КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 1 (29), 2025
ISSN 2663 - 4023
DOI 10.28925/2663-4023.2025.29.840
УДК 004.056

Яковлев Максиміліан
здобувач ступеня доктора філософії
Сумський державний університет, Суми, Україна
ORCID ID: 0000-0002-1196-4062
maksimilianyakovlev@gmail.com

Любчак Володимир
кандидат фізико-математичних наук, доцент, завідувач кафедри кібербезпеки
Сумський державний університет, Суми, Україна
ORCID ID: 0000-0002-7335-6716
v.liubchak@dcs.sumdu.edu.ua

МОЖЛИВОСТІ ШТУЧНОГО ІНТЕЛЕКТУ У ВИЯВЛЕННІ ТА ЗАПОБІГАННІ ФІШИНГУ Й КІБЕРАТАКАМ

Анотація. У статті досліджується подвійна роль штучного інтелекту (ШІ) у сучасному середовищі кібербезпеки, що проявляється як у посиленні кібератак, так і в розробці ефективних механізмів захисту інформаційних систем. Проаналізовано зростаючу складність загроз, зумовлених розвитком технологій машинного навчання, обробки природної мови та генеративного ШІ (GenAI), які дозволяють зловмисникам автоматизувати, підвищувати точність та маскування атак, включаючи фішинг, створення шкідливого програмного забезпечення та використання діпфейків. У цьому дослідженні розглядаються як рішення на основі ШІ можуть покращити кібербезпеку, виявляючи можливі загрози реального часу за допомогою методів машинного навчання, обробки природної мови та розпізнавання образів. Особливу увагу приділено необхідності інтеграції ШІ з контролем з боку людини, підкреслюючи важливість поєднання автоматизованих відповідей з експертним аналізом для ефективного зниження ризиків та адаптації до нових викликів. Розглянуто низку сучасних інструментів та підходів, що застосовуються для здійснення кібератак за допомогою ШІ, зокрема ChatGPT, WormGPT, FraudGPT та Morris 2.0, демонструючи їхні можливості у створенні переконливих шахрайських сценаріїв та адаптивного шкідливого ПЗ. Паралельно вивчено інструментально-технічні рішення на основі ШІ, призначені для виявлення та протидії кіберзагрозам у реальному часі. Описано функціональні можливості таких систем, як Darktrace Antigena, Cylance Endpoint Security, Splunk, Exabeam, IBM QRadar, Microsoft Sentinel, які використовують поведінковий аналіз, машинне навчання та розпізнавання образів для проактивного виявлення аномалій, автоматизованого реагування на інциденти та підвищення загального рівня безпеки. Дослідження демонструє, як ШІ трансформує кібербезпеку, пропонуючи адаптивну та проактивну стратегію для боротьби з виникаючими кіберзагрозами. Вивчаючи сучасне використання ШІ це дослідження демонструє, як він може трансформувати кібербезпеку, пропонуючи проактивну та адаптивну стратегію для боротьби з кібератаками та захисту конфіденційних даних.

Ключові слова: шкідливе програмне забезпечення; генеративний штучний інтелект; кіберзагрози; фішингові атаки; автоматизоване реагування.

ВСТУП

Розробка базової системи захисту інформації за допомогою штучного інтелекту ведеться багатьма впливовими міжнародними організаціями, урядами країн та приватними компаніями. ЄС розробляє всеосяжну законодавчу базу з широким, міжсекторальним охопленням, включаючи такі регламенти, як Загальний регламент про захист даних, Закон про цифрові послуги, Закон про цифрові ринки та Акт про ШІ, з

метою створити послідовне регуляторне середовище для всіх держав-членів. В рамках Акту про ШІ ведеться інвестування у проекти, які використовують ШІ для виявлення та запобігання кібератакам [14].

Відомства застосовують надійні заходи, такі як шифрування і передові технології, а також співпрацюють з міжнародними партнерами для протидії новим загрозам, забезпечуючи захист національних інтересів і збереження суспільної довіри. Для покриття широкого спектру можливих вразливостей, найчастіше, використовується генеративний штучний інтелект.

Постановка проблеми. Подвійна природа GenAI в кібербезпеці включає в себе як корисні, так і потенційно шкідливі програми, що викликає значні занепокоєння. З розвитком технологій штучного інтелекту в злочинців з'являється все більше можливостей завдання шкоди інформаційно-комунікаційним системам. Саме тому необхідно знайти дієві способи боротьби з такими типами атак.

Генеративный ШІ революціонізує кібербезпеку, пропонуючи передові рішення для виявлення, запобігання та реагування на загрози. Використовуючи такі моделі, як GAN, GenAI імітує кібератаки, щоб покращити системи виявлення загроз, як навчають в Інституті глибокого навчання Nvidia. При виявленні фішингу та шахрайства GenAI генерує реалістичні сценарії фішингу, значно покращуючи показники виявлення, як показало дослідження [1] Плімутського університету.

Генеративний ШІ здатний створювати дуже переконливі фішингові електронні листи, генерувати адаптивне шкідливе програмне забезпечення, автоматизувати створення варіантів шкідливого програмного забезпечення та виробляти глибокі підробки для кампаній шпигунського фішингу, створюючи тим самим серйозні загрози інформаційній безпеці. Він також може генерувати реалістичні фейкові новини, що підкреслює його потенціал для зловживань у кампаніях з дезінформації [2].

Необхідно дослідити, які технології та інструменти застосовуються для кібератак, а також які існують засоби для їх виявлення та протидії.

Аналіз останніх досліджень і публікацій. Дослідження впливу штучного інтелекту на глобальну світову безпеку розкрито в роботі [3], потенційні катастрофічні наслідки які можуть виникнути внаслідок кібератак на критично важливі вузли державних та всесвітньо відомих світових компаній.

Деякі дослідження зосереджені на інструментах GenAI та їхній продуктивності. Наприклад, Браун та ін. розширили можливості Обробки Природньої Мови — обробки шляхом навчання GPT-3, авторегресійної мовної моделі з 175 мільярдами параметрів, що свідчить про виняткову здатність до навчання з кількох спроб [4] .В даному дослідженні виявлено, що без спеціального навчання ця модель добре справляється з різними завданнями обробки природної мови, такими як переклад і відповіді на запитання.

Нещодавнє дослідження [5] представляє новий підхід до оцінки потенційно серйозних небезпек, пов'язаних з моделями GenAI, таких як обман, маніпуляції та функції кіберзлочинів Щоб розробники ШІ могли приймати обґрунтовані рішення щодо навчання, розгортання і застосування стандартів кібербезпеки, запропонована методологія підкреслює необхідність збільшення критеріїв оцінки для точної оцінки шкідливих можливостей і узгодження систем ШІ.

Інша робота [6] містить ретельний аналіз, який надихнув на комплексне застосування ChatGPT у цифрових судових розслідуваннях, вказуючи на важливі обмеження і перспективні можливості, які притаманні GenAI в його нинішньому вигляді. Використовуючи методичні експерименти, вони окреслюють тонку грань, що відокремлює винахідницький внесок ШІ від життєво важливої вимоги професійного

людського нагляду в процедурах кібербезпеки, відкриваючи двері для додаткових досліджень в галузі інтеграції великих мовних моделей (ВММ), таких як GPT-4, в цифрову криміналістику і кібербезпеку.

Мета статті. Метою цієї статті є огляд досліджень та інструментів, що використовуються для здійснення кібератак, а також засобів захисту від них, зосереджуючи увагу на ролі штучного інтелекту як у створенні загроз, так і в розробці ефективних механізмів кібербезпеки.

СУЧАСНІ ПІДХОДИ ДО ЗДАЙСНАННЯ КІБЕРАТАК З ВИКОРИСТАННЯМ ШТУЧНОГО ІНТЕЛЕКТУ

Кіберзлочинці дедалі активніше використовують штучний інтелект для масштабування атак і отримання фінансової вигоди. Серед ключових інструментів — генерація шкідливого програмного забезпечення, автоматизовані фішингові кампанії та самонавчальні ШІ-хробаки.

У 2022 році шахраї спробували зв'язатися з одним із засновників компанії Baykar defence, яка виробляє знамениті безпілотники Байрактар, Халюком Байрактаром. Розмова велась від імені українського прем'єр-міністра Дениса Шмигаля, відеозображення якого, власне, було створено за допомогою технології deepfake [7]. Метою цієї акції була спроба дискредитація співробітництва між Україною та Туреччиною. Людина не завжди може визначити чи це є згенероване відео чи реальне. Технологія deepfake стрімко розвивається, тому надалі буде все важче відрізнити реальні відео від них неозброєним оком.

З розвитком генеративних моделей зловмисники пристосували ці технології для своїх потреб. За останні роки з'явилися неконтрольовані AI-чат-боти та навіть спеціалізована велика мовна модель створена для підтримки кіберзлочинної діяльності. Такі інструменти відкривають доступ до інформації, яку традиційні ШІ-системи зазвичай блокують з міркувань етики та безпеки.

Поєднання подібних технологій з відкритими персональними даними та хакерськими ресурсами з даркнету значно спрощує вхід до сфери кіберзлочинності. Це дозволяє новачкам швидко почати діяльність, а досвідченим хакерам — удосконалювати свої методи.

ChatGPT. Здатність ChatGPT генерувати відповіді, подібні до людських, і розуміти контекст зробила його популярним інструментом для створення розмовних агентів, контенту, аналізу даних, а також для досліджень та інновацій. Однак його ефективність і легкодоступність також роблять його привабливою ціллю для створення шкідливого контенту, зокрема фішингових атак, які можуть наражати користувачв на небезпеку.

Бегу та інші [8] досліджують роль ChatGPT у розвитку фішингових атак, оцінюючи його здатність автоматизувати розробку складних фішингових кампаній. У дослідженні вивчається, як ChatGPT може генерувати різні компоненти фішингової атаки, включаючи клонування веб-сайтів, інтеграцію коду для крадіжки облікових даних, заплутування коду, автоматизоване розгортання, реєстрацію доменів та інтеграцію зворотного проксі-сервера. Автори пропонують модель загроз, яка використовує ChatGPT, оснащений базовими навичками Python і доступом до моделей OpenAI Codex, для спрощення розгортання фішингової інфраструктури. Вони демонструють потенціал ChatGPT для прискорення операцій зловмисників і наводять приклад фішингового сайту, що імітує LinkedIn.

У роботі [9] дослідники виявили кілька шкідливих запитів, які можна використати в ChatGPT для генерації функціональних фішингових вебсайтів. Використовуючи ітеративний підхід, вони з'ясували, що ці сайти можуть імітувати відомі бренди та

відтворювати кілька тактик ухилення, які вже відомі своєю здатністю обходити виявлення антифішинговими системами. Атаки були згенеровані за допомогою звичайного ChatGPT без необхідності у попередньому зломії або використанні шкідливих атакувальних програм (експлойтів).

WormGPT. WormGPT забезпечує миттєві відповіді та не обмежує довжину повідомлень, що дозволяє користувачам безперешкодно взаємодіяти з чатботом. Серед можливостей WormGPT:

- Злам облікових записів у соцмережах — підтримуються атаки на Facebook, Instagram, TikTok, Telegram, Viber, WhatsApp та інші сервіси.
- Створення шкідливих програм — здатний генерувати віруси, кейлогери та інше шкідливе ПЗ.
- Компрометація пристроїв — дозволяє атакувати ПК, ноутбуки, смартфони (Android, iOS), пристрої IoT та камери спостереження.
- Фішинг і соціальна інженерія — можна організовувати спам-розсилки, писати фішингові листи та створювати маніпулятивні схеми.
- Злам вебресурсів і DDoS-атаки — використовується для атак на сайти, додатки та для проведення DDoS.

Користувачі можуть обирати серед різних AI-моделей для загального чи спеціального призначення, а також зберігати та переглядати історію спілкування у будь-який момент. Додатково у бета-версії WormGPT доступні розширені функції, такі як «контекстна пам’ять» для підтримки тривалих діалогів та «форматування коду» для зручної структурування і використання шкідливих скриптів [15].

FraudGPT. FraudGPT — прорив у сфері зловмисних кібератак: FraudGPT — це новий генеративний ШІ-інструмент на основі підписки, створений для того, щоб вийти за межі передбаченого використання технологій і обійти обмеження, відкриваючи шлях до створення надзвичайно переконливих фішингових листів та шахрайських вебсайтів. Цей інструмент було виявлено командою дослідників загроз компанії Netenrich у липні 2023 року в каналах Telegram даркнету [16]. FraudGPT являє собою зсув парадигми у техніках проведення атак і потенційно демократизує використання генеративного ШІ як зброї у широкому масштабі.

Його розроблено для автоматизації широкого спектру завдань: від написання шкідливого коду та створення невиявного шкідливого ПЗ до складання переконливих фішингових повідомлень. FraudGPT дає змогу навіть недосвідченим зловмисникам запускати атаки. Цей інструмент поєднує перевірені інструменти атак, включаючи індивідуальні інструкції з хакерства, пошук вразливостей і використання експлойтів нульового дня — без потреби у високих технічних знаннях.

FraudGPT — це тонко налаштована велика мовна модель типу GPT-3.5 або GPT-4, спеціально адаптована для:

- генерування шкідливого контенту (PowerShell, Ptyhon, JavaScript);
- обходу етичних обмежень звичайних великих мовних моделей;
- допомоги в кібершахрайстві.

В даному інструменті використовується модель із відкритим кодом LLaMA або GPT-J, натренована на витоках з форумів, інструкціях по написанню шкідливого коду та коді відкритого доступу з GitHub/ExploitDB.

Завдяки своїй здатності позбавляти будь-яких засобів захисту чи етичних бар’єрів, ця технологія може бути використана для різноманітних зловмисних цілей. Можливості FraudGPT охоплюють створення фішингових листів і матеріалів соціальної інженерії,

розробку експлойтів, шкідливого ПЗ та інструментів для тестування на проникнення або зловмисного впливу на системи, пошук вразливостей, скомпрометованих облікових даних і вразливих сайтів, а також надання інструкцій з технік вторгнення в системи та кіберзлочинної діяльності [16].

Morris 2.0. У березні група дослідників представила Morris 2.0 — перший у світі самореплікативний вірус, створений на основі штучного інтелекту. Цей програмний агент здатен автономно проникати в електронні поштові системи, аналізувати вміст повідомлень та поширювати шкідливе програмне забезпечення без необхідності взаємодії з користувачем. Morris 2.0 є продовженням або оновленням попередніх великих мовних моделей і зосереджена на високій ефективності, мультимодальності та точному слідуванні інструкціям. Технічні характеристики моделі свідчать про її орієнтацію на реальні застосування — як у хмарі, так і на пристроях [17].

Проєкт отримав назву на честь оригінального хробака Morris, розробленого студентом Корнельського університету Робертом Т. Моррісом у 1988 році. На відміну від класичних комп'ютерних вірусів, які діють за жорстко закладеними алгоритмами, Morris 2.0 використовує генеративні можливості ШІ для адаптації до цільового середовища. Основним каналом розповсюдження виступають скомпрометовані ШІ-системи, які забезпечують сприятливі умови для ініціації й ескалації атаки. Завдяки цьому даний інструмент здатен проникати в мережі з високою ефективністю.

Після інфікування системи Morris 2.0 застосовує засоби штучного інтелекту для структурного аналізу архітектури, виявлення вразливостей та формування унікальних векторів атаки. Такий адаптивний підхід значно ускладнює виявлення загрози й робить її стійкою до традиційних методів захисту.

Архітектура моделі Morris 2.0 грунтується на нейронній мережі, що використовує механізм уваги (self-attention) для ефективного опрацювання послідовностей даних. У моделі реалізовано методи розрідженого розподілу уваги або поєднання спеціалізованих обчислювальних блоків, що дає змогу зменшити обчислювальні витрати без втрати точності. Для стабільної роботи з довгими текстовими входами застосовується обертальне позиційне кодування. В обчислювальних шарах використовуються вдосконалені функції активації, які сприяють кращому навчанню та узагальненню. Крім того, модель здатна обробляти як текстову, так і візуальну інформацію, що забезпечується поєднанням кодувальних компонентів для зображень із текстовим представленням у спільному просторі ознак. Така архітектура забезпечує високу гнучкість і пристосованість моделі до широкого спектра прикладних завдань.

Попри те, що Morris 2.0 було створено виключно в контрольованому середовищі з дослідницькою метою, потенційна загроза його поширення є предметом серйозного занепокоєння. Демонстраційні експерименти проводились із використанням популярних генеративних моделей, зокрема ChatGPT та Gemini, що підкреслює актуальність викликів, пов'язаних із безпекою ШІ-технологій.

ІНСТРУМЕНТИ ТА ТЕХНОЛОГІЇ З ВИКОРИСТАННЯМ ШТУЧНОГО ІНТЕЛЕКТУ ДЛЯ ВИЯВЛЕННЯ ТА ПРОТИДІЇ КІБЕРЗАГРОЗАМ

Більшість компаній, що працюють у сфері кібербезпеки, схиляються до використання аналогічних технологій для боротьби з кіберзагрозами — зокрема, залучення моделей штучного інтелекту. Деякі дослідницькі групи, які представляють

спільноту так званих «етичних» хакерів, розробляють контрольовані зловмисні ШІ-моделі для вивчення їхнього потенціалу та ризиків.

Зокрема, компанія Mithril Security створила інструмент під назвою **PoisonGPT**, що слугує експериментальним засобом для аналізу здатності генеративного ШІ поширювати дезінформацію та фейкові новини в межах масштабних інформаційних кампаній. У майбутньому компанія планує запустити AICert — відкриту платформу, призначену для створення криптографічно захищених ідентифікаційних карток моделей ШІ. Це рішення дозволятиме встановлювати зв'язок між конкретною моделлю, її кодом і навчальним набором даних, аналогічно до принципів функціонування захищеного апаратного забезпечення.

Іншим прикладом є проєкт **HackingBuddyGPT**, спрямований на підтримку дослідників у сфері безпеки у вивченні потенційних векторів атак за допомогою великих мовних моделей [18]. Інструмент дає змогу здійснювати контрольовані симуляції атак і водночас аналізувати специфіку генерації шкідливих шаблонів LLM у порівнянні з діями людських операторів. Усі зібрані в межах проєкту дані є публічно доступними, що сприяє відкритості досліджень і дозволяє фахівцям з кіберзахисту адаптувати або вдосконалювати власні захисні стратегії на основі отриманих результатів [13].

Розширене виявлення загроз і розпізнавання аномалій. Системи штучного інтелекту докорінно змінюють сферу кібербезпеки, застосовуючи машинне навчання для виявлення шкідливого ПЗ, програм-вимагачів та інших загроз, а також для ідентифікації нетипової мережевої активності, що може свідчити про порушення безпеки. Наприклад, система **Antigena** від Darktrace використовує навчання без учителя для виявлення аномальної мережевої поведінки, такої як діяльність ботнетів або ексфільтрація даних у режимі реального часу. Інструменти на зразок **Cylance** та **CrowdStrike Falcon** застосовують глибинне навчання для розпізнавання нових загроз шкідливого ПЗ [10].

Cylance Endpoint Security — це уніфіковане рішення для захисту кінцевих точок, розроблене з урахуванням нових реалій. Воно об'єднує найкращі доступні інструменти на основі штучного інтелекту для виявлення, захисту та усунення загроз на кожній кінцевій точці. Сучасні кіберзлочинці використовують штучний інтелект (ШІ) для створення все більш досконалих загроз, що максимізують охоплення та вплив їхніх атак. Сучасні рішення також повинні використовувати можливості машинного навчання та ШІ. Cylance Endpoint Security пропонує рішення на основі штучного інтелекту для Zero Trust для всього спектру пристроїв, мереж, додатків та людей. Підхід Zero Trust модернізує мережеву безпеку, одночасно покращуючи та вдосконалюючи мережевий досвід для кінцевих користувачів. Модель безпеки Zero Trust за замовчуванням не довіряє нічому і нікому, включаючи користувачів всередині робочої мережі. Кожен користувач, кінцева точка та мережа вважаються потенційно ворожими. У системі безпеки Zero Trust жоден користувач не може отримати доступ до чого-небудь, поки не доведе, хто він є, що його доступ авторизований, що мережа, до якої він підключений, не скомпрометована, і що він сам або шкідливе програмне забезпечення, яке ховається на його пристрої, не діють зловмисно [19].

Платформи, як-от **Splunk**, аналізують журнали безпеки для виявлення аномалій, наприклад, невдалих спроб входу в систему, тоді як **Exabeam** і **SentinelOne** формують базові профілі поведінки користувачів з метою виявлення нетипових дій, що можуть свідчити відповідно про атаки програм-вимагачів або загрози типу APT (Advanced Persistent Threats).

Сервіс **Recorded Future** збирає дані про загрози для їхнього проактивного виявлення. Проте надмірна залежність від ШІ може призводити до ігнорування тонких, малопомітних загроз, що підкреслює необхідність збалансованого підходу до кіберзахисту. Технології,

такі як **TensorFlow** та **PyTorch**, додатково підсилюють ці системи, даючи змогу розробляти індивідуальні моделі для вивчення шаблонів у даних і оперативного реагування на нові загрози.

Проактивне виявлення загроз. Штучний інтелект та автоматизація змінюють підхід до виявлення загроз, забезпечуючи автоматичне виявлення активів, формування динамічних базових ліній і виявлення аномалій. Ці технології аналізують великі обсяги даних, щоб виділити шкідливі події та ідентифікувати індикатори компрометації. Такі компанії, як IBM, Palo Alto Networks і Huntress, розширюють можливості проактивного пошуку загроз, сприяючи ранньому їхньому виявленню та нейтралізації.

Інструменти на зразок **Stealthwatch** від **Cisco** створюють карти активів і потоків даних, виявляючи вразливі місця. Платформи штучного інтелекту, як-от **Vectra Cognito** і **Darktrace**, формують базові моделі звичайної поведінки, що дає змогу виявляти аномалії, які можуть свідчити про наявність загроз [11].

Система **QRadar** від IBM автоматизує процес пошуку загроз, аналізуючи потоки даних на наявність шкідливих подій. Інструменти поведінкового аналізу, такі як **Exabeam**, використовують машинне навчання для виявлення шаблонів поведінки користувачів та об’єктів, сприяючи проактивному виявленню загроз. Платформи, керовані ШІ, як-от **Cortex XDR** від Palo Alto Networks та Huntress, інтегрують і аналізують інформацію про загрози з метою їхнього пріоритетного опрацювання та підтримки розслідувань інцидентів.

Дані системи збирають лог-файли, мережеві потоки, дані з кінцевих точок, корелюють події та аналізують поведінку, щоб надати повну картину безпекових інцидентів.

На прикладі QRadar розберемо структуру подібних рішень. Основні компоненти таких систем [21]:

- Збір та агрегація даних — це фундамент будь-якої системи виявлення загроз. Чим більше різноманітних та якісних даних збирається, тим повнішу картину можна отримати.
- Нормалізація та збагачення даних — зібрані дані часто надходять у різних форматах і потребують уніфікації для ефективного аналізу. Збагачення додає цінний контекст.
- Зберігання даних — великі обсяги даних потребують ефективного та масштабованого зберігання.
- Аналіз даних та виявлення загроз — використовується широкий спектр методів для виявлення аномалій та шкідливої активності. За основу беруться технології SIEM (Security Information and Event Management) та SOAR (Security Orchestration, Automation and Response)

Автоматизоване реагування на інциденти. Штучний інтелект трансформує реагування на інциденти у сфері кібербезпеки, забезпечуючи автономне виконання дій з локалізації та усунення загроз, зокрема — відключення скомпрометованих облікових записів, відкликання облікових даних, ізоляцію інфікованих кінцевих точок та блокування підозрілих IP-адрес.

Ключові гравці, такі як Microsoft, FireEye та Fortinet, посилюють ці можливості, скорочуючи час виявлення та реагування, а також покращуючи загальну безпеку систем. Наприклад, **Cortex XSOAR** від Palo Alto Networks автоматизує ізоляцію кінцевих точок; **QRadar** від IBM аналізує інциденти та пропонує дії для реагування; а **FortiWeb** від Fortinet блокує підозрілі IP-адреси в реальному часі [12].

Крім того, **Falcon** від CrowdStrike аналізує поведінку користувачів для виявлення загроз; Helix від FireEye інтегрує інформацію про загрози для автоматизованого реагування; а **Azure Sentinel** від Microsoft оптимізує робочі процеси реагування на інциденти.

На прикладі Azure Sentinel розберемо як працюють механізми реагування на інциденти. Правила автоматизації застосовуються до таких категорій випадків використання [20]:

- Виконання основних завдань автоматизації для обробки інцидентів без використання посібників. Наприклад:
  - Додавання завдань з інцидентами, які повинні виконувати аналітики.
  - Придушення шумних інцидентів.
  - Сортування нових інцидентів шляхом зміни їх статусу з «Новий» на «Активний» та призначення власника.
  - Позначення інцидентів тегами для їх класифікації.
  - Ескалація інциденту шляхом призначення нового власника.
  - Закриття вирішених інцидентів із зазначенням причини та додаванням коментарів.
- Автоматизація відповідей для декількох правил аналізу одночасно.
- Контроль порядку виконання дій.
- Перевірка вмісту інциденту (сповіщення, об’єкти та інші властивості) і вжиття подальших заходів шляхом виклику сценарію дій.
- Правила автоматизації також можуть бути механізмом, за допомогою якого ви запускаєте сценарій дій у відповідь на сповіщення, не пов’язане з інцидентом.

Правила автоматизації оптимізують використання автоматизації в Microsoft Sentinel, дозволяючи спростити складні робочі процеси для процесів координації реагування на загрози.

Правила автоматизації складаються з трьох основних компонентів:

- **Тригери**, які визначають, який тип інциденту викликає виконання правила, залежно від умов.
- **Умови**, які визначають точні обставини, за яких правило виконується і здійснює дії.
- **Дії**, які змінюють інцидент певним чином або викликають плейбук, який виконує більш складні дії та взаємодіє з іншими службами.

Виявлення та реагування в реальному часі. Штучний інтелект підвищує рівень безпеки організацій завдяки можливостям виявлення та реагування в реальному часі, що дає змогу швидко ідентифікувати нетипову активність і нейтралізувати загрози.

Системи виявлення вторгнень, посилені ШІ, як-от **Stealthwatch** від Cisco, аналізують мережевий трафік, тоді як рішення для захисту кінцевих точок, як-от **Singularity** від SentinelOne, здійснюють моніторинг поведінки пристроїв і реагують на аномалії.

Підсилені ШІ системи управління інформацією та подіями безпеки, як-от **Splunk Enterprise Security**, здійснюють аналіз подій у реальному часі, виявляючи загрози та автоматично реагуючи на них. Платформи, як-от **Cortex XDR** від Palo Alto Networks, здійснюють пошук загроз у мережах, на кінцевих точках і в хмарних середовищах, забезпечуючи виявлення та реагування в реальному часі.

Інструменти автоматизації безпеки та реагування на інциденти, наприклад, **Resilient** від IBM, автоматизують заходи з кіберзахисту, а рішення, такі як **Exabeam**, використовують машинне навчання для виявлення відхилень у поведінці користувачів і генерації сповіщень. Ці технології значно покращують здатність швидко реагувати на кіберзагрози, підвищуючи загальний рівень безпеки.

Виклики впровадження ШІ для захисту. Попри значний потенціал, інтеграція штучного інтелекту в системи кібербезпеки стикається з низкою суттєвих викликів. Однією

з головних проблем є високий рівень хибних спрацьовувань, які ШІ-моделі можуть генерувати. Це призводить до перевантаження аналітиків безпеки нерелевантними сповіщеннями, знижуючи їхню ефективність. Для мінімізації цих хибних спрацьовувань необхідне постійне тонке налаштування моделей, що вимагає значних ресурсів та експертних знань, а також ефективна інтеграція з людським аналізом. Крім того, для ефективного навчання та роботи ШІ-моделей потрібні величезні обсяги якісних, різноманітних та репрезентативних даних. Збір, зберігання та обробка таких даних є складним завданням, особливо для невеликих організацій, а також піднімає питання конфіденційності та відповідності регуляторним нормам, таким як GDPR [22]. Третім важливим викликом є адаптація до нових загроз. Кіберзлочинці постійно змінюють свої тактики та інструменти, створюючи експлойти та раніше невідомі види шкідливого програмного забезпечення. ШІ-системи повинні бути достатньо гнучкими та здатними до швидкого перенавчання, щоб ефективно виявляти ці нові загрози, не покладаючись виключно на відомі сигнатури. Нарешті, існує ризик «отруєння» даних, коли зловмисники свідомо маніпулюють навчальними даними ШІ-моделей захисту, щоб знизити їхню ефективність, спричинити хибні спрацьовування або змусити їх пропускати атаки. Це вимагає розробки та впровадження стійких до таких маніпуляцій алгоритмів та механізмів верифікації навчальних даних.

ВИСНОВОКИ

Проведене дослідження глибоко аналізує дуальну природу використання штучного інтелекту (ШІ) у сучасному середовищі кібербезпеки. Було показано, що ШІ став потужним інструментом як для підвищення ефективності кібератак, так і для розробки складних механізмів їхнього виявлення та запобігання. З одного боку, генеративні моделі ШІ, такі як ChatGPT, WormGPT, FraudGPT, та самореплікативні агенти на кшталт Morris 2.0, значно розширюють арсенал зловмисників, дозволяючи створювати надзвичайно переконливі фішингові кампанії, адаптивне шкідливе програмне забезпечення та діпфейки, що ставить нові виклики перед існуючими системами захисту.

З іншого боку, у відповідь на ці загрози, рішення на основі ШІ активно застосовуються для посилення кібербезпеки. Методи машинного навчання, обробки природної мови та розпізнавання образів дозволяють ефективно виявляти аномалії в мережевому трафіку, розпізнавати шкідливі URL-адреси та підозрілу поведінку в реальному часі, що часто залишається непоміченим для традиційних підходів. Системи, такі як Darktrace, Cylance, Splunk, IBM QRadar та Microsoft Sentinel, демонструють здатність до проактивного виявлення загроз, автоматизованого реагування на інциденти та комплексного моніторингу, забезпечуючи захист усього спектру пристроїв та мереж.

Ключовим результатом дослідження є наголос на необхідності інтеграції автоматизованих можливостей ШІ з експертним аналізом фахівців. Цей гібридний підхід є критично важливим для забезпечення ефективного зниження ризиків та адаптації до нових, еволюціонуючих загроз, які можуть виникати внаслідок постійного розвитку ШІ. Таким чином, ШІ не просто трансформує підходи до кібербезпеки, а пропонує проактивну та адаптивну стратегію, яка є незамінною для боротьби з кібератаками та захисту конфіденційних даних у сучасному цифровому світі.

Підготовлено в межах виконання теми № 54.16.04.01-25/27.ЗФ-01

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Zellers R., Holtzman A., Bisk Y., Farhadi A., Choi Y. Defending Against Neural Fake News // Advances in Neural Information Processing Systems (NeurIPS 2019). – 2019. – С. 9054–9065. – URL: https://papers.nips.cc/paper/9106-defending-against-neural-fake-news.pdf
2. Bostrom N., Yudkowsky E. (eds.) Global Catastrophic Risks. – Oxford: Oxford University Press, 2008. – С. 308–345.
3. Kenneth R. Feinberg Center for Catastrophic Risk Management and Compensation. Insuring Catastrophic Cyber Risk. – 2025.
4. Shevlane T. An Early Warning System for Novel AI Risks // Google DeepMind. – 2024. – URL: https://deepmind.google/discover/blog/an-early-warning-system-for-novel-ai-risks/
5. Scanlon M., Breitinger F., Hargreaves C., Hilgert J.-N., Sheppard J. ChatGPT for Digital Forensic Investigation: The Good, the Bad, and the Unknown // Forensic Science International: Digital Investigation. – 2023. – Vol. 46. – Article 301609.
6. Tihanyi N., Ferrag M.A., Jain R., Debbah M. CyberMetric: A Benchmark Dataset for Evaluating Large Language Models Knowledge in Cybersecurity. – 2024.
7. Головне управління розвідки МО України. ГУР МО зірвало провокацію російських пранкерів проти генерального директора компанії Baykar Defence Халюка Байрактара. – URL: https://gur.gov.ua/content/hur-mo-zirvalo-provokatsiiu-rosiiskykh-prankeriv-proty-heneralnoho-dyrektora-kompanii-baykar-defence-khaliuka-bairaktara.html
8. Begou N., Vinoy J., Duda A., Korczynski M. Exploring the Dark Side of AI: Advanced Phishing Attack Design and Deployment Using ChatGPT. – 2023.
9. Saha Roy S., Naragam K.V., Nilizadeh S. Generating Phishing Attacks Using ChatGPT. – 2023.
10. Islam R. Generative AI, Cybersecurity, and Ethics. – George Mason University, Fairfax, Virginia, United States, 2025.
11. Piconese F., Hakkala A., Virtanen S., Crispo B. Deployment of Next Generation Intrusion Detection Systems against Internal Threats in a Medium-sized Enterprise, 2020.
12. Šuškalo D., Morić Z., Redžepagić J., Regvart D. Comparative Analysis of IBM QRadar and Wazuh for Security Information and Event Management, 2023.
13. Happe A., Cito J. Getting Pwn'd by AI: Penetration Testing with Large Language Models
14. The EU Artificial Intelligence Act. – URL: https://artificialintelligenceact.eu/
15. Kelley D. WormGPT – The Generative AI Tool Cybercriminals Are Using to Launch Business Email Compromise Attacks // SlashNext. – 2023.
16. Krishnan R. FraudGPT: The Villain Avatar of ChatGPT // Neterich. – 2023.
17. Хробак Morris II. Що відомо про комп'ютерні віруси в епоху ШІ // ForkLog UA. – 2024.
18. Helping Ethical Hackers Use AI in 50 Lines of Code or Less. – URL: https://hackingbuddy.ai/
19. What is Cylance Endpoint Security? – URL: https://docs.blackberry.com/en/unified-endpoint-security/blackberry-ues/overview/What-is-Unified-Endpoint-Security
20. Automate Threat Response in Microsoft Sentinel with Automation Rules. – URL: https://learn.microsoft.com/en-us/azure/sentinel/automate-incident-handling-with-automation-rules?tabs=onboarded
21. QRadar overview. – URL: https://www.ibm.com/docs/en/qsip/7.5?topic=started-qradar-overview
22. Dalalah D., Dalalah M. The false positives and false negatives of generative AI detection tools in education and academic research: The case of ChatGPT. – 2023.

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 4 (28), 2025
ISSN 2663 - 4023

Maksymilian Yakovlev
Ggraduate student
Sumy State University, Sumy, Ukraine
ORCID ID: 0000-0002-1196-4062
maksimilianyakovlev@gmail.com

Volodymyr Lubchak
PhD in Physics and Mathematics
Associate Professor, Professor of the Department of Computer Science
Sumy State University, Sumy, Ukraine
ORCID ID: 0000-0002-7335-6716
v.liubchak@dcs.sumdu.edu.ua

POTENTIALS OF ARTIFICIAL INTELLIGENCE IN DETECTING AND PREVENTING PHISHING AND CYBER ATTACKS

Abstract. This paper examines the dual role of artificial intelligence (AI) in today's cybersecurity landscape, highlighting its capacity to both enhance cyberattacks and support the development of effective defense mechanisms for information systems. The study analyzes the increasing complexity of threats driven by advancements in machine learning, natural language processing, and generative AI (GenAI), which enable attackers to automate, improve the accuracy of, and disguise attacks, including phishing, malware creation, and the use of deepfakes. This research explores how AI-based solutions can strengthen cybersecurity by detecting potential threats in real time through machine learning, NLP, and image recognition techniques. Special attention is given to the necessity of integrating AI with human oversight, emphasizing the importance of combining automated responses with expert analysis to effectively mitigate risks and adapt to emerging challenges. The paper reviews a range of modern tools and techniques used to execute AI-assisted cyberattacks, such as ChatGPT, WormGPT, FraudGPT, and Morris 2.0, demonstrating their capabilities in crafting convincing fraudulent scenarios and adaptive malicious software. In parallel, it examines AI-powered technological solutions designed to detect and counter cyber threats in real time. It describes the functionality of systems like Darktrace Antigena, Cylance Endpoint Security, Splunk, Exabeam, IBM QRadar, and Microsoft Sentinel, which leverage behavioral analysis, machine learning, and image recognition for proactive anomaly detection, automated incident response, and overall security enhancement. This study illustrates how AI is transforming cybersecurity by providing adaptive and proactive strategies to combat emerging cyber threats. By exploring contemporary AI applications, it demonstrates how AI can reshape cybersecurity by offering proactive and adaptive approaches to counter cyberattacks and protect sensitive data.

Keywords: malware; generative artificial intelligence; cyber threats; phishing attacks; automated response.

REFERENCES (TRANSLATED AND TRANSLITERATED)

1. Zellers R., Holtzman A., Bisk Y., Farhadi A., Choi Y. Defending Against Neural Fake News // Advances in Neural Information Processing Systems (NeurIPS 2019). – 2019. – С. 9054–9065. – URL: https://papers.nips.cc/paper/9106-defending-against-neural-fake-news.pdf
2. Bostrom N., Yudkowsky E. (eds.) Global Catastrophic Risks. – Oxford: Oxford University Press, 2008. – С. 308–345.
3. Kenneth R. Feinberg Center for Catastrophic Risk Management and Compensation. Insuring Catastrophic Cyber Risk. – 2025.
4. Shevlane T. An Early Warning System for Novel AI Risks // Google DeepMind. – 2024. – URL: https://deepmind.google/discover/blog/an-early-warning-system-for-novel-ai-risks/
5. Scanlon M., Breitinger F., Hargreaves C., Hilgert J.-N., Sheppard J. ChatGPT for Digital Forensic Investigation: The Good, the Bad, and the Unknown // Forensic Science International: Digital Investigation. – 2023. – Vol. 46. – Article 301609.
6. Tihanyi N., Ferrag M.A., Jain R., Debbah M. CyberMetric: A Benchmark Dataset for Evaluating Large Language Models Knowledge in Cybersecurity. – 2024.

7. Main Directorate of Intelligence of the Ministry of Defence of Ukraine. Defence Intelligence of Ukraine Thwarted a Provocation by Russian Pranksters Against Baykar Defence CEO Haluk Bayraktar. – URL: https://gur.gov.ua/content/hur-mo-zirvalo-provokatsiiu-rosiiskykh-prankeriv-proty-heneralnoho-dyrektora-kompanii-baykar-defence-khaliuka-bairaktara.html
8. Begou N., Vinoy J., Duda A., Korczynski M. Exploring the Dark Side of AI: Advanced Phishing Attack Design and Deployment Using ChatGPT. – 2023.
9. Saha Roy S., Naragam K.V., Nilizadeh S. Generating Phishing Attacks Using ChatGPT. – 2023.
10. Islam R. Generative AI, Cybersecurity, and Ethics. – George Mason University, Fairfax, Virginia, United States, 2025.
11. Piconese F., Hakkala A., Virtanen S., Crispo B. Deployment of Next Generation Intrusion Detection Systems against Internal Threats in a Medium-sized Enterprise, 2020.
12. Šuškalo D., Morić Z., Redžepagić J., Regvart D. Comparative Analysis of IBM QRadar and Wazuh for Security Information and Event Management, 2023.
13. Happe A., Cito J. Getting Pwn'd by AI: Penetration Testing with Large Language Models
14. The EU Artificial Intelligence Act. – URL: https://artificialintelligenceact.eu/
15. Kelley D. WormGPT – The Generative AI Tool Cybercriminals Are Using to Launch Business Email Compromise Attacks // SlashNext. – 2023.
16. Krishnan R. FraudGPT: The Villain Avatar of ChatGPT // Neterich. – 2023.
17. Morris Worm II. What Is Known About Computer Viruses in the Age of AI // ForkLog UA. – 2024.
18. Helping Ethical Hackers Use AI in 50 Lines of Code or Less. – URL: https://hackingbuddy.ai/
19. What is Cylance Endpoint Security? – URL: https://docs.blackberry.com/en/unified-endpoint-security/blackberry-ues/overview/What-is-Unified-Endpoint-Security
20. Automate Threat Response in Microsoft Sentinel with Automation Rules. – URL: https://learn.microsoft.com/en-us/azure/sentinel/automate-incident-handling-with-automation-rules?tabs=onboarded
21. QRadar overview. – URL: https://www.ibm.com/docs/en/qsip/7.5?topic=started-qradar-overview
22. Dalalah D., Dalalah M. The false positives and false negatives of generative AI detection tools in education and academic research: The case of ChatGPT. – 2023.