ISSN 2519-2310 (Online)

КОМП'ЮТЕРНІ НАУКИ ТА КІБЕРБЕЗПЕКА, 2024, № 2 (26)

COMPUTER SCIENCE AND CYBERSECURITY, 2024, Iss. 2 (26)

DOI: https://doi.org/10.26565/2519-2310-2024-2-05
УДК 004.77

ДОСЛІДЖЕННЯ ПОТОЧНОГО СТАНУ ТА ПЕРСПЕКТИВИ ЗАСТОСУВАННЯ ШТУЧНОГО ІНТЕЛЕКТУ В КІБЕРБЕЗПЕЦІ

Юрій Голіков[1], e-mail: yuriy@devbrother.com
Єлизавета Острянська[2], молодший науковий співробітник, e-mail: antelizza@gmail.com,
ORCID: https://orcid.org/0000-0003-1412-8470
[1]*DevBrother tech company, США*
[2]*Харківський національний університет імені В.Н. Каразіна,*
майдан Свободи, 4, Харків, 61022, Україна
Рукопис надійшов 1 листопада 2024 р. Отримано після рецензування 2 грудня 2024 р.
Прийнято 23 грудня 2024 р.

Анотація: В сучасному світі, з розвитком нових технологій, штучний інтелекту (ШІ) у сфері кібербезпеки є невід'ємною складовою, тому вивчення його переваг, ризиків та можливі сценарії використання є актуальною темою дослідження. В сучасному цифровому середовищі, де кіберзагрози стають дедалі складнішими, впровадження технологій ШІ значно підвищує ефективність систем захисту, дозволяючи автоматизувати виявлення та реагування на атаки. Тому було розглянуто основні напрями застосування ШІ в кібербезпеці, зокрема виявлення загроз, аналіз шкідливого програмного забезпечення, посилення криптографії, захист від фішингових атак та прогнозування атак. Одним із ключових аспектів є інтеграція ШІ в системи управління інформаційною безпекою (SIEM), які аналізують величезні обсяги даних та допомагають виявляти аномалії. Такі системи значно знижують навантаження на команди безпеки та підвищують точність і швидкість реагування. Окрему увагу приділено аналізу сучасних антивірусних рішень, що використовують штучний інтелект, а саме Microsoft Defender for Endpoint та Darktrace. Вони базуються на алгоритмах поведінкового аналізу та машинного навчання, що дозволяє ефективніше виявляти складні загрози та запобігати інцидентам. Microsoft Defender забезпечує високий рівень захисту кінцевих точок. Натомість Darktrace використовує самонавчальні моделі для аналізу мережевого трафіку, що дозволяє виявляти загрози нульового дня та внутрішні загрози в організаціях. Розглянуто основні ризики, пов'язані із використанням ШІ у кіберзлочинності. ШІ активно застосовується зловмисниками для автоматизації атак, що значно підвищує їхню ефективність і складність виявлення. Розглянуто основні типи загроз на основі ШІ, а саме такі атаки, як атаки отруєння даних, атаки ухилення, атаки швидкого впровадження та соціальна інженерія на основі ШІ. Для зменшення ризиків пропонується розробка стійких моделей ШІ, що менше піддаються атакам, підвищення прозорості алгоритмів, а також впровадження міжнародних стандартів регулювання ШІ, чим активно займаються великі команди дослідників, серед яких і NIST. Також рекомендується підвищення обізнаності користувачів та спеціалістів з кібербезпеки, оскільки людський фактор залишається однією з найбільших вразливостей систем. У підсумку, наголошується, що штучний інтелект є ключовим фактором розвитку кібербезпеки, який може значно покращити захист інформації та критичних систем. Водночас, без належного регулювання та захисних заходів, ШІ може стати потужним інструментом для зловмисників, що створює нові виклики

для безпеки в цифрову епоху. Баланс між інноваціями, етичними нормами та безпекою стане ключовим фактором у формуванні стратегії ефективного використання ШІ у майбутньому.

Ключові слова: *штучний інтелект, кібербезпека, машинне навчання, SIEM, IDS, антивірус на основі ШІ*

Як цитувати: Голіков Ю., Острянська Є. Дослідження поточного стану та перспективи застосування штучного інтелекту в кібербезпеці. *Комп’ютерні науки та кібербезпека.* 2024; № 2(26): С. 51–65. https://doi.org/10.26565/2519-2310- 2024-2-05

In cites: Golikov Yu., Ostrianska Ye. (2024). Research on the current state and prospects of the application of artificial intelligence in cybersecurity. *Computer Science and Cybersecurity.* 2(26): 51–65. https://doi.org/10.26565/2519-2310-2024-2-05 (in Ukrainian)

1. Вступ

Штучний інтелект (ШІ) вже більше десятиліття змінює простір кібербезпеки, завдяки машинному навчанню (ML), яке прискорює виявлення загроз і виявляє аномальну поведінку користувачів і об’єктів. І тому ШІ поступово стає невід’ємною частиною сучасних систем кібербезпеки, допомагаючи ідентифікувати загрози, автоматизувати процеси та підвищувати рівень захисту інформації.

Одним із важливих викликів у використанні ШІ для кібербезпеки є формування довіри. Дані, які використовуються для навчання моделей AI/ML, керують виходом моделей. Якщо навчальні дані не відображають «реальний світ», модель може спотворити її здатність забезпечувати очікувані результати. Деякі дані, такі як інформація про загрози, «хороші» та «погані» характеристики файлів, індикатори компрометації тощо, є для всіх.

Ще одна серйозна проблема – безпека даних. Важливо визначити та контролювати, якими навчальними даними можна ділитися, а які дані залишаються секретними в середині організацій. У чужих руках ці дані можуть допомогти зловмисникам у їхніх атаках, щоб підірвати здатність AI/ML ідентифікувати їхні файли, програми та поведінку як недійсні. У зв’язку з цим державам і комерційним структурам необхідно розробити нормативні акти, стандарти та найкращі практики, щоб запобігти новим загрозам із застосуванням ШІ.

Так, наприклад, NIST [17, 18] вже очолює та бере участь у розробці технічних стандартів, включаючи міжнародні стандарти, які сприяють інноваціям і громадській довірі до систем, які використовують ШІ. Широкий спектр стандартів для даних штучного інтелекту, продуктивності та управління є – і все більше буде – пріоритетом для надійного та відповідального ШІ.

NIST розробив план глобальної взаємодії з просуванням і розробкою стандартів ШІ. Мета полягає в тому, щоб стимулювати розробку та впровадження консенсусних стандартів, пов’язаних зі штучним інтелектом, співпрацю та координацію, а також обмін інформацією. Відображаючи внесок державного та приватного секторів. 26 липня 2024 року, після розгляду коментарів громадськості щодо плану проєкту, NIST опублікував План глобальної взаємодії зі стандартами ШІ [17].

Розвиток технологій ШІ приніс не лише нові можливості, але й нові ризики та небезпеку. Спільнота науковців та дослідників по всьому світу стурбована та попереджають про потенційні проблеми поширення ШІ. Зокрема, нещодавній звіт Національного центру кібербезпеки Великої Британії (NCSC) [14] застерігає, що протягом найближчих двох років технології ШІ напевно збільшать динаміку кібератак та посилять їхній вплив на існуючі

криптосистеми та засоби захисту інформації. За оцінками Центру [14], ШІ відкриває широкі можливості у цьому напрямку навіть для тих криптоаналітиків, що не мають відповідних технічних навичок. І вони стверджують, що вже після 2025 року і далі, коли ШІ пройде достатньо успішні навчання, ШІ, майже напевно покращиться, забезпечуючи швидші й точніші кібероперації.

У найближчі роки роль ШІ лише зростатиме. Наприклад, ШІ здатен виявляти загрози у реальному часі, тобто може аналізувати величезні обсяги даних, виявляючи аномальні дії та потенційні загрози швидше, ніж це здатна зробити людина. Також алгоритми ШІ можуть не тільки ідентифікувати загрози, а й миттєво блокувати шкідливий трафік або змінювати правила доступу до мережі. Що також важливо, автоматизовані рішення допомагають уникнути помилок, які можуть виникати через людську неуважність або недостатню кваліфікацію.

Тому, метою цієї статті є дослідження поточного стану використання ШІ в сфері кібербезпеки, а також загрози та ризики пов’язані з його використанням. В статті розглядаються сучасні антивірусні рішення, найпопулярніші атаки з використанням ШІ та методи протидії. А також перспективи використання ШІ та доведення того, що в сучасному світі ШІ є невід’ємною частиною кіберзахисту.

2. Основні напрямки використання ШІ в сфері кібербезпеки

Нижче розглянемо шість основних можливих варіантів використання ШІ в сфері кібербезпеки (рис. 1):

1. Автоматизоване виявлення загроз та реагування. Традиційні системи кібербезпеки потребують постійного моніторингу та аналізу великих обсягів даних, що складно реалізувати вручну. ШІ може виконувати наступні задачі:

- Виявляти аномальну активність у мережі за допомогою алгоритмів машинного навчання.
- Аналізувати поведінку користувачів та пристроїв для виявлення підозрілих дій.
- Автоматично реагувати на загрози, блокуючи потенційно небезпечні дії без втручання людини.

Прикладом таких систем є, наприклад, система SIEM (Security Information and Event Management) з підтримкою ШІ, що аналізує події в реальному часі та попереджає про можливі атаки. Більш детально системи SIEM будуть розглянуті в розділі 3 цієї статті.

2. Використання ШІ в аналізі шкідливого програмного забезпечення. Традиційні антивіруси використовують підписи для виявлення шкідливого ПЗ, що робить їх менш ефективними проти нових атак. Натомість ШІ дозволяє:

- Використовувати поведінковий аналіз для виявлення нових вірусів та троянських програм.
- Розпізнавати різні види атак, які змінюють свій код для обходу антивірусів.
- Автоматично створювати та оновлювати бази загроз без необхідності ручного внесення змін.

Наприклад, зараз набувають популярності ШІ-антивіруси, такі як Microsoft Defender ATP або Darktrace, які аналізують поведінку програм і виявляють загрози без використання втручання людини. Більш детально дані методи ми розглядаємо в розділі 4 цієї статті.

3. Підсилення криптографії та постквантова безпека. Як відомо, з розвитком квантових комп’ютерів традиційні криптографічні алгоритми можуть стати вразливими. ШІ в свою чергу може допомогти у:

- Створенні адаптивних криптографічних рішень, які зможуть змінювати ключі та алгоритми в реальному часі.
- Оптимізації шифрування та розшифрування, що зробить криптографічні системи ефективнішими.
- Розробці нових постквантових криптографічних алгоритмів, які будуть стійкими до атак квантових комп’ютерів.

Тому ШІ активно використовується для тестування нових стійких шифрів. Оптимізація шифрування та управління ключами – алгоритми машинного навчання можуть покращувати ефективність криптографічних протоколів, забезпечуючи швидше та надійніше шифрування. Наприклад, NIST активно досліджує алгоритми шифрування та підписів постквантової криптографії [1], а ШІ може прискорити їх тестування та впровадження [18].

4. ШІ в протидії фішинговим атакам. Фішинг – це одна з найпоширеніших кіберзагроз, яка стає все складнішою. ШІ може допомогти у:

- Автоматичному розпізнаванні підозрілих електронних листів та URL-адрес.
- Аналізі мови та структури повідомлень для виявлення шахрайських схем.
- Покращенні механізмів аутентифікації, наприклад, через поведінкову біометрію або динамічні капчі.

Одним з найвідоміших прикладів використання ШІ в протидії фішингу є Google, що використовує ШІ в Gmail для фільтрації фішингових листів, що дозволяє блокувати понад 99% шкідливих повідомлень.

5. Використання ШІ для прогнозування атак. Штучний інтелект може допомогти не тільки реагувати на атаки, але й передбачати їх, аналізуючи величезні масиви даних:

- Використання аналізу великих даних для ідентифікації трендів у кібератаках.
- Створення моделей-прогнозів, які дозволяють передбачати потенційні вразливості систем.
- Автоматичний аналіз темної мережі (Dark Web) для виявлення нових загроз та хакерських інструментів.

Наприклад, IBM Watson for Cyber Security [7] використовує когнітивний аналіз для ідентифікації нових загроз, аналізуючи мільйони статей, форумів та повідомлень у даркнеті.

Рис. 1 – Використання ШІ в кібербезпеці
Fig. 1 – Using AI in cybersecurity

3. Захист на основі ШІ та підвищення ефективності кібербезпеки

Системи виявлення вторгнень (IDS) на базі штучного інтелекту (ШІ) є сучасними інструментами кібербезпеки, які використовують можливості машинного навчання та аналізу даних для ідентифікації та запобігання несанкціонованим діям у мережах та системах. Вони здатні адаптуватися до нових загроз, аналізуючи великі обсяги даних та виявляючи аномалії, що можуть свідчити про потенційні атаки.

Існують два основні методи, які використовуються в IDS:

1) Виявлення на основі відомих алгоритмів. Цей метод передбачає порівняння вхідного трафіку з базою даних відомих атак. Якщо знайдено збіг, система генерує сповіщення. Однак цей підхід обмежений у виявленні нових або модифікованих атак, які ще не мають відповідних алгоритмів.

2) Виявлення на основі аномалій: Цей метод використовує моделі нормальної поведінки системи. Відхилення від цієї норми розглядаються як потенційні загрози. ШІ відіграє ключову роль у побудові та оновленні таких моделей, що дозволяє виявляти навіть невідомі раніше атаки.

Зрозуміло, що інтеграція штучного інтелекту в системи виявлення вторгнень значно підвищує ефективність кібербезпеки, дозволяючи виявляти та запобігати як відомим, так і новим загрозам. Однак для максимального використання потенціалу таких систем необхідно враховувати можливі обмеження та етичні аспекти їх застосування, бо, наприклад, використання ШІ може порушувати питання конфіденційності, особливо якщо системи аналізують персональні дані без належного контролю.

4. Система SIEM з підтримкою ШІ

Управління інформацією про безпеку та подіями, або SIEM (Security Information and Event Management), – це рішення безпеки, яке допомагає організаціям розпізнавати й усувати потенційні загрози та вразливості безпеки до того, як у них з’явиться шанс порушити бізнес-операції [7].

Системи SIEM допомагають групам безпеки підприємства виявляти аномалії поведінки користувачів і використовувати штучний інтелект (ШІ) для автоматизації багатьох ручних процесів, пов’язаних із виявленням загроз і реагуванням на інциденти. Наразі, найвідомішими сучасними системами SIEM є такі програмні забезпечення: Splunk Enterprise Security (Splunk), Elastic Security, IBM QRadar SIEM, Wazuh SIEM, Microsoft Sentinel.

Початкові платформи SIEM були інструментами керування журналами. Вони поєднали функції керування інформацією про безпеку (SIM) і керування подіями безпеки (SEM). Ці платформи забезпечували моніторинг і аналіз подій, пов’язаних із безпекою, у реальному часі. Термін SIEM було введено у 2005 році для поєднання технологій SIM і SEM.

Крім того, вони полегшили відстеження та реєстрацію даних безпеки для цілей відповідності або аудиту. Протягом багатьох років програмне забезпечення SIEM розвивалося, щоб включити аналітику поведінки користувачів і об’єктів, а також інші передові засоби аналітики безпеки, штучний інтелект і можливості машинного навчання для виявлення аномальної поведінки та індикаторів передових загроз. Сьогодні SIEM став основним елементом сучасних центрів безпеки (SOC) для моніторингу безпеки та управління відповідністю.

Етапи роботи систем SIEM (рис. 2) можна поділити наступним чином [6]:

- Збір даних – усі джерела інформації про мережеву безпеку, наприклад сервери, операційні системи, брандмауери, антивірусне програмне забезпечення та системи запобігання вторгненням, налаштовані на передачу даних про події в інструмент SIEM.

Більшість сучасних інструментів SIEM використовують агенти для збору журналів подій із систем підприємства, які потім обробляються, фільтруються та надсилаються до SIEM. Деякі SIEM дозволяють збирати дані без агентів. Наприклад, Splunk [8] пропонує безагентний збір даних у Windows за допомогою WMI.

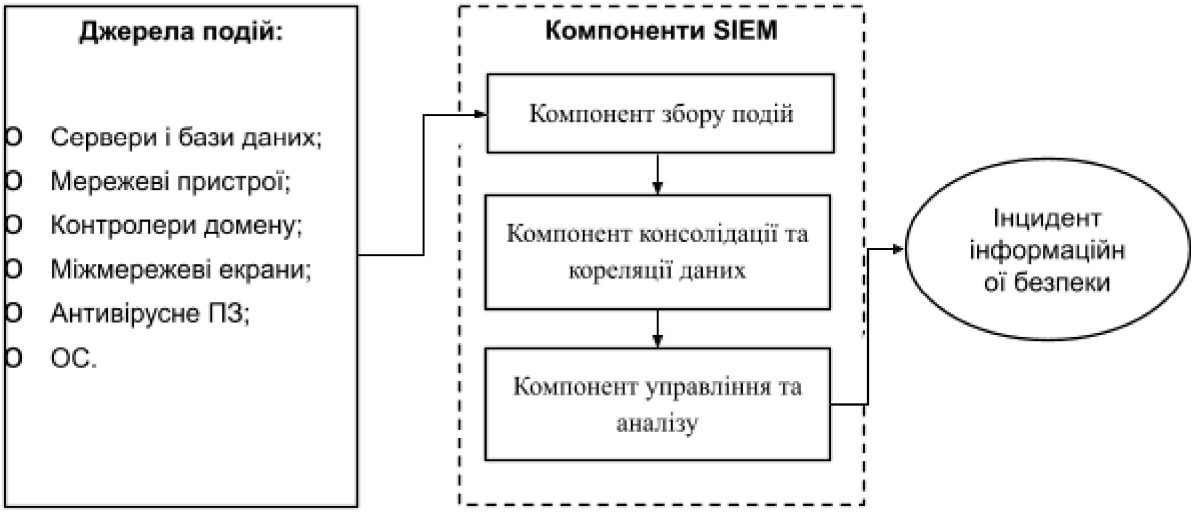

Рис. 2 – Схема роботи системи SIEM
Fig. 2 – SIEM system operation diagram

Системи AI SIEM починаються з агрегування даних із різних джерел, таких як мережеві пристрої, сервери, бази даних і програми. Після прийому необроблені дані перетворюються в стандартизований формат, що забезпечує послідовність і точність аналізу даних незалежно від джерела. AI та ML значно автоматизують ці процеси, підвищуючи швидкість і інтелект, з якими дані безпеки агрегуються та нормалізуються, скорочуючи ручні зусилля та час [9].

- Політики – адміністратор SIEM створює профіль, який визначає поведінку систем підприємства як за звичайних умов, так і під час попередньо визначених інцидентів безпеки. SIEM надають стандартні правила, сповіщення, звіти та інформаційні панелі, які можна налаштовувати відповідно до конкретних потреб безпеки.
- Консолідація та кореляція даних – рішення SIEM консолідують, аналізують та аналізують файли журналів. Потім події класифікуються на основі необроблених даних і застосовуються правила кореляції, які поєднують окремі події даних у значущі проблеми безпеки.
- Системи AI SIEM використовують прогнозовану аналітику для прогнозування потенційних майбутніх загроз шляхом аналізу архівних даних безпеки та виявлення закономірностей. Ця можливість дозволяє організаціям проактивно захищати свої системи, а не реагувати на загрози, щойно вони виникають. Ця база знань дозволяє моделям штучного інтелекту, які є основою рішення, створювати дедалі точніші реакції безпеки та підходи до запобігання інцидентам із часом і накопиченням нових даних.

Постійне вивчення проблем минулого підвищує точність і надійність систем SIEM на основі штучного інтелекту проти дедалі потужніших кіберзагроз. Зрештою, SIEM на основі штучного інтелекту об’єднує різні компоненти, такі як ШІ, ML, глибоке навчання, NLP і UEBA, які розширюють можливості традиційної SIEM. Ця інтеграція веде до більш розумних,

ефективних і проактивних заходів кібербезпеки, що має вирішальне значення в середовищі кіберзагроз, що постійно змінюється [9].

Сповіщення – якщо подія або набір подій запускає правило SIEM, система сповіщає персонал служби безпеки. У разі виявлення загрози ШІ надає системам SIEM можливість автоматизувати частини процесу реагування на інцидент. Це включає в себе автоматичний запуск сповіщень, впровадження попередньо визначених дій відповіді або організацію складних робочих процесів відповіді. Одним із таких прикладів є автоматизований динамічний робочий процес, де робочий процес, що встановлюється після потенційної загрози, адаптований до відповідної загрози.

Без допомоги SIEM кількість часу, який знадобиться аналітикам безпеки для достовірного виявлення підозрілих дій шляхом кореляції журналів між різними типами пристроїв, буде дуже великим з огляду на складність більшості мереж. Рідко вдається вчасно виявити та відреагувати на будь-яку загрозу чи атаку на їхні інфраструктури, щоб запобігти будь-якій шкоді. Крім того, рішення SIEM може розширити можливості використання зібраної інформації [3].

Очевидно, що ШІ ставатиме все більш важливим у майбутньому SIEM, оскільки когнітивні можливості покращуватимуть здатність системи приймати рішення. Це також дозволить системам адаптуватися та розвиватися зі збільшенням кількості кінцевих точок. Як IoT, хмарні обчислення, мобільні та інші технології збільшують обсяг даних, який повинен споживати інструмент SIEM. ШІ пропонує потенціал для рішення, яке підтримує більше типів даних і комплексне розуміння загроз у їх розвитку.

5. Антивірусні рішення на основі ШІ

Штучний інтелект відіграє ключову роль у сучасних антивірусних рішеннях, таких як, наприклад, Microsoft Defender Advanced Threat Protection (ATP) [11] та Darktrace [12]. Ці системи використовують можливості ШІ для покращення виявлення та реагування на кіберзагрози.

Microsoft Defender ATP [11] є корпоративною платформою безпеки кінцевих точок, розробленою для запобігання, виявлення, дослідження та реагування на загрози. Вона інтегрує сенсори поведінки кінцевих точок, аналітику безпеки в хмарі та розвідку загроз для забезпечення комплексного захисту. Сенсори, вбудовані в Windows 10, збирають та обробляють поведінкові сигнали, які надсилаються до ізольованого хмарного середовища Microsoft Defender для подальшого аналізу.

Основні можливості Microsoft Defender включають [11]:

- Зниження поверхні атаки: мінімізація векторів атак для зменшення можливостей проникнення зловмисників.
- Захист нового покоління: використання передових методів машинного навчання для виявлення складних атак.
- Виявлення та реагування на кінцевих точках (EDR): забезпечення глибокого аналізу та візуалізації загроз у реальному часі.
- Автоматизоване розслідування та реагування: зменшення навантаження на команди безпеки шляхом автоматизації процесів.

Серед переваг рішення Microsoft Defender можна виділити те, що система має глибоку інтеграцію з Windows, Microsoft 365 та Azure, але в той же час це можна розглядати і як недолік, оскільки вона не покриває весь мережевий трафік так ефективно, як, наприклад, Daktrace [12], що буде розглянуто далі. Серед переваг також – автоматичне дослідження загроз та вбудовані

можливості SIEM та SOAR через Microsoft Sentinel. Далі розглянемо рішення Daktrace.

Darktrace [12] – це компанія з кібербезпеки, яка використовує штучний інтелект та машинне навчання для виявлення кібератак та вразливостей у комп'ютерних системах. Її технологія спрямована на швидке та ефективне виявлення загроз, використовуючи поведінковий аналіз для ідентифікації аномалій, що можуть свідчити про атаки.

Darktrace застосовує самонавчальний ШІ для забезпечення проактивного кіберзахисту. Основний принцип роботи Darktrace – це використання машинного навчання для аналізу поведінки користувачів, пристроїв і мережевих потоків. Її технологія здатна:

- Виявляти загрози в реальному часі: аналізуючи поведінку мережі та користувачів для виявлення аномалій. І що не менш важливо – виявляти нові, невідомі загрози, тобто без використання відомих алгоритмів або баз відомих вірусів.
- Автономно реагувати на загрози: здійснювати автоматичні дії для нейтралізації атак без втручання людини.
- Забезпечувати захист у різних середовищах: мережах, електронній пошті, хмарних сервісах, операційних технологіях та кінцевих точках, а також зовнішніх платформ, наприклад, AWS, Azure, Google Cloud.

Таким чином, бачимо, що обидві системи і Microsoft Defender for Endpoint [11], і Darktrace [12] використовують штучний інтелект (ШІ) для покращення кібербезпеки, але вони мають різні підходи та сфери застосування. Розглянемо їх у порівняльному аналізі в таблиці 1.

Таблиця 1 – Порівняльний аналіз сучасних антивірусних рішень на основі ШІ
Table 1 – Comparative analysis of modern AI-based antivirus solutions

Характеристика	Microsoft Defender ATP	Darktrace
Тип рішення	Endpoint Detection and Response (EDR), SIEM/SOAR	Network Detection and Response (NDR), Автономна безпека ШІ
Основний підхід	Використання поведінкового аналізу та баз загроз для захисту кінцевих точок	Самонавчальний ШІ для аналізу аномалій у мережі, електронній пошті та хмарних сервісах
Основні можливості	Виявлення загроз на кінцевих точках, аналіз атак, автоматизоване реагування	Виявлення мережевих загроз, автономне реагування через Darktrace Antigena
Захист від програм-вимагачів	Блокує відомі атаки, аналізує поведінку, зупиняє підозрілі процеси	Виявляє аномальну активність у реальному часі, блокує шкідливі дії на рівні мережі
Фокус на загрози	Віруси, шкідливе ПЗ, експлойти, атакуючі скрипти, компрометація облікових записів	Аномалії в мережевому трафіку, внутрішні загрози, атаки нульового дня
Інтеграція	Глибока інтеграція з екосистемою Microsoft 365, Azure, Defender XDR	Підтримка гібридних середовищ: мережа, електронна пошта, хмара, промислові системи
Хмарна підтримка	Microsoft 365, Azure	AWS, Google Cloud, Azure, локальні мережі

Розглянемо для кого ж найкраще підходить кожне рішення.

Microsoft Defender for Endpoint:

- Великі та середні компанії, які працюють в екосистемі Microsoft.
- Організації, що шукають потужний захист кінцевих точок із вбудованою SIEM-аналітикою.
- Компанії з IT-командою, яка готова керувати безпекою через Microsoft Security Center.

Darktrace краще буде інтегрувати у:

- Підприємства з розгалуженими мережами, що потребують глибокого аналізу трафіку.
- Організації, які хочуть виявляти аномалії в реальному часі та автоматично на них реагувати.
- Компанії, які використовують змішані середовища (локальні сервери, хмарні платформи, IoT).

Ці приклади демонструють, як інтеграція ШІ в антивірусні рішення підвищує ефективність виявлення та реагування на сучасні кіберзагрози, забезпечуючи проактивний захист інформаційних систем.

6. ШІ як інструмент зловмисника – риски та загрози, пов'язані з ШІ

В попередніх розділах було розглянуто переваги використання ШІ в кібербезпеці, але безперечно є і багато викликів щодо його використання, оскільки ШІ не лише відкриває нові можливості для розвитку, але й стає інструментом у руках зловмисників, що створює нові ризики та загрози. Найпоширенішими серед таких загроз у сучасному світі є:

- Атаки на основі ШІ: Кіберзлочинці можуть використовувати ШІ для проведення більш складних та цілеспрямованих атак. Це підвищує ефективність їхніх дій та ускладнює виявлення загроз.
- Дезінформація та глибокі підробки (Deepfake): ШІ дозволяє створювати реалістичні фальшиві відео, аудіо та тексти, що можуть підривати довіру до авторитетних джерел, маніпулювати громадською думкою та втручатися у виборчі процеси.
- Автоматизація кібератак: Зловмисники можуть використовувати ШІ для автоматизації атак, що дозволяє проводити їх швидше та з меншою ймовірністю бути виявленими.
- Атаки на системи ШІ: Зловмисники можуть маніпулювати даними, на яких навчаються системи ШІ, що призводить до неправильних рішень та підвищує вразливість систем.

Зловмисники використовують ШІ для автоматизації та підвищення ефективності атак соціальної інженерії. Наприклад, нейромережі можуть автоматично створювати дуже правдоподібні фішингові повідомлення, які переконують користувачів поділитися своїми паролями або іншою важливою інформацією. Це дозволяє зловмисникам отримати доступ до системи без необхідності проводити технічні атаки, обходячи багато рівнів захисту. Також зловмисники активно використовують ШІ для проведення різноманітних атак, що вимагає від фахівців з безпеки розробки нових стратегій захисту. Тому далі розглянемо основні типи атак на основі ШІ.

6.1. Атаки отруєння даних

Перші атаки отруєння даних (Data Poisoning) [15] були проведені в кібербезпеці ще в 2006 [19] та 2008 [20] роках і з того часу набули популярності серед зловмисників. Дані атаки відбуваються на етапі навчання або донавчання моделі ШІ. Зловмисники вводять шкідливі дані в навчальну базу, що призводить до неправильного функціонування системи та генерації хибних результатів. Це може спричинити помилки в класифікації або прийнятті рішень. Наприклад, мережі GAN можуть створювати штучні дані, які виглядають легітимно, але призначені для введення в оману або псування моделей машинного навчання, що впливає на їх продуктивність і надійність.

Отруєння даними також може посилити існуючі упередження в системах ШІ [16]. Зловмисники можуть націлитися на конкретні підмножини даних, наприклад на певну

демографію, щоб ввести упереджені дані. Через це модель штучного інтелекту може працювати нечесно або неточно. Наприклад, моделі розпізнавання обличь, навчені з упередженими або фальшивими даними, можуть неправильно ідентифікувати людей із певних груп, що призведе до дискримінаційних результатів. Ці типи атак можуть вплинути як на справедливість, так і на точність моделей ML у різних програмах.

Отруєння даних також може відкрити двері для більш складних атак [16], таких як інверсійні атаки, під час яких хакери намагаються реконструювати навчальні дані моделі. Після того, як зловмисник успішно отруює навчальні дані, він може далі використовувати ці вразливості для запуску більш серйозних атак. У системах, розроблених для чутливих завдань, таких як кібербезпека, ці ризики можуть бути особливо небезпечними.

Щоб захиститися від атак з отруєнням даних, організації можуть впроваджувати стратегії, які допоможуть забезпечити цілісність навчальних наборів даних, покращити надійність моделі та постійно контролювати моделі ШІ.

6.2. Атаки ухилення

Атаки ухилення (Evasion Attacks) [18] полягають у створенні спеціальних вхідних даних, які вводять модель ШІ в оману, змушуючи її робити неправильні прогнози або класифікації. Такі атаки можуть бути спрямовані на обхід систем виявлення шкідливого програмного забезпечення або інших захисних механізмів.

Відкриття атак ухилення від моделей машинного навчання викликало підвищений інтерес до змагального машинного навчання, що призвело до значного зростання цього дослідницького простору за останнє десятиліття. Під час атаки з ухиленням метою зловмисника є створення змагальних прикладів, які визначаються як тестові зразки [21].

У програмах кібербезпеки змагальні приклади повинні поважати обмеження, накладені семантикою програми та представленням функцій кіберданих, таких як мережевий трафік або бінарні файли програми [18].

FENCE – це загальна структура для створення атак ухилення білої скриньки з використанням градієнтної оптимізації в дискретних областях і підтримує низку лінійних і статистичних залежностей характеристик [22]. FENCE було застосовано до двох програм безпеки мережі: виявлення зловмисного домену та класифікація шкідливого мережевого трафіку. В [23] цю техніку було застосовано для виявлення мережевих вторгнень і класифікаторів фішингу. В документі [23] зазначено, що атаки з безперервних доменів не можуть бути легко застосовані в обмежених середовищах, оскільки вони призводять до нездійсненних прикладів змагання. Пієрацці та ін. [24] обговорюють труднощі встановлення можливих атак ухилення в кібербезпеці через обмеження в просторі функцій та формалізують атаки ухилення в проблемному просторі та створюють можливі змагальні приклади для шкідливих програм Android.

6.3. Атаки швидкого впровадження

Під час атаки швидкого впровадження (Prompt Injection) [18] зловмисники вставляють шкідливі інструкції в запити до моделей ШІ, змушуючи їх виконувати небажані дії або надавати конфіденційну інформацію, тобто обманюють модель для повернення неочікуваної відповіді та змушуючи додаток діяти незапланованими способами. Успішне впровадження може призвести до витоку конфіденційних даних, знищення інформації та інших типів шкоди залежно від застосунку.

6.4. Атаки соціальної інженерії з використанням ШІ

ШІ використовується для автоматизації та підвищення ефективності атак соціальної інженерії, таких як фішинг або маніпуляції, що ускладнює їх виявлення.

Зловмисники використовують методи соціальної інженерії для приховування своєї справжньої «особистості», представляючись жертвам надійними організаціями або особами. Ці атаки спрямовані на отримання особистої інформації для доступу до цільової мережі шляхом обману та маніпуляцій. Соціальна інженерія використовується як перший етап великої кібератаки для проникнення в систему, установки шкідливого ПЗ, розкриття конфіденційних даних тощо.

Наприклад, у випадку з фішингом [18], раніше було продемонстровано, що великі мовні моделі (LLM) можуть створювати переконливі шахрайства, такі як фішингові електронні листи [25]. Тепер, коли LLM можуть легше інтегруватися з додатками, вони можуть не тільки створювати шахрайські дії, але й широко поширювати такі атаки [26]. Користувачі, швидше за все, будуть більш сприйнятливі до цих нових атак, на відміну від фішингових електронних листів, оскільки їм бракує досвіду та обізнаності про цю нову техніку загроз.

Також сама LLM діє як комп’ютер, на якому працює та поширюється шкідливий код. Наприклад, інструмент автоматичної обробки повідомлень, який може читати та створювати електронні листи та переглядати особисті дані користувачів, може поширити ін’єкцію на інші моделі, які можуть читати ці вхідні повідомлення [27].

Тож, розглянемо заходи протидії атакам на основі ШІ. Щоб убезпечитися від атак з використанням ШІ необхідно зробити наступні дії:

- Покращення якості даних: Забезпечення чистоти та надійності даних для навчання моделей ШІ допомагає зменшити ризик отруєння даних.
- Розробка стійких моделей: Створення моделей ШІ, стійких до атак, шляхом впровадження методів захисту та регулярного тестування.
- Моніторинг та аудит: Постійний нагляд за роботою моделей ШІ та проведення аудитів для виявлення можливих вразливостей.
- Навчання персоналу: Підвищення обізнаності працівників щодо можливих атак на основі ШІ та методів їх виявлення.

У сучасному цифровому середовищі важливо бути готовим до нових викликів, пов'язаних з використанням ШІ, та впроваджувати відповідні заходи для забезпечення кібербезпеки.

7. Висновки

1. У цій статті було проведено всебічний аналіз поточного стану та перспектив застосування штучного інтелекту (ШІ) у сфері кібербезпеки. Розглянуто як переваги впровадження ШІ в системи захисту, так і ризики, пов’язані з його використанням.

2. ШІ дозволяє автоматизувати процеси виявлення та реагування на загрози, що значно підвищує ефективність кіберзахисту. Використання алгоритмів машинного навчання сприяє швидкому аналізу великих обсягів даних та ідентифікації аномалій у поведінці користувачів і систем.

3. Сучасні системи кібербезпеки, такі як SIEM (Security Information and Event Management), значно виграють від інтеграції ШІ, оскільки здатні аналізувати події в реальному часі та попереджати про можливі атаки. Постійне вивчення проблем минулого підвищує точність і надійність систем SIEM на основі штучного інтелекту проти дедалі потужніших кіберзагроз. Зрештою, SIEM на основі штучного інтелекту об’єднує різні компоненти, такі як

AI, ML, глибоке навчання, NLP і UEBA. Ця інтеграція веде до більш розумних, ефективних і проактивних заходів кібербезпеки, що має вирішальне значення в середовищі кіберзагроз, що постійно змінюється. При цьому, велика кількість інтеграцій з різноманітними системами, дозволяє системам SIEM стежити та накопичувати дані щодо поточного стану сфери кіберзахисту інформаційної інфраструктури щодо певних міжнародних та національних стандартів, таких як ISO 27001, GDPR чи PCI DSS.

4. ШІ може аналізувати величезні обсяги даних і виявляти закономірності, що вказують на потенційні загрози, дозволяючи організаціям випереджати хакерів. А також, ШІ здатний швидше і точніше ідентифікувати загрози, а також автоматично блокувати шкідливий трафік без втручання людини. Завдяки поведінковому аналізу, ШІ-антивіруси (наприклад, Microsoft Defender ATP та Darktrace) можуть ефективніше виявляти загрози, ніж традиційні антивірусні програми.

5. Щодо використання систем захисту виявлення атак, то якщо компанія активно використовує Microsoft 365, Azure та Windows – краще обрати Microsoft Defender for Endpoint. Він забезпечить глибоку інтеграцію, автоматизоване реагування та поведінковий аналіз загроз на кінцевих пристроях. Якщо потрібен гнучкий та автономний захист усіх рівнів мережі – варто розглянути Darktrace. Він підійде організаціям, які хочуть виявляти аномалії, аналізувати кіберзагрози в реальному часі та реагувати на них без втручання людей. Але ідеальний варіант – поєднання обох рішень: Microsoft Defender for Endpoint для захисту кінцевих точок, а Darktrace – для мережевого аналізу та автоматичного реагування на загрози.

6. Використання ШІ не лише для захисту, а й для атак становить значну загрозу. Зловмисники можуть застосовувати ШІ для автоматизації атак, маніпуляцій та соціальної інженерії. Зловмисники все частіше використовують ШІ для створення складного шкідливого програмного забезпечення, яке може адаптуватися до захисних заходів. Експерти з безпеки занепокоєні потенційними автономними атаками ШІ, що змушує компанії готуватися вже зараз. Організаціям потрібно впроваджувати комплексні стратегії, щоб скористатися перевагами ШІ, одночасно мінімізуючи його потенційні загрози.

7. Перспективами розвитку та рекомендації щодо використання ШІ в кібербезпеці є:

- Підвищення прозорості алгоритмів ШІ та впровадження етичних стандартів є критично важливими для довіри до технологій ШІ у сфері безпеки.
- Розробка стійких моделей ШІ, які будуть менш вразливими до атак з боку зловмисників.
- Інвестування в дослідження постквантової криптографії та новітніх методів аутентифікації для зміцнення систем кібербезпеки.
- Посилення міжнародного співробітництва у сфері стандартів та регулювання ШІ, зокрема завдяки ініціативам таких організацій, як NIST.

8. Таким чином, штучний інтелект має потенціал повністю змінити підхід до кібербезпеки. Його здатність швидко аналізувати великі масиви даних, передбачати загрози та автоматично реагувати на атаки робить його ключовим елементом захисту у цифровому світі. Проте, слід пам’ятати, що кіберзлочинці також використовують ШІ, тому майбутнє кібербезпеки залежатиме від балансу між інноваціями у захисті та загрозами з боку злочинних угруповань.

9. Майбутнє кібербезпеки – це симбіоз людини та штучного інтелекту, де аналітики та експерти з безпеки співпрацюють із розумними системами для створення надійного кіберпростору.

Конфлікт інтересів

Автори повідомляють про відсутність конфлікту інтересів.

References

1. NIST standardization process “Post-Quantum Cryptography: Digital Signature Schemes”. URL: https://csrc.nist.gov/Projects/pqc-dig-sig/round-1-additional-signatures
2. TAO, Feng; Akhtar, Muhammad Shoaib; Jiayuan, Zhang. (2021) The future of artificial intelligence in cybersecurity: A comprehensive survey. *EAI Endorsed Transactions on Creative Technologies*, 8.28: e3-e3. DOI: https://doi.org/10.4108/eai.7-7-2021.170285
3. Leung, B.K. (2021). Security Information and Event Management (SIEM) Evaluation Report. ScholarWorks. URL: https://scholarworks.calstate.edu/downloads/41687p49q
4. González-Granadillo, G., González-Zarzosa, S., Diaz, R. (2021). Security Information and Event Management (SIEM): Analysis, Trends, and Usage in Critical Infrastructures. *Sensors*, 21(14). DOI: https://doi.org/10.3390/s21144759
5. Muhammad, S., et al. (2023). Effective Security Monitoring Using Efficient SIEM Architecture. *Human-centric Computing and Information Sciences*, 13. DOI https://doi.org/10.22967/HCIS.2023.13.017
6. What is SIEM. Security Information and Event Management Tools. (n.d.). Imperva. URL: https://www.imperva.com/learn/application-security/siem/
7. IBM Security QRadar. What is security information and event management (SIEM)? URL https://www.ibm.com/think/topics/siem
8. Splunk. The Splunk SIEM. URL: https://www.splunk.com/en_us/products/enterprise-security.html
9. Stellar Cyber. AI SIEM: The 6 Components of AI-Based SIEM. URL: https://stellarcyber.ai/learn/ai-driven-siem/
10. ISO/IEC 27001: (2022). Information technology – Security techniques – Information security management systems – Requirements. International standard. 3 Edition.
11. Microsoft Defender for Endpoint. (2024). URL: https://learn.microsoft.com/uk-ua/defender-endpoint/microsoft-defender-endpoint
12. Darktrace. Official website (2024). URL: https://darktrace.com/
13. Mauri L., Damiani E. Modeling Threats to AI-ML Systems Using STRIDE. Sensors (2022), (17), 6662; DOI: https://doi.org/10.3390/s22176662
14. The near-term impact of AI on the cyber threat. URL: https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat
15. Nihad Hassan. What is data poisoning (AI poisoning) and how does it work? Search Enterprise AI, TechTarget, (2024). URL: https://www.techtarget.com/searchenterpriseai/definition/data-poisoning-AI-poisoning
16. Tom Krantz, Alexandra Jonker. What is data poisoning? IBM. URL: https://www.ibm.com/think/topics/data-poisoning
17. NIST Trustworthy and Responsible AI NIST AI 100-5. A Plan for Global Engagement on AI Standards. DOI: https://doi.org/10.6028/NIST.AI.100-5
18. Vassilev A, Oprea A, Fordyce A, Anderson H (2024) Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. (National Institute of Standards and Technology, Gaithersburg, MD) *NIST Artificial Intelligence (AI) Report, NIST Trustworthy and Responsible AI NIST AI* 100-2e2023. DOI: https://doi.org/10.6028/NIST.AI.100-2e2023
19. R. Perdisci, D. Dagon, Wenke Lee, P. Fogla, and M. Sharif. (2006) Misleading worm signature generators using deliberate noise injection. *In 2006 IEEE Symposium on Security and Privacy* (S&P’06), Berkeley/Oakland, CA, IEEE. DOI: https://doi.org/10.1109/SP.2006.26
20. Blaine Nelson, Marco Barreno, Fuching Jack Chi, Anthony D. Joseph, Benjamin I.P. Rubinstein, Udam Saini, Charles Sutton, and Kai Xia. (2008) Exploiting machine learning to subvert your spam filter. *In First USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET 08)*, San Francisco, CA, USENIX Association. URL: https://people.eecs.berkeley.edu/~tygar/papers/SML/Spam_filter.pdf
21. Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. (2014) Intriguing properties of neural networks. *In International Conference on Learning Representations.* DOI: https://doi.org/10.48550/arXiv.1312.6199

22. Alesia Chernikova and Alina Oprea. (2022) FENCE: Feasible evasion attacks on neural networks in constrained environments. *ACM Transactions on Privacy and Security (TOPS) Journal.* DOI: https://doi.org/10.48550/arXiv.1909.10480
23. Ryan Sheatsley, Blaine Hoak, Eric Pauley, Yohan Beugin, Michael J. Weisman, and Patrick McDaniel (2021). On the robustness of domain constraints. *In Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, CCS '21*, page 495–515, New York, NY, USA. Association for Computing Machinery. DOI: https://doi.org/10.48550/arXiv.2105.08619
24. Fabio Pierazzi, Feargus Pendlebury, Jacopo Cortellazzi, and Lorenzo Cavallaro (2020). Intriguing properties of adversarial ML attacks in the problem space. *In 2020 IEEE Symposium on Security and Privacy (S&P)*, pages 1308–1325. IEEE Computer Society. DOI: https://doi.org/10.48550/arXiv.1911.02142
25. Daniel Kang, Xuechen Li, Ion Stoica, Carlos Guestrin, Matei Zaharia, and Tatsunori Hashimoto (2023). Exploiting Programmatic Behavior of LLMs: Dual-use through standard security attacks. DOI: https://doi.org/10.48550/arXiv.2302.05733
26. Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz (2023). Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. DOI: https://doi.org/10.48550/arXiv.2302.12173

RESEARCH ON THE CURRENT STATE AND PROSPECTS OF THE APPLICATION OF ARTIFICIAL INTELLIGENCE IN CYBERSECURITY

Yuriy Golikov[1], e-mail: yuriy@devbrother.com;
Yelyzaveta Ostrianska[2], junior researcher; e-mail: antelizza@gmail.com;
ORCID: https://orcid.org/0000-0003-1412-8470
[1]*DevBrother tech company, USA*
[2]*V. N. Karazin Kharkiv National University, Ukraine*
Manuscript was received November 1, 2024; Received after review December 2, 2024;
Accepted December 23, 2024

Abstract. In the modern world, with the development of new technologies, artificial intelligence (AI) in cybersecurity has become an integral component. Therefore, studying its advantages, risks, and potential use cases is a highly relevant research topic. In today's digital environment, where cyber threats are becoming increasingly sophisticated, the implementation of AI technologies significantly enhances the effectiveness of security systems by enabling automated threat detection and response. In this study the main applications of AI in cybersecurity were examined, including threat detection, malware analysis, cryptographic security enhancement, phishing protection, and attack prediction. One of the key aspects is the integration of AI into Security Information and Event Management (SIEM) systems, which analyze vast amounts of data and help detect anomalies. Such systems reduce the workload on security teams and improve the accuracy and speed of threat response. Special attention is given to the analysis of modern AI-powered antivirus solutions, particularly Microsoft Defender for Endpoint and Darktrace. These solutions are based on behavioral analysis algorithms and machine learning, allowing for more effective detection of complex threats and incident prevention. Microsoft Defender provide a high level of endpoint protection. Meanwhile, Darktrace utilizes self-learning models to analyze network traffic, enabling the detection of zero-day threats and internal risks within organizations. The study also learns the major risks associated with the use of AI in cybercrime. AI is increasingly leveraged by malicious actors to automate attacks, significantly increasing their effectiveness and making detection more challenging. The primary

AI-based cyber threats discussed include Data Poisoning attacks, Evasion Attacks, Prompt Injection Attacks, and AI-based social engineering. To mitigate these risks, the development of robust AI models resistant to adversarial attacks, increased algorithm transparency, and the implementation of international AI regulation standards is recommended, including NIST. Additionally, raising awareness among users and cybersecurity specialists is crucial, as the human factor remains one of the most significant vulnerabilities in security systems. In conclusion, it is said that AI is a key factor in the advancement of cybersecurity, offering significant improvements in protecting information and critical systems. However, without proper regulation and protective measures, AI can become a powerful tool for cybercriminals, posing new security challenges in the digital age. Striking a balance between innovation, ethical standards, and security will be essential in shaping the future strategy for the effective use of AI.

Keywords: *artificial intelligence, cybersecurity, machine learning, SIEM, IDS, AI-based antivirus*

Conflicts of Interest: the authors declare no conflict of interest.