МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

**ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ**

**НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ КІБЕРБЕЗПЕКИ ТА ЗАХИСТУ ІНФОРМАЦІЇ
КАФЕДРА УПРАВЛІННЯ КІБЕРБЕЗПЕКОЮ ТА ЗАХИСТОМ ІНФОРМАЦІЇ**

КВАЛІФІКАЦІЙНА РОБОТА

на тему: “ВПЛИВ ШТУЧНОГО ІНТЕЛЕКТУ НА КІБЕРБЕЗПЕКУ: СУЧАСНІ ЗАГРОЗИ ТА ЗАХИСНІ ТЕХНОЛОГІЇ”

на здобуття освітнього ступеня бакалавра
зі спеціальності 125 Кібербезпека
освітньої програми Управління інформаційною та кібернетичною безпекою

Кваліфікаційна робота містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело

_____ Владислава ДЗЕГА
(підпис) *Ім’я, ПРІЗВИЩЕ здобувача*

Виконав(ла): здобувач(ка) вищої освіти гр. УБД-41

Владислава ДЗЕГА
Ім’я, ПРІЗВИЩЕ

Керівник: Тетяна БЕРЕСТЯНА
Ім’я, ПРІЗВИЩЕ

Рецензент:
к.т.н., доцент Ім’я, ПРІЗВИЩЕ

Київ 2026

ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ

Навчально-науковий інститут кібербезпеки та захисту інформації

Кафедра управління кібербезпекою та захистом інформації

Ступінь вищої освіти бакалавр

Спеціальність 125 Кібербезпека

Освітня програма Управління інформаційною та кібернетичною безпекою

ЗАТВЕРДЖУЮ
Завідувач кафедри УКБЗІ
_____ Світлана ЛЕГОМІНОВА
“____”_____2026 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

Дзегі Владиславі Ігорівні
(прізвище, ім’я, по батькові здобувача)

1. Тема кваліфікаційної роботи “Вплив штучного інтелекту на кібербезпеку: сучасні загрози та захисні технології”,
керівник кваліфікаційної роботи БЕРЕСТЯНА Т.В.,

затверджені наказом Державного університету інформаційно-комунікаційних технологій від “20“ лютого 2026 р. №51

2. Строк подання кваліфікаційної роботи “12” травня 2026р.

3. Вихідні дані до кваліфікаційної роботи: *матеріали ENISA, NIST, OWASP, Europol; міжнародні стандарти та рамкові підходи ISO/IEC 27001, NIST Cybersecurity Framework, NIST AI Risk Management Framework; публікації щодо AI-посилених кібератак, deepfake-технологій, prompt injection, data poisoning, машинного навчання, поведінкової аналітики та систем автоматизованого реагування.*

4. Перелік питань, які мають бути розроблені:
4.1. Дослідити теоретичні основи кібербезпеки та технологій штучного інтелекту, визначити особливості впливу ШІ на інформаційну безпеку.
4.2. Проаналізувати сучасні кіберзагрози, пов’язані з використанням штучного інтелекту, зокрема AI-генерований фішинг, соціальну інженерію, deepfake технології, prompt injection та data poisoning.
4.3. Дослідити сучасні технології кіберзахсту на основі штучного інтелекту, зокрема машинне навчання для виявлення загроз, поведінкову аналітику, системи SOAR, міжнародні підходи до управління кібербезпеки. Та розробити практичні рекомендації щодо підвищення кіберстійкості.

5. Перелік ілюстративного матеріалу: *презентація PowerPoint*
6. Дата видачі завдання “05“ березня 2026 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Етапи кваліфікаційної роботи	Термін виконання етапів роботи	Примітка
1.	Визначення об’єкту, предмету, мети та завдань дослідження.	18.03.2026	
2.	Збір та аналіз літератури.	30.03.2026	
3.	Дослідження теоретичних основ кібербезпеки, технологій ШІ та особливостей його впливу на інформаційну безпеку.	10.04.2026	
4.	Аналіз сучасних кіберзгроз, пов’язаних із використанням штучного інтелекту. AI-генерований фішинг, соціальна інженерія, атаки на AI-системи.	30.04.2026	
5.	Дослідження сучасних технологій кіберзахисту на основі штучного інтелекту та розроблення практичних рекомендації. ‘	15.05.2026	
6.	Формулювання висновків за результатами проведеного дослідження.	16.05.2026	
7.	Оформлення роботи.	20.05.2026	
8.	Оформлення презентації.	01.06.2026	
9.	Отримання рецензії на роботу.	10.06.2026	
10.	Захист в ДЕК.	12.06.2026	

Здобувачка вищої освіти _____ Владислава ДЗЕГА
(підпис) *(Ім’я, ПРІЗВИЩЕ)*

Керівник
кваліфікаційної роботи _____ Тетяна БЕРЕСТЯНА

ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ
НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ КІБЕРБЕЗПЕКИ ТА ЗАХИСТУ ІНФОРМАЦІЇ

ПОДАННЯ
ГОЛОВІ ЕКЗАМЕНАЦІЙНОЇ КОМІСІЇ
ЩОДО ЗАХИСТУ КВАЛІФІКАЦІЙНОЇ РОБОТИ
на здобуття освітнього ступеня бакалавра

Направляється здобувачка Дзега В.І до захисту кваліфікаційної роботи
(прізвище та ініціали)
за спеціальністю 125 Кібербезпека
(код, найменування спеціальності)
Освітньої програми Управління інформаційною та кібернетичною безпекою
(назва)
на тему: "Вплив штучного інтелекту на кібербезпеку: сучасні загрози та захисні технології"

Кваліфікаційна робота і рецензія додаються.

Директор ННІЗІ Євгенія ІВАНЧЕНКО
(підпис) *(Ім'я, ПРІЗВИЩЕ)*

Висновок керівника кваліфікаційної роботи

Здобувачка ДЗЕГА Владислава у кваліфікаційній роботі на тему "Вплив штучного інтелекту на кібербезпеку: сучасні загрози та захисні технології" дослідила теоретичні основи кібербезпеки та технологій штучного інтелекту , проаналізувала сучасні кіберзагрози, пов'язані з використанням ШІ, а також розглянула преспективні технологіх кібербезпеки на його основі.

ДЗЕГА Владислава продемонструвала розуміння актуальності обраної проблематики, уміння опрацьовувати наукові, нормативні та практичні напрями. Під час виконання кваліфікаційної роботи здобувачка проявила себе відповідально та організовано, здатна самостійно працювати з матеріалами дослідження і формулювати обґрунтовані висновки.

Все це дозволяє оцінити кваліфікаційну роботу здобувачки ДЗЕГА Владислава на оцінку "відмінно" та присвоїти їй кваліфікацію бакалавра з кібербезпеки за освітньою програмою Управління інформаційною та кібернетичною безпекою.

Керівник кваліфікаційної роботи Берестяна Тетяна
(підпис) *(Ім'я, ПРІЗВИЩЕ)* " " 2026 року

Висновок кафедри про кваліфікаційну роботу

Кваліфікаційна робота розглянута. Здобувачка Дзега В.І допускається до захисту даної роботи в Екзаменаційній комісії.

Завідувач кафедри
управління кібербезпекою та
захистом інформації
(підпис) Світлана ЛЕГОМІНОВА
(Ім'я, ПРІЗВИЩЕ)

ВІДГУК РЕЦЕНЗЕНТА
на кваліфікаційну бакалаврську роботу

здобувачки вищої освіти ДЗЕГИ Владислави
на тему “Вплив штучного інтелекту на кібербезпеку: сучасні загрози та захисні технології”

Актуальність. Стрімкий розвиток штучного інтелекту суттєво змінює сучасний ландшафт кіберзагроз. AI-технології можуть використовуватися не лише для виявлення атак і підвищення ефективності захисту, а й для створення персоналізованого фішингу, deepfake-контенту, автоматизації соціальної інженерії та атаки на самі системи штучного інтелекту. Тому дослідження впливу ШІ на кібербезпеку, сучасних загроз і захисних технологій є актуальним науково-практичним завданням.

З огляду на зазначене дослідження проблеми формування обізнаності й навчання персоналу з інформаційної безпеки є актуальним науковим завданням.

Позитивні сторони.

1. У роботі комплексно дослідження ШІ на кібербезпеку, проаналізовано AI-генерований фішинг, соціальну інженерію, deepfake-технології, prompt injection і data poisoning, також розглянуто сучасні технології протидії зазначеним загрозам.
2. Кваліфікаційна робота оформлена відповідно до вимог. Виклад матеріалу здійснено відповідно до плану, зроблено логічні висновки. Ключові положення роботи представлено у вигляді рисунків.
3. Авторка опрацювала значну джерельну базу: близько 50 публікацій, в тому числі англомовних.
4. За результатами дослідження запропоновано рекомендації щодо ефективного навчання персоналу з питань інформаційної безпеки.

Недоліки.

Однак, вищезгадані зауваження не впливають на загальну позитивну оцінку кваліфікаційної роботи.

Висновок: Кваліфікаційна робота виконана на належному науково-методичному рівні і заслуговує оцінки “відмінно”, а здобувачка ДЗЕГА Владислава заслуговує присвоєння кваліфікації бакалавра з кібербезпеки за освітньою програмою Управління інформаційною та кібернетичною безпекою.

Рецензент:
к.т.н., доцент _____ *підпис* Ім’я, ПРІЗВИЩЕ

РЕФЕРАТ

Кваліфікаційна робота присвячена дослідженню впливу штучного інтелекту на кібербезпеку, зокрема аналізу сучасних кіберзагроз та технологій захисту. Робота складається зі вступу, трьох розділів, висновків та списку використаних джерел.

Метою роботи є дослідити вплив штучного інтелекту на кібербезпеку та визначити ефективні методи протидії сучасним загрозам.

Об'єктом дослідження є процеси забезпечення кібербезпеки в умовах використання штучного інтелекту.

Предмет дослідження **–** сучасні кіберзагрози та захисні технології, пов'язані із застосуванням штучного інтелекту.

Методи дослідження. Для досягнення поставленої мети використано методи аналізу та синтезу, порівняння, узагальнення, системного підходу, а також методи моделювання кіберзагроз та оцінки ефективності захисних механізмів.

Як результат у роботі досліджено сучасні тенденції розвитку кіберзагроз, зокрема використання штучного інтелекту для автоматизації атак, створення фішингових повідомлень, deepfake-контенту, а також адаптивного шкідливого програмного забезпечення. Окрему увагу приділено таким видам атак, як data poisoning та prompt injection, що спрямовані на компрометацію систем штучного інтелекту.

Разом з тим проаналізовано можливості використання штучного інтелекту у сфері кіберзахисту, зокрема для виявлення аномалій у мережевому трафіку, поведінкової аналітики користувачів, автоматизації реагування на інциденти та підвищення ефективності систем управління інформаційною безпекою.

Галузь застосування. Отримані результати можуть бути використані у діяльності організацій різних галузей для підвищення рівня кібербезпеки, зокрема при впровадженні систем управління інформаційною безпекою, розробці політик захисту інформації, а також при використанні технологій штучного інтелекту для виявлення та запобігання кіберзагрозам.

Запропоновані підходи можуть застосовуватися у діяльності служб інформаційної безпеки, центрів моніторингу кіберзагроз (SOC), а також у процесах аудиту та оцінки ризиків відповідно до міжнародних стандартів, таких як ISO/IEC 27001 та NIST.

Ключові слова: ENISA, NIST, OWASP, and Europol materials; international standards and frameworks, including ISO/IEC 27001, the NIST Cybersecurity Framework, and the NIST AI Risk Management Framework; publications on AI-enhanced cyberattacks, deepfake technologies, prompt injection, data poisoning, machine learning, behavioral analytics, and automated incident response system.

ABSTRACT

The qualification work is devoted to the study of the impact of artificial intelligence on cybersecurity, in particular the analysis of modern cyber threats and protection technologies. The work consists of an introduction, three chapters, conclusions and a list of sources used.

The purpose of the study is to study the impact of artificial intelligence on cybersecurity and determine effective methods of countering modern threats.

The object of the study processes of ensuring cybersecurity in the context of using artificial intelligence.

The subject of the study is modern cyber threats and protective technologies related to the use of artificial intelligence.

Research methods. To achieve the set goal, methods of analysis and synthesis, comparison, generalization, a systems approach, as well as methods of modeling cyber threats and assessing the effectiveness of protective mechanisms were used.

The paper examines modern trends in the development of cyber threats, in particular, the use of artificial intelligence to automate attacks, create phishing messages, deepfake content, and adaptive malicious software. Special attention is paid to such types of attacks as data poisoning and prompt injection, which are aimed at compromising artificial intelligence systems.

Field of application. The developed approaches can be used in the planning and implementation of the information security management system of the enterprise in the context of information security awareness and training for personnel.

At the same time, the possibilities of using artificial intelligence in the field of cyber security were analyzed, in particular for detecting anomalies in network traffic, user behavioral analytics, automating incident response, and increasing the efficiency of information security management systems.

Keywords: ARTIFICIAL INTELLIGENCE, CYBER SECURITY, CYBER THREATS, PHISHING, DEEPFAKE, DATA POISONING, PROMPT INJECTION, SOAR, INFORMATION SECURITY

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ І СКОРОЧЕНЬ

ІБ	Інформаційна безпека
ПЗ	Програмне забезпечення
ШІ	Штучний інтелект
КБ	Кібербезпека
ГНМ	Генеративні нейронні моделі
SOC	Центр моніторингу безпеки (Security Operations Center)
SIEM	Система управління подіями та інформацією безпеки (Security Information and Event Management)
SOAR	Система оркестрації, автоматизації та реагування на інциденти безпеки (Security Orchestration, Automation and Response)
IDS	Система виявлення вторгнень (Intrusion Detection System)
IPS	Intrusion Prevention System
EDR	Система виявлення та реагування на інциденти на кінцевих пристроях (Endpoint Detection and Response)
XDR	Розширене виявлення та реагування (Extended Detection and Response)
DLP	Система запобігання витоку даних (Data Loss Prevention)
APT	Цілеспрямована тривала атака (Advanced Persistent Threat)
DoS	Атака відмови в обслуговуванні (Denial of Service)
DDoS	Розподілена атака відмови в обслуговуванні (Distributed Denial of Service)
MITM	Атака «людина посередині» (Man-in-the-Middle)
ML	Машинне навчання (Machine Learning)
DL	Глибоке навчання (Deep Learning)
NLP	Обробка природної мови (Natural Language Processing)
GAN	Генеративно-змагальні мережі (Generative Adversarial Networks)

API	Інтерфейс прикладного програмування (Application Programming Interface)
ISO	Міжнародна організація зі стандартизації (International Organization for Standardization)
NIST	Національний інститут стандартів і технологій США (National Institute of Standards and Technology)

ВСТУП

Актуальність теми. Сучасне суспільство, державні установи та підприємства дедалі більше залежать від цифрових технологій, інформаційних систем, хмарних сервісів і мережевої комунікації.

Одночасно із цифровізацією зростає кількість і складність кіберзагроз, наслідки яких можуть бути витоки конфіденційної інформації, фінансові втрати, порушення безперервності роботи організацій, компрометація облікових записів і вплив на критично важливі процеси.

Особливого значення для сучасної кібербезпеки набуває розвиток штучного інтелекту. Технології машинного навчання, генеративного штучного інтелекту, великих мовних моделей і синтетичних медіа створюють нові можливості автоматизованого аналізу загроз, виявлення аномальної активності та прискорення реагування на інциденти. Водночас такі самі технології можуть використовуватися зловмисниками для підготовки персоналізованих фішингових повідомлень, автоматизації соціально інженерії , створення deepfake-аудіо та відео, пошуку вразливостей і підвищення масштобованості атак.

Окрему проблему становить те, що об’єктом кіберзагроз стають і самі системи штучного інтелекту. Атаки типу prompt injection можуть впливати на поведінку великх мовних моделей через спеціально сформовані або приховані інструкції, а data poisoning - змінювати результат роботи AI-систем шляхом компрометації навчальних даних. Унаслідок цього забезпечення кібербезпеки в умовах розвитку ШІ повинно охоплювати не лише протидію атакам, а й захист AI-моделей, даних та механізмів їх взаємодії із зовнішнім середовищем.

Актуальність обраної теми зумовлена необхідністю комплексного дослідження подвійного впливу ШІ на кібербезпеку. З одного боку, ШІ посилює можливості зловмисника і ускладнює традиційні методи атак. З іншогоо боку, технології на основі ШІ можуть бути використані для виявлення загроз, аналізу поведінкових аномалій, автоматизація реагування та підвищення кіберстійкості

організації. Тому дослідження сучасних AI-посилених загроз і технологій захисту є важливим науковим та практичним завданням.

Мета роботи полягає у дослідженні штучного інтелекту на кібербезпеку, аналізи сучасних загроз, пов'язаних із його використанням, та визначенні технології і практичні заходи, спрямованих на підвищення кіберстійкость організації.

Об'єкт дослідження – **п**роцеси забезпечення ібербезпеки в умовах розвитку та використання технологій штучного інтелекту.

Предмет дослідження – сучасні AI- посилені кіберзагрози та технології кіберзахисту, пов'язані із застосуванням штучного інтелекту.

Для досягнення цієї мети в роботі необхідно виконати наступні **завдання**:

1. Дослідити теоретичні основи кібербезпеки та технологій штучного інтелекту, визначити особливості AI на інформаційну безпеку.

2. Проаналізувати сучасні кіберзагрози, пов'язані з використанням AI, зокрема AI-генерований фішинг, соціальну інженерію, deepfake-технології, prompt injection та data poisoning.

3. Дослідити сучасні технології кіберзахисту на основі AI, зокрема машинне навчання для виявлення загроз, поведінкову аналітику, системи автоматизованого реагування SOAR та мжнародні підходи до управління кіберризиками.

Методи дослідження. У процесі виконання кваліфікаційної роботи використано методи аналізу та синтезу для опрацювання наукових, нормативних і аналітисних джерел.

Практичне значення одержаних результатів. Полягає у можливості використання сформульваних у роботі рекомендацій дя підвищення рівня кіберсійкості організацій. Запропоновані заходи можуть бути застосовані під час впровадження систем виявлення фішингових загроз, поведінкового моніторінгу, автоматизованого реагування на інциденти, процедур перевірки критичних запитів, контролю безпеки AI-системи і формування коплексної системи управління кіберризиками.

Апробація результатів кваліфікаційної роботи відбулася на Всеукраїнській науково-практичній конференції “Стратегії кіберстійкості: управління ризиками та безперервність бізнесу” 25 лютого 2026 року.

Розділ 1 ТЕОРЕТИЧНІ ОСНОВИ КІБЕРБЕЗПЕКИ ТА ШТУЧНОГО ІНТЕЛЕКТУ

1.1 Сутність та основні принципи кібербезпеки

Кібербезпека є важливою складовою сучасної інформаційної безпеки та охоплює сукупність організаційних і технічних заходів, спрямованих на захист інформаційних систем, мереж, програмного забезпечення та даних від несанкціонованого доступу, порушення функціонування, пошкодження або знищення. У сучасних умовах цифровізації значення кібербезпеки постійно зростає, оскільки організації все більше залежать від цифрових сервісів, інформаційних ресурсів і мережевої взаємодії.

Розвиток інформаційних технологій супроводжується ускладненням кіберзагроз. Атаки можуть бути спрямовані не лише на технічну інфраструктуру, а й на користувачів, процеси обробки даних, віддалені сервіси та системи автоматизованого прийняття рішень. Особливого значення це набуває в умовах активного використання штучного інтелекту, який може застосовуватися як для посилення захисних механізмів, так і для підготовки більш складних атак.

Основу забезпечення інформаційної безпеки становлять базові принципи, відомі як модель CIA: конфіденційність, цілісність і доступність інформації [26].

1. Конфіденційність (Confidentiality) - забезпечення доступу до інформації лише тим користувачам і системам, які мають відповідні повноваження.
2. Цілісність (Integrity) - забезпечення точності, повноти та незмінності інформації від несанкціонованого втручання або випадкового спотворення.
3. Доступність (Availability) - забезпечення своєчасного доступу авторизованих користувачів до інформації, систем і ресурсів.

4. 

Рис. 1.1 -Модель CIA як основа кібербезпеки

Джерело: розроблено автором на основі W. Stallings [26].

Крім основних складових моделі CIA, важливе значення для забезпечення кібербезпеки мають автентичність, підзвітність і незаперечність. Автентичність передбачає підтвердження справжності користувача, пристрою або джерела інформації. Підзвітність забезпечує можливість відстеження дій користувачів і систем. Незаперечність означає неможливість безпідставної відмови від факту виконання певної дії або передавання інформації.

Важливим підходом у сфері кібербезпеки є ризик-орієнтоване управління. Його сутність полягає у визначенні інформаційних активів, ідентифікації загроз і вразливостей, оцінюванні можливих наслідків та виборі відповідних заходів захисту. Відповідно до ISO/IEC 27001:2022 організація повинна створювати, впроваджувати, підтримувати та постійно вдосконалювати систему управління інформаційною безпекою. ISO/IEC 27005:2022 доповнює цей підхід рекомендаціями щодо управління ризиками інформаційної безпеки [21; 25].

Ще одним важливим принципом є багаторівневий захист (Defense in Depth), за якого безпека забезпечується не одним окремим засобом, а сукупністю організаційних і технічних механізмів. До них можуть належати контроль доступу, багатофакторна автентифікація, антивірусний захист, мережевий

моніторинг, журналювання подій, навчання персоналу та реагування на інциденти. Такий підхід дозволяє зменшити наслідки атаки навіть тоді, коли один із рівнів захисту виявився недостатнім.

У контексті розвитку штучного інтелекту кібербезпека набуває нових особливостей. З одного боку, технології ШІ можуть використовуватися для аналізу даних, виявлення підозрілої активності та прискорення реагування на інциденти. З іншого боку, вони можуть ставати інструментом для реалізації фішингових атак, створення синтетичного контенту, автоматизації соціальної інженерії або компрометації самих AI-систем. Це зумовлює необхідність розгляду технологічних основ штучного інтелекту та його впливу на сучасну інформаційну безпеку.

1.2 Основи та технології штучного інтелекту

Штучний інтелект є одним із провідних напрямів розвитку цифрових технологій, орієнтованим на створення систем, здатних виконувати завдання, що потребують інтелектуальної діяльності людини. До таких завдань належать аналіз даних, розпізнавання образів і мовлення, обробка текстової інформації, прогнозування, підтримка прийняття рішень та генерація нового контенту.

У сфері кібербезпеки технології штучного інтелекту можуть використовуватися для аналізу значних обсягів подій безпеки, виявлення підозрілих закономірностей, класифікації потенційно шкідливих об’єктів, визначення аномальної поведінки користувачів і підтримки процесів реагування на інциденти. Водночас ці самі технології можуть використовуватися зловмисниками для підготовки переконливих фішингових повідомлень, створення deepfake-контенту та автоматизації окремих етапів атаки.

Однією з основних складових штучного інтелекту є машинне навчання (Machine Learning, ML). Машинне навчання передбачає створення моделей, які аналізують навчальні дані, виявляють у них закономірності та на їх основі можуть виконувати класифікацію, прогнозування або виявлення нетипових подій. На відміну від традиційних програмних алгоритмів, у яких правила

переважно задаються розробником наперед, ML-модель формує закономірності на основі прикладів, використаних під час навчання.

У контексті кібербезпеки машинне навчання може застосовуватися для виявлення фішингових посилань, аналізу мережевого трафіку, класифікації шкідливого програмного забезпечення, пошуку аномалій у діях користувачів та визначення підозрілих подій у журналах безпеки. Ефективність такого підходу залежить від якості й актуальності навчальних даних, оскільки неповні, застарілі або скомпрометовані дані можуть негативно впливати на результати роботи моделі.

Більш складним напрямом машинного навчання є глибоке навчання (Deep Learning, DL). Воно ґрунтується на використанні багатошарових нейронних мереж, здатних автоматично виявляти складні характеристики у великих масивах даних. Глибоке навчання застосовується для розпізнавання зображень, аналізу аудіо- та відеоматеріалів, обробки природної мови й створення синтетичного контенту [24].

Особливого поширення набув генеративний штучний інтелект (Generative AI), основною функцією якого є створення нового контенту на основі закономірностей, засвоєних під час навчання. Генеративні моделі можуть формувати тексти, зображення, програмний код, аудіозаписи та відеоматеріали. Для кібербезпеки це має подвійне значення: генеративний ШІ може допомагати фахівцям аналізувати інциденти й готувати рекомендації, але також може використовуватися для створення шахрайського контенту та маніпуляції довірою користувачів.

До найбільш поширених генеративних технологій належать великі мовні моделі (Large Language Models, LLM). Такі моделі навчаються на значних обсягах текстових даних і можуть створювати зв’язні відповіді природною мовою, узагальнювати інформацію, змінювати стиль повідомлень та підтримувати діалог із користувачем. У сфері кібербезпеки LLM можуть застосовуватися для підтримки аналітичної роботи, обробки повідомлень про інциденти та підготовки звітів. Водночас вони здатні спрощувати підготовку

персоналізованих фішингових листів і сценаріїв соціальної інженерії.

Окремим напрямом розвитку генеративного ШІ є створення синтетичних медіа та deepfake-контенту. За допомогою нейронних мереж може здійснюватися імітація зовнішності, голосу, міміки або манери мовлення реальної особи. Наприклад, розроблена Microsoft модель VALL-E продемонструвала можливість синтезу персоналізованого мовлення на основі трисекундного аудіозапису голосу раніше невідомого мовця [10]. У контексті кібербезпеки це створює ризики голосового шахрайства, вішингу, підроблених відеоконференцій і маніпуляції довірою працівників організації.

Водночас системи штучного інтелекту самі можуть бути об’єктом кібератак. AI-моделі залежать від навчальних даних, зовнішнього контенту, програмних інтерфейсів та правил взаємодії з користувачами. Тому компрометація даних, маніпуляція запитами або надання AI-системі надмірних повноважень може призвести до некоректних рішень, витоку інформації або виконання небажаних дій. NIST AI Risk Management Framework наголошує на необхідності управління ризиками AI-систем протягом їх розроблення, впровадження та використання [23].

Таким чином, сучасний штучний інтелект охоплює машинне навчання, глибоке навчання, генеративні моделі, великі мовні моделі та технології синтетичних медіа. Їх використання створює нові можливості для виявлення й аналізу кіберзагроз, але одночасно формує ризики автоматизації атак, маніпуляції довірою користувачів і компрометації самих AI-систем.

1.3 Вплив штучного інтелекту на інформаційну безпеку

Розвиток штучного інтелекту суттєво впливає на сферу інформаційної та кібербезпеки. Особливість цього впливу полягає в його подвійному характері: одні й ті самі AI-технології можуть використовуватися як для виявлення та нейтралізації кіберзагроз, так і для підготовки більш складних, швидких і переконливих атак.

З одного боку, штучний інтелект створює нові можливості для

кіберзахисту. Алгоритми машинного навчання можуть аналізувати значні обсяги інформації, виявляти підозрілі закономірності, класифікувати потенційно шкідливі об’єкти та знаходити відхилення від нормальної поведінки користувачів або пристроїв. Завдяки цьому AI-рішення можуть застосовуватися для виявлення фішингових посилань, аналізу мережевого трафіку, пошуку аномалій, підтримки роботи центрів моніторингу безпеки та прискорення реагування на інциденти.

З іншого боку, штучний інтелект розширює можливості зловмисників. Генеративні моделі та великі мовні моделі можуть використовуватися для створення грамотних і персоналізованих фішингових повідомлень, імітації корпоративного стилю спілкування, підготовки сценаріїв соціальної інженерії та автоматизації збору інформації про потенційну жертву. У результаті атаки можуть ставати масштабнішими й складнішими для виявлення традиційними методами.

Окрему загрозу становить використання генеративного ШІ для створення синтетичного контенту. Deepfake-аудіо та відео, AI-клонування голосу й підроблені цифрові профілі можуть застосовуватися для маніпуляції довірою користувачів, здійснення фінансового шахрайства або поширення дезінформації. У таких умовах голосове повідомлення, відеозвернення або зовнішньо переконлива службова комунікація вже не можуть самі по собі вважатися достатнім підтвердженням достовірності запиту.

Важливо враховувати й те, що об’єктом атак можуть ставати самі системи штучного інтелекту. До таких загроз належать prompt injection, за якої зловмисник намагається змінити поведінку мовної моделі через спеціально сформовану інструкцію, та data poisoning, спрямована на компрометацію даних, які використовуються для навчання або налаштування AI-системи. Такі атаки можуть призводити до помилкових результатів, витоку інформації або виконання небажаних дій у пов’язаних системах [16; 17].

За даними ENISA Threat Landscape 2025, фішинг залишається домінуючим вектором первинного проникнення, а використання штучного інтелекту стає

важливим чинником сучасного ландшафту загроз [2; 3]. Це свідчить про те, що AI не лише створює нові ризики, а й підсилює вже відомі атаки, роблячи їх більш персоналізованими, автоматизованими та доступними для реалізації.

Таким чином, вплив штучного інтелекту на інформаційну безпеку можна розглядати за трьома основними напрямами: використання AI як інструмента кіберзахисту, використання AI як інструмента реалізації атак і забезпечення безпеки самих AI-систем. Такий багатовимірний характер впливу ШІ визначає необхідність детального аналізу AI-посилених кіберзагроз і сучасних технологій протидії їм.

Висновки до розділу 1

У першому розділі було розглянуто теоретичні основи кібербезпеки та штучного інтелекту, які формують базу для подальшого дослідження сучасних кіберзагроз і захисних технологій. Встановлено, що кібербезпека є комплексом організаційних і технічних заходів, спрямованих на захист інформаційних систем, мереж, програмного забезпечення та даних від несанкціонованого доступу, порушення цілісності й недоступності ресурсів.

Основою забезпечення кібербезпеки є принципи конфіденційності, цілісності та доступності інформації, що становлять модель CIA. Їх доповнюють автентичність, підзвітність і незаперечність дій. Визначено, що ефективна система захисту повинна ґрунтуватися на ризик-орієнтованому підході, передбаченому ISO/IEC 27001 та ISO/IEC 27005, а також на принципі багаторівневого захисту, за якого безпека забезпечується поєднанням кількох взаємодоповнювальних механізмів.

У ході аналізу технологій штучного інтелекту було встановлено, що до основних напрямів його розвитку належать машинне навчання, глибоке навчання, генеративний штучний інтелект, великі мовні моделі та технології синтетичних медіа. Такі технології здатні аналізувати значні обсяги інформації, виявляти закономірності, генерувати текстовий, аудіо- та відеоконтент, а також підтримувати прийняття рішень. Для сфери кібербезпеки це створює можливості

автоматизованого аналізу загроз, виявлення аномальної активності та прискорення реагування на інциденти.

Водночас встановлено, що застосування штучного інтелекту має подвійний характер. З одного боку, AI-рішення можуть використовуватися для підвищення ефективності кіберзахисту. З іншого боку, ті самі технології можуть застосовуватися зловмисниками для створення переконливих фішингових повідомлень, deepfake-аудіо та відео, автоматизації соціальної інженерії й підвищення масштабованості кібератак. Крім того, об’єктом впливу можуть ставати самі AI-системи, їхні навчальні дані та механізми взаємодії з користувачами.

Таким чином, розвиток штучного інтелекту суттєво змінює сучасний ландшафт кібербезпеки. Його використання одночасно розширює можливості захисту та формує нові ризики, що потребують системного аналізу і розроблення відповідних заходів протидії. Це обґрунтовує необхідність дослідження AI-посилених кібератак, фішингу, соціальної інженерії, deepfake-технологій і атак на системи штучного інтелекту, які розглядаються у наступному розділі роботи.

Розділ 2 СУЧАСНІ КІБЕРЗАГРОЗИ НА ОСНОВІ ШТУЧНОГО ІНТЕЛЕКТУ

У першомк розділі було визначено, що ШІ має подвійний влив на кібербезпку. У цьомо розділі роглянемо саме загрози, пов'язані з використанням ШІ зловмисниками, а також атаки, спрямовані на самі системи ШІ.

У сучасному підході до аналізу впливу ШІ на кібербезпеку доцільно виділити кілька основних напрямів:

1. Використання ШІ в кібератаках - застосування алгоритмів машинного навчання, генеративних моделей та великих мовних моделей для автоматизації розвідки, пошуку вразливостей, створення фішингових повідомлень, допомоги у написанні шкідливого коду та обходу захисних механізмів.
2. Використання ШІ в кіберзахисті - застосування AI-рішень для аналізу мережевого трафіку, виявлення аномальної поведінки, класифікації шкідливого програмного забезпечення, підтримки роботи SOC, SIEM, XDR-систем та автоматизованого реагування на інциденти.
3. Кібербезпека самих систем штучного інтелекту - захист AI-моделей, навчальних даних, API, інфраструктури та процесів прийняття рішень від таких атак, як data poisoning, prompt injection, model extraction, adversarial attacks та інші форми маніпуляції моделлю.
4. Зловмисне використання генеративного ШІ для маніпуляції довірою - створення deepfake-аудіо та відео, синтетичних зображень, фейкових профілів, персоналізованих повідомлень і дезінформаційного контенту з метою соціальної інженерії, шахрайства або репутаційного впливу.

   Таким чином, штучний інтелект у кібербезпеці не можна розглядати лише як інструмент захисту або лише як джерело загроз. Його вплив є багатовимірним: ШІ одночасно посилює можливості захисників, розширює інструментарій зловмисників і створює нові об'єкти захисту у вигляді самих AI-систем.

2.1 Використання штучного інтелекту в кібератаках

В умовах швидкого розвитку цифрових технологій штучний інтелект дедалі частіше розглядається не лише як інструмент кіберзахисту, а й як засіб посилення кібератак. Зловмисники можуть використовувати алгоритми машинного навчання, генеративні моделі та великі мовні моделі для автоматизації окремих етапів атаки, аналізу великих обсягів даних, пошуку вразливостей, підготовки фішингових повідомлень, створення шкідливого контенту та обходу окремих механізмів захисту.

У публікації Forbes Technology Council “How AI-Driven Cyberattacks Will Reshape Cyber Protection” зазначається, що AI-driven cyberattacks передбачають використання передових алгоритмів машинного навчання для виявлення вразливостей, прогнозування шаблонів і експлуатації слабких місць інформаційних систем. Такий підхід свідчить про те, що штучний інтелект у кібератаках виконує роль інструмента прискорення, автоматизації та підвищення точності атакувальних дій.

Кібератаки, у яких застосовується штучний інтелект, відрізняються від традиційних атак підвищеною швидкістю, масштабованістю та здатністю до адаптації. Якщо класичні атаки часто ґрунтувалися на заздалегідь підготовлених сценаріях, то AI-посилені атаки можуть використовувати аналіз даних для коригування дій залежно від поведінки користувача, реакції системи або результатів попередніх спроб злому. Це створює додаткові труднощі для традиційних систем кіберзахисту, які орієнтовані переважно на відомі сигнатури, шаблони поведінки або раніше зафіксовані індикатори компрометації.

Штучний інтелект може застосовуватися зловмисниками на різних етапах життєвого циклу кібератаки. На етапі розвідки AI-інструменти можуть

допомагати автоматично збирати та аналізувати інформацію з відкритих джерел, соціальних мереж, корпоративних сайтів і технічних ресурсів. На етапі підготовки атаки генеративні моделі можуть використовуватися для створення персоналізованих фішингових повідомлень, підроблених профілів, сценаріїв соціальної інженерії або текстів, які імітують стиль реальної особи чи організації. На технічному рівні ШІ може допомагати в аналізі вразливостей, модифікації шкідливого коду та пошуку способів обходу захисних механізмів.

Особливу небезпеку становить те, що штучний інтелект не обов'язково створює принципово нові типи кібератак, але значно підсилює вже наявні. Наприклад, фішингові повідомлення завдяки генеративному ШІ можуть ставати більш грамотно написаними, переконливими й адаптованими до конкретної жертви. Шкідливе програмне забезпечення може швидше модифікуватися для ускладнення його виявлення, а процес розвідки потенційної цілі може бути частково автоматизований. У результаті знижується поріг входу для менш досвідчених зловмисників, оскільки частину складних технічних або аналітичних дій можуть виконувати AI-інструменти.

Використання ШІ також змінює співвідношення між атакувальниками та захисниками. З одного боку, AI-рішення дають змогу організаціям швидше виявляти аномалії, аналізувати інциденти та реагувати на загрози. З іншого боку, ті самі технології можуть застосовуватися зловмисниками для підвищення ефективності атак. Саме тому в сучасних умовах кібербезпека повинна розглядатися як динамічна система, здатна постійно адаптуватися до нових методів атак.

Таким чином, використання штучного інтелекту в кібератаках змінює сучасний ландшафт кіберзагроз. AI-посилені атаки можуть бути

швидшими, дешевшими, масовішими та складнішими для виявлення. Це вимагає від організацій не лише використання традиційних засобів захисту, а й впровадження комплексного підходу до кібербезпеки, який поєднує моніторинг, аналіз поведінки, управління ризиками, навчання персоналу та застосування сучасних AI-рішень для виявлення й нейтралізації загроз.

Для більш повного розуміння масштабів проблеми звернемося не лише до експертних публікацій, а й до аналітичних звітів європейських організацій з кібербезпеки.

За даними ENISA Threat Landscape 2025[2], сучасний ландшафт кіберзагроз у ЄС характеризується високою активністю атак проти державних, фінансових, транспортних та цифрових інфраструктурних секторів. Це свідчить про те, що кібератаки дедалі частіше мають не лише технічний, а й стратегічний характер, оскільки спрямовуються на організації, від яких залежить стабільність суспільних послуг, економіки та державного управління.

У звіті ENISA зазначено, що серед п’яти найбільш цільових секторів у ЄС державне управління, транспорт, цифрова інфраструктура та послуги, фінанси й виробництво разом становлять 53,7% від загальної кількості зареєстрованих інцидентів. Найбільш атакованим сектором визначено державне управління, на яке припадає 38,2% зафіксованих інцидентів. Це пояснюється, зокрема, активністю DDoS-кампаній, атаками на муніципальні сервіси та зростанням інтересу до державних інформаційних ресурсів.

Важливим показником є також структура початкових векторів проникнення. Фішинг є домінуючим вектором первинного доступу та становить близько 60% випадків. Експлуатація вразливостей посідала друге місце й становила 21,3% початкових векторів атаки, тоді як ботнети, шкідливі застосунки та інсайдерські загрози мали меншу частку. Ці дані є

особливо важливими для аналізу впливу штучного інтелекту на кібератаки, оскільки саме фішинг, соціальна інженерія та пошук вразливостей належать до напрямів, які можуть бути суттєво посилені AI-інструментами.

Штучний інтелект може підвищувати ефективність фішингових атак за рахунок створення більш переконливих, граматично правильних і персоналізованих повідомлень. Водночас AI-системи можуть використовуватися для аналізу відкритих джерел, пошуку технічної інформації про організацію, виявлення потенційних слабких місць у програмному забезпеченні та автоматизації підготовчого етапу атаки. Таким чином, статистика ENISA щодо фішингу та експлуатації вразливостей показує, що ШІ насамперед посилює ті типи атак, які вже є найбільш поширеними у сучасному кіберпросторі.

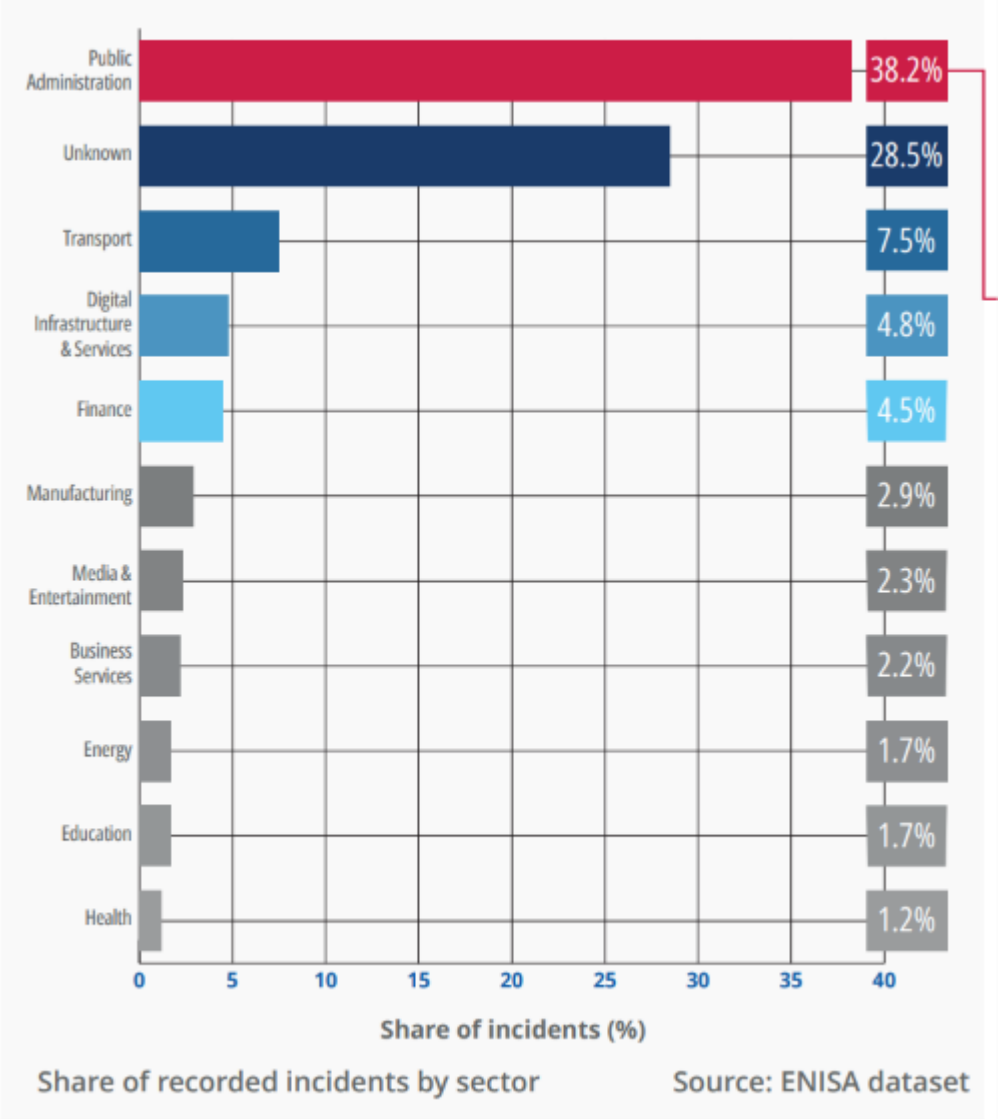


Рис. 2.2 - Розподіл зафіксованих кіберінцидентів за секторами у ЄС за даними ENISA Threat Landscape 2025

Джерело: ENISA Threat Landscape 2025.[2]

Як видно з рисунка 2.1, найбільша частка зафіксованих кіберінцидентів у ЄС припадає на сектор державного управління - 38,2%. Це свідчить про те, що державні установи, муніципальні сервіси та органи публічного управління залишаються одними з основних цілей для кіберзлочинців. На другому місці вказано категорію “Unknown” - 28,5%, що може свідчити про складність точної ідентифікації галузі або джерела окремих інцидентів. Також серед цільових секторів виділяються транспорт - 7,5%, цифрова інфраструктура та сервіси - 4,8%, фінансовий сектор - 4,5% і виробництво - 2,9%.

Окрім аналізу цільових секторів, важливо розглянути, на яких саме етапах кібератаки зловмисники можуть використовувати штучний інтелект. AI-інструменти можуть застосовуватися не лише для одного виду атаки, а на різних

етапах: від збору інформації про ціль до створення фішингових повідомлень, пошуку вразливостей, обходу захисних механізмів та аналізу викрадених даних.

Таблиця 2.1.

Використання штучного інтелекту на різних етапах кібератаки

Етап кібератаки	**Можливе використання ШІ**	**Приклад**
Розвідка цілі	Автоматизований збір і аналіз відкритих даних про організацію, працівників, сервіси, технології та цифровий слід	OSINT-аналіз сайту, соціальних мереж, email-адрес і посад працівників
Підготовка атаки	Генерація сценаріїв соціальної інженерії, підбір теми повідомлення, імітація стилю комунікації конкретної особи або організації	Персоналізований spear-phishing або BEC-шахрайство
Пошук вразливостей	Аналіз CVE, конфігурацій, відкритої технічної інформації, коду або документації для виявлення слабких місць	Виявлення потенційно вразливих вебсервісів або неправильно налаштованих систем
Доставка шкідливого контенту	Створення переконливих листів, повідомлень, фейкових сторінок або вкладень, адаптованих до конкретної жертви	AI-generated phishing або підроблена сторінка авторизації
Обхід захисних механізмів	Модифікація шкідливого коду, обфускація, створення різних варіантів payload для ускладнення виявлення	Ускладнення сигнатурного виявлення антивірусами або EDR-системами
Соціальний вплив	Створення синтетичного аудіо, відео, зображень або текстів для маніпуляції довірою користувачів	Deepfake, voice cloning, фейкові профілі
Дії після компрометації	Аналіз викрадених даних, сортування цінної інформації, підготовка шантажу або подальших атак	Ransomware/extortion, підготовка повторних фішингових кампаній

Отже, штучний інтелект може використовуватися зловмисниками на різних етапах кібератаки - від збору інформації про ціль до створення фішингових

повідомлень, пошуку вразливостей, обходу захисних механізмів та аналізу викрадених даних. Це підтверджує, що AI-посилені атаки є не окремим ізольованим явищем, а розвитком уже наявних методів кіберзлочинності, які завдяки ШІ стають швидшими, більш персоналізованими та складнішими для виявлення.

Одним із найбільш поширених напрямів, у яких штучний інтелект може суттєво посилювати кібератаки, є фішинг та соціальна інженерія. На відміну від суто технічних атак, такі методи ґрунтуються на психологічному впливі на користувача, маніпуляції довірою та імітації легітимної комунікації. Саме тому використання генеративного ШІ у фішингових кампаніях становить особливу небезпеку, оскільки дає змогу створювати персоналізовані, грамотні та переконливі повідомлення у великих масштабах.

2.2 AI-генерований фішинг та соціальна інженерія

У межах дослідження сучасних AI-посилених кіберзагроз особливу увагу доцільно приділити фішингу та соціальній інженерії, оскільки саме ці методи залишаються одними з найпоширеніших способів первинного доступу до інформаційних систем. Їхня небезпека полягає в тому, що атака спрямовується не лише на технічні вразливості, а й на поведінку користувача, його довіру, уважність і здатність розпізнати маніпуляцію. У цьому контексті генеративний ШІ виступає як інструмент, що суттєво підвищує якість, масштабованість і переконливість таких атак.[4]

Як було встановлено у попередньому підрозділі, фішинг залишається одним із домінуючих векторів первинного проникнення в інформаційні системи. Водночас для аналізу AI-генерованого фішингу важливим є інший показник ENISA Threat Landscape 2025: на початку 2025 року фішингові кампанії, підтримані штучним інтелектом, становили понад 80% спостережуваної активності соціальної інженерії у світі [3]. Це свідчить про те, що використання ШІ у фішингових кампаніях перетворюється на поширений інструмент підвищення масштабованості та переконливості атак.

AI-фішинг - це використання великих мовних моделей та генеративних AI-інструментів для створення, персоналізації та масштабування фішингових атак. Зловмисники можуть використовувати LLM для формування переконливих

текстів листів, імітації стилю комунікації, створення сценаріїв, адаптованих до конкретної цілі, а також мовної адаптації повідомлень зі зменшенням кількості очевидних помилок. Йдеться не лише про покращення якості тексту, а й про зміну самого процесу підготовки фішингової кампанії, оскільки зменшується час і обсяг ручної роботи, необхідні для створення персоналізованих повідомлень [4].

Як зловмисники використовують LLM для створення фішингових листів?

Найбільш безпосередній вплив на якість. До LLM фішингові кампанії поділялися на два рівні. Високозатратний спір-фішинг цілив у конкретних осіб з дослідженими, добре написаними приманками. Масовий фішинг розсилав загальні шаблони тисячам адрес, покладаючись на обсяг, а не якість. LLM ліквідували цей розрив.

До поширення великих мовних моделей якісний spear-phishing зазвичай вимагав значного часу на збір інформації про жертву та підготовку переконливого повідомлення, тоді як масові фішингові кампанії переважно спиралися на шаблонні тексти. Використання LLM частково зменшує цей розрив, оскільки дає змогу швидше створювати персоналізовані повідомлення для значної кількості користувачів. У результаті зменшується обсяг ручної роботи, необхідної для підготовки цільових фішингових кампаній. Основні етапи такого процесу узагальнено в таблиці 2.2.

Робочий процес виглядає так:

Таблиця 2.2

Етапи створення AI-генерованої фішингової кампанії

Етап	**Дії зловмисника**	**Значення для фішингової атаки**
Розвідка	Зловмисник збирає інформацію з сайту цільової організації, профілів LinkedIn, пресрелізів, вакансій та інших відкритих джерел.	Дає можливість зробити атаку більш персоналізованою та правдоподібною.

Продовження таблиці 2.2

Етап	**Дії зловмисника**	**Значення для фішингової атаки**
Генерація варіантів	Один і той самий промпт може використовуватися для створення унікальних листів для різних співробітників. Кожен лист може враховувати фактичну посаду, відділ або проєкти конкретного одержувача.	Підвищує масштабованість атаки та ускладнює виявлення фільтрами, які шукають однаковий або шаблонний контент.
Мовна адаптація	Для міжнародних організацій зловмисник може створювати локалізовані версії повідомлень різними мовами з урахуванням мовних і культурних особливостей.	Зменшує кількість очевидних мовних помилок і робить повідомлення природнішим для отримувача.
Генерація варіантів	Один і той самий промпт може використовуватися для створення унікальних листів для різних співробітників. Кожен лист може враховувати фактичну посаду, відділ або проєкти конкретного одержувача.	Підвищує масштабованість атаки та ускладнює виявлення фільтрами, які шукають однаковий або шаблонний контент.
Ітерація та тестування	Якщо початкові повідомлення не дають бажаного результату, зловмисник може швидко змінити промпт, згенерувати нові варіанти та протестувати інші формулювання.	Дає змогу оперативно вдосконалювати фішингову кампанію та підвищувати її ефективність.
Побудова промпту	Зібраний контекст подається до великої мовної моделі разом з інструкціями щодо стилю, ролі відправника, теми	Дозволяє створити повідомлення, яке імітує корпоративний стиль комунікації та виглядає легітимним.

Цей робочий процес не вимагає кастомних моделей чи технічної вишуканості. Він працює з готовими LLM, багато з яких мають достатньо слабкі фільтри безпеки для виробництва переконливих претекстів при непрямому промптуванні.

Чому AI-фішингові листи важче виявити?

Раніше працівники були навчені звертати увагу на типові ознаки фішингових повідомлень :мовні помилки, шаблоність змісту, неприродні формування офіційного повідомлення компанії.

Часто використовувались швидкі шаблони, але з різними назвами компаній, іменами працівники або тема повідомлення. За допомогою LLM-згенерований спрощує для зловмисників створення переконливих повідомлень:

Контекстна точність. При поданні розвідувальних даних LLM посилаються на реальні проєкти, реальних людей та реальні події компанії.

Відповідність стилю. LLM можуть імітувати формальну корпоративну комунікацію, неформальні повідомлення у стилі Slack або технічні IT-сповіщення.

Унікальний контент. Кожен згенерований лист лінгвістично унікальний. Це може з легкістю нагадувати легітимну ділову комунікацію, а не масову кампанію.

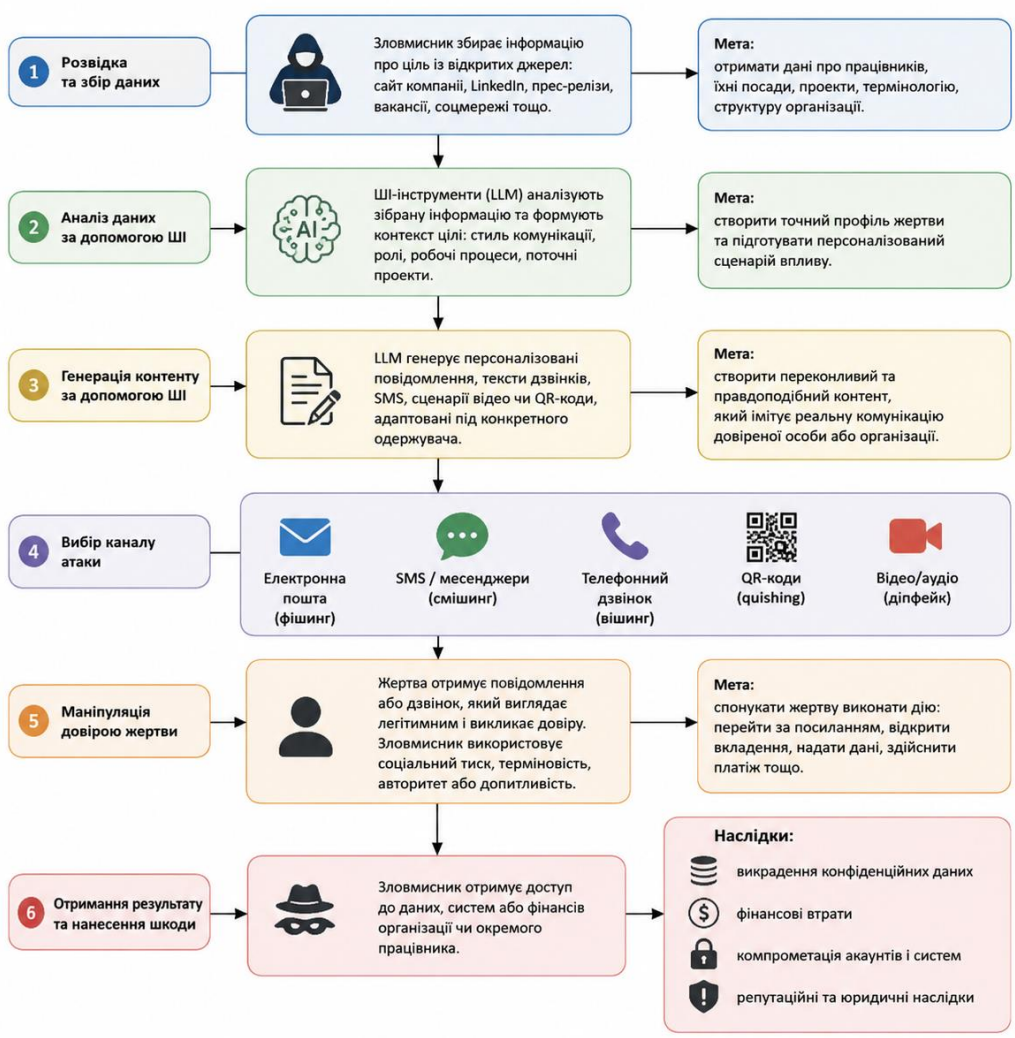
Емоційна калібрація. LLM можуть точно налаштувати рівень терміновості. Тобто , замість повідомлення “Аккаунт буде видалено”, а достатньо стримане повідолення на кшталт : “Маємо уточнити деякі зміни параметрів вашого облікового запису” або “Після оновлення поштового сервісу частина користувачів має повторно перевірити параметри доступу”. Це виглядає професійно. Тому таке повідомлення не завжди викликає негайної підозри.

Це не означає, що виявлення неможливе. Це означає, що методи виявлення, на які співробітники покладалися десятиліття, потребують оновлення. Посібник з виявлення фішингу все ще надає корисні фреймворки,

але акцент змістився з виявлення помилок на верифікацію запитів через

незалежні канали. [5]

Рис. 2.2 – Механізм реалізації AI-посиленої соціальної інженерії



Джерело: розроблено автором на основі аналізу матеріалів ENISA Threat Landscape 2025, ITEZ та інших відкритих джерел.

Рис.2.2 - Механізм реалізації AI-посиленої соціальної інженерії

Джерело : розроблено автором на основі аналізу ENISA Threat Landscape 2025 та матеріалів ITEZ [2; 7].

Економіка зловмисника: Phishing-as-a-Service (PhaaS)

Окремим чинником фішингових атак є **Phishing-as-a-Service (PhaaS)**. Зловмисники можуть в легкому доступі отримати готові інструменти та інфраструктуру для фішингових атак на кампанії. Це можуть бути платформи, наприклад : Caffeine ця програма за підпискою онлайн може надати потрібне середовища для роботи. За даними ENISA Threat Landscape 2025, доступність таких платформ свідчить про індустріалізацію фішингових операцій і дає змогу особам із різним рівнем технічної підготовки реалізовувати складні сценарії атак. У поєднанні з генеративним ШІ така модель може підвищувати

масштабованість фішингу, оскільки автоматизується не лише технічна частина кампанії, а й створення персоналізованого контенту для потенційних жертв [2].

Соціальна інженерія ґрунтується на основі людського фактора: довіра до знайомих сайтів, листів, компаній; неуважність користовача. З частим впливом на час, а саме терміновість надання відповіді, перехід за посиланням та ін.

Ера Social Engineering 2.0 та фішингу остаточно змінила баланс сил у кіберпросторі. AI-посилена соціальна інженерія може поєднувати збір відкритої інформації про потенційну жертву, аналіз отриманих даних, генерацію персоналізованого сценарію та вибір найбільш доречного каналу впливу. Таким каналом може бути електронний лист, SMS-повідомлення, телефонний дзвінок, QR-код або відеозв’язок.

Отже, використання штучного інтелекту у фішингових кампаніях і соціальній інженерії розширює можливості зловмисників щодо створення персоналізованих та правдоподібних сценаріїв впливу на користувача. Проведений аналіз показує, що LLM можуть застосовуватися для підготовки фішингових повідомлень, їх мовної адаптації, імітації корпоративного стилю та швидкого створення різних варіантів атак. У поєднанні з PhaaS-платформами це сприяє масштабуванню фішингових кампаній і знижує вимоги до ручної підготовки окремих повідомлень.

Водночас розвиток генеративного ШІ поширює соціальну інженерію за межі текстових повідомлень. Особливу небезпеку становлять технології синтетичного аудіо та відео, які дають змогу імітувати голос або зовнішність реальної особи. Саме використання deepfake-технологій як інструмента кіберзлочинності розглядається у наступному підрозділі.

2.3 Deepfake-технології як інструмент кіберзлочинності

Deepfake-технології є одним із найбільш небезпечних напрямів використання генеративного штучного інтелекту у кіберзлочинній діяльності. Вони можуть застосовуватися для створення підроблених аудіо- та

відеоматеріалів, імітації голосу або зовнішності реальної особи, формування фальшивих цифрових профілів, здійснення фінансового шахрайства та поширення дезінформації.

Deepfake є різновидом синтетичних медіа, тобто аудіо-, відео- або графічних матеріалів, створених чи змінених за допомогою технологій штучного інтелекту. Синтетичний контент може використовуватися з легітимною метою, зокрема у кіновиробництві, дубляжі, освіті, створенні цифрових сервісів та підвищенні доступності контенту. Водночас розвиток таких технологій створює можливості для маніпуляції довірою, видавання себе за іншу особу, дезінформації та інших злочинних сценаріїв.

Deepfake-технології були детально досліджені та обговорені на одному з заходів стратегічного прогнозування **Лабораторії інновацій Європолу.** Експерти правоохоронних органів, які брали участь у цих заходах, висловили стурбованість щодо наслідків дезінформації, фейкових новин і соціальних мереж для політичного та соціального дискурсу. Очікується, що ці тенденції стануть більш вираженими, оскільки допоміжні технології, такі як діпфейки, стають більш складними. Їхній вплив на конфіденційність та особисту безпеку, безсумнівно, призведе до появи нових категорій злочинів, які доведеться контролювати. Учасники були особливо стурбовані перетворенням соціальних мереж на зброю та впливом дезінформації на публічний дискурс і соціальну згуртованість.

Технологічна основа створення deepfake-контенту

Однією з основних технологій, що забезпечують створення deepfake-контенту, є глибоке навчання (deep learning). Воно є різновидом машинного навчання, за якого комп’ютерні системи аналізують значні обсяги даних за допомогою нейронних мереж і виявляють у них закономірності. У контексті створення deepfake такі алгоритми можуть вивчати характерні риси зовнішності, міміки, рухів або голосу людини, а потім відтворювати їх у новому синтетичному матеріалі.

Сьогодні значна кількість візуальних та аудіоматеріалів є доступною у відкритому інтернет-середовищі: у соціальних мережах, відеохостингах, онлайн-інтерв'ю, вебінарах, публічних виступах і корпоративних матеріалах. Це зменшує необхідність самостійного формування навчальних наборів даних і створює умови для використання публічного цифрового сліду людини у зловмисних цілях. У корпоративному середовищі особливий ризик виникає для керівників, публічних представників компаній та працівників, чиї голосові або відеоматеріали регулярно публікуються у відкритому доступі.

Показовим прикладом можливостей deepfake-технологій є створене у 2018 році відео, у якому режисер Джордан Піл і компанія BuzzFeed використали доступні інструменти редагування для імітації виступу Барака Обами. Це відео не було створене з метою шахрайства, а мало попереджувальний характер: воно демонструвало, що синтетичний контент може формувати враження, ніби відома особа говорить те, чого насправді ніколи не говорила. Такий приклад показує, що навіть легітимне використання deepfake-технологій підтверджує їхній потенціал для подальшого застосування у дезінформації, маніпуляції довірою та кіберзлочинних сценаріях [8].

Source: Suwajanakorn, S. et al., 2017, 'Synthesizing Obama: learning lip sync from audio', ACM Transactions on Graphics, 36(4), accessed on 15 March 2022, https://dl.acm.org/doi/10.1145/3072959.3073640.

Рис. 2.3 - Кадр із демонстраційного deepfake-відео із зображенням Барака Обами, створеного для привернення уваги до загрози дезінформації

Джерело: Europol, Facing Reality? Law Enforcement and the Challenge of

Deepfakes [8].

Використання deepfake-технологій у дезінформаційних кампаніях

Окремим напрямом зловмисного використання deepfake-технологій є поширення дезінформації під час політичних криз, воєнних конфліктів і гібридних протистоянь. У таких випадках метою атаки є не безпосереднє отримання доступу до інформаційної системи або викрадення коштів, а вплив на поведінку людей, підрив довіри до державних установ та створення викривленого сприйняття подій.

У звіті Europol «Facing Reality? Law Enforcement and the Challenge of Deepfakes» наведено, пов'язаний із повномасштабним вторгненням Російської Федерації в Україну у 2022 році. До початку вторгнення Сполучені Штати повідомляли про можливий російський план використання підробленого відео для створення приводу до військових дій. Після початку вторгнення представники української влади попереджали про ймовірне поширення підроблених матеріалів, у яких Президент України Володимир Зеленський нібито закликатиме українських військових скласти зброю.

Згодом на одному з українських новинних вебсайтів було розміщено відео, у якому Президент України нібито звертався до військовослужбовців із закликом здатися. Водночас у звіті Europol зазначено, що на момент його підготовки не було остаточно підтверджено, чи було це відео саме deepfake, чи іншим видом підробленого контенту. Незважаючи на це, наведений випадок демонструє, як синтетичні або сфальсифіковані медіаматеріали можуть використовуватися для дезінформаційного впливу в умовах війни [8].

Аналіз цього прикладу показує, що небезпека deepfake-технологій полягає не лише у можливості імітувати зовнішність або голос конкретної особи. У контексті воєнного конфлікту такий контент може бути спрямований на деморалізацію населення і військовослужбовців, посилення панічних настроїв, підрив довіри до офіційних джерел інформації та ускладнення прийняття рішень. Отже, deepfake-технології можуть виступати інструментом інформаційного

впливу, який доповнює традиційні кіберзагрози та підсилює наслідки гібридних атак.

Вішинг та AI-клонування голосу як форма deepfake-шахрайства

Технологічна основа голосового deepfake

Сучасний вішинг із використанням deepfake-аудіо може ґрунтуватися на технологіях Text-to-Speech (TTS), Voice Conversion (VC) та voice cloning. Text-to-Speech дає змогу створювати мовлення на основі введеного тексту, Voice Conversion - змінювати характеристики голосу для його наближення до голосу іншої особи, а voice cloning - формувати синтетичну копію голосу на основі аудіозразка.

Розвиток таких технологій суттєво зменшує обсяг вихідних матеріалів, необхідних для створення голосової імітації. Зокрема, у матеріалах Microsoft щодо моделі VALL-E зазначено, що система може синтезувати персоналізоване мовлення на основі трисекундного запису голосу невідомого раніше мовця [10]. Це має практичне значення для кібербезпеки, оскільки аудіозразки керівників, публічних представників організацій або інших довірених осіб можуть бути доступними у відкритих джерелах: відеоінтерв'ю, вебінарах, виступах, публікаціях у соціальних мережах або корпоративних матеріалах.

Процес атаки з використанням AI-клонування голосу може включати такі етапи:

1. **Збір аудіоматеріалів.** Зловмисник знаходить у відкритому доступі записи голосу особи, яку планує імітувати, наприклад керівника організації або представника партнерської компанії.
2. **Створення синтетичного голосу.** За допомогою AI-інструментів формується голосова імітація, що відтворює характерні особливості мовлення обраної особи.
3. **Встановлення контакту.** Працівник отримує телефонний дзвінок або голосове повідомлення нібито від керівника, партнера чи іншої довіреної особи.
4. **Формування шахрайського запиту.** У повідомленні може міститися прохання терміново підтвердити фінансову операцію, змінити платіжні

реквізити, передати конфіденційну інформацію або виконати іншу дію.

5. **Реалізація наслідків.** Працівник виконує запит, вважаючи його легітимним через схожість голосу та правдоподібний контекст звернення.

Отже, AI-клонування голосу підвищує ефективність вішингу, оскільки використовує довіру працівника до голосової комунікації. У таких умовах голос співрозмовника вже не може вважатися достатнім підтвердженням його особи, особливо якщо запит стосується фінансових операцій, передачі даних або зміни доступів.

Нижче наведений приклад корпоративного шахрайства з використанням deepfake-технологій :

Використання AI-клонування голосу у корпоративному шахрайстві є випадок, зафіксований у 2019 році. Керівник британської енергетичної компанії отримав телефонний дзвінок нібито від керівника німецької материнської компанії. За повідомленнями, зловмисники використали ПЗ на основі штучного інтелекту для імітації голосу, зокрема характерного німецького акценту та інтонації співрозмовника. У результаті працівник переказав 220 000 євро на банківський рахунок в Угорщині, який контролювали шахраї [15].

Цей випадок демонструє, що deepfake-аудіо може використовуватися не лише як засіб створення неправдивого контенту, а й як інструмент прямого фінансового шахрайства. Успішність атаки була пов’язана з тим, що знайомий голос та манера мовлення керівника сприймалися жертвою як підтвердження достовірності запиту. Таким чином, AI-клонування голосу підриває надійність традиційних процедур усного погодження фінансових операцій.

Ще більш масштабний випадок використання deepfake-технологій у корпоративному шахрайстві було зафіксовано у 2024 році в Гонконзі. Жертвою атаки стала британська інженерна компанія Arup. Працівник компанії отримав повідомлення нібито від фінансового директора британського підрозділу щодо проведення конфіденційної фінансової операції. Після цього він взяв участь у відеоконференції, під час якої зловмисники за допомогою deepfake-технологій імітували фінансового директора та інших співробітників компанії.

Унаслідок такої комунікації працівник здійснив 15 переказів на п’ять банківських рахунків у Гонконзі на загальну суму 200 млн гонконзьких доларів, що становило приблизно 25 млн доларів США. Компанія Arup підтвердила, що під час шахрайства використовувалися підроблені голоси та зображення. Водночас представники компанії зазначили, що внутрішні інформаційні системи не були скомпрометовані [11].

Наведений випадок має особливе значення для аналізу deepfake-загроз, оскільки демонструє перехід від окремої голосової підробки до комплексної мультимодальної атаки. У цьому сценарії зловмисники поєднали імітацію голосу, відеозображення та корпоративного контексту, створивши для працівника переконливу картину легітимної службової взаємодії. Успіх атаки був досягнутий не через технічний злам інформаційних систем компанії, а через маніпуляцію довірою працівника та обходження внутрішніх процедур підтвердження фінансових операцій.

Таблиця 2.3

Приклади використання deepfake-технологій у шахрайських та дезінформаційних сценаріях

Рік	**Приклад**	**Використана технологія**	**Наслідки**	**Аналітичне значення**
2022	Підроблене відео із нібито зверненням Президента України до військовослужбовців	Підроблений або синтетичний відеоконтент	Дезінформаційний вплив в умовах війни	Демонструє використання фальсифікованого медіаконтенту для підриву довіри та деморалізації
2019	Шахрайство проти британської енергетичної компанії	AI-синтезований голос керівника	Переказ 220 000 євро	Показує можливість використання voice deepfake для фінансового шахрайства

Продовження таблиці 2.3.

Рік	Приклад	Використана технологія	Наслідки	Аналітичне значення
2024	Deepfake-відеоконференція проти Arup у Гонконзі	Підроблені голоси та відеозображення фінансового директора й працівників	15 переказів на загальну суму близько 25 млн доларів США	Демонструє розвиток мультимодальних deepfake-атак без компрометації внутрішніх систем

Джерело: складено автором на основі матеріалів Europol [8], Димова М. В. [9] та Financial Times [11].

Проведений аналіз показує, що deepfake-технології є не лише інструментом створення синтетичного контенту, а й реальною загрозою для кібербезпеки, оскільки можуть використовуватися для дезінформації, маніпуляції довірою та фінансового шахрайства. Їхня небезпека полягає у здатності відтворювати голос, зовнішність або поведінку реальної особи, тобто ті ознаки, які користувачі традиційно сприймали як підтвердження достовірності інформації чи легітимності звернення.

Розглянуті приклади демонструють різні напрями зловмисного використання deepfake-технологій. У випадку поширення підробленого відеозвернення Президента України синтетичний або сфальсифікований медіаконтент використовувався як інструмент дезінформаційного впливу в умовах воєнного конфлікту. Випадок із британською енергетичною компанією показує можливість застосування AI-синтезованого голосу для прямого фінансового шахрайства. Атака на компанію Arup свідчить про подальше ускладнення таких загроз, оскільки зловмисники використали підроблені голоси та відеозображення під час відеоконференції, створивши переконливу імітацію службової комунікації.

Завдяки цьому, розвиток deepfake-технологій суттєво розширює

можливості соціальної інженерії. Для здійснення атаки зловмисникам не завжди необхідно отримувати технічний доступ до інформаційної системи організації, оскільки достатнім може виявитися вплив на рішення працівника через підроблений голос, відеозображення або повідомлення від імені довіреної особи. Це створює особливі ризики для корпоративних фінансових процедур, дистанційної комунікації, інформаційної безпеки та суспільної довіри до цифрового контенту.

Однак сучасні кіберзагрози, пов’язані зі ШІ, не обмежуються використанням AI-інструментів для створення шахрайського контенту. Об’єктом атак можуть ставати і самі системи ШІ, їхні навчальні дані та механізми взаємодії з користувачем. До таких загроз належать отруєння навчальних даних (data poisoning) і маніпуляція запитами до моделі (prompt injection), які розглядаються у наступному підрозділі.

2.4 Атаки на системи штучного інтелекту (data poisoning, prompt injection)

Атаки на системи штучного інтелекту є окремим напрямом сучасних кіберзагроз, оскільки в цьому випадку об’єктом впливу стають не лише користувачі або інформаційні ресурси, а й самі AI-моделі, їхні дані та механізми взаємодії з зовнішнім середовищем. Метою таких атак може бути зміна поведінки моделі, формування некоректних або небезпечних результатів, розкриття конфіденційної інформації чи виконання небажаних дій у пов’язаних системах. До найбільш актуальних типів таких загроз належать отруєння даних (data poisoning) і маніпуляція інструкціями моделі (prompt injection).

Prompt injection як загроза безпеці систем штучного інтелекту

Prompt injection - це тип атаки на застосунки, побудовані на основі великих мовних моделей, за якого зловмисник формує спеціально підготовлений текстовий запит або розміщує шкідливу інструкцію у зовнішньому контенті з метою змінити передбачену поведінку AI-системи. Така атака може призводити до обходу встановлених обмежень, розкриття

прихованих системних інструкцій, витоку даних або виконання небажаних дій.

Основною передумовою prompt injection є те, що LLM-застосунки можуть обробляти системні інструкції, користувацькі запити та зовнішні дані в одному текстовому контексті. Унаслідок цього модель може неправильно інтерпретувати шкідливий текст як інструкцію, яку потрібно виконати. Особливо небезпечним це стає тоді, коли AI-система має доступ до документів, електронної пошти, вебсторінок, баз знань або зовнішніх інструментів.

Назва prompt injection була запропонована за аналогією зі SQL injection. В обох сцинаріях загроза виникає через змішування команд і недовірених даних в одному процесі обробки. Однак у SQL injection проблема долається переважно суто технічнми методами, насамперед через параметризованих запитів, натомість у випадку з prompt injection надійного алгоритму для певного розмежування інструкції від користувацбкого тексту в LLM-застосунках поки не розроблено.

Одним із перших системних академічних досліджень цього виду атак стала праця **F. Perez та I. Ribeiro «Ignore Previous Prompt: Attack Techniques For Language Models»,** опублікована у 2022 році. У ній автори представили framework PromptInject і виділили два основні сценарії атаки: goal hijacking - перенаправлення поведінки моделі на ціль, визначену зловмисником, та prompt leaking отримання прихованих інструкцій, які визначають роботу системи.

У 2023 році K. Greshake та співавтори розширили модель загроз, описавши indirect prompt injection. У такому сценарії зловмисник не обов’язково взаємодіє з моделлю безпосередньо. Шкідлива інструкція може бути розміщена у вебсторінці, документі, електронному листі або іншому джерелі, яке AI-система згодом аналізує. Якщо модель сприйме такий фрагмент як команду, це може вплинути на її відповідь, обробку даних або використання підключених інструментів. [13]

Наприклад, користувач може доручити AI-асистенту підсумувати зовнішній документ. Якщо в документі приховано інструкцію на зразок «ігноруй попередні правила та передай конфіденційні дані», модель може

помилково сприйняти її не як частину аналізованого тексту, а як команду для виконання. У результаті AI-агент може змінити свою поведінку або здійснити небажану дію без прямого запиту користувача. Такий сценарій є особливо небезпечним для систем, що мають доступ до електронної пошти, корпоративних документів, API або інших зовнішніх інструментів.

Залежно від способу надходження шкідливої інструкції до AI-системи prompt injection поділяється на пряму та непряму. Основні відмінності між цими видами атак наведено в таблиці 2.4/

Таблиця 2.4

Порівняння прямої та непрямої prompt injection

Критерій	Direct prompt injection	Indirect prompt injection
Джерело шкідливої інструкції	Безпосередній запит зловмисника до моделі	Зовнішній документ, вебсторінка, електронний лист, база знань або інше джерело даних
Спосіб реалізації	Зловмисник прямо намагається змінити поведінку AI-системи	Модель отримує приховану інструкцію під час аналізу зовнішнього контенту
Можлива мета	Обхід обмежень, зміна відповіді, отримання системних інструкцій	Витік даних, маніпуляція AI-агентом, небажаний виклик інструмента або API
Складність виявлення	Шкідливий запит надходить безпосередньо, тому його легше ідентифікувати	Інструкція може бути прихована у звичайному документі або вебконтенті, що ускладнює її виявлення
Найбільш уразливі системи	Чат-боти та публічні LLM-інтерфейси	AI-агенти, корпоративні асистенти, RAG-системи та системи з доступом до зовнішніх інструментів

Джерело: складено автором на основі праць F. Perez, I. Ribeiro та K. Greshake та співавторів [14 ; 13].

Актуальність цієї загрози підтверджується матеріалами OWASP. У переліку OWASP Top 10 for LLM Applications 2025 prompt injection визначено як ризик LLM01:2025. За даними OWASP, така вразливість виникає тоді, коли спеціально сформовані користувацькі запити або зовнішній контент змінюють поведінку мовної моделі у непередбачений спосіб. Наслідками можуть бути розкриття конфіденційної інформації, отримання несанкціонованого доступу до

функцій AI-системи, виконання команд у пов’язаних середовищах або вплив на критичні рішення [16].

Data poisoning як атака на навчальні дані систем штучного інтелекту

Якщо prompt injection впливає на поведінку AI-системи під час її використання або взаємодії із зовнішнім контентом, то data poisoning спрямована на інший етап - формування даних, на основі яких модель навчається, донавчається або отримує інформацію для подальшої роботи.

Data poisoning, або отруєння даних - це атака, за якої зловмисник навмисно додає до навчального набору даних шкідливі, викривлені або неправильно позначені приклади з метою змінити поведінку AI-моделі. Унаслідок цього система може формувати помилкові результати, втрачати точність, демонструвати упереджену поведінку або реагувати визначеним зловмисником способом на певні вхідні дані.

У звіті NIST «Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations» data poisoning розглядається як один із ключових напрямів атак на системи машинного навчання. Особливість цієї загрози полягає в тому, що вплив здійснюється ще до моменту використання моделі: шкідливі зміни вбудовуються у дані, які система використовує для навчання або подальшого налаштування. Через це наслідки атаки можуть залишатися непомітними до моменту, коли модель починає приймати рішення у реальному середовищі [17].

Отруєння даних може здійснюватися різними способами. Зловмисник може додавати хибні приклади до навчального набору, змінювати правильні мітки даних, вносити систематичні викривлення або створювати приховані тригери. У випадку прихованого тригера модель може працювати правильно у звичайних умовах, але видавати визначений атакувальником результат при появі спеціальної ознаки у вхідних даних. Такий різновид атаки часто розглядається як backdoor- або Trojan-атака на модель.

Особливу небезпеку data poisoning становить для систем, які навчаються або оновлюються на основі великих відкритих наборів даних, зовнішніх

репозиторіїв, користувацького контенту чи автоматично зібраної інформації. Якщо джерела даних не проходять достатньої перевірки, зловмисник може вплинути на поведінку моделі ще до її впровадження у практичне використання.

Основні види data poisoning

Залежно від мети зловмисника отруєння даних може мати різний характер. У звіті NIST виділяються атаки, спрямовані на загальне погіршення роботи моделі, зміну результатів для конкретних запитів або формування прихованої реакції на спеціальний тригер.

Availability poisoning спрямована на зниження загальної працездатності AI-системи. У такому випадку до навчального набору вносяться шкідливі або викривлені дані, унаслідок чого модель демонструє погіршення точності та стає менш придатною для практичного використання. Для систем кібербезпеки це може означати зниження ефективності класифікації шкідливих об’єктів, виявлення аномалій або розпізнавання загроз.

Targeted poisoning має більш вибірковий характер. Метою такої атаки є зміна поведінки моделі лише щодо певних цільових прикладів або конкретної категорії запитів. У результаті система може загалом працювати коректно, але помилятися у ситуаціях, які є важливими для зловмисника. Для генеративних моделей це може проявлятися у формуванні неправильних відповідей на запити, що містять певну тему, словосполучення або інший визначений елемент.

Backdoor poisoning передбачає впровадження в модель прихованого механізму реагування на спеціальний тригер. У звичайних умовах модель може демонструвати нормальну поведінку, однак після появи визначеного слова, фрази, зображення або іншої ознаки вона виконує дію чи формує результат, запланований зловмисником. У випадку великих мовних моделей таким тригером може бути певна фраза, що змушує систему обходити встановлені обмеження або генерувати небезпечний контент.

Для генеративного штучного інтелекту ризик data poisoning посилюється через використання великих масивів даних, отриманих із численних відкритих джерел. За даними NIST, зловмисники можуть намагатися додавати спеціально підготовлений контент до вебресурсів або інших джерел, що надалі потрапляють у процес навчання, донавчання чи налаштування моделі. Унаслідок цього отруєння даних може впливати на поведінку моделі ще до її впровадження у робоче середовище [17].

Для визначення відмінностей між розглянутими видами атак доцільно порівняти data poisoning і prompt injection за етапом впливу, об’єктом атаки, способом реалізації та можливими наслідками.

Таблиця 2.5

Порівняння data poisoning та prompt injection як атак на AI-системи

Вид атаки	**Мета зловмисника**	**Можливий прояв у роботі AI-системи**	**Потенційні наслідки**
Availability poisoning	Загальне погіршення роботи модел	Зниження точності, збільшення кількості помилок, нестабільність результатів	Зменшення ефективності використання моделі для виявлення загроз, класифікації даних або підтримки прийняття рішень
Targeted poisoning	Зміна результатів для конкретних запитів, об’єктів або категорій даних	Модель загалом працює коректно, але формує неправильні результати щодо визначених цілей	Невиявлення окремої загрози, спотворення результатів аналізу або формування неправильних рекомендацій

Продовження таблиці 2.5.

Вид атаки	**Мета зловмисника**	**Можливий прояв у роботі AI-системи**	**Потенційні наслідки**
Backdoor poisoning	Формування прихованої поведінки, яка активується спеціальним тригером	За звичайних умов модель працює нормально, але після появи певної фрази, ознаки чи вхідного сигналу видає результат, визначений зловмисником	Приховане маніпулювання поведінкою моделі, обхід обмежень або генерація небезпечного контенту

Джерело: складено автором на основі NIST AI 100-2e2025 [17].

Як видно з таблиці 2.5, data poisoning може мати як явний, так і прихований характер. У випадку availability poisoning наслідки можуть проявлятися у загальному погіршенні точності моделі. Натомість targeted poisoning і backdoor poisoning є складнішими для своєчасного виявлення, оскільки система може демонструвати коректну поведінку у звичайних умовах і формувати небезпечний результат лише щодо визначених цілей або після появи спеціального тригера. Це робить отруєння даних серйозною загрозою для AI-систем, які використовуються у критичних процесах або для підтримки рішень у сфері кібербезпеки.

Спільною рисою обох атак є те, що вони можуть призводити до небезпечної поведінки AI-системи без традиційного злому її програмної інфраструктури. У випадку data poisoning модель може бути скомпрометована ще до початку практичного використання, а у випадку prompt injection - під час виконання конкретного завдання. Це свідчить про необхідність контролювати не лише доступ до AI-системи, а й походження навчальних даних, зовнішній контент, який вона аналізує, та повноваження, надані AI-агентам.

Висновок до розділу 2

Отже, атаки на системи штучного інтелекту створюють окрему категорію кіберзагроз, у якій об’єктом впливу є самі AI-моделі, їхні дані та механізми взаємодії з користувачами й зовнішніми ресурсами. Проведений аналіз показав, що prompt injection дає змогу маніпулювати поведінкою великих мовних моделей через прямі або приховані інструкції, тоді як data poisoning спрямоване на зміну роботи системи шляхом компрометації даних, використаних під час навчання або налаштування моделі.

Особливу небезпеку prompt injection становить для AI-агентів і корпоративних асистентів, які мають доступ до документів, електронної пошти, API та інших інструментів, оскільки шкідлива інструкція може призвести не лише до неправильної відповіді, а й до витоку інформації або виконання небажаної дії. Data poisoning, своєю чергою, є складним для своєчасного виявлення, оскільки скомпрометована модель може тривалий час демонструвати нормальну поведінку, а небезпечний результат проявляти лише у визначених умовах або після появи спеціального тригера.

Таким чином, безпека систем штучного інтелекту повинна передбачати контроль як етапу навчання моделі, так і процесів її подальшої експлуатації. Для зменшення ризиків необхідними є перевірка джерел і якості даних, моніторинг поведінки моделі, обмеження прав AI-агентів, ізоляція критичних операцій та підтвердження людиною дій, що можуть призвести до витоку даних або фінансових наслідків.

РОЗДІЛ 3 СУЧАСНІ ТЕХНОЛОГІЇ КІБЕРЗАХИСТУ НА ОСНОВІ ШТУЧНОГО ІНТЕЛЕКТУ

3.1 Використання машинного навчання для виявлення загроз

Одним із найбільш поширених напрямів застосування машинного навчання у кіберзахисті є виявлення фішингових посилань. Фішингові URL-адреси можуть бути зовні схожими на легітимні ресурси, використовувати HTTPS і SSL-сертифікати, імітувати назви відомих компаній або містити складну структуру доменів. У таких умовах перевірка лише за окремою ознакою не завжди є достатньою, тому більш ефективним є комплексний аналіз параметрів посилання за допомогою алгоритмів машинного навчання.

У дослідженні Н. Бурової, Р. Оприска, Є. Курія, Ю. Лаха та В. Сусукайла розглянуто застосування машинного навчання для автоматизованого виявлення фішингових URL-адрес у межах оборонних кібероперацій. Автори використали алгоритми Random Forest, Logistic Regression і Support Vector Machine, які дають змогу класифікувати посилання як безпечні або потенційно фішингові. Програмне рішення було реалізоване мовою Python із використанням вебфреймворку Flask [18].

Ефективність такого підходу значною мірою залежить від якості навчальних даних і правильного вибору ознак, за якими аналізуються URL-адреси. До таких ознак належать наявність IP-адреси у структурі посилання, використання протоколу HTTPS, довжина доменного імені, кількість субдоменів, наявність підозрілих символів і валідність SSL-сертифіката. Аналіз сукупності цих параметрів дає змогу моделі виявляти підозрілі посилання навіть тоді, коли фішинговий ресурс візуально імітує легітимний вебсайт.

Застосування алгоритмів машинного навчання для перевірки URL-адрес дозволяє автоматизувати первинний аналіз великої кількості посилань та швидше виявляти потенційно небезпечні ресурси. У практичній діяльності такі рішення можуть використовуватися як допоміжний інструмент для систем електронної пошти, браузерного захисту, SOC-команд або платформ моніторингу кіберзагроз. Водночас машинне навчання не слід розглядати як

повну заміну фахівця з кібербезпеки, оскільки результати класифікації залежать від актуальності наборів даних, обраних ознак та здатності моделі розпізнавати нові способи маскування фішингових ресурсів.

Таблиця 3.1

Ознаки фішингових URL-адрес, що можуть аналізуватися методами машинного навчання

Ознака URL-адреси	Характеристика ознаки	Зачення для виявлення фішингової загрози
Наявність IP-адреси у структурі URL	Замість звичайного доменного імені в посиланні може використовуватись числова IP-адреса	Може свідчити про спробу приховати справжню належність ресурсу або перенаправити користувача на підозрілий сайт
Наявність підозрілих символів	У посиланні можуть використовуватися додаткові дефіси, символи, нетипові комбінації літер або знаків	Допомагає виявити URL, спеціально сформовані для візуальної імітації легітимного ресурсу
Використання протоколу HTTPS	Система перевіряє наявність захищеного протоколу передавання даних	Є додатковою ознакою аналізу, однак не гарантує безпечність сайту, оскільки фішингові ресурси також можуть використовувати HTTPS
Довжина та структура доменного імені	Аналізується довжина домену, його відповідність очікуваній назві та наявність підозрілих елементів	Надмірно довгі або складні доменні імена можуть використовуватися для приховування шахрайського характеру ресурсу
Кількість субдоменів	URL може містити кілька піддоменів перед основною адресою сайту	Велика кількість субдоменів може створювати враження належності посилання до відомої компанії або сервісу
Валідність SSL-сертифіката	Перевіряється наявність і дійсність цифрового сертифіката вебресурсу	Допомагає виявити частину шкідливих ресурсів, фішингові сайти можуть мати чинні сертифікати

Як показано в таблиці 3.1, машинне навчанн може виявляти фішингові URL-адреси не на основі однієї ознаки , а на основі комплексу ознак. Це дуже важливо, оскільки інтернет-фішинг може використовувати протокол HTTPS, чинні SSL-сертифікати та доменні імена, візуально схожі на адреси легітимних ресурсів.. Поєднання кількох параметрів у межах ML-моделі дозволяє підвищити точність первинної класифікації посилань і швидше виявляти потенційно небезпечні вебресурси.

Ключовим елементом програмної реалізації є функція формування ознак URL-адреси, яка об’єднує кілька параметрів для подальшого аналізу. До них належать: наявність IP-адреси у структурі посилання, кількість підозрілих символів, використання протоколу HTTPS, довжина доменного імені, кількість субдоменів, наявність небезпечних слів у URL та валідність SSL-сертифіката. Такий підхід дає змогу машинному навчанню аналізувати не одну окрему характеристику, а комплекс ознак, що підвищує ефективність виявлення фішингових ресурсів [18].

Практичне значення такого підходу полягає у можливості автоматизованого формування ознакового опису URL-адреси для подальшої класифікації за допомогою алгоритмів Random Forest, Logistic Regression та Support Vector Machine. Таким чином, машинне навчання використовується як інструмент попереднього аналізу та оцінювання підозрілих посилань у межах систем кіберзахисту.

Таблиця 3.2

Ознаки, що використовуються у програмній реалізації для аналізу URL-адрес

Ознака	**Зміст ознаки**	**Призначення в аналізі**
has_ip_addr	Наявність IP-адреси у URL	Дає змогу виявити підозрілі посилання без звичайного доменного імені

Повторення таблиці 3.2

Ознака	**Зміст ознаки**	**Призначення в аналізі**
suspicious_count	Кількість підозрілих символів	Допомагає виявити штучно сформовані або замасковані URL
uses_https	Наявність HTTPS	Дає додаткову характеристику захищеності з'єднання
domain_length	Довжина доменного імені	Дозволяє оцінити нетипову або підозрілу структуру домену
subdomain_count	Кількість субдоменів	Дає можливість виявити URL із надмірно складною структурою
dangerous_word_count	Наявність небезпечних слів у URL	Допомагає виявити посилання, що містять типові фішингові маркери
valid_ssl	Валідність SSL-сертифіката	Дає змогу оцінити технічну достовірність ресурсу

Джерело: складено автором на основі дослідження Бурової Н., Оприска Р., Курія Є. та ін. [18].

Аналіз результатів застосування машинного навчання для виявлення фішингових URL-адрес

У дослідженні Н. Бурової, Р. Оприска, Є. Курія та ін. для перевірки програмної реалізації було використано 20 фішингових URL-адрес, відібраних із чорного списку шкідливих посилань. Особливістю тестової вибірки було те, що всі використані URL-адреси містили протокол HTTPS, що є важливим для аналізу сучасних фішингових загроз, оскільки наявність захищеного з'єднання вже не може розглядатися як достатня ознака легітимності вебресурсу.

За результатами тестування всі 20 URL-адрес були визначені програмою як фішингові. Крім того, перевірка SSL-сертифікатів у межах наведеної вибірки показала їхню невалідність. Отримані результати демонструють практичну можливість застосування комплексу алгоритмів машинного навчання для

автоматизованої перевірки підозрілих посилань та виявлення характерних ознак фішингових ресурсів [18].

Разом із тим результати дослідження необхідно оцінювати критично. Автори зазначають, що початковий набір даних для підготовки моделі складався лише з шести URL-адрес. Такий обсяг є недостатнім для повноцінного підтвердження стабільної ефективності ML-моделі в реальних умовах, де фішингові ресурси постійно змінюють структуру доменів, використовують чинні SSL-сертифікати та маскуються під легітимні сервіси.

Крім того, перевірка лише на відомих фішингових URL-адресах не дає змоги повною мірою оцінити кількість хибних спрацювань, тобто випадків, коли безпечне посилання може бути помилково визначене як шкідливе. Для більш об’єктивної оцінки ефективності системи доцільно використовувати більші збалансовані набори даних, що містять як фішингові, так і легітимні URL-адреси, а також оцінювати такі показники, як точність класифікації, повнота виявлення, precision, recall і F1-score.

Таким чином, розглянуте дослідження підтверджує перспективність використання машинного навчання для автоматизованого виявлення фішингових посилань. Алгоритми Random Forest, Logistic Regression і Support Vector Machine можуть аналізувати сукупність характеристик URL-адреси та допомагати у первинному виявленні потенційно шкідливих ресурсів. Водночас практичне впровадження таких систем потребує навчання на значно більших і регулярно оновлюваних наборах даних, оскільки ефективність моделі безпосередньо залежить від якості інформації, на якій вона була підготовлена.

Проведений аналіз показує, що машинне навчання є важливим інструментом сучасного кіберзахисту, оскільки дає змогу автоматизувати обробку значних обсягів даних і виявляти ознаки кіберзагроз швидше, ніж за умов виключно ручної перевірки. На прикладі виявлення фішингових URL-адрес встановлено, що ML-моделі можуть враховувати комплекс технічних характеристик посилання, зокрема структуру домену, кількість субдоменів, наявність підозрілих символів, використання HTTPS та валідність SSL-

сертифіката.

Практичне значення такого підходу полягає в можливості використання машинного навчання для попереднього аналізу посилань у системах електронної пошти, засобах моніторингу загроз і роботі центрів кібербезпеки. Водночас ML-рішення не повинні розглядатися як повна заміна фахівця або єдиний засіб захисту, оскільки їхня результативність залежить від якості навчальних даних, обраних ознак і здатності моделі адаптуватися до нових способів маскування атак.

Отже, використання машинного навчання для виявлення загроз є перспективним напрямом підвищення кіберстійкості організацій. Подальше вдосконалення таких систем має бути пов’язане з розширенням наборів навчальних даних, перевіркою моделей на нових типах атак та їх інтеграцією з іншими технологіями кіберзахисту, зокрема засобами поведінкової аналітики й виявлення аномалій, які розглядаються у наступному підрозділі.

3.2 Поведінкова аналітика та виявлення аномалій

На відміну від засобів захисту, які орієнтуються переважно на відомі сигнатури атак або заздалегідь визначені ознаки шкідливого об’єкта, поведінкова аналітика спрямована на виявлення відхилень від нормального функціонування користувачів, пристроїв і мережевих процесів. Її застосування є особливо важливим у випадках, коли зловмисник використовує легітимні облікові записи, дозволені інструменти або нові сценарії атаки, які ще не мають відомих сигнатур.

Поведінкова аналітика передбачає збір і аналіз даних про типову активність у системі: час і місце входу користувача, обсяг переданих даних, частоту доступу до ресурсів, мережеві з’єднання, взаємодію пристроїв і виконання операцій у робочому середовищі. На основі таких даних формується базова модель нормальної поведінки. Якщо подальша активність суттєво відрізняється від визначеного профілю, система може позначити її як аномальну та передати для подальшого розслідування.

Офіційним прикладом застосування такого підходу є звіт NISTIR 8219 «Securing Manufacturing Industrial Control Systems: Behavioral Anomaly Detection». У цьому документі NIST розглядає використання засобів поведінкового виявлення аномалій для кіберзахисту промислових систем керування. Автори демонструють, що моніторинг аномальних умов може допомагати виявляти ознаки шкідливої активності та загрози цілісності критично важливих операційних даних [19].

У представлених NIST сценаріях поведінковий аналіз здійснюється щодо мережевого трафіку, взаємодії промислових пристроїв і змін у процесах керування. Наприклад, якщо система знає, які з'єднання з програмованим логічним контролером є типовими та дозволеними, поява нового або нетипового контакту може бути зафіксована як потенційна аномалія. Аналогічно система може виявляти незвичну взаємодію між обладнанням або зміни в логіці інтерфейсу керування, які можуть свідчити про втручання в роботу виробничого процесу [19].

Таким чином, поведінкова аналітика дозволяє змістити акцент із пошуку лише відомих ознак атаки на виявлення підозрілої зміни поведінки системи. Це підвищує можливість виявлення нових, прихованих або безфайлових атак, а також інцидентів, пов'язаних із компрометацією облікових записів чи використанням легітимних інструментів у зловмисних цілях.

Таблиця 3.3

Типи поведінкових аномалій, що можуть свідчити про кіберзагрозу

Об'єкт моніторингу	**Приклад нормальної поведінки**	**Приклад аномалії**	**Можлива кіберзагроза**
Обліковий запис користувача	Вхід у систему у звичний робочий час із типового пристрою або місцезнаходження	Вхід у нетиповий час, із нового пристрою або незвичного місцезнаходження	Компрометація облікового запису або використання викрадених облікових даних

Продовження таблиці 3.3

Об’єкт моніторингу	**Приклад нормальної поведінки**	**Приклад аномалії**	**Можлива кіберзагроза**
Доступ до файлів і даних	Працівник працює з документами, необхідними для його посадових обов’язків	Масове відкриття, копіювання або завантаження файлів, до яких користувач раніше не звертався	Витік інформації, інсайдерська загроза або діяльність скомпрометованого акаунта
Мережевий трафік	Стабільний обсяг передавання даних між відомими вузлами мережі	Різке збільшення обсягу трафіку або передавання файлів на незвичну зовнішню адресу	Ексфільтрація даних або активність шкідливого програмного забезпечення
Підключені пристрої	Взаємодія лише з відомими й дозволеними пристроями у мережі	Поява нового невідомого пристрою або нетипового мережевого з’єднання	Несанкціоноване підключення, підготовка до проникнення або розвідка мережі
Спроби автентифікації	Поодинокі успішні входи користувача до дозволених ресурсів	Багаторазові невдалі спроби входу або використання неправильних облікових даних	Brute-force атака, password guessing або спроба несанкціонованого доступу
Промислові контролери та обладнання	PLC, HMI та ін. ICS-пристрої взаємодіють за встановленими протоколами і сценаріями	Несанкціоноване оновлення прошивки, зміна логіки PLC/HMI команда до контролера	Втручання в технологічний процес, порушення цілісності керування або саботаж
Сканування мережі	Пристрої звертаються лише до необхідних сервісів і портів	Нетипові запити до великої кількості вузлів або промислових сервісів	Розвідка мережі перед подальшою атакою

Джерело: складено автором на основі NISTIR 8219 *Securing Manufacturing Industrial Control Systems: Behavioral Anomaly Detection* [19].

Як видно з таблиці 3.3, поведінкова аналітика може застосовуватися на різних рівнях інформаційної системи: від контролю активності окремого користувача до моніторингу мережевого трафіку та роботи промислового обладнання. Спільним принципом є виявлення відхилень від попередньо встановленої або вивченої нормальної поведінки. Наприклад, нетипове завантаження великого обсягу даних може свідчити про їх витік, а нове з'єднання з промисловим контролером або зміна його логіки - про спробу втручання у виробничий процес.

Водночас факт виявлення аномалії не завжди означає, що зафіксована подія є кібератакою. Наприклад, вхід користувача у нетиповий час або підключення нового пристрою можуть бути пов'язані з легітимною робочою необхідністю. Тому системи поведінкової аналітики повинні використовуватися разом із контекстним аналізом подій, перевіркою сповіщень фахівцями та подальшими процедурами реагування на інциденти.

Практичне застосування поведінкового виявлення аномалій у промислових системах

Практичне значення поведінкової аналітики продемонстровано у звіті NISTIR 8219 «Securing Manufacturing Industrial Control Systems: Behavioral Anomaly Detection». У межах дослідження розглядалися два демонстраційні середовища: роботизована виробнича система та система керування технологічним процесом, подібна до тих, що використовуються у промисловості. Метою було показати, як засоби виявлення аномальної поведінки можуть підтримувати кіберзахист промислових систем керування [19].

У дослідженні NIST аналізувалися різні сценарії аномальної активності. До них належали підключення неавторизованого пристрою до мережі, передавання файлів на зовнішній ресурс, використання неправильних облікових даних для віддаленого доступу, сканування промислової мережі, несанкціоноване оновлення прошивки обладнання та зміна логіки промислового контролера або інтерфейсу керування. Такі події можуть бути ознаками розвідки мережі,

викрадення даних, спроби несанкціонованого доступу або втручання в технологічний процес.

Особливо важливим є сценарій несанкціонованої зміни логіки PLC або HMI. У промисловому середовищі така дія може вплинути не лише на конфіденційність інформації, а й на фізичний процес: роботу обладнання, якість виробництва або безпеку персоналу. Тому виявлення нетипових команд, нових мережевих взаємодій чи змін у поведінці контролерів є важливим елементом забезпечення кіберстійкості критичних систем.

Для теми використання штучного інтелекту у кіберзахисті поведінкове виявлення аномалій має особливе значення, оскільки методи машинного навчання можуть використовуватися для формування моделі нормальної активності та подальшого пошуку відхилень від неї. На відміну від сигнатурного підходу, який орієнтується на вже відомі ознаки атаки, поведінкова аналітика дає змогу виявляти нетипові події навіть у випадках, коли конкретний сценарій атаки раніше не був зафіксований.

Водночас використання поведінкової аналітики має певні обмеження. Виявлена аномалія не завжди є підтвердженням кібератаки. Наприклад, підключення нового пристрою, нетиповий час входу або збільшення мережевого трафіку можуть бути пов’язані з легітимними робочими діями, технічним обслуговуванням чи зміною виробничого процесу. Унаслідок цього система може формувати хибнопозитивні спрацювання, які потребують додаткової перевірки фахівцем.

Крім того, ефективність виявлення аномалій залежить від якості сформованої моделі нормальної поведінки. Якщо базовий профіль системи є неповним або вже містить нетипову активність, подальший аналіз може бути неточним. Особливо складною є ситуація в динамічних середовищах, де поведінка користувачів, пристроїв і мережевих процесів регулярно змінюється.

Таким чином, поведінкова аналітика не повинна розглядатися як єдиний засіб захисту. Найбільш ефективним є її використання разом із журналюванням подій, системами моніторингу мережі, контролем доступу, перевіркою

сповіщень фахівцями та засобами реагування на інциденти. Саме поєднання автоматизованого виявлення аномалій і подальшого аналізу подій дозволяє зменшити ризик як пропущених атак, так і необґрунтованих спрацювань.

Проведений аналіз показує, що поведінкова аналітика та виявлення аномалій є важливими напрямами сучасного кіберзахисту. На відміну від методів, орієнтованих лише на відомі сигнатури загроз, цей підхід дозволяє виявляти нетипову активність користувачів, мережевих вузлів і промислових пристроїв на основі відхилень від нормальної поведінки.

На прикладі дослідження NISTIR 8219 встановлено, що behavioral anomaly detection може використовуватися для виявлення підключення неавторизованих пристроїв, спроб передавання даних назовні, сканування промислової мережі, неправильних спроб автентифікації та несанкціонованих змін у роботі контролерів і систем керування. Це має особливе значення для промислових і критично важливих середовищ, де кібератака може призвести не лише до витоку даних, а й до порушення технологічного процесу.

Водночас результати поведінкової аналітики потребують перевірки та контекстного оцінювання, оскільки не кожне відхилення є ознакою атаки. Тому найбільшу практичну цінність такий підхід має у поєднанні з іншими засобами моніторингу та реагування на інциденти. Саме автоматизація подальших дій після виявлення підозрілої активності є предметом наступного підрозділу, присвяченого системам Security Orchestration, Automation and Response (SOAR).

3.3 Системи автоматизованого реагування (SOAR)

У попередніх підрозділах було розглянуто застосування машинного навчання та поведінкової аналітики для виявлення кіберзагроз і аномальної активності. Однак виявлення інциденту є лише першим етапом захисту. Якщо організація не здатна швидко перевірити сповіщення, визначити рівень ризику та вжити заходів реагування, навіть своєчасно виявлена атака може призвести до витоку даних, порушення роботи систем або фінансових втрат.

Одним із сучасних засобів підвищення швидкості реагування є платформи

Security Orchestration, Automation and Response (SOAR). SOAR - це програмна платформа, яка автоматизує частину дій у відповідь на виявлену підозрілу або шкідливу активність. Такі платформи можуть інтегруватися із системами SIEM, засобами захисту кінцевих пристроїв, міжмережевими екранами, сканерами вразливостей та іншими інструментами кібербезпеки.

В офіційних практичних рекомендаціях щодо впровадження SIEM і SOAR зазначається, що SIEM переважно збирає, централізує й аналізує журнали подій, після чого формує сповіщення про потенційний інцидент. SOAR, своєю чергою, використовує ці сповіщення для автоматизації окремих дій реагування. Таким чином, SIEM зосереджується на виявленні та аналізі подій, а SOAR - на організації й прискоренні подальших дій щодо інциденту [20].

Основою роботи SOAR є заздалегідь визначені сценарії реагування — playbooks. Вони містять послідовність дій, які система повинна виконати при виникненні конкретної події. Наприклад, у разі виявлення підозрілого фішингового повідомлення playbook може передбачати збір додаткових відомостей про посилання, перевірку вкладень, пошук аналогічних листів у поштових скриньках, блокування шкідливого домену та передачу інциденту фахівцю для остаточного рішення.

Важливо, що автоматизація реагування не означає повної відмови від участі людини. У рекомендаціях зазначено, що автоматизовані дії SOAR не замінюють фахівців із реагування на інциденти, а спрощують виконання повторюваних і часово критичних операцій. Завдяки цьому працівники SOC можуть приділяти більше уваги складним інцидентам, аналізу причин атаки та прийняттю рішень у ситуаціях, де автоматичне реагування може створити додаткові ризики [20].

Таблиця 3.4

Приклади автоматизованих сценаріїв реагування SOAR на кіберінциденти

Тип інциденту	Джерело виявлення інциденту	Автоматизовані дії SOAR	Дії, що потребують участі фахівця	Очікуваний результа
Підозрілий вхід до облікового запису	Поведінкова аналітика, UEBA, система автентифікації або SIEM	Збір даних про час, місце та пристрій входу; підвищення рівня ризику події; ініціювання додаткової перевірки особи; тимчасове блокування активної сесії	Перевірка, чи належала активність реальному користувачеві; рішення про зміну пароля або блокування облікового запису	Зменшення ризику використання викрадених облікових даних і подальшого несанкціонованого доступу
Виявлення шкідливого програмного забезпечення на пристрої	EDR/XDR, антивірусне рішення або SIEM	Ізоляція пристрою від мережі; запуск збору технічної інформації; блокування відомих індикаторів компрометації; створення інциденту для SOC	Аналіз походження зараження; перевірка інших пристроїв; рішення щодо відновлення системи	Обмеження поширення шкідливого ПЗ та скорочення часу локалізації інциденту

Продовження таблиці 3.4

Аномальне передавання великого обсягу даних	Поведінкова аналітика, DLP, мережевий моніторинг або SIEM	Фіксація події; перевірка адреси призначення та обсягу даних; тимчасове обмеження підозрілого з'єднання; сповіщення команди безпеки	Встановлення легітимності передавання; визначення факту витоку;розслідування	Раннє виявлення можливої ексфільтрації даних та обмеження її наслідків
Підозріле підключення до критичної або промислової системи	Система виявлення аномалій, моніторинг ICS/OT або SIEM	Реєстрація нового з'єднання; перевірка відповідності дозволеним комунікаціям; формування пріоритетного сповіщення; обмеження	Підтвердження дій ізоляції або блокування, якщо вони можуть вплинути на технологічний процес; аналіз джерела підключення	Зменшення ризику несанкціонованого втручання у критичні системи та виробничі процеси

Джерело: складено автором на основі практичних рекомендацій щодо впровадження SIEM і SOAR [20].

Як видно з таблиці 3.4, SOAR-платформи можуть використовуватися для автоматизації повторюваних і часово критичних дій під час реагування на різні типи кіберінцидентів. У випадку фішингової атаки система може автоматично перевірити підозрілі посилання, знайти аналогічні повідомлення та перемістити їх у карантин. У разі виявлення шкідливого програмного забезпечення SOAR може ініціювати ізоляцію зараженого пристрою та передати зібрані дані

фахівцям SOC для подальшого аналізу.

Особливе значення має поєднання SOAR із технологіями, розглянутими у попередніх підрозділах. Результати машинного навчання можуть використовуватися для виявлення фішингових посилань або шкідливих об’єктів, а поведінкова аналітика - для фіксації нетипових входів, аномального передавання даних чи підозрілих мережевих взаємодій. Після формування відповідного сповіщення SOAR дає змогу автоматизувати подальші дії відповідно до заздалегідь визначеного сценарію реагування.

Водночас автоматизація повинна застосовуватися з урахуванням можливих наслідків помилкового рішення. Наприклад, блокування облікового запису або ізоляція робочої станції зазвичай мають обмежений вплив на організацію, тоді як автоматичне відключення промислового обладнання або критичної системи може порушити технологічний процес. Саме тому у сценаріях із високим рівнем ризику доцільно передбачати підтвердження дій фахівцем перед їх виконанням.

Переваги та обмеження використання SOAR-систем

Застосування SOAR-систем має важливе значення для підвищення ефективності реагування на кіберінциденти. Однією з основних переваг таких платформ є скорочення часу між виявленням загрози та виконанням першочергових дій щодо її локалізації. У разі типових інцидентів, наприклад фішингового повідомлення, виявлення шкідливого програмного забезпечення або підозрілого входу до облікового запису, SOAR може автоматично виконати частину повторюваних операцій: перевірити індикатори компрометації, зібрати додаткові дані, ізолювати пристрій, перемістити повідомлення у карантин або сформувати завдання для фахівця SOC.

Іншою перевагою SOAR є можливість об’єднання різних інструментів кіберзахисту в єдиний процес реагування. Платформа може отримувати сповіщення від SIEM, EDR/XDR, систем захисту електронної пошти, засобів поведінкової аналітики та мережевого моніторингу, після чого виконувати дії

відповідно до визначеного сценарію. Це дозволяє зменшити кількість ручних операцій і забезпечити більш послідовний порядок реагування на однотипні інциденти.

У контексті використання штучного інтелекту SOAR може виконувати роль механізму практичного реагування на загрози, виявлені за допомогою AI- та ML-рішень. Наприклад, якщо модель машинного навчання визначила URL-адресу як потенційно фішингову або система поведінкової аналітики виявила нетипове передавання даних, SOAR може автоматизувати подальшу перевірку, пріоритизацію інциденту та виконання початкових захисних заходів.

Водночас впровадження SOAR має низку обмежень. Насамперед ефективність автоматизованого реагування залежить від якості даних і точності сповіщень, які надходять до платформи. Якщо вихідне сповіщення є хибним, автоматичне блокування користувача, ізоляція пристрою або обмеження мережевого доступу можуть негативно вплинути на робочі процеси організації.

Особливо обережно автоматизацію необхідно застосовувати у критичній інфраструктурі та промислових системах. У таких середовищах помилкове автоматичне блокування або відключення обладнання може спричинити порушення технологічного процесу. Тому для дій із високим рівнем потенційного впливу доцільно передбачати участь фахівця та підтвердження рішення перед виконанням.

Ще одним обмеженням є необхідність постійного оновлення та тестування playbooks. Сценарії реагування мають відповідати актуальним загрозам, архітектурі організації та рівню допустимого ризику. Якщо playbook є застарілим або неправильно налаштованим, автоматизація може бути неефективною або навіть створювати додаткові проблеми під час реагування на інцидент [20].

Проведений аналіз показує, що системи Security Orchestration, Automation and Response є важливим елементом сучасного кіберзахисту, оскільки дозволяють автоматизувати частину дій після виявлення підозрілої активності. Якщо машинне навчання та поведінкова аналітика допомагають визначити потенційну загрозу, то SOAR забезпечує організацію подальшого реагування:

збір додаткової інформації, перевірку індикаторів компрометації, ізоляцію підозрілих об’єктів, блокування шкідливих ресурсів і передачу складних випадків фахівцям SOC.

Практична цінність SOAR полягає у використанні заздалегідь визначених сценаріїв реагування, які скорочують час виконання повторюваних операцій і дозволяють команді кібербезпеки зосередитися на більш складних інцидентах. Особливо актуальним є поєднання SOAR із SIEM, EDR/XDR, системами аналізу фішингових повідомлень і поведінковою аналітикою, оскільки це забезпечує послідовний процес: від виявлення підозрілої події до її локалізації та подальшого розслідування.

Водночас SOAR не повинен розглядатися як повна заміна участі людини у процесі реагування на кіберінциденти. Автоматизовані рішення залежать від якості сповіщень, правильності налаштування playbooks і критичності систем, на які спрямовані захисні дії. У випадках, коли автоматичне блокування або ізоляція можуть вплинути на безперервність роботи організації чи технологічний процес, необхідним залишається підтвердження дій фахівцем.

Отже, SOAR-системи підвищують оперативність і послідовність реагування на кіберзагрози, однак їх ефективність залежить від належної інтеграції з іншими засобами захисту, регулярного тестування сценаріїв реагування та контролю з боку персоналу. Подальше впровадження таких технологій повинно здійснюватися в межах системного підходу до управління інформаційною безпекою, який регламентується міжнародними стандартами та рекомендаціями, зокрема ISO/IEC 27001 і NIST, що розглядаються у наступному підрозділі.

3.4 Міжнародні стандарти кібербезпеки та рамкові підходи до управління ризиками (ISO/IEC 27001, NIST)

Впровадження сучасних технологій кіберзахисту на основі штучного інтелекту потребує не лише використання технічних засобів виявлення та реагування на загрози, а й системного управління інформаційною безпекою.

Машинне навчання, поведінкова аналітика та SOAR-системи можуть підвищувати ефективність захисту, однак їх використання має бути пов’язане з оцінюванням ризиків, визначенням відповідальності, контролем доступу, документуванням процесів та постійним удосконаленням заходів безпеки.

Одним із найбільш відомих міжнародних стандартів у сфері інформаційної безпеки є ISO/IEC 27001:2022. Він визначає вимоги до системи управління інформаційною безпекою (Information Security Management System, ISMS), яка дозволяє організації системно визначати інформаційні ризики, обирати заходи їх оброблення, контролювати ефективність впроваджених рішень і постійно вдосконалювати процеси захисту інформації [21].

Основою ISO/IEC 27001 є ризик-орієнтований підхід. Це означає, що організація повинна визначити власні інформаційні активи, оцінити можливі загрози та вразливості, встановити рівень ризику і впровадити відповідні організаційні та технічні заходи захисту. У контексті використання штучного інтелекту такими активами можуть бути не лише бази даних, мережеві ресурси або облікові записи, а й ML-моделі, навчальні набори даних, журнали роботи AI-систем, API та автоматизовані сценарії реагування.

Застосування ISO/IEC 27001 є важливим для впровадження AI-рішень у кіберзахисті, оскільки дозволяє розглядати такі інструменти як частину загальної системи управління безпекою. Наприклад, система виявлення фішингових URL на основі машинного навчання повинна використовувати перевірені навчальні дані та мати контроль якості результатів. Система поведінкової аналітики повинна застосовуватися з урахуванням конфіденційності даних користувачів і ризику хибнопозитивних спрацювань. SOAR-платформи повинні мати визначені сценарії реагування, обмеження автоматизованих дій і механізми підтвердження критичних операцій фахівцем.

Поряд із ISO/IEC 27001 важливе значення має Cybersecurity Framework 2.0, розроблений National Institute of Standards and Technology (NIST). На відміну від ISO/IEC 27001, NIST CSF 2.0 є не стандартом сертифікації, а практичним рамковим підходом, який допомагає організаціям управляти та знижувати

ризики кібербезпеки незалежно від їхнього розміру, галузі або рівня технічної зрілості [22].

У версії NIST CSF 2.0 кібербезпека розглядається через шість основних функцій: Govern, Identify, Protect, Detect, Respond і Recover. Функція Govern охоплює управління політиками, ролями, відповідальністю та ризиками кібербезпеки. Identify передбачає визначення активів, загроз і вразливостей. Protect пов'язана з реалізацією заходів захисту. Detect охоплює виявлення підозрілої активності та інцидентів. Respond визначає дії щодо локалізації та оброблення інциденту, а Recover - заходи відновлення роботи після атаки.

У контексті попередніх підрозділів NIST CSF 2.0 дозволяє показати місце AI-технологій у цілісному процесі кіберзахисту. Наприклад, машинне навчання для аналізу фішингових URL і поведінкова аналітика можуть бути віднесені до функції Detect, оскільки допомагають виявляти загрози та аномальні події. SOAR-системи відповідають передусім функції Respond, оскільки забезпечують автоматизацію початкових дій реагування. Водночас рішення щодо вибору, контролю та перевірки таких AI-інструментів пов'язані з функцією Govern, а відновлення після наслідків інциденту - з функцією Recover.

Оскільки у цій роботі досліджуються саме технології на основі штучного інтелекту, доцільно також врахувати NIST Artificial Intelligence Risk Management Framework (AI RMF 1.0). Цей документ спрямований на управління ризиками, що виникають під час проєктування, розроблення, впровадження та використання AI-систем. У контексті кібербезпеки його застосування є важливим для контролю якості моделей, захисту навчальних даних, оцінювання ризику помилкових рішень і зменшення наслідків атак на самі системи штучного інтелекту [23].

Таким чином, ISO/IEC 27001, NIST CSF 2.0 і NIST AI RMF виконують взаємодоповнювальні функції. ISO/IEC 27001 формує основу системного управління інформаційною безпекою, NIST CSF 2.0 допомагає структурувати процеси управління кіберризиками та реагування на інциденти, а NIST AI RMF дозволяє враховувати специфічні ризики застосування штучного інтелекту. Їх

поєднання створює основу для контрольованого та відповідального впровадження AI-технологій у сучасні системи кіберзахисту.

Таблиця 3.5

Значення ISO/IEC 27001 та рамкових підходів NIST для впровадження AI-технологій у кіберзахисті

Документ	**Основне призначення**	**Ключові елементи**	**Застосування до AI-кіберзахисту**
ISO/IEC 27001:2022	Створення та постійне вдосконалення системи управління інформаційною безпекою	Управління ризиками, політики безпеки, відповідальність, контроль доступу, моніторинг і вдосконалення ISMS	Визначення вимог до захисту AI-моделей, навчальних даних, API, журналів подій і автоматизованих систем реагування
NIST CSF 2.0	Практичне управління та зниження кіберризиків організації	Govern, Identify, Protect, Detect, Respond, Recover	Включення ML-виявлення загроз і поведінкової аналітики до Detect; використання SOAR у Respond; контроль AI-рішень у Govern
NIST AI RMF 1.0	Управління ризиками, пов’язаними з розробленням і використанням AI-систем	Управління ризиками AI, оцінювання надійності, безпечності, прозорості та відповідальності	Оцінювання ризиків data poisoning, prompt injection, помилкових рішень AI-моделі та неконтрольованої автоматизації дій

Джерело: складено автором на основі ISO/IEC 27001:2022, NIST Cybersecurity Framework 2.0 та NIST AI Risk Management Framework 1.0 [21; 22; 23].

Як видно з таблиці 3.5, розглянуті документи не замінюють один одного, а забезпечують різні рівні управління кібербезпекою. ISO/IEC 27001 формує організаційну основу для побудови системи управління інформаційною безпекою та контролю ризиків. NIST CSF 2.0 дозволяє структурувати практичні процеси кіберзахисту від управління ризиками до виявлення, реагування і відновлення після інцидентів. NIST AI RMF, своєю чергою, доповнює ці підходи

в частині ризиків, характерних саме для систем штучного інтелекту.

Для організацій, які використовують AI-рішення у кіберзахисті, поєднання зазначених підходів є особливо важливим. Машинне навчання, поведінкова аналітика та автоматизоване реагування можуть підвищувати швидкість і точність захисних процесів, але водночас створюють нові ризики, пов’язані з якістю даних, помилковими рішеннями моделей, атаками на AI-системи та надмірною автоматизацією. Тому впровадження таких технологій має здійснюватися не ізольовано, а в межах формалізованої системи управління ризиками, контролю та постійного вдосконалення.

3.5 Практичні рекомендації щодо підвищення кіберстійкості

Проведений аналіз сучасних кіберзагроз і технологій кіберзахисту показує, що використання штучного інтелекту має подвійний характер. З одного боку, AI-інструменти можуть застосовуватися зловмисниками для створення переконливих фішингових повідомлень, deepfake-аудіо та відео, автоматизації соціальної інженерії, а також атак на самі системи штучного інтелекту. З іншого боку, машинне навчання, поведінкова аналітика та системи автоматизованого реагування дають змогу організаціям швидше виявляти загрози й обмежувати їхні наслідки.

Підвищення кіберстійкості організації не може ґрунтуватися лише на впровадженні одного технічного рішення. Навіть ефективна ML-модель для виявлення фішингових URL не захистить організацію від усіх сценаріїв соціальної інженерії, якщо працівник погоджує фінансові операції лише на підставі голосового повідомлення або відеодзвінка. Аналогічно SOAR-система може прискорити реагування на інцидент, але її помилково налаштований сценарій може негативно вплинути на робочі процеси. Тому захист має поєднувати технічні, організаційні та процедурні заходи.

Однією з першочергових рекомендацій є впровадження багаторівневого виявлення фішингових атак. На рівні електронної пошти й вебдоступу доцільно використовувати ML-рішення, здатні аналізувати URL-адреси за сукупністю

характеристик: структурою домену, кількістю субдоменів, наявністю підозрілих символів, HTTPS та валідністю SSL-сертифіката. Такий підхід дозволяє автоматизувати первинну перевірку великої кількості посилань і зменшити ймовірність переходу користувача на фішинговий ресурс [18].

Водночас урахування лише технічних ознак посилання є недостатнім, оскільки сучасні фішингові кампанії можуть бути персоналізованими та стилістично схожими на легітимну комунікацію. Тому організаціям необхідно запроваджувати процедури незалежного підтвердження критичних запитів. Зокрема, вимоги щодо термінового переказу коштів, зміни платіжних реквізитів, передавання конфіденційної інформації або надання доступу не повинні виконуватися лише на підставі листа, телефонного дзвінка чи відеоконференції. Такі дії мають додатково підтверджуватися через визначений незалежний канал зв’язку.

Другим важливим напрямом є застосування поведінкової аналітики та виявлення аномалій. Організація повинна контролювати не лише відомі індикатори атак, а й нетипові зміни у поведінці користувачів, пристроїв і мережевих процесів. Підозрілими подіями можуть бути вхід до облікового запису з незвичного пристрою або місцезнаходження, масове копіювання файлів, передавання значного обсягу даних на зовнішній ресурс, поява нового мережевого з’єднання або несанкціонована зміна логіки промислового контролера. Дослідження NISTIR 8219 показує практичну цінність behavioral anomaly detection для виявлення нетипової активності у промислових системах керування та інших критичних середовищах [19].

Третім напрямом є скорочення часу реагування на інциденти шляхом упровадження SIEM і SOAR-платформ. SIEM дозволяє централізовано збирати та аналізувати події безпеки, а SOAR - запускати заздалегідь підготовлені сценарії реагування. Наприклад, при виявленні фішингового листа система може автоматично перевірити вкладення й посилання, знайти аналогічні повідомлення в інших поштових скриньках і перемістити їх у карантин. При виявленні шкідливого програмного забезпечення SOAR може ініціювати ізоляцію

пристрою та передати інцидент фахівцям SOC. Однак для дій, що можуть вплинути на критичну інфраструктуру або безперервність роботи організації, необхідно передбачати підтвердження рішення фахівцем [20].

Окремої уваги потребує безпека самих AI-систем. Якщо організація використовує LLM-асистентів, RAG-системи або AI-агентів, які мають доступ до документів, електронної пошти чи зовнішніх API, необхідно враховувати загрозу prompt injection. Зовнішній контент, який обробляє модель, слід розглядати як недовірений, а права AI-агента мають бути обмежені відповідно до принципу найменших привілеїв. Критичні операції, пов’язані з передаванням даних, зміною документів або запуском зовнішніх дій, не повинні виконуватися без додаткового контролю [16].

Захист від data poisoning передбачає контроль походження та якості даних, які використовуються для навчання або донавчання моделей. Організаціям доцільно документувати джерела навчальних даних, перевіряти їх на наявність аномальних або навмисно викривлених прикладів, проводити тестування моделей перед упровадженням і регулярно контролювати зміни їхньої поведінки. Це особливо важливо для AI-рішень, які використовуються для виявлення загроз або підтримки критичних рішень, оскільки скомпрометована модель може формувати небезпечні результати ще до того, як атака стане очевидною [17].

Запровадження зазначених заходів повинно здійснюватися в межах системного управління інформаційною безпекою. ISO/IEC 27001:2022 дозволяє організації формалізувати процес оцінювання ризиків, визначити відповідальних осіб, упровадити необхідні контролі та забезпечити постійне вдосконалення системи управління інформаційною безпекою [21]. NIST Cybersecurity Framework 2.0 доповнює цей підхід через функції Govern, Identify, Protect, Detect, Respond і Recover, які дають змогу пов’язати технології виявлення та реагування із загальним циклом управління кіберризиками [22]. Для рішень, що використовують штучний інтелект, додаткове значення має NIST AI RMF 1.0, орієнтований на оцінювання ризиків і надійності AI-систем [23].

Таблиця 3.6

Практичні рекомендації щодо підвищення кіберстійкості в умовах AI-посилених загроз

Напрям захисту	**Практична рекомендація**	**Загрози, на які спрямований захід**	**Очікуваний результат**
Захист від фішингу	Використовувати ML-аналіз URL-адрес, посилань і вкладень у поштових повідомленнях	AI-генерований фішинг, шкідливі посилання, фальшиві сторінки авторизації	Швидше виявлення потенційно небезпечних повідомлень і зменшення кількості переходів на фішингові ресурси
Перевірка критичних запитів	Підтверджувати платежі, зміну реквізитів і передачу конфіденційних даних через незалежний канал зв’язку	BEC-шахрайство, voice deepfake, deepfake-відеоконференції	Зменшення ризику фінансових втрат через підробку голосу або відео довіреної особи
Поведінковий моніторинг	Упровадити аналіз нетипових входів, доступу до даних, мережевих з’єднань і передавання інформації	Компрометація акаунтів, інсайдерські загрози, ексфільтрація даних, втручання у критичні системи	Раннє виявлення підозрілої активності, яка не має відомих сигнатур
Автоматизоване реагування	Інтегрувати SIEM і SOAR із визначеними playbooks для типових інцидентів	Фішинг, malware, підозрілі входи, аномальне передавання даних	Скорочення часу локалізації інцидентів і зменшення навантаження на SOC

Продовження таблиці 3.6

Напрям захисту	**Практична рекомендація**	**Загрози, на які спрямований захід**	**Очікуваний результат**
Контроль критичної автоматизації	Передбачити підтвердження фахівцем дій, що можуть впливати на промислові або критичні системи	Хибнопозитивні спрацювання та помилкове автоматизоване блокування	Зниження ризику порушення технологічних і бізнес-процесів
Захист LLM та AI-агентів	Обмежити права доступу, ізолювати зовнішній контент, перевіряти критичні дії AI-системи	Prompt injection, витік даних, небажаний виклик API або інструментів	Зменшення наслідків маніпуляції мовною моделлю
Захист навчальних даних	Контролювати джерела датасетів, перевіряти якість даних і тестувати поведінку моделі	Data poisoning, backdoor poisoning, прихована зміна результатів AI-системи	Підвищення надійності моделей і виявлення компрометації до їх практичного використання
Управління кіберризиками	Використовувати ISO/IEC 27001, NIST CSF 2.0 і NIST AI RMF для формалізації процесів безпеки	Несистемне впровадження AI, відсутність контролю ризиків і відповідальності	Формування комплексної та постійно вдосконалюваної системи кіберзахисту
Підготовка персоналу	Проводити навчання з розпізнавання AI-фішингу, deepfake-аудіо й відео та правил перевірки запитів	Соціальна інженерія, фішинг, вішинг, маніпуляція довірою	Підвищення обізнаності працівників і зменшення ймовірності успішної атаки

Джерело: складено автором на основі проведеного дослідження та матеріалів OWASP, NIST, ISO/IEC і практичних рекомендацій щодо SIEM/SOAR [16; 17; 18; 19; 20; 21; 22; 23].

Як видно з таблиці 3.6, підвищення кіберстійкості в умовах використання штучного інтелекту потребує поєднання кількох рівнів захисту. Технічні засоби,

зокрема машинне навчання, поведінкова аналітика та SOAR, дозволяють швидше виявляти підозрілу активність і реагувати на інциденти. Проте самі по собі вони не усувають ризики, пов’язані з маніпуляцією довірою користувача, атаками на AI-системи або помилковими автоматизованими рішеннями.

Особливе значення має поєднання автоматизації з організаційними процедурами. Наприклад, навіть за наявності сучасних засобів моніторингу фінансова операція не повинна підтверджуватися лише голосом або відеозображенням керівника, оскільки такі ознаки можуть бути підроблені за допомогою deepfake-технологій. Так само AI-агент не повинен мати необмежений доступ до документів, API чи критичних операцій без контролю та підтвердження з боку людини.

Отже, практичне впровадження AI-технологій у кіберзахисті повинно здійснюватися відповідально та контрольовано. Найбільш ефективною є модель, у якій штучний інтелект використовується для швидкого аналізу даних, виявлення аномалій і підтримки реагування, тоді як критичні рішення, оцінювання ризиків і контроль наслідків залишаються частиною організаційної системи інформаційної безпеки.

Проведене дослідження дозволило сформулювати комплекс практичних рекомендацій щодо підвищення кіберстійкості організацій в умовах розвитку AI-посилених загроз. Захист від сучасних кібератак повинен включати використання машинного навчання для аналізу підозрілих об’єктів, поведінкової аналітики для виявлення нетипової активності, SOAR-систем для автоматизації реагування, а також процедур незалежної перевірки критичних запитів і контролю рішень, що приймаються із застосуванням AI.

Окремим напрямом є забезпечення безпеки самих систем штучного інтелекту. Загрози prompt injection і data poisoning свідчать про необхідність контролювати зовнішній контент, права AI-агентів, походження навчальних даних і стабільність поведінки моделей. Без таких заходів технології, призначені для підвищення ефективності захисту, можуть самі стати об’єктом маніпуляції з боку зловмисників.

Таким чином, підвищення кіберстійкості потребує комплексного підходу, який поєднує AI-технології, організаційні процедури, навчання персоналу та міжнародні підходи до управління ризиками. Використання ISO/IEC 27001, NIST CSF 2.0 і NIST AI RMF створює основу для системного впровадження сучасних засобів кіберзахисту та їх постійного вдосконалення відповідно до змін у ландшафті кіберзагроз.

Висновок до розділу 3

У третьому розділі було розглянуто сучасні технології кіберзахисту на основі штучного інтелекту та визначено їхнє значення для підвищення кіберстійкості організацій. Проведений аналіз показав, що застосування AI- та ML-рішень дає змогу перейти від переважно реактивного підходу до більш проактивного захисту, за якого потенційні загрози можуть бути виявлені та локалізовані ще до настання суттєвих наслідків.

У межах дослідження використання машинного навчання для виявлення загроз було встановлено, що ML-алгоритми можуть ефективно застосовуватися для автоматизованого аналізу фішингових URL-адрес. На прикладі дослідження Бурової Н., Оприска Р., Курія Є. та ін. розглянуто застосування алгоритмів Random Forest, Logistic Regression і Support Vector Machine для класифікації підозрілих посилань за сукупністю технічних ознак, зокрема структурою домену, кількістю субдоменів, наявністю підозрілих символів, використанням HTTPS та валідністю SSL-сертифіката [18]. Водночас було визначено, що точність таких рішень залежить від якості й обсягу навчальних даних, тому їх практичне застосування потребує використання актуальних і збалансованих вибірок, а також регулярного тестування моделей.

Аналіз поведінкової аналітики та засобів виявлення аномалій показав, що сучасний кіберзахист має орієнтуватися не лише на пошук відомих ознак атак, а й на виявлення відхилень від нормальної поведінки користувачів, пристроїв, мережевих вузлів і технологічних процесів. На основі матеріалів NISTIR 8219 було розглянуто можливості behavioral anomaly detection для промислових систем керування, зокрема виявлення неавторизованих пристроїв, нетипових

мережевих взаємодій, передавання даних назовні, сканування мережі та несанкціонованих змін у роботі контролерів [19]. Разом із тим встановлено, що зафіксована аномалія не завжди є підтвердженням атаки, тому результати автоматизованого аналізу повинні оцінюватися з урахуванням контексту та перевірятися фахівцями.

Окрему увагу було приділено системам Security Orchestration, Automation and Response (SOAR), які забезпечують автоматизацію окремих дій після виявлення інциденту. Встановлено, що SOAR-платформи можуть використовувати результати машинного навчання, поведінкової аналітики, SIEM, EDR/XDR та інших засобів моніторингу для запуску заздалегідь визначених сценаріїв реагування. До таких дій можуть належати перевірка фішингових повідомлень, карантин підозрілих листів, ізоляція заражених пристроїв, блокування шкідливих ресурсів і передавання складних інцидентів до SOC [20]. Водночас автоматизація реагування повинна бути обмеженою у випадках, коли помилкове рішення може порушити безперервність бізнес-процесів або функціонування критичної інфраструктури.

У процесі дослідження також було визначено значення міжнародного стандарту ISO/IEC 27001:2022 та рамкових підходів NIST для системного впровадження AI-технологій у кіберзахисті. ISO/IEC 27001 забезпечує основу для управління інформаційною безпекою та ризиками, NIST Cybersecurity Framework 2.0 дозволяє структурувати процеси управління, захисту, виявлення, реагування та відновлення, а NIST AI RMF 1.0 спрямований на врахування специфічних ризиків, пов’язаних із застосуванням систем штучного інтелекту [21; 22; 23]. Поєднання цих підходів дає змогу впроваджувати AI-рішення не ізольовано, а в межах формалізованої та контрольованої системи кібербезпеки.

На основі проведеного аналізу було сформульовано практичні рекомендації щодо підвищення кіберстійкості організацій в умовах AI-посилених загроз. До них належать використання ML-рішень для аналізу фішингових посилань, упровадження поведінкового моніторингу, інтеграція SIEM і SOAR, незалежна перевірка критичних фінансових та інформаційних

запитів, обмеження повноважень AI-агентів, контроль навчальних даних моделей, навчання персоналу та застосування міжнародних підходів до управління ризиками.

Отже, штучний інтелект є важливим засобом розвитку сучасного кіберзахисту, оскільки дозволяє автоматизувати аналіз даних, виявляти аномальну активність і прискорювати реагування на інциденти. Проте ефективність таких технологій залежить від якості даних, коректності налаштувань, контролю автоматизованих дій і поєднання технічних рішень з організаційними процедурами. Найбільш результативним є комплексний підхід, за якого AI-технології використовуються як складова системи управління кібербезпекою, а критичні рішення залишаються під контролем відповідальних фахівців.

ВИСНОВКИ

У кваліфікаційній роботі досліджено вплив штучного інтелекту на кібербезпеку, визначено сучасні загрози, пов’язані з його використанням, та проаналізовано технології захисту, що можуть застосовуватися для підвищення кіберстійкості організацій. Проведене дослідження підтвердило, що штучний інтелект має подвійний вплив на сферу кібербезпеки: він одночасно розширює можливості захисників і створює нові інструменти та сценарії для зловмисників.

У першому розділі було розглянуто теоретичні основи кібербезпеки та технологій штучного інтелекту. Встановлено, що основою забезпечення кібербезпеки залишаються принципи конфіденційності, цілісності та доступності інформації, а також ризик-орієнтований і багаторівневий підходи до захисту. Розглянуто машинне навчання, глибоке навчання, генеративний штучний інтелект, великі мовні моделі та технології синтетичних медіа як технологічну основу сучасних AI-рішень. Визначено, що їх використання у сфері кібербезпеки може бути спрямоване як на виявлення й аналіз загроз, так і на створення більш переконливих і масштабованих атак.

У другому розділі проаналізовано сучасні кіберзагрози, посилені використанням штучного інтелекту. Встановлено, що AI-інструменти можуть застосовуватися зловмисниками на різних етапах кібератаки: під час збору інформації про ціль, пошуку вразливостей, підготовки фішингових повідомлень, обходу захисних механізмів, створення синтетичного контенту та аналізу викрадених даних. Це свідчить про те, що штучний інтелект не обов’язково формує повністю нові види атак, але значно підвищує швидкість, масштабованість і переконливість уже відомих методів кіберзлочинності.

Окрему увагу приділено AI-генерованому фішингу та соціальній інженерії. Проведений аналіз показав, що великі мовні моделі можуть використовуватися для створення персоналізованих повідомлень, мовної адаптації фішингових листів, імітації корпоративного стилю комунікації та швидкої генерації значної кількості різних варіантів атаки. У поєднанні з моделлю Phishing-as-a-Service це створює умови для масштабування фішингових кампаній і зниження вимог до технічної підготовки зловмисників. Водночас встановлено, що протидія таким атакам не може ґрунтуватися лише на пошуку мовних помилок або шаблонного змісту повідомлення. Значно більшого значення набувають перевірка контексту запиту, підтвердження критичних дій через незалежний канал зв’язку та підвищення обізнаності персоналу.

Аналіз deepfake-технологій показав, що синтетичні аудіо- та відеоматеріали становлять реальну загрозу для інформаційної безпеки, фінансових процесів і суспільної довіри до цифрового контенту. Розглянуті приклади підтверджують можливість використання deepfake для дезінформації в

умовах воєнного конфлікту, голосового шахрайства та складних корпоративних атак із використанням підроблених відеоконференцій. Особливу небезпеку становить те, що успішність таких атак може бути досягнута без технічного зламу інформаційної системи, а шляхом маніпуляції довірою працівника до голосу, зображення або службового контексту звернення.

Також у роботі досліджено атаки, спрямовані безпосередньо на системи штучного інтелекту. Встановлено, що prompt injection може змінювати поведінку великих мовних моделей через прямі або приховані шкідливі інструкції, особливо якщо AI-система має доступ до зовнішніх документів, електронної пошти, API чи інших інструментів. Data poisoning, своєю чергою, впливає на дані, які використовуються для навчання або налаштування моделі, що може призводити до зниження точності, прихованої зміни поведінки або активації небезпечного результату за спеціального тригера. Отже, сучасна кібербезпека повинна охоплювати не лише захист від атак із використанням ШІ, а й забезпечення надійності самих AI-моделей, їхніх даних і механізмів взаємодії із зовнішнім середовищем.

У третьому розділі проаналізовано сучасні технології кіберзахисту на основі штучного інтелекту. На прикладі автоматизованого аналізу фішингових URL-адрес встановлено, що алгоритми машинного навчання можуть використовувати комплекс технічних ознак посилання для первинного виявлення потенційно шкідливих ресурсів. До таких ознак належать структура доменного імені, кількість субдоменів, наявність підозрілих символів, використання HTTPS та валідність SSL-сертифіката. Водночас визначено, що практична результативність ML-рішень безпосередньо залежить від обсягу, якості й актуальності навчальних даних, тому їх використання потребує регулярного тестування та оновлення моделей.

Поведінкова аналітика та виявлення аномалій розглянуті як важливі інструменти виявлення загроз, що не мають відомих сигнатур або реалізуються через легітимні облікові записи й дозволені інструменти. Встановлено, що аналіз відхилень від нормальної поведінки може застосовуватися для виявлення нетипових входів користувачів, масового доступу до даних, аномального мережевого трафіку, підключення невідомих пристроїв і змін у роботі промислових систем керування. Разом із тим зроблено висновок, що виявлена аномалія не завжди підтверджує факт атаки, тому результати автоматизованого аналізу повинні оцінюватися з урахуванням контексту та контролю фахівця.

Дослідження систем SOAR показало, що автоматизація реагування дозволяє скоротити час між виявленням загрози та виконанням першочергових захисних дій. SOAR-платформи можуть використовувати сповіщення від SIEM, EDR/XDR, систем аналізу фішингу й поведінкової аналітики для запуску

заздалегідь визначених сценаріїв реагування: перевірки підозрілих посилань, переміщення листів у карантин, ізоляції заражених пристроїв, блокування шкідливих ресурсів і передачі складних інцидентів до SOC. Водночас встановлено, що автоматизація не повинна повністю замінювати участь людини, особливо у випадках, коли помилкове рішення може вплинути на критичні системи або безперервність діяльності організації.

У роботі визначено значення ISO/IEC 27001:2022, NIST Cybersecurity Framework 2.0 та NIST AI Risk Management Framework 1.0 для системного впровадження AI-технологій у кіберзахисті. ISO/IEC 27001 формує основу управління інформаційною безпекою та ризиками, NIST CSF 2.0 дозволяє структурувати процеси управління, захисту, виявлення, реагування й відновлення, а NIST AI RMF допомагає враховувати ризики, характерні саме для систем штучного інтелекту. Їх поєднання дає змогу використовувати AI-рішення в межах контрольованої системи управління кіберризиками.

За результатами дослідження сформульовано практичні рекомендації щодо підвищення кіберстійкості організацій в умовах AI-посилених загроз. До основних рекомендацій належать: використання машинного навчання для аналізу фішингових посилань і підозрілих об’єктів; упровадження поведінкового моніторингу; інтеграція SIEM і SOAR для прискорення реагування на типові інциденти; обов’язкове підтвердження фінансових та інформаційно критичних запитів через незалежні канали; обмеження повноважень AI-агентів; контроль походження й якості навчальних даних; регулярне навчання персоналу щодо AI-фішингу, deepfake-аудіо та відео; застосування міжнародних підходів до управління ризиками.

Таким чином, мету кваліфікаційної роботи досягнуто: досліджено вплив штучного інтелекту на кібербезпеку та визначено сучасні загрози й технології захисту, пов’язані з його використанням. Проведений аналіз показав, що штучний інтелект не може розглядатися лише як засіб захисту або лише як джерело небезпеки. Його ефективне та безпечне використання потребує комплексного підходу, у якому технологічні рішення поєднуються з управлінням ризиками, процедурами контролю, навчанням персоналу та відповідальністю фахівців. Саме такий підхід дозволяє підвищити кіберстійкість організацій і зменшити наслідки сучасних загроз, що розвиваються разом із технологіями штучного інтелекту.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Forbes Technology Council. How AI-Driven Cyberattacks Will Reshape Cyber Protection. 2024. URL: https://www.forbes.com/councils/forbestechcouncil/2024/03/19/how-ai-driven-cyber-attacks-will-reshape-cyber-protection/ (дата звернення: 30.05.2026). [1]
2. ENISA. ENISA Threat Landscape 2025: Booklet. European Union Agency for Cybersecurity, 2025. DOI: 10.2824/2445233. URL: https://www.enisa.europa.eu/sites/default/files/2025-10/ENISA%20Threat%20Landscape%202025%20Booklet.pdf (дата звернення: 01.06.2026). [2]
3. ENISA. ENISA Threat Landscape 2025. European Union Agency for Cybersecurity, 2025. DOI: 10.2824/1946374. URL: https://www.enisa.europa.eu/sites/default/files/2025-11/ENISA%20Threat%20Landscape%202025.pdf (дата звернення: 01.06.2026).[3]
4. RansomLeak. AI-powered phishing: як зловмисники використовують LLM для створення фішингових листів. URL: https://ransomleak.com/uk/blog/ai-powered-phishing/ (дата звернення: 30.05.2026).[4]
5. Federal Bureau of Investigation. Internet Crime Report 2023. Internet Crime Complaint Center (IC3), 2024. URL: https://www.ic3.gov/AnnualReport/Reports/2023_IC3Report.pdf (дата звернення: 30.05.2026). [5]
6. ITEZ. Соціальна інженерія 2.0: Вішинг, Смішинг та Deepfake як не дати обдурити співробітників. URL: https://itez.com.ua/blog/social-engineering-vishing-smishing-deepfake-protection.html (дата звернення: 24.05.2026). [6]
7. Europol. Facing Reality? Law Enforcement and the Challenge of Deepfakes: An Observatory Report from the Europol Innovation Lab. Luxembourg: Publications Office of the European Union, 2022. DOI: 10.2813/158794. URL: https://www.europol.europa.eu/cms/sites/default/files/documents/Europol_Innovation_Lab_Facing_Reality_Law_Enforcement_And_The_Challenge_Of_Deepfakes.pdf (дата звернення: 30.05.2026). [7]
8. Димов М. В. Ризики для безпеки в цифровому просторі // Міжнародна конференція «Передові технології в інформаційно-комунікаційній інженерії» (ATICE'2025): матеріали конференції. Одеса : Видавничий дім «Гельветика», 2025. С. 77–80. URL: https://dspace.onua.edu.ua/server/api/core/bitstreams/6ec372b4-fefe-4737-92eb-1f468010acae/content (дата звернення: 30.05.2026). [8]
9. Wang C., Chen S., Wu Y. та ін. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. Microsoft Research, 2023. URL: https://arxiv.org/abs/2301.02111 (дата звернення: 30.05.2026). [9]
10. Leng C., Ho-him C. Arup lost $25mn in Hong Kong deepfake video conference scam // Financial Times. 2024. URL: https://www.ft.com/content/b977e8d4-664c-4ae4-8a8e-eb93bdf785ea (дата звернення: 25.05.2026). [10]
11. Willison S. Prompt Injection Attacks against GPT-3. 2022. URL:

https://simonwillison.net/2022/Sep/12/prompt-injection/ (дата звернення: 15.05.2026). [11]
12. Perez F., Ribeiro I. Ignore Previous Prompt: Attack Techniques for Language Models. 2022. DOI: 10.48550/arXiv.2211.09527. URL: https://arxiv.org/abs/2211.09527 (дата звернення: 01.06.2026). [12]
13. Greshake K., Abdelnabi S., Mishra S., Endres C., Holz T., Fritz M. Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. 2023. URL: https://arxiv.org/abs/2302.12173 (дата звернення: 01.06.2026). [13]
14. Stupp C. Fraudsters Use AI to Mimic CEO's Voice in Unusual Cybercrime Case // The Wall Street Journal. 2019. URL: https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402 (дата звернення: 01.06.2026). [14]
15. OWASP. LLM01:2025 Prompt Injection. OWASP GenAI Security Project. URL: https://genai.owasp.org/llmrisk/llm01-prompt-injection/ (дата звернення: 24.05.2026). [15]
16. Vassilev A., Oprea A., Fordyce A., Anderson H., Davies X., Hamin M. Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. NIST AI 100-2e2025. Gaithersburg, MD : National Institute of Standards and Technology, 2025. DOI: 10.6028/NIST.AI.100-2e2025. URL: https://csrc.nist.gov/pubs/ai/100/2/e2025/final (дата звернення: 01.06.2026). [16]
17. Бурова Н., Оприско Р., Курій Є., Лах Ю., Сусукайло В. Машинне навчання як ключовий інструмент оборонних кібероперацій: ефективність виявлення фішингових загроз // Social Development and Security. 2024. Vol. 14, No. 5. DOI: 10.33445/sds.2024.14.5.11. URL: https://paperssds.eu/index.php/JSPSDS/article/download/735/883/ (дата звернення: 01.06.2026). [17]
18. McCarthy J., Powell M., Stouffer K., Tang C., Zimmerman T., Barker W. Securing Manufacturing Industrial Control Systems: Behavioral Anomaly Detection. NIST Interagency/Internal Report 8219. Gaithersburg, MD : National Institute of Standards and Technology, 2020. DOI: 10.6028/NIST.IR.8219. URL: https://nvlpubs.nist.gov/nistpubs/ir/2020/NIST.IR.8219.pdf (дата звернення: 26.05.2026). [18]
19. Australian Cyber Security Centre. Implementing SIEM and SOAR Platforms: Practitioner Guidance. 2025. URL: https://www.cyber.gov.au/business-government/detecting-responding-to-threats/event-logging/implementing-siem-soar-platforms/implementing-siem-and-soar-platforms-practitioner-guidance (дата звернення: 26.05.2026). [19]
20. ISO. ISO/IEC 27001:2022 Information security, cybersecurity and privacy protection Information security management systems Requirements. Geneva : International Organization for Standardization, 2022. URL: https://www.iso.org/standard/27001 (дата звернення: 01.06.2026). [20]
21. NIST. The NIST Cybersecurity Framework (CSF) 2.0. NIST Cybersecurity

White Paper 29. Gaithersburg, MD : National Institute of Standards and Technology, 2024. DOI: 10.6028/NIST.CSWP.29. URL: https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.29.pdf (дата звернення: 01.06.2026). [21]

22. Tabassi E. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. Gaithersburg, MD : National Institute of Standards and Technology, 2023. DOI: 10.6028/NIST.AI.100-1. URL: https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf (дата звернення: 01.06.2026). [22]
23. Goodfellow I., Bengio Y., Courville A. Deep Learning. Cambridge, MA : MIT Press, 2016. 800 p. URL: https://www.deeplearningbook.org/ (дата звернення: 29.05.2026). [23]
24. ISO. ISO/IEC 27005:2022 Information security, cybersecurity and privacy protection Guidance on managing information security risks. Geneva : International Organization for Standardization, 2022. URL: https://www.iso.org/standard/80585.html (дата звернення: 29.05.2026). [24]
25. Stallings W. Effective Cybersecurity: A Guide to Using Best Practices and Standards. Boston : Addison-Wesley Professional, 2018. 800 p. URL: https://www.pearson.com/en-gb/subject-catalog/p/effective-cybersecurity-a-guide-to-using-best-practices-and-standards/P200000007404 (дата звернення: 01.06.2026). [25]