів і обов’язкових журналів.

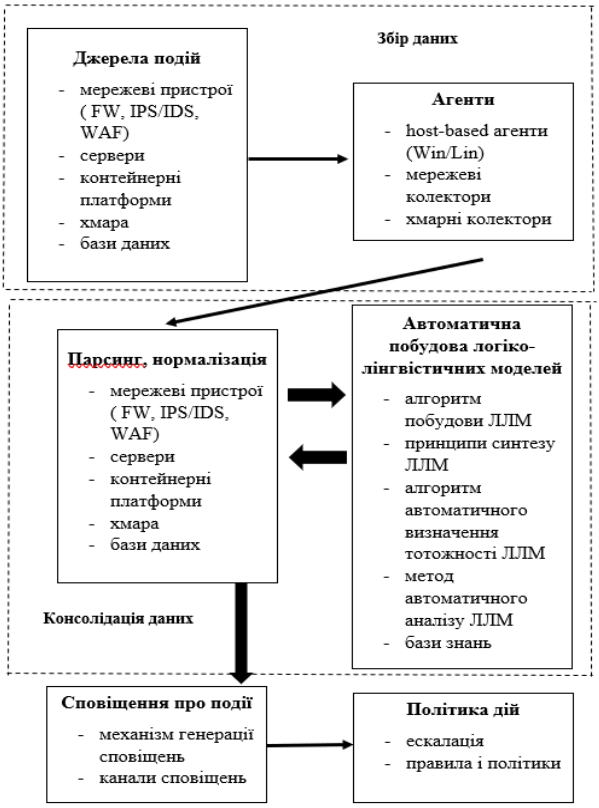

Рисунок 1 – Архітектура SIEM-системи з інтелектуальною компонентою пошуку та аналізу інформації

Блок «Агенти» – це перший активний шар SIEM-системи, який забезпечує захоплення, попередню обробку та надійну доставку подій від джерела до аналітичного ядра. Від якості цього шару залежить повнота даних, достовірність того, що події не будуть підмінені або пошкоджені, та на швидкість реакції на затримки між подіями та виявленням інциденту.

Консолідація даних – це етап аналізу зібраної інформації. Щоб було зручніше як для системи, так і для спеціалістів, SIEM автоматично сортує вхідну інформацію**.** Потім дані групуються за категоріями та визначається кореляція подій. Цієї інформації вже достатньо, щоб зробити висновки та простежити причинно-наслідкові зв'язки. Останнє надзвичайно важливо, коли йдеться про розслідування інциденту кібербезпеки. Саме на цьому етапі пропонується залучити блок логіко-лінгвістичного моделювання вхідних повідомлень.

Блок «Парсинг, нормалізація» відповідає за прийняття потоку даних від агентів-колекторів для їх подальшої структуризації, тобто приведення полів до єдиного стандарту (нормалізації), узгодження часових міток та формату, а також відкидання зайвого шуму та дублювання. Саме блок парсингу та нормалізації працює в SIEM-системах не коректно, отримує хаотичні логи, які важко аналізувати, і навіть найкращі аналітичні правила дають багато помилкових спрацювань. Тому пропонується підсилити етап консолідації даних блоком автоматичної побудови логіко-лінгвістичних моделей електронних текстових документів.

Лінгвістичний процесор для автоматизованого формування логіко-лінгвістичних моделей представляє собою транслятор, що встановлює відповідність між текстом та його змістом і здатен працювати з природною мовою у всьому її об'ємі. Дві основні функції лінгвістичного процесору полягають у вилученні змісту із заданої текстової інформації та вираженні отриманого змісту мовою логіки предикатів для можливості подальшого порівняння сформованих моделей

Використання моделей інтелектуального виявлення загроз дозволяє знизити навантаження на персонал безпеки та підвищити їхню продуктивність. Це особливо важливо для великих організацій з розгалуженою інфраструктурою, де щодня генеруються мільйони подій безпеки. Інтелектуальні механізми забезпечують можливість адаптивної моделі реагування на нові типи атак. Машинне навчання дозволяє системам SIEM та IDS/IPS постійно вдосконалювати свої алгоритми виявлення шляхом самонавчання на основі нових зразків загроз і поведінкових аномалій [3].

Отже, інтеграція інтелектуальних механізмів, таких як логіко-лінгвістичне моделювання, у SIEM-системи та системи виявлення вторгнень є стратегічно необхідною умовою для забезпечення належного рівня кібербезпеки сучасних організацій.

1. Вавіленкова А. І. Аналіз і синтез логіко-лінгвістичних моделей речень природної мови : монографія / А. І. Вавіленкова. – Київ : ТОВ «СІК ГРУП УКРАЇНА», 2017. – 152 с.
2. Вавіленкова А. І. Механізми здійснення кібератак та їх аналітичного виявлення / А. І. Вавіленкова, О. І. Скіцько, А. А. Півень // International Science Journal of engineering & Agriculture. – 2023. – №6. – С.1–7.
3. Вавіленкова А. І. Методи і засоби захисту від кіберзагроз : навч. посіб. / А. І. Вавіленкова, Ю. Є. Добришин, О. Ф. Шатирко. – Київ : Нац. акад. СБУ, 2025. – 175 с.

С.М. Сидоренко

ВПЛИВ ШТУЧНОГО ІНТЕЛЕКТУ НА КОНФІДЕНЦІЙНІСТЬ, ДЕСКРИМІНАЦІЮ ТА НАГЛЯД

Стрімкий розвиток технологій штучного інтелекту (ШІ) у XXI столітті зумовив глибокі трансформації у всіх сферах суспільного життя. Алгоритми машинного навчання, обробка великих даних, комп'ютерний зір, генеративні моделі та інтелектуальні системи прийняття рішень формують нову інфраструктуру цифрової взаємодії. Хоча ШІ створює безпрецедентні можливості для економічного зростання, автоматизації та соціального розвитку, він одночасно стає джерелом суттєвих викликів у сфері конфіденційності, дискримінації та цифрового нагляду. Суспільства стикаються з необхідністю збалансувати інновації з гарантуванням прав людини, прозорістю технологій та етичними принципами їх застосування.

У науковій та нормативній літературі підкреслюється, що впровадження ШІ у сфери критичної інфраструктури, такі як охорона здоров'я, правоохоронна діяльність, фінансовий сектор, державне управління, вимагає системного аналізу потенційних ризиків. Зростає потреба у глибокому розумінні того, як саме алгоритми обробляють дані, які обмеження вони мають, та як машинні моделі можуть ненавмисно формувати нові або посилювати існуючі форми нерівності [1].

Дослідження показують, що алгоритми збору інформації та обробки великих масивів даних можуть об'єднувати та аналізувати розрізнені набори даних, відновлюючи унікальні профілі користувачів. Зокрема, навіть шаблони руху мишкою або ритм натискань на клавіатуру здатні виступати у ролі поведінкових біометричних характеристик. Це підвищує ризики несанкціонованого профілювання, стеження та викрадення даних.

Біометричні дані (відбитки пальців, розпізнавання обличчя, голосових характеристик) належать до найчутливіших категорій інформації. На відміну від паролів, які можна змінити, біометричні маркери залишаються незмінними протягом життя. Використання ШІ у розпізнаванні облич може призводити до масштабних витоків та зловживань: підроблення біометричних даних, атаки через 3D-моделі облич, автоматизований захват даних з відкритих джерел.

Поширення камер із функціями комп'ютерного зору створює інфраструктуру постійного моніторингу, що може використовуватися як приватними компаніями, так і структурами державної безпеки. Контроль за зберіганням та обробкою таких даних стає критично важливим аспектом захисту приватності.

Попри формальне існування механізмів згоди на обробку персональних даних, користувачі часто не усвідомлюють масштаби збору інформації та можливі наслідки. Таймлайни даних, приховані трекери, алгоритмічне

поєднання метаданих – усе це створює асиметрію між користувачами та технологічними компаніями. У практиці спостерігається явище «примусової згоди»: відмова від надання даних фактично унеможливлює користування сервісами.

Таким чином, конфіденційність в епоху ШІ набуває багатовимірного характеру, де потреба у захисті приватного життя зіштовхується з фактом повсюдної цифрової взаємодії [2].

Одним із ключових викликів сучасних інтелектуальних систем є алгоритмічна упередженість – систематичні помилки, пов'язані з нерівномірністю навчальних даних або некоректністю моделей. Відомо, що ШІ може відтворювати та підсилювати соціальні стереотипи, оскільки працює з історичними даними, які містять упередження за ознаками статі, етнічності, віку, соціального статусу.

Причини алгоритмічної дискримінації включають:

- небалансовані вибірки даних, що відображають перевагу певних груп;
- некоректні характеристики моделі, що можуть непрямо пов’язуватися з дискримінаційними ознаками;
- використання неконтрольованого машинного навчання, де моделі формують небажані патерни без людського втручання;
- непрозорість алгоритмів глибинного навчання, що унеможливлює аудит та пояснення рішень.

Розвиток ШІ радикально розширює можливості систем спостереження. Якщо раніше нагляд обмежувався фізичними камерами або ручною обробкою інформації, то сучасні системи комп’ютерного зору здатні [3]:

- ідентифікувати обличчя у реальному часі;
- аналізувати поведінкові патерни;
- прогнозувати потенційно підозрілу активність;
- розпізнавати емоції та психометричні характеристики.

Це формує нову парадигму «інтелектуального нагляду», яка поєднує моніторинг, аналіз та автоматизоване ухвалення рішень.

Системи нагляду на основі ШІ можуть використовуватися як інструменти:

- політичного контролю**,**
- придушення свободи слова**,**
- стеження за громадськими рухами**,**
- профілювання населення**.**

У деяких країнах такі системи інтегровані у державні механізми контролю, що породжує ризики масових порушень прав людини. Особливо небезпечним є створення інфраструктури «тотального цифрового обліку», коли кожна взаємодія громадянина у публічному просторі може бути зафіксована та проаналізована.

Технології штучного інтелекту відкривають широкі можливості для економічного і соціального розвитку, однак їх застосування супроводжується суттєвими викликами. Вплив ШІ на конфіденційність, дискримінацію та нагляд є багатовимірним і потребує глибокого міждисциплінарного аналізу. Алгоритмічні системи можуть як сприяти підвищенню ефективності й безпеки, так і створювати ризики масового стеження, профілювання, упередженості та обмеження прав людини [4].

Перспективи подальшого розвитку ШІ повинні базуватися на етичних принципах, прозорості рішень, захисті персональних даних та відповідальності розробників і організацій. Лише за умови забезпечення балансу між інноваціями та правами людини інтелектуальні технології можуть стати фундаментом сталого інформаційного суспільства.

1. European Commission. (2024). *Artificial Intelligence Act (AI Act).* Official Journal of the European Union.
2. Kosanović, M., & Marković, M. (2022). *Artificial Intelligence, Privacy, and Human Rights.* Journal of Cyber Policy.
3. Литвиненко О.Є. Системи штучного інтелекту: навч. посіб. / О. Є. Литвиненко – К.:НАУ. – 2023. – 168 с. [Електронний ресурс].
4. Троцько В.В. Методи штучного інтелекту: Навчально-методичний і практичний посібник. / В.В. Троцько – Київ: Університет економіки та права «КРОК», 2020 – 86 с.

А.Р. Коновалова

ШІ-КОНТЕНТ У СОЦМЕРЕЖАХ: ВИКЛИКИ МЕДІАГРАМОТНОСТІ ТА БЕЗПЕКИ

Стрімкий розвиток генеративного штучного інтелекту радикально змінює природу інформаційного споживання та виробництва контенту в соціальних мережах. У 2025 році ШІ став не лише інструментом для творчих індустрій, але й ключовим фактором трансформації цифрової екосистеми, де кордони між правдою, вигадкою та симуляцією дедалі більше розмиваються. Поява високореалістичних deepfake-відео, автоматизованих текстових та візуальних генерацій і "синтетичних інфлюенсерів" створює новий рівень інформаційної вразливості, з якою суспільство ще не стикалося в такому масштабі. З огляду на війну в Україні та гібридні загрози, поширення ШІ-контенту отримує додатковий вимір - воно впливає не лише на довіру аудиторії, а й на національну безпеку.

Соціальні мережі, що побудовані на алгоритмічних рекомендаціях і максимізації залучення, стали середовищем, де ШІ-контент здатен поширюватися зі швидкістю, яку неможливо контролювати традиційними засобами. З одного боку, це відкриває нові можливості для брендів, медіа та креаторів; з іншого - створює умови для маніпуляцій, фрагментації інформаційного поля та масових дезінформаційних кампаній. У цих умовах актуальним стає питання вироблення стандартів комунікаційної безпеки, визначення меж відповідального використання ШІ та формування нових компетентностей медіаграмотності, які дозволять суспільству адаптуватися до цифрової реальності, де синтетичний контент стає нормою.

У листопаді 2025 року українські медіадослідники фіксують хвилю системного використання deepfake-технологій у TikTok, спрямовану на маніпуляцію довірою до відомих журналісток. За даними розслідування Texty.org.ua, було проаналізовано 595 ШІ-згенерованих роликів, що загалом набрали понад 24 млн переглядів. У цих відео зловмисники створювали повністю штучні копії журналісток - від синтезованих голосів до повноцінних цифрових аватарів із згенерованою мімікою й інтонаціями. Особливо активно експлуатувався образ ведучої Соломії Вітвіцької: зафіксовано 100 роликів, де її "цифрова копія" озвучує вигадані повідомлення, часто узгоджені з пропагандистськими наративами. Більшість таких відео містили аудіоманіпуляції - реальні сюжети були переозвучені штучно створеним текстом, що імітує голоси відомих медійниць, посилюючи ефект достовірності. Водночас контент із повним підробленням і зображення, і звуку становив близько 45% вибірки, демонструючи, що deepfake-маніпуляції перестали бути технологічним експериментом і перетворилися на інструмент системного інформаційного впливу [1].

Масштабне поширення таких підроблених відео демонструє, наскільки швидко ШІ-контент здатен зруйнувати традиційні механізми довіри у соцмережах. Користувачі вже не можуть орієнтуватися на візуальні або аудіальні ознаки автентичності, адже якість підробок перевищує поріг людського сприйняття. Це створює нову комунікаційну реальність, у якій фактологічна точність змагається із швидкістю й вірусністю неправдивих повідомлень. У результаті аудиторія стикається з "ефектом інформаційного розмиття": кордон між журналістикою, розвагами та маніпуляцією стає дедалі нечіткішим.

Для брендів та медіакомпаній це означає появу нових норм комунікаційної безпеки, які ще кілька років тому здавалися зайвими. У 2025 році компанії змушені впроваджувати політики верифікації контенту, цифрового захисту облич та голосів своїх представників, а також навчати команди розпізнавати ШІ-маніпуляції. Бренди більше не можуть покладатися на репутаційний капітал як основний щит від фейків - навпаки, що впізнаваніша публічна особа, то легше її перетворити на «цифровий інструмент» для атак конкурентів або зовнішніх акторів. Це формує запит на нові стандарти прозорості, маркування контенту, партнерських перевірок та швидкого кризового реагування. У комунікаційному середовищі, де штучні образи поширюються швидше, ніж офіційні спростування, стратегічна медіаграмотність стає не тільки компетенцією аудиторії, а й ключовою частиною безпеки будь-якої організації.

Бренди сьогодні опиняються на передовій інформаційної безпеки: використання генеративного ШІ відкриває нові можливості для швидкої адаптації контенту та персоналізації, але одночасно породжує серйозні репутаційні ризики. Як зазначає Інститут масової інформації (IMI), редакції та маркетингові команди все частіше стикаються з потребою впроваджувати політики щодо використання ШІ - зокрема, чітко визначати, хто відповідає за результати роботи моделей, які завдання їм ставлять та як контролюється якість [2]. У ситуації, коли один фейковий AI-ролик із участю брендової фігури може спричинити значні втрати довіри, бізнесу необхідно не просто реагувати на кризу, а попереджати її через проактивне налаштування процесів.

Водночас комунікаційна політика брендів і медіа в Україні змушена адаптуватися до нових стандартів цифрової епохи: як підкреслює Центр демократії та верховенства права (ЦЕДЕМ), необхідно маркувати AI-контент, розкривати інформацію про використання систем ШІ, перевіряти легальність і етичність вибору моделей [3]. Крім того, ініціативи державних і приватних організацій (наприклад, програмні документи від Міністерство цифрової трансформації України) містять алгоритми оцінювання ризиків, що допомагають брендам та медіа обирати надійні системи і зменшувати можливі правові, фінансові та репутаційні втрати [4].

Водночас звіт “Фільтр” від МКІП констатує, що алгоритмічні упередження й викривлення інформації через ШІ значно ускладнюють медіаграмотність: штучний інтелект може підсилювати дезінформацію, змінювати способи таргетингу й маніпулювати споживанням контенту [5].

У таких умовах особливої ваги набуває оперативна ідентифікація фейків та швидке реагування, адже темп поширення ШІ-контенту значно випереджає можливості традиційних механізмів перевірки. Чимало українських фактчекінгових ініціатив уже створюють інструменти моніторингу, однак ключовим викликом залишається автоматизація процесів виявлення deepfake-відео, аудіоманіпуляцій і синтетичних зображень у реальному часі. Крім того, враховуючи, що значна частина маніпулятивних роликів використовує реальні персональні дані - обличчя, міміку, голоси - суспільству потрібно формувати нову культуру захисту цифрової особистості.

Не менш важливою є поява стандарту маркування власного AI-контенту, який поступово стає вимогою доброчесності. Відкритість щодо використання генеративних інструментів зменшує ризики маніпуляції, підвищує довіру до виробників контенту й допомагає аудиторії розрізняти, де починається художня реконструкція, а де - навмисне викривлення фактів. Розроблення цього стандарту неможливе без юридичних рамок: потребує врегулювання авторського права на синтетичні матеріали, визначення відповідальності за поширення фейкових відео та координації з державними структурами у випадках масових атак чи інформаційних диверсій.

Станом на 2025 рік ШІ-контент у соціальних мережах став ключовим фактором трансформації інформаційного простору, одночасно відкриваючи нові можливості та породжуючи безпрецедентні ризики для медіаграмотності, безпеки й довіри. Масове використання deepfake-технологій та алгоритмічних маніпуляцій, зокрема у TikTok, демонструє, наскільки легко сьогодні можна викривити реальність, створити фальшиві образи публічних осіб і впливати на громадську думку через емоційно насичений, візуально переконливий контент. У цих умовах межа між креативним використанням ШІ та маніпуляцією дедалі більше стирається, а відповідальність переноситься не лише на авторів контенту, а й на платформи, бренди та державу.

Водночас стрімкий розвиток інструментів ШІ формує нову екосистему відповідальності. Журналістам і користувачам потрібні навички ідентифікації синтетичних матеріалів, брендам - оновлена політика комунікаційної безпеки та захисту репутації, а державним органам - оперативні механізми реагування й оновлене законодавче поле. Чітке маркування AI-контенту, захист цифрових образів, прозорі алгоритми модерації та міжсекторальна координація стають базовими умовами стійкості до інформаційних атак.

Отже, головним викликом сьогодні є не стільки сам факт існування ШІ-технологій, скільки здатність суспільства виробити нові стандарти

комунікації, що поєднують інновації та безпеку. Лише формування комплексної системи - технічної, освітньої, правової й етичної — дозволить зберегти довіру, убезпечити інформаційний простір і забезпечити здорову взаємодію між людьми, медіа та штучним інтелектом у добу цифрової реальності.

1. Штучний інтелект і ТікТок: як відомих журналісток перетворюють на фейки / С. Міхальков та ін. *Texty.org.ua*. URL: https://texty.org.ua/projects/116211/shtuchnyj-intelekt-i-tiktok-yak-vidomyx-zhurnalistok-peretvoryuyut-na-fejky/ (дата звернення: 19.11.2025).
2. Турчина М. Штучний інтелект у роботі медійників: поради від Медіабази Хмельницький. *Інститут масової інформації*. URL: https://imi.org.ua/advices/shtuchnyj-intelekt-v-roboti-medijnykiv-porady-vid-mediabazy-hmelnytskyj-i69173 (дата звернення: 03.07.2025).
3. Штучний інтелект: можливості та виклики для медіа. *Центр демократії та верховенства права*. URL: https://cedem.org.ua/consultations/shtuchnyj-intelekt-mozhlyvosti-ta-vyklyky-dlya-media/ (дата звернення: 20.02.2025).
4. Поліковська Ю. Мінцифри представило рекомендації для українських медіа з відповідального використання ШІ. *ms.detector.media*. URL: https://ms.detector.media/internet/post/34061/2024-01-25-mintsyfry-predstavylo-rekomendatsii-dlya-ukrainskykh-media-z-vidpovidalnogo-vykorystannya-shi/ (дата звернення: 25.01.2024).
5. Вплив штучного інтелекту на розвиток медіаграмотності. Аналітичний звіт 2024 / О. Кравченко та ін. *Фільтр - Національний проєкт з медіаграмотності*. URL: https://filter.mcsc.gov.ua/news/vplyv-shtuchnoho-intelektu-na-rozvytok-mediahramotnosti-analitychnyy-zvit-2024/ (дата звернення: 06.01.2025).

І.В. Плетяний

РОЗВИТОК МЕТОДІВ АВТОМАТИЗАЦІЇ НАВЧАЛЬНО-МЕТОДИЧНОГО ЗАБЕЗПЕЧЕННЯ ТРЕНАЖЕРІВ З ВИКОРИСТАННЯМ ЕЛЕМЕНТІВ ШТУЧНОГО ІНТЕЛЕКТУ

Одним із основних напрямків наукових досліджень, які проводить відділ «Моделювання енергетичних процесів та систем» ІПМЕ ім. Г.Є. Пухова НАН України є розробка методів та комп'ютерних засобів конструювання тренажерних систем підготовки та контролю знань диспетчерського персоналу в енергетиці [1]. Цей напрямок досліджень передбачає розробку інформаційної технології побудови тренажерних систем для підготовки диспетчерського персоналу в енергетиці, яка базується на концепції сценарно-імітаційного моделювання енергетичного обладнання та робочої діяльності персоналу.

Основні переваги тренажерів, які реалізовані у відповідності з вище вказаною технологією, характеризуються не «ступенем подоби» або «глибиною моделювання», а сукупністю реалізованих тренажерних занять, які формують необхідні професійні навички щодо управління енергетичним обладнанням [2]. Тренажерне заняття реалізується у формі сценарію, основною якого є робоча діяльність персоналу, до якої добавлені елементи навчальної діяльності та відповідна реакція енергетичного обладнання на дії персоналу. Результатом проектування тренажера є набір тренажерних занять у вигляді графічних діаграм і задокументованих процесів на основі стандартизованих нотацій BPMN. Отримані діаграми – основа для програмної реалізації тренажера.

На основі оцінки складності моделі енергетичного обладнання або режимів його функціонування (моделі об'єкта тренажера) в сенсі забезпечення комфортного часу відклику, можливе використання імітаційних моделей фрагментів управління об'єктом для кожного тренажерного заняття замість єдиної моделі об'єкта тренажера. Ці імітаційні моделі можуть бути задані формулами або даними у вигляді зрізів режиму функціонування відповідно до сценарію виконання тренажерного заняття.

В основі імітаційного моделювання енергетичного обладнання тренажерів лежать принципи блочного модельного конструювання, відповідно з якими імітаційна модель розглядається як з'єднання в причинно – наслідкову мережу функціональних елементів, типових бібліотечних елементів, функціональних блоків та відповідних шаблонів.

Якщо припустити що імітаційна модель управління певним енергетичним об'єктом розроблена максимально наближеною до оригіналу і повністю задовольняє вимогам щодо математичного (імітаційного) моделювання для тренажерів, то наступним кроком, в рамках загальної концепції створення тренажерів, буде його впровадження в певну систему підготовки персоналу, тобто розробка методики його застосування. Наш досвід показує, що це далеко не простий процес і іноді навчально-методичне забезпечення тренажера, яке створене під час його впровадження, може

нівелювати певні показники глибини і обсягів моделювання і в результаті значно знизити ефективність застосування тренажера в цілому. Основний висновок має наступний зміст – наявність інструментів для розробки навчально-методичного забезпечення тренажера має значний вплив на ефективність застосування тренажера в навчальному процесі і посідає значне місце в інформаційній технології побудови тренажерних систем для підготовки диспетчерського персоналу в енергетиці.

Значний потенціал подальшого розвитку технології, на нашу думку, міститься в напрямку створення інструментів побудови і автоматизації навчально-методичного забезпечення (супроводження) тренажу з використанням елементів штучного інтелекту (ШІ) та елементів геймифікації.

Реалізація інформаційної технології побудови тренажерних систем для підготовки диспетчерського персоналу в енергетиці здійснюється нами на базі сучасних програмних систем, призначених, в першу чергу, для створення комп'ютерних ігор (Unity, GDevelop). Сучасний комп'ютерний тренажер є певною комп'ютерною грою і такі елементи геймифікації, як система балів, рівнів, досягнень і візуальний зворотній зв'язок, можуть сприяти підвищенню мотивації, залученості і стійкості засвоєння навчального матеріалу. Такий підхід має сприяти підвищенню ефективності тренажера як засобу підготовки.

Сучасні тенденції розвитку систем ШІ створюють умови і мотивацію для наукових досліджень в сфері використання елементів ШІ в тренажерних системах і створенні відповідних інструментів їх побудови. Вони дадуть змогу, на наш погляд, ухвалювати рішення на основі аналізу конкретних умов щодо успішності/неуспішності операторської діяльності, автоматичного аналізу помилок і формування індивідуальних траєкторій навчання. При цьому інтелектуальна компонента має будуватись на основі правил і містити механізм відстеження дій в реальному часі з можливістю оперативного виявлення помилок і корегування складності сценаріїв навчання. Такий підхід дозволить динамічне нарощування рівнів складності у відповідності з поточними навичками того хто навчається, а також забезпечить персоналізований розвиток компетенцій.

Таким чином, перелічені вище напрямки розвитку інформаційної технології побудови тренажерів для підготовки диспетчерського персоналу в енергетиці будуть враховані під час виконання науково-дослідної роботи «Тренажер-І», а саме при розробці елементів технології конструювання моделі інструктора, що включає імітаційну модель оцінювання навчальної діяльності для додавання її в сценарно-імітаційну модель тренажерного заняття.

1. Абрамович Р.П., Самойлов В.Д. Технології конструювання комп'ютерних систем підготовки персоналу в енергетиці // К.: «Три К», 2021. – 111 с. (ISBN 978-966-7690-58-8).
2. Плетяний І.В., Самойлов В.Д., Чьочь В.В. Проєктування локальних тренажерів на основі сценарно-імітаційного моделювання // «Ядерна та радіаційна безпека», №4(100) (2023). ISSN (print) 2073-6231. DOI:10.32918/nrs.2023.4(100).05.

ЗМІСТ

ЗБІРНИК МАТЕРІАЛІВ

НАУКОВО-ПРАКТИЧНОЇ КОНФЕРЕНЦІЇ
«ШТУЧНИЙ ІНТЕЛЕКТ І БЕЗПЕКА»
04 грудня 2025 року

З питаннями щодо конференції звертатися:

ІПМЕ ім. Г.Є. Пухова НАН України,

вул. Олега Мудрака (Генерала Наумова), 15,

кім. 303, Цуркан Оксана володимирівна, тел. 424-91-62,

068-014-57-22, e-mail: otsurkan24@gmail.com