**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: “ЕТИЧНІ ПІДХОДИ ВПРОВАДЖЕННЯ ТЕХНОЛОГІЙ ШТУЧНОГО ІНТЕЛЕКТУ ДЛЯ КІБЕРЗАХИСТУ КРИТИЧНИХ СИСТЕМ”

на здобуття освітнього ступеня магістра
зі спеціальності 125 Кібербезпека та захист інформації
освітньо-професійної програми Управління інформаційною та кібернетичною безпекою

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

_____ Олександр КОДИМСЬКИЙ

*(підпис)* *Ім’я, ПРІЗВИЩЕ здобувача*

Виконав: здобувач вищої освіти гр. УБДМ-61
Олександр КОДИМСЬКИЙ

Керівник:
*д-р іст. наук, професор* Володимир ШУЛЬГА
Ім’я, ПРІЗВИЩЕ

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

**Київ 2025**

**ДЕРЖАВНИЙ УНІВЕРСИТЕТ**
**ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ**

**Навчально-науковий інститут кібербезпеки та захисту інформації**

Кафедра Управління кібербезпекою та захистом інформації
Ступінь вищої освіти магістр
Спеціальність 125 Кібербезпека та захист інформації
Освітньо-професійна програма Управління інформаційною та кібернетичною безпекою

**ЗАТВЕРДЖУЮ**
Завідувач кафедрою УКБЗІ
_____ Світлана ЛЕГОМІНОВА
"____" _____ 2025 р.

# ЗАВДАННЯ
# НА КВАЛІФІКАЦІЙНУ РОБОТУ

Студенту Кодимському Олександру Михайловичу
*(прізвище, ім'я, по батькові здобувача)*

1. Тема кваліфікаційної роботи: "Етичні підходи впровадження технологій штучного інтелекту для кіберзахисту критичних систем"

керівник кваліфікаційної роботи Володимир ШУЛЬГА, д-р іст. наук, професор.
*(Ім'я, ПРІЗВИЩЕ, науковий ступінь, вчене звання)*
затверджені наказом Державного університету інформаційно-комунікаційних технологій від "30" жовтня 2025 р. № 467.

2. Строк подання кваліфікаційної роботи "12" грудня 2025 р.
3. Вихідні дані до кваліфікаційної роботи: нормативно-правова база та стандарти, моделі загроз, вразливості систем, методики обробки й знеособлення даних, потенційні сценарії зловживань AI.
4. Перелік питань, які потрібно розробити:
   1. Дослідити теоретико-методологічні основи застосування технологій штучного інтелекту у сфері кіберзахисту критичних систем.
   2. Проаналізувати етичні виклики та методи впровадження AI у кіберзахист критичних систем.
   3. Визначити методику впровадження етичних підходів у розробку та експлуатацію AI-систем і розробити рекомендації для організацій щодо розробки політик та процедур впровадження етичного AI у кіберзахист.
5. Перелік ілюстративного матеріалу: *презентація*
6. Дата видачі завдання "02" жовтня 2025 р.

# КАЛЕНДАРНИЙ ПЛАН

| № з/п | Назви етапів кваліфікаційної роботи | Строк виконання етапів роботи | Примітка |
|---|---|---|---|
| 1. | Визначення об’єкту, предмету, мети та завдань дослідження. | 10.10.2025 | |
| 2. | Збір та аналіз літератури. | 23.10.2025 | |
| 3. | Дослідження теоретико-методологічних основ застосування технологій штучного інтелекту у сфері кіберзахисту критичних систем. | 27.10.2025 | |
| 4. | Аналіз етичних викликів та методів впровадження AI у кіберзахист критичних систем. | 10.11.2025 | |
| 5. | Визначення методики впровадження етичних підходів у розробку та експлуатацію AI-систем і розробка рекомендацій для організацій щодо розробки політик та процедур впровадження етичного AI у кіберзахист. | 15.11.2025 | |
| 6. | Формулювання висновків за результатами дослідження. | 22.11.2025 | |
| 7. | Оформлення роботи. | 04.12.2025 | |
| 8. | Оформлення презентації. | 14.12.2025 | |
| 9. | Отримання рецензії на роботу. | 18.12.2025 | |
| 10. | Захист в ЕК. | __.01.2026 | |

Здобувач вищої освіти _____ Олександр КОДИМСЬКИЙ
*(підпис)* *(Ім’я, ПРІЗВИЩЕ)*

Керівник
кваліфікаційної роботи _____ Володимир ШУЛЬГА
*(підпис)* *(Ім’я, ПРІЗВИЩЕ)*

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

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

Направляється здобувач Кодимський О.М. до захисту кваліфікаційної роботи
*(прізвище та ініціали)*
за спеціальністю 125 Кібербезпека та захист інформації
*(код, найменування спеціальності)*
Освітньо-професійної програми Управління інформаційною та кібернетичною безпекою
*(назва)*
на тему: "Етичні підходи впровадження технологій штучного інтелекту для кіберзахисту критичних систем"
Кваліфікаційна робота і рецензія додаються.

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

**Висновок керівника кваліфікаційної роботи**

Здобувач **КОДИМСЬКИЙ Олександр** у кваліфікаційній роботі дослідив теоретико-методологічні основи застосування технологій штучного інтелекту у сфері кіберзахисту критичних систем, проаналізував етичні виклики та методи впровадження AI у кіберзахист критичних систем, а також вивчив методику впровадження етичних підходів у розробку та експлуатацію AI-систем і розробив рекомендації для організацій щодо розробки політик та процедур впровадження етичного AI у кіберзахист.

**КОДИМСЬКИЙ Олександр** показав високу теоретичну і практичну підготовку, володіння науково-дослідницькими методами, вміння самостійно знаходити шляхи вирішення проблеми дослідження. Результати дослідження апробовані на конференції "Актуальні проблеми безпеки інформаційно-комунікаційних систем" 05 листопада 2025 року.

Все це дозволяє оцінити кваліфікаційну роботу здобувача **КОДИМСЬКОГО Олександра** на оцінку "добре" та присвоїти йому кваліфікацію "Магістр з кібербезпеки та захисту інформації за освітньо-професійною програмою Управління інформаційною та кібернетичною безпекою".

Керівник кваліфікаційної роботи _____ Володимир ШУЛЬГА
*(підпис)* *(Ім'я, ПРІЗВИЩЕ)*

"___" _____2025 року

**Висновок кафедри про кваліфікаційну роботу**

Кваліфікаційна робота розглянута. Здобувач Кодимський О.М. допускається до захисту даної роботи в Екзаменаційній комісії.

Завідувач кафедрою
Управління кібербезпекою та захистом
інформації _____ Світлана ЛЕГОМІНОВА
*(підпис)* *(Ім'я, ПРІЗВИЩЕ)*

**ВІДГУК РЕЦЕНЗЕНТА**
**на кваліфікаційну магістерську роботу**

здобувача вищої освіти Кодимського Олександра Михайловича
на тему “ Етичні підходи впровадження технологій штучного інтелекту для кіберзахисту критичних систем”

**Актуальність** У зв’язку з широким впровадженням технологій штучного інтелекту в управління та моніторинг критичних інформаційних інфраструктур зростає імовірність як значного підвищення ефективності кіберзахисту, так і появи нових ризиків, пов’язаних із порушенням приватності, упередженістю алгоритмів та непрозорістю прийняття рішень. Це робить нагальним розроблення етичних підходів і практик, які забезпечать баланс між автоматизацією захисту, дотриманням прав людини та стійкістю критичних систем до помилок і зловживань. Тема магістерської роботи спрямована на формалізацію цих підходів та вироблення рекомендацій для безпечного і відповідального застосування ШІ у кіберзахисті критичних об’єктів.

**Позитивні сторони**

Автором ґрунтовно досліджено теоретико-методологічні основи застосування штучного інтелекту у сфері кібербезпеки критичної інфраструктури, що дозволило систематизувати сучасні підходи та визначити ключові напрями використання AI у процесах виявлення, аналізу та реагування на кіберзагрози.

У роботі проведено комплексний аналіз основних етичних викликів, пов’язаних із впровадженням AI у кіберзахист критичних систем, зокрема проблем прозорості алгоритмів, відповідальності за прийняті автоматизовані рішення, захисту персональних даних та мінімізації ризиків упередженості моделей. На основі цього обґрунтовано доцільність застосування етичних принципів як невід’ємного елемента життєвого циклу AI-систем у сфері кібербезпеки.

Робота логічно структурована, характеризується аргументованістю висновків і відповідністю поставленим завданням та меті дослідження. За змістом, рівнем опрацювання матеріалу та отриманими результатами кваліфікаційна робота відповідає вимогам, що ставляться до магістерських робіт

**Недоліки**

Доцільно було б візуалізувати порядок впровадження рекомендацій для організацій щодо розробки політик та процедур впровадження етичного AI у кіберзахист. Однак, зауваження не впливає на загальну позитивну оцінку кваліфікаційної роботи.

**Висновок:** Кваліфікаційна робота виконана на належному науково-методичному рівні і заслуговує позитивної оцінки, а здобувач КОДИМСЬКИЙ Олександр Михайлович заслуговує присвоєння кваліфікації “Магістр кібербезпеки за освітньо-професійною програмою Управління інформаційною безпекою”.

*підпис*

Рецензент: доцент кафедри
Технічних систем кіберзахисту

к.т.н, доцент _____ Юрій ПЕПА

# РЕФЕРАТ

Текстова частина кваліфікаційної роботи на здобуття освітнього ступеня магістра: 73 стор., 14 рис., 11 табл., 72 джерел.

***Метою роботи*** є розробка науково-обґрунтованих рекомендацій щодо впровадження AI-технологій у системи кіберзахисту критичних об'єктів із дотриманням етичних принципів та вимог безпеки.

***Об'єктом дослідження*** є системи кіберзахисту критичних інформаційних інфраструктур та процеси впровадження технологій штучного інтелекту у їх управління.

***Предмет дослідження*** – етичні підходи, методи та практики інтеграції технологій штучного інтелекту у системи кіберзахисту критичних об'єктів.

***Методи дослідження***. Для вирішення етичних підходів впровадження технологій штучного інтелекту для кіберзахисту критичних систем використовуються теоретичні, емпіричні, прикладні та візуалізаційні методи.

***Короткий зміст роботи.*** Як результат у роботі досліджено теоретико-методологічні основи застосування технологій штучного інтелекту у сфері кіберзахисту критичних систем, проаналізовано етичні виклики та методи впровадження AI у кіберзахист критичних систем, а також вивчено методику впровадження етичних підходів у розробку та експлуатацію AI-систем і розроблено рекомендації для організацій щодо розробки політик та процедур впровадження етичного AI у кіберзахист.

***Галузь застосування.*** Розроблені етичні підходи можуть бути застосовані при плануванні та впровадженні систем кіберзахисту критичних інформаційних інфраструктур підприємств та організацій. Вони дозволяють інтегрувати технології штучного інтелекту в процеси забезпечення інформаційної безпеки з урахуванням етичних принципів.

**КЛЮЧОВІ СЛОВА:** ЕТИЧНІ ПІДХОДИ, ШТУЧНИЙ ІНТЕЛЕКТ, КІБЕРЗАХИСТ, КРИТИЧНІ СИСТЕМИ, СИСТЕМА УПРАВЛІННЯ ІНФОРМАЦІЙНОЮ БЕЗПЕКОЮ.

## ABSTRACT

The text part of the qualification work for obtaining a master's degree: 73 pages, 14 figures, 11 tables, 72 sources.

The purpose of the work is to develop scientifically sound recommendations for the implementation of AI technologies in cyber protection systems for critical facilities in compliance with ethical principles and security requirements.

***Object of research*** is cyber protection systems for critical information infrastructures and the processes of implementing artificial intelligence technologies in their management.

***Subjec***t ***of research*** is the ethical approaches, methods and practices for integrating artificial intelligence technologies into cyber protection systems for critical facilities.

***Research methods*** Theoretical, empirical, applied, and visualisation methods are used to address ethical approaches to the implementation of artificial intelligence technologies for the cyber protection of critical systems.

***Brief content of research***. As a result, the work examines the theoretical and methodological foundations of the application of artificial intelligence technologies in the field of cyber protection of critical systems, analyses the ethical challenges and methods of implementing AI in the cyber protection of critical systems, and the methodology for implementing ethical approaches in the development and operation of AI systems is studied, and recommendations are developed for organisations on the development of policies and procedures for implementing ethical AI in cyber security.

***Field of research***. The developed ethical approaches can be applied in the planning and implementation of cyber protection systems for critical information infrastructures of enterprises and organisations. They allow the integration of artificial intelligence technologies into information security processes, taking into account ethical principles.



**KEYWORDS:** ETHICAL APPROACHES, ARTIFICIAL INTELLIGENCE,

CYBER SECURITY, CRITICAL SYSTEMS, INFORMATION SECURITY MANAGEMENT SYSTEM,

# ЗМІСТ

# ВСТУП

***Актуальність теми.*** Критичні інформаційні системи підприємств, державних установ та інфраструктур потребують надійного захисту від кіберзагроз. Розвиток технологій штучного інтелекту та автоматизації відкриває нові можливості для забезпечення кіберзахисту, проте одночасно ставить завдання дотримання етичних принципів: прозорості рішень, відповідальності за автоматизовані дії та захисту персональних і конфіденційних даних.

Зростання складності та масштабу кіберзагроз робить необхідним інтегрування штучного інтелекту в системи захисту критичних інфраструктур, при цьому важливим є не лише технічний ефект, але й етична безпека – уникнення потенційних ризиків для користувачів та суспільства в цілому. Тому дослідження етичних підходів до впровадження технологій штучного інтелекту в кіберзахист критичних систем є актуальним науковим завданням, що сприяє розробці безпечних, ефективних та відповідальних рішень у сфері інформаційної безпеки.

***Мета роботи*** полягає у розробці науково-обґрунтованих рекомендацій щодо впровадження AI-технологій у системи кіберзахисту критичних об’єктів із дотриманням етичних принципів та вимог безпеки.

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

1. Дослідити теоретико-методологічні основи застосування технологій штучного інтелекту у сфері кіберзахисту критичних систем.

2. Проаналізувати етичні виклики та методи впровадження AI у кіберзахист критичних систем.

3. Визначити методику впровадження етичних підходів у розробку та експлуатацію AI-систем і розробити рекомендації для організацій щодо розробки політик та процедур впровадження етичного AI у кіберзахист.

***Об’єкт дослідження*** **–** системи кіберзахисту критичних інформаційних інфраструктур та процеси впровадження технологій штучного інтелекту у їх управління.

***Предмет дослідження*** – етичні підходи, методи та практики інтеграції технологій штучного інтелекту у системи кіберзахисту критичних об’єктів.

***Методи дослідження.*** Аналіз і синтез – для опрацювання наукових публікацій, стандартів (ISO/IEC, NIST), рекомендацій ЄС та міжнародних організацій щодо використання штучного інтелекту у кібербезпеці та формування цілісного уявлення про етичні підходи.

Системний та структурно-функціональний аналіз – для дослідження ролі AI у кіберзахисті критичних систем, визначення взаємозв’язків між технічними, організаційними та етичними компонентами системи безпеки.

Порівняльний аналіз – для зіставлення існуючих етичних моделей, фреймворків та підходів до впровадження етичного AI (trustworthy AI, responsible AI) у різних галузях та їх адаптації до сфери кіберзахисту критичної інфраструктури.

Метод експертних оцінок – для визначення значущості окремих етичних принципів (прозорість, підзвітність, недискримінація, безпечність) у процесі розробки та експлуатації AI-систем кіберзахисту.

Моделювання – для формування методики впровадження етичних підходів у життєвий цикл AI-систем кіберзахисту, з урахуванням етапів проєктування, навчання, тестування, впровадження та експлуатації.

***Наукова новизна*** роботи полягає в тому, що запропоновано комплексний підхід до впровадження етичних принципів у системи кіберзахисту критичних об’єктів на основі технологій штучного інтелекту за рахунок удосконалення формування політик і процедур етичного використання AI у кіберзахисті критичних систем.

***Практичне значення одержаних результатів.*** Застосування розроблених етичних підходів та моделей впровадження технологій штучного інтелекту дозволить організаціям здійснювати обґрунтований вибір методів і інструментів для кіберзахисту критичних систем. Результати дослідження можуть бути використані для інтеграції AI-технологій у процеси управління доступом, виявлення та нейтралізації загроз, а також для забезпечення захисту

даних за допомогою сучасних механізмів шифрування та моніторингу.

Запропоновані підходи також сприятимуть розробці політик кібербезпеки, які відповідають міжнародним стандартам та етичним нормам, забезпечуючи надійний захист критичних систем з урахуванням технічних можливостей, ресурсів організації та етичних вимог щодо використання штучного інтелекту.

***Апробація результатів*** кваліфікаційної роботи відбулася на науково-практичній конференції “Актуальні проблеми безпеки інформаційно-комунікаційних систем” 05 листопада 2025 року.

# РОЗДІЛ 1
# ТЕОРЕТИКО-МЕТОДОЛОГІЧНІ ОСНОВИ ЗАСТОСУВАННЯ ТЕХНОЛОГІЙ ШТУЧНОГО ІНТЕЛЕКТУ У СФЕРІ КІБЕРЗАХИСТУ КРИТИЧНИХ СИСТЕМ

Розвиток цифрового простору та зростання складності кіберзагроз зумовлюють необхідність переосмислення традиційних підходів до захисту критично важливих систем. У цьому контексті технології штучного інтелекту (ШІ) стають ключовим інструментом, здатним забезпечити проактивне виявлення, прогнозування та нейтралізацію атак, що постійно еволюціонують. Теоретико-методологічні засади використання ШІ у сфері кіберзахисту формують наукове підґрунтя для створення інтелектуальних систем оборони, які поєднують автоматизований аналіз, адаптивні моделі поведінки та високий рівень автономності. Розгляд цих основ дозволяє не лише окреслити потенціал сучасних технологій, а й визначити напрями їх ефективного впровадження в інфраструктуру критичних об’єктів.

## 1.1. Поняття критичних інформаційних інфраструктур та їх роль у національній безпеці.

Критичні інформаційні інфраструктури становлять сукупність технологічних, програмних та організаційних компонентів, від безперебійного функціонування яких безпосередньо залежить стабільність держави, її економічний розвиток, обороноздатність та суспільна безпека. У сучасній цифровизованій екосистемі ці інфраструктури інтегрують інформаційні потоки між різними секторами, забезпечуючи стійкий обмін даними та координацію процесів, що мають стратегічне значення [1]. До їх складу зазвичай включають енергетичні системи управління, транспортні та логістичні мережі, фінансові платформи, телекомунікаційні вузли, автоматизовані системи управління

промисловими комплексами, а також державні ресурси, пов’язані з наданням адміністративних послуг. Однак визначальною їх ознакою є не перелік об’єктів, а ступінь залежності суспільства і держави від їх нормального функціонування, а також рівень потенційних збитків у разі їх порушення.

Сучасні концепції національної безпеки трактують критичні інформаційні інфраструктури як ядро життєдіяльності країни, адже інформація стала не лише інструментом управління, а й ресурсом стратегічного значення. Кіберзагрози, спрямовані на порушення роботи таких інфраструктур, можуть призвести до ефекту каскадних відмов, коли вихід з ладу одного елементу тягне за собою системні збої в інших секторах [2]. Це зумовлює необхідність комплексного розуміння внутрішніх зв’язків між різними рівнями інфраструктури, їх технологічної взаємозалежності та соціально-економічних наслідків можливих атак. У глобальному масштабі спостерігається тенденція до зростання кількості інцидентів, що спрямовані саме на критичні об’єкти, а їх характер еволюціонує від простих порушень доступності до складних багатовекторних операцій, орієнтованих на маніпулювання даними, фізичне ураження обладнання та підрив загальнонаціональних процесів.

З огляду на це, критичні інформаційні інфраструктури стають об’єктом не лише технічного захисту, а й стратегічного планування, що передбачає оцінювання ризиків, прогнозування загроз та формування політики стійкості. Їхня роль у національній безпеці полягає не лише у забезпеченні оперативності управлінських рішень, а й у створенні умов для стабільного функціонування держави в умовах зовнішнього тиску чи надзвичайних ситуацій. У багатьох країнах сформовано спеціалізовані органи, відповідальні за координацію діяльності у сфері захисту критичних інфраструктур, що підкреслює їхнє стратегічне значення та необхідність централізованого управління під час кризових ситуацій.

Важливим аспектом концептуального розуміння критичних інформаційних інфраструктур є їхня високодинамічна природа. У процесі цифрової трансформації відбувається постійне розширення меж таких

інфраструктур: нові технології, мережеві архітектури, хмарні рішення, промисловий інтернет речей та автономні системи ускладнюють структуру об’єктів і водночас створюють додаткові вектори ураження. Це змінює традиційні уявлення про безпеку, адже інциденти можуть виникати не лише внаслідок зовнішніх атак, а й через технічні несправності, несанкціоновані втручання всередині організацій, помилки персоналу чи недосконаліть політик доступу [3]. Отже, розуміння критичних інфраструктур потребує міждисциплінарного підходу, що поєднує кібернетику, інженерію, системний аналіз, правові норми та управлінські методики. Еволюція цих інфраструктур вказана на рис. 1.1.

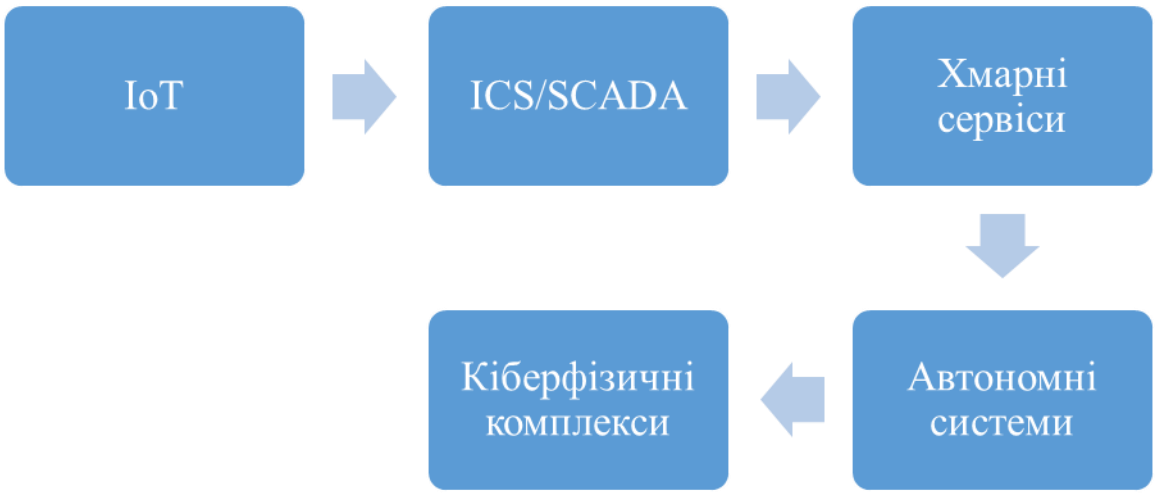


Рис. 1.1. Еволюція критичних інформаційних інфраструктур у процесі цифрової трансформації

Підвищення значущості критичних інформаційних інфраструктур також обумовлено тим, що їх стійкість стає фактором геополітичного впливу. Держави, здатні забезпечити безперебійний функціонал своїх систем, отримують конкурентні переваги у сфері міжнародних відносин, економічного розвитку та оборони [4]. У той же час уразливість критичних інфраструктур

може бути використана як інструмент тиску чи дестабілізації. Це особливо помітно в умовах гібридних конфліктів, де кібероперації стають повноцінним елементом стратегії впливу, а атаки на інформаційні системи можуть супроводжуватися інформаційно-психологічними, економічними чи політичними засобами впливу.

Критичні інформаційні інфраструктури є складними соціотехнічними системами, що не лише забезпечують функціонування держави, але й формують рівень її стійкості в умовах сучасних загроз [5]. Їхній захист становить один із фундаментальних елементів системи національної безпеки, а ефективне управління такими інфраструктурами потребує поєднання технологічних рішень, нормативно-правових механізмів, організаційних заходів та стратегічного прогнозування. Відповідно, глибоке розуміння їх ролі та специфіки функціонування є передумовою формування ефективної політики кіберзахисту в умовах зростаючої цифрової взаємозалежності та глобальної нестабільності. Модель управління стійкістю критичних інформаційних інфраструктур вказана на рис. 1.2.


Рис.1.2. Модель управління стійкістю критичних інформаційних інфраструктур

## 1.2. Сучасні загрози та виклики у сфері кіберзахисту критичних систем

Сучасний стан розвитку інформаційних технологій та інтеграції цифрових систем у критичні об’єкти державної та економічної інфраструктури створює нові умови для функціонування кіберпростору, що характеризуються високим рівнем взаємозалежності та швидкою трансформацією загроз. Високий рівень автоматизації та інтеграції інформаційних систем забезпечує ефективність управлінських процесів, проте одночасно збільшує вразливість критичних систем до різноманітних кібератак, що мають потенційно катастрофічні наслідки [6]. Розвиток технологій штучного інтелекту, Інтернету речей, хмарних сервісів та автономних систем не тільки підвищує функціональну складність інфраструктур, але й відкриває нові вектори атак, які вимагають комплексного аналізу та постійного вдосконалення стратегій кіберзахисту.

Однією з ключових проблем сучасного кіберзахисту є швидка еволюція загроз, що змінюються як за характером, так і за методами реалізації. Класичні загрози, пов’язані з несанкціонованим доступом до систем, заміщають більш складні багатовекторні атаки, які поєднують елементи соціальної інженерії, шкідливого програмного забезпечення, вразливостей промислових систем керування (ICS/SCADA) та мережевих архітектур [6]. Такі атаки не обмежуються локальними наслідками, а можуть мати масштабний каскадний ефект, порушуючи функціонування енергетичних, транспортних, фінансових та комунікаційних систем одночасно. Це ускладнює процес оперативного реагування та змушує розробників захисних рішень враховувати взаємозалежність різних компонентів критичної інфраструктури.

Особливу увагу у сучасних дослідженнях приділяють загрозам, що виникають унаслідок складних взаємодій кіберфізичних систем. Використання Інтернету речей у промислових і критично важливих об’єктах дозволяє підвищити ефективність виробничих процесів та управління ресурсами, проте одночасно створює нові уразливості [7]. Підключені пристрої, датчики,

виконавчі механізми та промислові контролери стають потенційними точками входу для кібератак, здатних змінювати фізичний стан об’єктів, порушувати технологічні процеси та спричиняти значні економічні та соціальні збитки. Відсутність стандартизованих протоколів безпеки, низький рівень оновлення програмного забезпечення та відсутність інтегрованих систем моніторингу роблять такі об’єкти особливо вразливими.

Крім технічних викликів, значну роль у формуванні сучасного ландшафту кіберзагроз відіграють соціальні та політичні фактори. Використання методів соціальної інженерії, фішингових атак, маніпуляцій персоналом та цільових інформаційних кампаній дозволяє зловмисникам обходити технічні бар’єри без прямого втручання у системи [8]. У поєднанні з кіберфізичними атаками це створює складні сценарії загроз, що потребують міждисциплінарного підходу до оцінки ризиків та розробки стратегій захисту. Геополітична нестабільність і розвиток гібридних конфліктів призводять до того, що критичні інформаційні системи стають об’єктом спеціалізованих атак, націлених на дестабілізацію державних структур та інфраструктурного забезпечення. Сучасні загрози вказані в табл. 1.1.

*Таблиця 1.1*

**Сучасні загрози, методи реалізації та потенційні наслідки**

| **Загроза** | **Метод реалізації** | **Потенційні наслідки** | **Приклад** |
|---|---|---|---|
| Класичні кібератаки | Несанкціонований доступ, шкідливе ПЗ | Порушення роботи систем, втрата даних | Атака на банки через фішинг |
| Кіберфізичні атаки | Вразливості ICS/SCADA, IoT | Збої технологічних процесів, аварії | Маніпуляції енергетичними системами |
| Соціальна інженерія | Фішинг, маніпуляції персоналом | Викриття конфіденційних даних | Злом електронної пошти співробітників |
| Гібридні атаки | Поєднання кібер і інформаційних кампаній | Дестабілізація інфраструктури та державних процесів | Кібероперації у гібридних конфліктах |
| Кібер-тероризм | Алгоритми AI для | Масштабні | Атака на критичні |

| | | | |
|---|---|---|---|
| | обходу захисту | економічні та соціальні збитки | об’єкти енергетики |

Окрему увагу приділяють атакам типу «кібер-тероризм», які спрямовані на досягнення політичних, економічних або соціальних цілей шляхом порушення роботи критичних систем. Ці загрози поєднують традиційні технічні методи вторгнення з інноваційними рішеннями на основі штучного інтелекту, алгоритмів машинного навчання та автономних систем. Наприклад, алгоритми самонавчання можуть використовуватися для обходу систем виявлення вторгнень, що підвищує ефективність атак і ускладнює процес їхнього прогнозування [9]. Водночас застосування штучного інтелекту в оборонних системах дозволяє виявляти патерни аномальної поведінки та забезпечує швидке реагування на загрози.

Серед ключових викликів сучасного кіберзахисту можна виділити проблему забезпечення цілісності, доступності та конфіденційності даних у складних мережевих середовищах. Прискорений розвиток технологій хмарних обчислень і централізованих систем зберігання даних створює необхідність у високотехнологічних методах шифрування, контролю доступу та моніторингу подій безпеки. Важливим аспектом є також інтеграція систем кіберзахисту з організаційними політиками управління ризиками та планами безперервності бізнес-процесів, що забезпечує комплексну стійкість критичних систем.

Сучасні підходи до протидії загрозам у критичних системах орієнтовані на використання багаторівневих стратегій захисту, включаючи превентивні, виявлювальні та реагувальні механізми. Превентивні заходи включають впровадження політик безпеки, оновлення систем, навчання персоналу та застосування стандартів кібербезпеки. Виявлювальні механізми базуються на аналітиці логів, моніторингу трафіку, виявленні аномалій у роботі систем та застосуванні алгоритмів машинного навчання для прогнозування атак. Реагувальні заходи забезпечують оперативне відновлення систем, локалізацію інцидентів та мінімізацію наслідків атак [10].

Не менш важливим аспектом сучасних викликів є проблема інтеграції кіберзахисту у багаторівневу інфраструктуру держави. Це включає координацію між різними відомствами, державними та приватними компаніями, операторами критичних систем, а також міжнародну співпрацю у сфері обміну інформацією про загрози. Високий рівень складності та взаємозалежності створює необхідність використання системного підходу, що поєднує технічні, організаційні та стратегічні компоненти.

Сучасні загрози у сфері кіберзахисту критичних систем мають комплексний характер і включають технічні, соціальні, політичні та організаційні аспекти. Вони характеризуються високою швидкістю еволюції, багатовекторністю та потенціалом створення масштабних наслідків. Ефективна протидія цим загрозам вимагає використання сучасних технологій, включно з алгоритмами штучного інтелекту, інтеграції кіберзахисних систем у стратегічне управління державою та постійного вдосконалення політик і стандартів безпеки. (рис. 1.3)


Рис. 1.3. Механізми багаторівневого захисту критичних систем

## 1.3. Можливості штучного інтелекту для підвищення ефективності кіберзахисту

Стрімке поширення цифрових технологій та зростання складності кіберзагроз обумовлюють необхідність використання інтелектуальних рішень для забезпечення високого рівня захисту критично важливих систем. Штучний інтелект дедалі частіше стає ключовим компонентом комплексної архітектури кібербезпеки, оскільки його алгоритми здатні аналізувати великі обсяги даних, виявляти приховані закономірності, прогнозувати потенційні атаки та забезпечувати автоматизоване реагування на інциденти [11]. Завдяки машинному навчанню, глибинним нейронним мережам та іншим інтелектуальним методам сучасні системи кіберзахисту отримують можливість швидко адаптуватися до змін середовища, що є критично важливим у контексті еволюційних і багатовекторних атак.

Одним із найважливіших напрямів застосування штучного інтелекту у сфері кіберзахисту є виявлення аномалій**.** Класичні системи, орієнтовані на сигнатурний аналіз, ефективно розпізнають лише відомі загрози, але виявляють значні обмеження у контексті нових або модифікованих атак [12]. Алгоритми машинного навчання дозволяють аналізувати нормальну поведінку системи та виявляти відмінності, що можуть свідчити про початок інциденту. Моделі кластеризації, глибинні автоенкодери, рекурентні та графові нейронні мережі застосовуються для аналізу мережевого трафіку, поведінкових патернів користувачів, системних журналів та інших джерел даних. У результаті системи виявлення аномалій здатні з високою точністю ідентифікувати загрози, які раніше не потрапляли до баз даних шкідливих сигнатур, що значно розширює можливості превентивного захисту критичних систем.

*Таблиця 1.2*

**Порівняння класичних і AI-орієнтованих систем кіберзахисту**

| Критерій | Класичні системи | Системи на основі ШІ |
|---|---|---|
| Виявлення нових загроз | Низька ефективність | Висока ефективність |
| Оновлення | Потребує сигнатур | Самонавчання |
| Швидкість аналізу | Середня | Дуже висока |
| Гнучкість | Обмежена | Адаптивна |
| Хибні спрацювання | Часті | Значно менше |
| Прогнозування атак | Відсутнє | Реалізоване |

Важливе місце у сучасних моделях безпеки займають інтелектуальні системи виявлення та запобігання вторгненням (IDS/IPS)**.** Завдяки поєднанню сигнатурних, евристичних та поведінкових методів вони здатні ефективно відслідковувати шкідливу активність і блокувати небезпечні дії у режимі реального часу. Використання штучного інтелекту дозволяє IDS/IPS системам не лише аналізувати аномалії, а й на основі історичних даних визначати потенційні сценарії розвитку атаки [13]. Це забезпечує можливість інтелектуального прийняття рішень, коли система може самостійно обирати оптимальний тип реагування: сповіщення адміністратора, блокування IP-адреси, ізоляцію сегмента мережі або запуск процедури відновлення. Поєднання алгоритмів навчання з учителем та без учителя дозволяє таким системам постійно вдосконалюватися, накопичувати знання та зменшувати кількість хибних спрацьовувань (рис. 1.4).

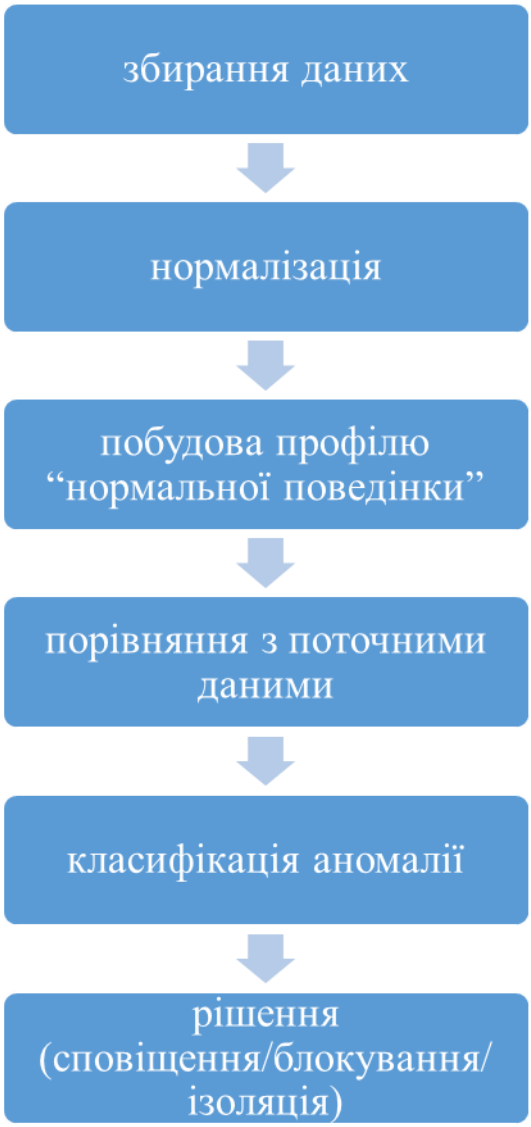


Рис. 1.4 Алгоритм роботи системи виявлення аномалій на основі ШІ

Не менш значущим напрямом застосування ШІ є прогнозування атак, що базується на аналізі трендів, історичних інцидентів, поведінкових особливостей зловмисників, а також моделюванні можливих сценаріїв розвитку подій. Методи прогнозної аналітики дають змогу створювати моделі атак, які враховують складну взаємодію різних векторів ураження [14]. Завдяки використанню стохастичних моделей, графових алгоритмів, байєсівських мереж та глибинних рекурентних мереж можливо визначати найбільш ймовірні точки входу атак, оцінювати рівень ризику та визначати пріоритети щодо

посилення захисту. В умовах критичних інфраструктур прогнозування є надзвичайно важливим, оскільки дозволяє завчасно підготуватися до потенційних інцидентів, оптимізувати ресурси та мінімізувати можливі наслідки.

Застосування штучного інтелекту у сфері кіберзахисту також пов'язане з активним використанням методів обробки великих даних. Критичні системи генерують мільярди подій щодня: системні логи, мережевий трафік, журнали доступу, телеметрію IoT, події контролерів промислової автоматизації. Традиційні аналітичні інструменти не здатні обробляти такі обсяги даних у режимі реального часу. Інтелектуальні системи на основі AI-HPC (високопродуктивних обчислень для ШІ) дозволяють виконувати складні операції за частки секунди, що забезпечує можливість швидкого реагування на інциденти в масштабі всієї інфраструктури [15].

Окремий напрям складають інструменти поведінкової аналітики (UEBA — User and Entity Behavior Analytics)**.** Вони використовують методи машинного навчання для побудови профілів звичайної поведінки користувачів, пристроїв і сервісів. Коли система фіксує відхилення від встановленої норми, вона визначає потенційний інцидент і ініціює захисні дії. У контексті критичних інфраструктур UEBA-платформи особливо ефективні для виявлення інсайдерських загроз, які традиційні засоби безпеки часто не здатні ідентифікувати. Поєднання UEBA з IDS/IPS формує комплексну екосистему захисту, в якій кожний рівень доповнює інший.

Значні можливості відкриває застосування методів глибинного навчання для аналізу мережевого трафіку**.** Глибокі нейронні мережі можуть автоматично класифікувати трафік, виявляти складні малопомітні патерни атак та розрізняти шкідливу й легітимну активність на рівні, недосяжному для класичних алгоритмів [16]. Моделі CNN, RNN, LSTM і трансформери добре зарекомендували себе у задачах виявлення DDoS-атак, сканування портів, мережевих вторгнень і ботнет-активності. Завдяки цьому зменшується рівень хибнопозитивних спрацювань і підвищується точність діагностики інцидентів.

Важливу роль відіграє також застосування штучного інтелекту в автоматизованому управлінні реагуванням на інциденти (SOAR — Security Orchestration, Automation and Response). Такі платформи дозволяють координувати дії різних підсистем безпеки, автоматизувати процеси збору доказів, аналізувати кореляцію подій та запускати попередньо визначені сценарії реагування. Інтеграція AI у SOAR-системи забезпечує складну логіку прийняття рішень, що дозволяє системі самостійно розв'язувати інциденти низької та середньої складності, зменшуючи навантаження на аналітиків.

Окремим викликом є проблема пояснюваності моделей штучного інтелекту. Для застосування AI у критичних інфраструктурах необхідно забезпечити високий рівень прозорості алгоритмів, можливість перевірки результатів та відповідність вимогам нормативно-правового регулювання. Методи Explainable AI (XAI) дозволяють пояснити логіку роботи моделі та обґрунтувати рішення, що є важливим у процесах аудиту та правового контролю [17]. Підсумуємо всі ці інструменти в табл. 1.3.

*Таблиця 1.3*

**AI-інструменти кіберзахисту та їх призначення**

| Інструмент | Технологія | Призначення |
|---|---|---|
| UEBA | ML + поведінкові моделі | Виявлення інсайдерів |
| SOAR | AI-координація | Автоматизація реагування |
| AI-Traffic Analyzer | DL-моделі | Аналіз трафіку |
| XAI | Explainable AI | Пояснення рішень моделей |
| HPC-AI | Високопродуктивне ML | Аналіз великих масивів даних |

Незважаючи на значний потенціал штучного інтелекту, його застосування у сфері кіберзахисту супроводжується певними проблемами. До основних із них належать необхідність значних обчислювальних ресурсів, складність побудови якісних навчальних вибірок, ризик атак на самі моделі машинного навчання (adversarial attacks), а також загрози некоректного використання AI з боку зловмисників [18]. Це підтверджує необхідність створення багаторівневої

системи безпеки, орієнтованої не лише на застосування алгоритмів ШІ, а й на забезпечення їхнього захисту, перевірки та регулярного оновлення.

Штучний інтелект стає ключовим інструментом у підвищенні ефективності кіберзахисту критичних систем. Його можливості охоплюють виявлення аномалій, запобігання вторгненням, прогнозування атак, поведінкову аналітику, автоматизацію реагування та глибинний аналіз мережевого трафіку. Інтеграція цих інструментів у комплексну модель безпеки формує високорівневу адаптивну архітектуру, здатну протистояти сучасним та перспективним загрозам.

## 1.4. Нормативно-правова та етична база застосування AI у сфері кібербезпеки

Стрімке впровадження технологій штучного інтелекту в системи захисту критичної інформаційної інфраструктури зумовлює необхідність формування комплексної нормативно-правової та етичної бази, здатної врегулювати питання безпеки, відповідальності, прозорості та захисту прав людини [19]. На рівні держав і наддержавних об’єднань формується широка рамка регуляцій, що визначає допустимі практики використання алгоритмів у сфері кібербезпеки, вимоги до їхнього функціонування та принципи контролю. Особливої ваги ці питання набувають у секторі критичних систем, де використання ШІ прямо впливає на сталий розвиток, економічну стабільність і національну безпеку. Регуляції мають запобігти зловживанням, забезпечити баланс між інноваційністю та ризиками, а також створити однакові підходи до відповідальності за наслідки роботи автоматизованих систем.

Країни Європейського Союзу демонструють найбільш системний підхід до регулювання штучного інтелекту, поєднуючи правові механізми з високими стандартами етичного контролю [20]. Центральним актом у цій сфері є AI Act, який запроваджує ризик-орієнтовану модель, визначаючи чіткі категорії застосування алгоритмів: від мінімального до неприйнятного ризику. У

контексті кіберзахисту критичних інфраструктур особливого значення набувають системи високого ризику, які мають відповідати суворим вимогам щодо прозорості, контролю якості, надійності та кіберстійкості. Від операторів таких систем вимагається забезпечення верифікації моделей, документування процесів навчання, оцінювання ризиків та наявність механізмів людського контролю над ключовими рішеннями. У поєднанні з директивами NIS2 та CER, що регулюють кіберстійкість критичних секторів, AI Act формує завершену нормативну рамку, де штучний інтелект виступає інструментом безпеки, а не джерелом додаткових загроз.

Сполучені Штати Америки демонструють іншу модель регулювання, орієнтовану на гнучкість, інновації та міжагентну координацію. На федеральному рівні ключовим документом є AI Executive Order 2023, який встановлює загальні принципи розвитку безпечного ШІ, включно з вимогами до тестування моделей, захисту даних, удосконалення кібербезпеки та зменшення потенційних зловживань [21]. Значну роль відіграє NIST AI Risk Management Framework, який визначає методи оцінювання ризиків, критерії надійності моделей та правила їх застосування у сфері національної безпеки. На відміну від ЄС, США віддають перевагу саморегуляції та стандартизації, стимулюючи ринок до впровадження безпечних технологій без жорсткої централізованої заборонної моделі. Особливий акцент робиться на захисті критичних інфраструктур через застосування AI-орієнтованих рішень для проактивного виявлення загроз, раннього попередження і реагування на атаки на основі глибокого аналізу даних.

Україна перебуває на етапі формування власної нормативно-правової бази у сфері штучного інтелекту, орієнтуючись на стандарти ЄС і розвиваючи інституційні механізми у сфері кібербезпеки. Прийняття Стратегії розвитку штучного інтелекту в Україні, а також оновленого законодавства у сфері кіберзахисту критичної інфраструктури створює фундамент для інтеграції технологій ШІ у державні системи безпеки [22]. У межах гармонізації з європейським законодавством Україна розробляє норми, що враховують ризик-

орієнтований підхід, захист даних, вимоги до прозорості алгоритмів та їх сертифікації. У зв'язку з веденням кібервоєн сучасного типу, значна увага приділяється можливості застосування ШІ у військових, оборонних та стратегічно важливих цивільних інфраструктурах. Це вимагає створення спеціальних режимів контролю, які забезпечать надійність моделей, стійкість до зовнішніх впливів та недопущення некоректної поведінки автономних систем у критичних умовах (рис. 1.5).

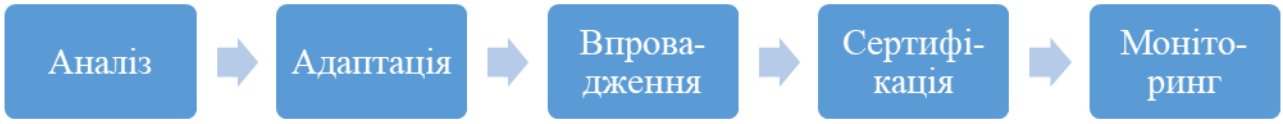


Рис. 1.5 Модель гармонізації українського законодавства з AI Act

Етичні аспекти впровадження штучного інтелекту виступають невід'ємним елементом правового регулювання, оскільки алгоритмізація процесів прийняття рішень може впливати на права людини, конфіденційність інформації та справедливість. Європейський Союз зосереджується на принципах відповідальності, пояснюваності та недискримінації алгоритмів, що закріплені у Етичних керівних принципах з ШІ [23]. У сфері кібербезпеки ці принципи означають обов'язок уникати надмірної автоматизації, яка може призвести до помилкових рішень, а також обов'язок забезпечувати контроль оператора над критичними функціями. У США етичні вимоги мають рекомендаційний характер і впроваджуються через стандарти NIST, що підкреслюють важливість зниження алгоритмічних упереджень, забезпечення людського контролю та підвищення прозорості у всіх сферах, де прийняті автоматизовані рішення впливають на безпеку громадян.

В Україні етичні питання лише починають інтегруватися у нормативну сферу. Наявні документи окреслюють загальні принципи безпечного застосування ШІ, проте спеціалізовані етичні норми для сектору кіберзахисту

перебувають на етапі розробки [24]. Основними викликами є недостатність стандартизованих підходів до оцінювання ризиків, відсутність національної системи аудиту алгоритмів та потреба в гармонізації з AI Act і міжнародними рекомендаціями OECD. Проте в умовах війни та постійних масштабних кібератак Україна швидко рухається в напрямі створення комплексного підходу, що включатиме регулювання автономних систем оборонного типу, спеціальні вимоги до верифікації моделей та посилення прозорості алгоритмів, які застосовуються державними органами.

Загалом, аналіз досвіду ЄС, США та України демонструє наявність двох основних моделей регулювання: європейська жорстка модель із централізованим контролем та американська гнучка модель зі стандартами, що рекомендуються ринку [25]. Україна, у свою чергу, поступово вибудовує власний гібридний підхід, що поєднує європейську нормативність та американську динамічність, що є оптимальним шляхом у контексті сучасних безпекових викликів.

Порівняння всіх цих підходів різних країн вказано в табл. 1.4.

*Таблиця 1.4*

**Порівняння регуляторних підходів ЄС, США та України у сфері AI для кібербезпеки**

| Параметр | ЄС | США | Україна |
|---|---|---|---|
| **Тип регулювання** | Жорстке, централізоване | Гнучке, стандартне | Гібридне |
| **Основні акти** | AI Act, NIS2 | EO 2023, NIST AI RMF | Стратегія ШІ, закони з КІ |
| **Режим контролю** | Високий | Середній | Наближається до ЄС |
| **Орієнтація** | Безпека та етичність | Інновації та гнучкість | Синтез |
| **Сфери підвищеної уваги** | КІІ, права людини | Нацбезпека, оборона | Кібервійна і КІІ |

## Висновки до розділу 1

У першому розділі була сформована цілісна теоретико-методологічна основа, необхідна для розуміння особливостей застосування технологій штучного інтелекту в кіберзахисті критичних інформаційних інфраструктур. Аналіз проведених підпунктів дозволив окреслити низку ключових закономірностей, тенденцій та викликів, що визначають сучасну парадигму захисту критично важливих систем у цифрову епоху.

По-перше, обґрунтовано, що критичні інформаційні інфраструктури є фундаментальними елементами життєдіяльності держави, а їх стабільність прямо впливає на функціонування економіки, державного управління, оборонної системи та суспільної безпеки. У зв’язку з цим вони потребують спеціалізованого підходу до кіберзахисту, який враховує унікальну архітектуру, високу взаємозалежність компонентів та потенційно катастрофічні наслідки будь-яких порушень. Важливо, що саме складність цих систем і глибина їх інтеграції у всі сфери державності створюють умови, за яких традиційні інструменти кіберзахисту виявляються недостатніми.

По-друге, дослідження сучасних загроз показало, що вектор атак проти критичних систем стрімко еволюціонує. Зростає роль багатовекторних, високотехнологічних і гібридних операцій, що поєднують технічні та інформаційно-психологічні інструменти впливу. Кіберзлочинні угруповання та державні актори активно використовують автоматизацію, машинне навчання та інтелектуальні механізми оптимізації атак, що вимагає від держав нових підходів до стратегічної оборони. Особливої небезпеки набувають атаки на промислові системи керування, енергетичні мережі, системи зв’язку та інші ключові домени, де навіть короткочасне порушення може спричинити каскадні ефекти.

По-третє, можливості штучного інтелекту формують новий рівень кіберстійкості критичних інфраструктур. Алгоритми виявлення аномалій, інтелектуальні IDS/IPS, системи прогнозування атак та моделі поведінкового

аналізу забезпечують значно більшу швидкість реагування, точність і адаптивність порівняно з класичними підходами. ШІ дозволяє переходити до проактивної моделі кіберзахисту, де загрози не лише виявляються, а й прогнозуються задовго до реалізації. Проте впровадження таких систем потребує високого рівня підготовки, надійної інфраструктури даних і механізмів контролю за поведінкою моделей.

По-четверте, нормативно-правова та етична складові відіграють визначальну роль у формуванні безпечних умов для використання штучного інтелекту в кібербезпеці. ЄС розвиває сувору ризик-орієнтовану модель регулювання, США роблять ставку на гнучкі стандарти та інновації, тоді як Україна вибудовує власну гібридну систему, орієнтовану на європейські вимоги та потреби оборонного сектору. Усі ці підходи підкреслюють важливість людського контролю, прозорості алгоритмів, відповідальності операторів та захисту прав людини. Без узгодженої нормативної бази впровадження ШІ у критичні інфраструктури може породжувати додаткові ризики, тому регуляція є такою ж важливою, як і технічні інструменти.

Таким чином, ефективне застосування штучного інтелекту у кіберзахисті можливе лише за умов поєднання трьох ключових факторів: глибокого розуміння специфіки критичних інфраструктур, розвитку сучасних інтелектуальних технологій та формування відповідальної нормативної й етичної рамки. Тільки комплексний підхід дозволить забезпечити стійкість державних та приватних систем перед загрозами, що постійно ускладнюються, а також створити підґрунтя для формування адаптивної, динамічної та випереджувальної системи кібербезпеки.

# РОЗДІЛ 2
# ЕТИЧНІ ВИКЛИКИ ТА МЕТОДИ ВПРОВАДЖЕННЯ AI У КІБЕРЗАХИСТ КРИТИЧНИХ СИСТЕМ

Інтеграція систем штучного інтелекту (ШІ) у сферу кіберзахисту критичних інфраструктур стає не лише технічною, а й етичною проблемою. Використання ШІ у забезпеченні безпеки критичних систем передбачає комплексне вирішення питань конфіденційності, прозорості алгоритмів, відповідальності за прийняті рішення та мінімізації ризиків для суспільства. У цьому контексті особлива увага приділяється визначенню етичних викликів та розробці методів впровадження технологій ШІ, які гарантують баланс між ефективністю кіберзахисту та дотриманням фундаментальних етичних принципів.

## 2.1. Етичні проблеми застосування AI у сфері кібербезпеки

Застосування штучного інтелекту у сфері кібербезпеки критичних систем відкриває нові горизонти ефективного виявлення та запобігання загрозам, проте одночасно створює комплекс етичних проблем, що потребують системного аналізу та регуляторного врегулювання. Етичні питання пов’язані як із технологічною природою ШІ, так і з його інтеграцією у середовища, що мають високий рівень критичності та значний вплив на суспільство та економіку.

Одним із ключових етичних викликів є забезпечення конфіденційності та безпеки персональних та організаційних даних, що обробляються системами ШІ. Для ефективного навчання моделей часто використовуються великі обсяги даних, включно з чутливою інформацією про користувачів та операційні процеси критичних інфраструктур [26]. Неправомірне використання або витік таких даних може мати серйозні наслідки, включно з порушенням права на приватність, фінансовими втратами та репутаційними ризиками.

До основних проблем належать:

1. Непрозорість алгоритмів – складність контролю за тим, як алгоритм обробляє персональні дані, і які рішення він приймає.

2. Можливість несанкціонованого доступу – уразливості у програмному забезпеченні можуть призвести до зловмисного використання даних.

3. Дискримінаційні ризики – неправильне або необ'єктивне навчання моделей на некоректних даних може призвести до несправедливого обмеження доступу або блокування певних користувачів. (табл. 2.1)

Таблиця 2.1

**Основні етичні ризики щодо конфіденційності у застосуванні AI**

| **Етична проблема** | **Потенційні наслідки** | **Приклади критичних систем** |
|---|---|---|
| Непрозорість алгоритмів | Неможливість перевірки рішень AI | Енергетичні мережі, транспорт |
| Несанкціонований доступ | Витік персональних даних, кібератаки | Банківські системи, дата-центри |
| Дискримінація | Несправедливе обмеження доступу | Системи управління трафіком, медичні бази даних |

Ще однією значною етичною проблемою є прозорість процесу прийняття рішень системами ШІ та відповідальність за ці рішення. У критичних системах, де будь-яка помилка може призвести до серйозних наслідків, необхідно чітко визначати, хто несе відповідальність за дії алгоритму: розробник, оператор системи чи організація, яка її експлуатує [27].

Основні аспекти включають:

- "Чорний ящик" алгоритмів – складність інтерпретації рішень AI.
- автоматизація рішень без контролю людини – ризик прийняття некоректних або небезпечних рішень без людського втручання.

- юридичні та моральні колізії – відсутність чітких правових рамок для відповідальності в разі негативних наслідків використання AI.

ШІ у кіберзахисті може бути використаний як інструмент для протидії загрозам, але одночасно існує ризик його використання для шкідливих цілей. Наприклад, автоматизовані системи виявлення вторгнень можуть бути обмануті методами атак на саму модель (adversarial attacks), що призводить до серйозних порушень безпеки [28].

Ключові етичні аспекти:

- маніпулювання навчальними даними – введення шкідливих даних у тренувальні вибірки;
- непередбачувана поведінка моделей – моделі можуть ухвалювати рішення, що не передбачені розробником;
- дилеми безпеки та приватності – баланс між захистом системи та вторгненням у приватні дані користувачів. (табл.2.2)

Таблиця 2.2

**Види маніпуляцій та їхні наслідки для критичних систем**

| **Тип маніпуляції** | **Механізм дії** | **Наслідки для системи** |
|---|---|---|
| Атаки на дані (data poisoning) | Шкідливе коригування тренувальних даних | Зниження ефективності AI, помилкові спрацювання |
| Adversarial attacks | Введення спеціально створених даних | Обхід систем захисту, порушення безпеки |
| Неконтрольоване навчання | Модель адаптується до небажаних шаблонів | Непередбачувана поведінка, ризики для критичних процесів |

Впровадження AI у критичні системи впливає не лише на технічний рівень, але й на соціальний та моральний [29]. Автоматизація прийняття рішень може призвести до зменшення ролі людини у процесах контролю, що викликає етичні запитання щодо довіри до систем та збереження людського фактора.

До основних проблем відносяться:

- розмиття відповідальності – людський контроль зменшується, що може знизити готовність персоналу брати відповідальність;
- залежність від технологій – суспільство може стати надмірно залежним від рішень AI, навіть у ситуаціях критичної важливості;
- порушення моральних норм – рішення, прийняті системою, можуть суперечити етичним стандартам або законодавчим нормам (рис.2.1).

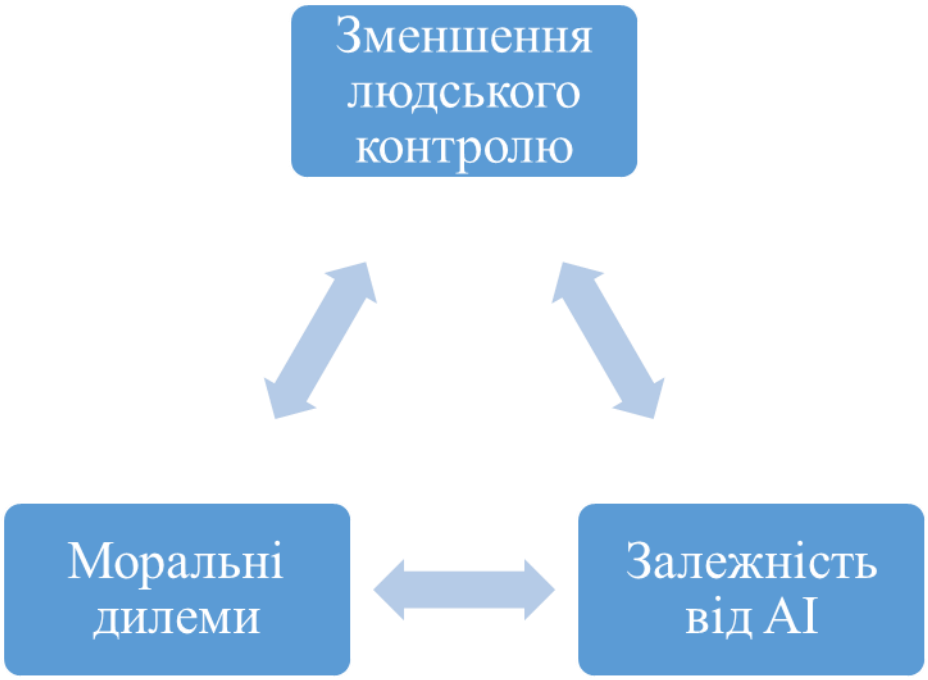


Рис. 2.1 Етичні ризики автоматизації у критичних системах

Етичні проблеми застосування AI у сфері кібербезпеки критичних систем є багаторівневими та включають технічні, юридичні, соціальні та моральні аспекти. Їхнє системне врахування є необхідною умовою безпечного впровадження технологій, а також формування довіри користувачів і суспільства до рішень, прийнятих на основі ШІ [30]. Подальші дослідження мають зосередитися на розробці стандартів етичного дизайну AI та механізмів підзвітності для зниження потенційних ризиків.

## 2.2. Ризики зловживання та можливі негативні наслідки впровадження AI у критичних системах

Застосування штучного інтелекту у критичних системах відкриває нові можливості для підвищення ефективності захисту та оптимізації процесів, проте одночасно створює потенційні ризики зловживань, які можуть мати значні негативні наслідки. Одним із основних ризиків є використання AI зловмисниками для обходу захисних механізмів або проведення складних кібероперацій [31]. Системи, що автоматично аналізують великі обсяги даних, можуть стати інструментом атак, якщо до них отримати контроль, що ставить під загрозу функціонування критичних інфраструктур та безпеку держави в цілому.

Таблиця 2.3

**Потенційні способи зловживання AI у критичних системах та їх наслідки**

| Спосіб зловживання AI | Механізм дії | Потенційні наслідки для критичних систем |
|---|---|---|
| Контроль над алгоритмами | Несанкціоноване отримання доступу до AI | Втрата контролю над системою, збій процесів |
| Маніпуляції даними (data poisoning) | Введення шкідливих даних у тренувальні вибірки | Зниження точності, хибні спрацювання, вразливість |
| Adversarial attacks | Використання спеціально створених вхідних даних | Обхід систем захисту, порушення безпеки |
| Автоматизовані помилкові рішення | Неконтрольоване навчання або неправильна конфігурація | Непередбачувана поведінка, порушення процесів |
| Соціальна маніпуляція | Використання AI для впливу на поведінку персоналу | Порушення довіри, саботаж або неправильне використання систем |

Іншим важливим аспектом є можливість непередбачуваної поведінки AI через недосконаліть алгоритмів або некоректне навчання моделей [32]. Навіть невеликі помилки у алгоритмах можуть призвести до серйозних відмов систем, блокування життєво важливих процесів або некоректного прийняття рішень,

що вимагає ретельного контролю та розробки механізмів корекції. Така непередбачуваність стає особливо критичною у сферах енергетики, транспорту та охорони здоров'я, де будь-яка збійна ситуація може спричинити значні соціальні та економічні збитки.

Зловживання AI також може проявлятися у маніпуляціях даними, на яких працюють моделі. Використання шкідливих або спотворених даних під час навчання систем здатне спричинити їхню функціональну неспроможність у реальному середовищі [33]. Такі атаки можуть мати як спеціалізований характер, спрямований на окремі критичні процеси, так і масовий, впливаючи на великі сегменти інфраструктури. Внаслідок цього зростає ймовірність помилкових спрацьовувань систем безпеки або, навпаки, невиявлення реальних загроз (рис.2.2).

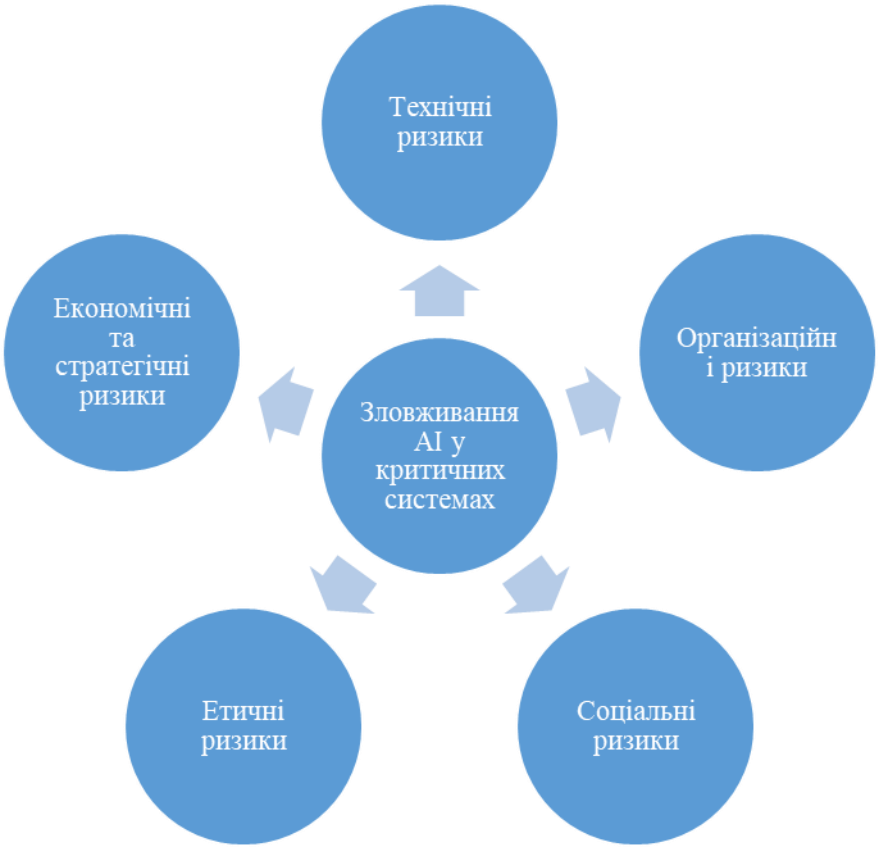


Рис. 2.2 Ризики, які виникають внаслідок зловживання AI і негативних наслідків

Особливу увагу слід приділяти питанням соціальної та організаційної безпеки. Автоматизація рішень, яка зменшує участь людини у процесах контролю, може спричинити формування залежності від технологій та зниження рівня професійної підготовки персоналу [34]. Надмірна довіра до AI у критичних системах здатна призвести до ігнорування потенційних загроз або відтермінування реагування на них, що підвищує ризики катастрофічних наслідків.

Крім того, існує ризик комбінації технологічних та етичних проблем, коли AI використовується для прийняття рішень, що мають прямий вплив на життя та безпеку людей. Неконтрольовані дії систем можуть призводити до порушення законодавчих норм, моральних принципів та фундаментальних прав громадян. Відсутність чітких правил регулювання та механізмів підзвітності підвищує ймовірність таких негативних наслідків.

Важливим аспектом є також економічний та стратегічний вплив потенційних зловживань AI. Критичні системи державного та приватного сектору, що інтегрують штучний інтелект, стають привабливою ціллю для кіберзлочинців та державних акторів, зацікавлених у дестабілізації інфраструктури або отриманні стратегічної переваги. Наслідки таких зловживань можуть варіюватися від локальних фінансових втрат до масштабних соціально-економічних криз, що вимагає комплексного підходу до оцінки ризиків і розробки систем протидії.

Крім безпосередніх атак, ризики включають також непрямі наслідки впровадження AI, пов’язані з порушенням довіри користувачів до систем. Якщо автоматизовані рішення будуть сприйматися як непрозорі або потенційно небезпечні, це може призвести до відмови від використання технологій або саботажу їх інтеграції в критичні процеси [35]. Така втрата довіри може знизити ефективність захисту, створюючи додаткові вразливості для зловмисних дій.

Ризики зловживання та можливі негативні наслідки впровадження AI у критичних системах носять комплексний характер, включаючи технічні, організаційні, соціальні, етичні та економічні аспекти. Для мінімізації цих ризиків необхідне системне оцінювання потенційних загроз, розробка нормативних рамок, впровадження механізмів контролю та моніторингу, а також постійне підвищення кваліфікації персоналу, що взаємодіє з AI. Тільки інтегрований підхід дозволяє забезпечити баланс між ефективністю технологій та безпекою критичних систем, зберігаючи при цьому дотримання етичних та соціальних норм (рис.2.3).


Рис.2.3 Цикл ризику через автоматизацію та залежність від AI

## 2.3. Порівняльний аналіз міжнародних практик етичного регулювання AI для безпеки критичної інфраструктури

У світовій практиці регулювання штучного інтелекту для забезпечення безпеки критичної інфраструктури спостерігається значна варіативність підходів, що відображає різницю у правових, культурних та технічних традиціях окремих країн. Країни Європейського Союзу приділяють особливу увагу етичним аспектам застосування AI, зокрема прозорості алгоритмів,

захисту персональних даних та підзвітності рішень. Регламент ЄС з етичного використання AI встановлює обов’язкові критерії для високоризикових систем, до яких належить критична інфраструктура, передбачаючи аудит алгоритмів, оцінку ризиків та забезпечення відповідності принципам справедливості та недискримінації [36].

У США підхід до регулювання AI для критичної інфраструктури більше орієнтований на стандарти безпеки та технічні протоколи, ніж на етичні принципи як такі. Важливим елементом є розробка керівних принципів для federal agencies та private sector, зосереджених на забезпеченні кіберзахисту, надійності систем та управлінні ризиками [37]. Окрему увагу приділяють питанням кібергромадянської безпеки та захисту критичних сервісів, проте стандарти прозорості та етичної підзвітності у США поки що менш суворі, ніж у ЄС.

У Китаї та низці країн Азії підходи до регулювання AI поєднують державний контроль і активне впровадження технологій у сферу безпеки. Регулювання часто передбачає жорсткий контроль доступу до даних та централізоване управління системами, при цьому питання етичної прозорості та участі громадськості не завжди є пріоритетними. Проте акцент на надійності та швидкості реагування на кіберзагрози дозволяє ефективно мінімізувати технічні ризики, хоча соціальні та моральні аспекти можуть залишатися недостатньо врегульованими.

Порівняння міжнародних практик етичного регулювання AI у критичній інфраструктурі різних країн продемонстровано в табл. 2.4.

Таблиця 2.4

**Порівняння міжнародних практик етичного регулювання AI у критичній інфраструктурі**

| Країна/Регіон | Основні підходи до регулювання | Пріоритети | Вразливі аспекти |
|---|---|---|---|
| ЄС | Регламентування, аудит алгоритмів, оцінка ризиків | Прозорість, права людини, підзвітність | Тривалі процедури впровадження, бюрократія |
| США | Стандарти безпеки та надійності, керівні принципи для агенцій | Технічна безпека, ефективність | Менше уваги етичній прозорості та участі громадськості |
| Китай та Азія | Централізований контроль, швидке впровадження технологій | Надійність, швидкість реагування | Соціальні та моральні аспекти менш врегульовані |
| Міжнародні організації (OECD, ISO) | Рекомендації, стандарти, гармонізація підходів | Безпека, етика, глобальна сумісність | Потреба в адаптації до національних особливостей |

Порівняльний аналіз міжнародних практик свідчить про те, що успішне впровадження AI у критичну інфраструктуру неможливе без комплексного підходу, який поєднує технічні стандарти, етичні принципи та правове регулювання. Ключові етичні критерії, які застосовуються у різних країнах, включають захист приватності, прозорість алгоритмів, підзвітність рішень, мінімізацію дискримінації та оцінку соціальних ризиків. Різниця між підходами проявляється у пріоритетах: у ЄС акцент робиться на права людини та етичну відповідальність, у США – на безпеку та ефективність, в Азії – на централізоване управління та швидкість реагування [38].

Також варто зазначити, що міжнародні організації, такі як OECD та ISO, розробляють рекомендації та стандарти для етичного використання AI, які можна адаптувати у національних політиках. Вони спрямовані на гармонізацію підходів, забезпечення взаємодії між країнами та створення умов для безпечного та етичного використання AI у критичних системах на глобальному

рівні [39]. Водночас практика показує, що стандарти потребують постійного оновлення відповідно до технологічного розвитку та появи нових загроз, що робить порівняльний аналіз надзвичайно важливим для формування ефективної національної політики (рис. 2.4).


Рис. 2.4 Міжнародні підходи до етичного регулювання AI у критичних системах

**Висновки до розділу 2**

Аналіз етичних аспектів впровадження штучного інтелекту у кіберзахист критичних систем показав, що застосування AI є багатовимірним процесом, що включає технічні, соціальні, моральні та правові компоненти. Етичні проблеми, пов’язані з конфіденційністю даних, прозорістю алгоритмів, підзвітністю рішень та потенційною дискримінацією, вимагають системного підходу до розробки та впровадження технологій, який забезпечує баланс між ефективністю кіберзахисту та дотриманням фундаментальних принципів етики.

Дослідження ризиків зловживання AI у критичних системах демонструє наявність потенційно небезпечних сценаріїв, які можуть виникнути як у результаті зовнішніх атак, так і через неконтрольоване або неправильно налаштоване функціонування алгоритмів. Негативні наслідки включають технічні відмови, організаційні проблеми, соціальну недовіру, порушення етичних норм та економічні втрати, що підкреслює необхідність комплексних механізмів моніторингу, контролю та підвищення кваліфікації персоналу.

Порівняльний аналіз міжнародних практик свідчить про різницю у підходах до етичного регулювання AI, що зумовлено культурними, правовими та технічними особливостями країн. Європейський Союз фокусується на захисті прав людини, прозорості та підзвітності, США – на технічній безпеці та ефективності систем, а країни Азії – на централізованому контролі та швидкості реагування. Універсальні принципи, які пропонують міжнародні організації, такі як OECD та ISO, спрямовані на гармонізацію практик і створення стандартів для безпечного та етичного використання AI у критичних інфраструктурах.

Загалом, впровадження AI у кіберзахист критичних систем неможливе без одночасного вирішення технічних, етичних та організаційних проблем. Досягнення балансу між ефективністю систем, дотриманням етичних принципів та мінімізацією ризиків зловживання є ключовим завданням для науковців, розробників і регуляторів. Системний підхід, який поєднує міжнародний досвід, стандарти та локальні практики, дозволяє забезпечити стійкість критичних систем, захист персональних та організаційних даних і формування довіри користувачів та суспільства до рішень, прийнятих на основі AI.

# РОЗДІЛ 3
# МЕТОДИКА ВПРОВАДЖЕННЯ ЕТИЧНИХ ПІДХОДІВ У РОЗРОБКУ ТА ЕКСПЛУАТАЦІЮ AI-СИСТЕМ КІБЕРЗАХИСТУ КРИТИЧНИХ ОБ'ЄКТІВ

## 3.1. Формування критеріїв оцінювання етичності та ефективності AI-рішень у сфері кіберзахисту

Формування критеріїв оцінювання етичності та ефективності AI-рішень у сфері кіберзахисту потребує системного підходу, що враховує як технологічні властивості алгоритмів, так і їхню поведінкову специфіку в умовах реальних загроз [40]. Основне завдання полягає у визначенні такого переліку вимірюваних характеристик, які дозволяють об'єктивно встановити, чи відповідає система вимогам безпечності, надійності та передбачуваності, а також чи не створює вона додаткових ризиків для критичної інфраструктури. Критерії мають базуватися не на абстрактних принципах, а на конкретних параметрах моделі, процесах прийняття рішень та особливостях її взаємодії з технологічним середовищем.

Першим кроком є визначення кола функцій, які система виконує у процесі кіберзахисту, та потенційних наслідків помилкових рішень. Це дозволяє оцінити ризикову зону, у межах якої етичні вимоги набувають статусу технічних обмежень. Наприклад, системи раннього виявлення загроз вимагають чітких критеріїв щодо рівня допустимого відсотка хибнопозитивних сигналів, оскільки надмірне навантаження на операторів може спричинити пропуск справжніх атак [41]. Своєю чергою, моделі автоматичного блокування трафіку потребують встановлення порогів автономії, що не допускають некоректного відключення критичних сервісів. Формування критеріїв на цьому етапі передбачає аналіз технологічного профілю системи та розроблення набору параметрів, які відображають її вплив на роботу об'єкта.

Наступним елементом є специфікація морально-етичних вимог у вимірювану форму. Етичність AI-системи має бути представлена не декларативно, а у вигляді показників, які можна перевірити експериментально. До таких належать рівень пояснюваності застосованих моделей, можливість відтворення механізму прийняття рішення, а також ступінь контрольованості автономної поведінки системи [42]. Для критичних об’єктів важливо, щоб система надавала оператору обґрунтовані підстави для кожного рішення, особливо у випадках, коли алгоритм виявляє аномалії, що не мають очевидних ознак атаки. Таким чином, до критеріїв входять параметри доступності логів рішень, структурованості аргументації та відповідності пояснень встановленим протоколам кіберзахисту.

Особливе значення має забезпечення відповідності системи вимогам щодо недискримінаційності та непрямої упередженості моделей. Хоча в кіберзахисті упередженість не пов’язана із соціальними групами, вона може проявлятися у вигляді нерівномірної чутливості до різних типів інцидентів або прив’язаності до конкретних патернів трафіку, що формує вразливі зони у системі моніторингу [43]. Тому до критеріїв включаються показники збалансованості навчальних вибірок, стабільності рішення моделі за умов варіативності даних та коректності її роботи при зустрічі з новими типами загроз. Для оцінки цих параметрів формуються тестові корпуси даних, що моделюють ситуації з невизначеними або нечіткими ознаками атаки.

Показники ефективності системи визначаються окремо, але в тісному зв’язку з етичними вимогами. Ефективність не може оцінюватись лише через точність, оскільки в контексті критичних об’єктів важливими є також швидкість обробки подій, стабільність роботи при пікових навантаженнях, відсутність деградації під час довготривалої експлуатації та здатність до адаптації [44]. Формування критеріїв ефективності передбачає розроблення кількісних метрик, таких як середній час реакції, коефіцієнт виявлення складних багатовекторних атак, час відновлення після внутрішніх збоїв та показники масштабованості при збільшенні обсягів вхідних даних.

Важливим етапом є також встановлення критеріїв оцінки безперервності кіберзахисних процесів у разі некоректної роботи моделі. До них належать стійкість системи до помилок, здатність ізолювати неправильно класифіковані події, а також наявність процедур резервного управління, що активуються у випадку збою алгоритму [45]. Ці вимоги не є суто технічними, оскільки вони визначають рівень допустимого ризику при застосуванні штучного інтелекту на об’єктах, де помилка може призвести до порушення роботи енергетичних, транспортних чи комунікаційних систем.

Додатковим критерієм є відповідність рішення нормативним стандартам, що регулюють кіберзахист критичних інфраструктур. Це включає дотримання вимог до збереження журналів подій, протоколів комунікації, регламентів криптографічного захисту та політик доступу [46]. Етичність у цьому контексті проявляється у неприйнятності будь-яких рішень моделі, що суперечать регламентам, навіть якщо вони з технічної точки зору оптимізують захисний процес. Тому критерії передбачають перевірку системи на відповідність нормативним рамкам через формалізовані чек-листи.

Усі зазначені критерії класифікуються на структурні групи: поведінкові, функціональні, нормативні, ризикові та операційні. Для кожної групи встановлюється набір вимірюваних метрик та правила їхньої інтерпретації. Особлива увага приділяється розробленню порогових значень, що визначають межі допустимої роботи AI-системи [47]. Це забезпечує можливість не лише оцінити якість роботи моделі під час тестування, але й контролювати її поведінку у реальному середовищі (рис. 3.1).


Рис. 3.1. Критерії етичності та ефективності

Оцінка критеріїв здійснюється у два етапи: лабораторне тестування та експлуатаційний аудит. На лабораторному етапі моделі перевіряються на репрезентативних наборах даних, що відображають типові та нетипові сценарії кіберзагроз. Тут формується попередній профіль етичності: визначається, чи здатна система пояснити свої рішення, чи стабільні її відповіді, чи не демонструє вона ознак неконтрольованої автономності [48]. На експлуатаційному етапі критерії застосовуються як інструмент моніторингу, що дозволяє виявляти небажані зміни поведінки алгоритмів, які можуть виникати внаслідок дрейфу даних чи ненавмисних взаємодій із системами об’єкта (табл. 3.1).

Таблиця 3.1

**Приклади вимірюваних показників для кожної групи критеріїв**

| Група критеріїв | Показник | Опис застосування |
|---|---|---|
| Поведінкові | Пояснюваність рішення | Чи містить рішення алгоритму обґрунтування |
| Функціональні | Час реакції системи | Швидкість обробки інцидентів |
| Нормативні | Відповідність протоколам доступу | Дотримання встановлених регламентів |
| Ризикові | Рівень автономності | Межі самостійних дій AI без оператора |
| Операційні | Стабільність під навантаженням | Поведінка в пікових режимах |

Завершальним елементом процесу є формування інтегральної оцінки, що поєднує окремі показники у зведений індекс етичності та ефективності. Цей індекс використовується для прийняття рішень про допустимість розгортання системи на критичному об’єкті або необхідність її доопрацювання [49]. Інтегральна оцінка також дозволяє порівнювати різні версії AI-моделей у процесі їхнього розвитку та визначати, чи збереглася етична відповідність після модифікацій або додаткового перенавчання.

## 3.2. Прикладна реалізація: аналіз обраного кейсу застосування AI у захисті критичної системи

Прикладна реалізація критеріїв оцінювання етичності та ефективності AI-систем набуває практичного значення лише тоді, коли вони застосовуються до конкретної ситуації, що відображає реальні умови функціонування критичної інфраструктури [50]. Для прикладу розглядається кейс впровадження системи інтелектуального виявлення та нейтралізації аномалій у сегменті керування електроенергетичною підстанцією середнього класу навантаження, яка містить автоматизований комплекс SCADA, підсистему телеметрії та засоби резервного

контролю [51]. Потреба у використанні AI зумовлена значним збільшенням обсягів телеметричних даних та складністю мануального аналізу поведінкових патернів обладнання, що унеможливлює своєчасну ідентифікацію прихованих атак на комунікаційні протоколи та канали передавання команд керування.

Під час проєктування системи було обрано модель гібридного аналізу аномалій, що поєднує методи розподіленої кластеризації та нейронні моделі прогнозування динаміки параметрів обладнання [52]. Основною функцією системи є раннє виявлення підозрілих відхилень у роботі окремих модулів підстанції, зокрема біля трансформаторних комірок, лінійних вимикачів та модулів регулювання напруги. На цьому етапі формуються вимоги до того, які саме етичні показники мають бути інтегровані у механізми прийняття рішень, щоб система не лише фіксувала загрозу, але й виконувала свої функції без створення додаткових ризиків [53]. Оскільки будь-яке хибне блокування може вплинути на балансування енергомережі та спричинити масштабні аварії, було встановлено жорсткі порогові значення автономних дій (рис. 3.2).


Рис. 3.2 Архітектура AI-системи для аналізу аномалій електропідстанції

Основним елементом прикладної оцінки стала модель сценарного тестування, яка включала опрацювання контрольованих інцидентів у середовищі цифрового двійника підстанції. До тестів були включені чотири базові сценарії: низькоінтенсивні зміни телеметрії, що імітують природні

коливання; багатовекторні атаки типу false data injection; перехоплення команд керування; та високочастотні флуктуації сигналу, що вимагають надзвичайно чутливого аналізу [54]. Під час кожного сценарію фіксувались реакції системи, рівень пояснюваності рішень, час обробки сигналів, стабільність вихідних класифікацій та ступінь відповідності нормативним вимогам.

Таблиця 3.2

**Сценарії тестування та очікувані реакції системи**

| Сценарій | Опис | Очікувана реакція |
|---|---|---|
| Низькоінтенсивні зміни | Природні коливання телеметрії | Відсутність хибних тривог |
| False data injection | Заміна реальних показників | Детекція і передача оператору |
| Перехоплення команд | Спроба зміни команд керування | Блокування та логування |
| Високочастотні зміни | Швидкі флуктуації параметрів | Стабільна класифікація без сповільнень |

Особливу увагу приділено аналізу поведінки системи у випадках, коли модель допускає невизначеність, тобто коли рівень впевненості нижчий за встановлений поріг. За цим критерієм було визначено, чи здатна система коректно ініціювати процедуру контролю оператора та чи не демонструє вона ознак небажаної автономності [55]. Також оцінювалось, чи не намагається система компенсувати невизначеність завищенням рівня ризику, що могло б призвести до необґрунтованих втручань у роботу обладнання [56]. На цьому етапі виникла потреба у введенні механізму багаторівневих пояснень: від стислих індикаторів для диспетчера до детальних логів для технічного аудиту (рис. 3.3).

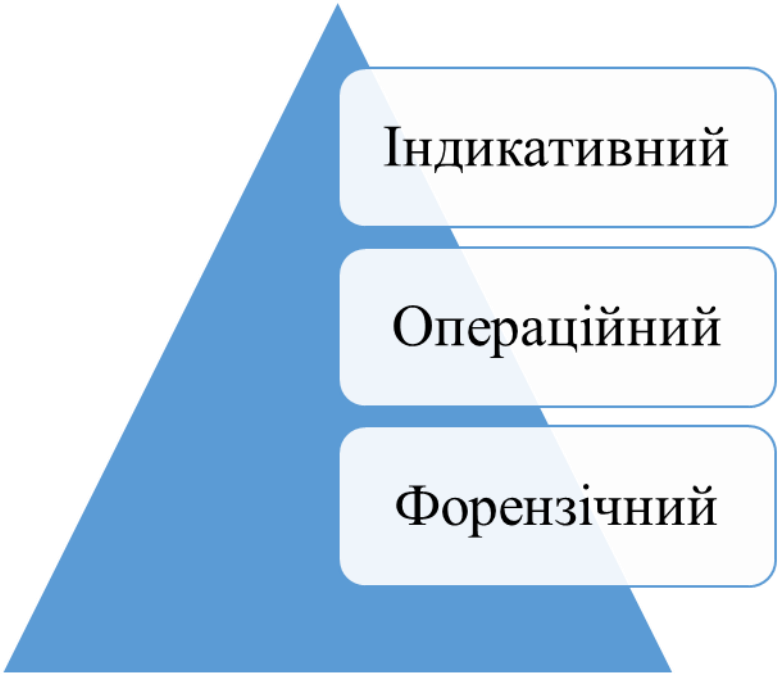


Рис. 3.3 Механізм багаторівневих пояснень

Практична реалізація критеріїв недискримінаційності моделі була проведена шляхом аналізу її здатності однаково обробляти сигнали з різних секцій підстанції. У ході тестування встановлено, що у вихідній версії алгоритму спостерігається підвищена чутливість до даних з лінійних вимикачів порівняно з трансформаторними вузлами [57]. Це призводило до понаднормової кількості хибнопозитивних попереджень. Причиною виявилось нерівномірне наповнення навчального датасету. Цей результат дозволив використати критерії перевірки збалансованості даних на практиці й підготувати оновлену навчальну вибірку з рівномірним представленням усіх типів подій.

Оцінка ефективності моделі здійснювалась на основі фактичних вимірювань із цифрового двійника та реальних даних підстанції [58]. Основними метриками стали середній час реакції, відсоток коректно визначених аномалій, стійкість під час пікових навантажень та збереження точності при довготривалому застосуванні. Результати засвідчили, що середній час обробки подій становить 130 мс, що відповідає нормативним вимогам енергетичного сектору. Проте у режимі високочастотних змін параметрів

спостерігалось збільшення часу до 210 мс, що потребувало оптимізації архітектури моделі. Це дозволило перевірити критерії операційної стійкості, визначені попередньо у методиці [59].

Важливою частиною прикладної реалізації стало тестування відповідності системи нормативним рамкам. Було створено набір аудиторських вимог, що відображають українські та міжнародні стандарти кіберзахисту для енергетичного сектору [60]. До них увійшли вимоги щодо журналювання, доступності логів, процедур форензіки та недопустимості модифікації службових команд без належної авторизації. Під час перевірки з'ясовано, що система не зберігала повних журналів проміжних рішень нейронної моделі. Це становило ризик для подальшого технічного розслідування інцидентів і було класифіковано як невідповідність нормативному критерію прозорості. Після доопрацювання система була оновлена так, щоб зберігати детальні журнали у форматі, придатному для відтворення логіки класифікації (рис. 3.4).


Рис. 3.4. Набір аудиторських вимог, які відображають українські та міжнародні стандарти кіберзахисту для енергетичного сектору

Результати всебічного аналізу дозволили сформувати інтегральну оцінку відповідності моделі етичним та ефективнісним критеріям [61]. Було встановлено, що система забезпечує високий рівень точності та задовільну

пояснюваність у більшості сценаріїв, демонструє стійкість під час навантаження, проте потребує подальшого вдосконалення щодо прозорості та рівномірності поведінки в усіх сегментах підстанції [62]. Інтегральний показник становив 0,83 за шкалою 0–1, що є достатнім для обмеженого експлуатаційного впровадження з вимогою посиленого моніторингу.

Цей кейс показує, що застосування методики оцінювання дозволяє фіксувати конкретні недоліки системи і спрямовувати подальший розвиток моделі у сторону зменшення ризиків [63]. Окрім того, процес автоматизації аудиту та формування інтегральної оцінки дає змогу контролювати вплив модифікацій системи у часі та запобігати регресії за етичними параметрами. У результаті створюється підхід, що поєднує технологічну ефективність з унеможливленням небажаної автономності та неконтрольованих рішень у критичних сферах.

### 3.3. Рекомендації для організацій щодо розробки політик та процедур впровадження етичного AI у кіберзахист

Розроблення організаційних політик та процедур впровадження етичного штучного інтелекту в системи кіберзахисту є обов’язковою складовою стратегічного управління безпекою критичних інфраструктур. Для того щоб інтеграція технологій штучного інтелекту не призводила до виникнення нових ризиків, організації повинні забезпечити суворе регламентування кожного етапу життєвого циклу AI-рішень: від проєктування та збору даних до експлуатації, моніторингу, аудиту та виведення з експлуатації [64]. Рекомендації в межах цього розділу спрямовані на формування комплексної системи управління етичністю AI, яка поєднує технологічні, нормативні та управлінські механізми. Така система має враховувати специфіку об’єкта, характер загроз, ступінь автономності моделей, вимоги до операційної безперервності та рівень допустимого впливу алгоритмів на процеси керування критичними технологіями [65].

Першим базовим елементом є створення організаційної структури, що відповідає за нагляд за етичними аспектами використання AI [66]. Організації доцільно формувати мультидисциплінарні наглядові комітети або робочі групи, до складу яких входять фахівці з кіберзахисту, експерти з етичних ризиків, інженери з машинного навчання, спеціалісти з технічного аудиту та юридичні консультанти. Їхнім завданням є узгодження принципів, формування політик, визначення меж відповідальності та встановлення критеріїв оцінки етичності рішень [67]. Такі комітети мають отримувати доступ до всієї технічної документації, журналів навчання моделей, даних про тестування й результати зовнішніх аудитів. Це забезпечує незалежну позицію при оцінюванні рішень та виключає можливість конфлікту інтересів, який може виникати у розробників або операторів системи.

Другим важливим напрямом є формування політики управління даними, оскільки будь-яка модель штучного інтелекту ґрунтується на даних, що можуть містити характерні закономірності, потенційні упередження або технічні аномалії [68]. Організації повинні розробити процедури верифікації джерел даних, встановити вимоги до репрезентативності та повноти навчальних вибірок, передбачити механізми перевірки відповідності наборів даних операційній специфіці системи. Особливе значення має регулярна перевірка старіння даних та контроль дрейфу характеристик, що можуть виникати у результаті зміни мережевого середовища або поведінки користувачів. У межах політики управління даними також визначаються вимоги до анонімізації, мінімізації обсягів даних, обмеження доступу та процедури документування операцій із датасетами [69]. Організація має забезпечити, щоб жоден елемент даних, який використовується для навчання або прогнозування, не створював загрозу інформаційній безпеці чи цілісності технологічних процесів.

Таблиця 3.3

**Ключові вимоги політики управління даними**

| Компонент | Вимога | Приклад реалізації |
|---|---|---|
| Репрезентативність | Повнота вибірок | Баланс типів подій |
| Анонімізація | Мінімізація конфіденційних даних | Видалення службових ідентифікаторів |
| Контроль дрейфу | Періодичний аналіз стабільності | Щомісячний аудит даних |

Наступною рекомендацією є обов’язковість розроблення технічних протоколів пояснюваності моделей. Пояснюваність у контексті критичних інфраструктур не є факультативною характеристикою — вона є операційною вимогою для здійснення форензіки, аудиту та прийняття рішень. Організація має визначити мінімальний набір параметрів, які модель повинна пояснювати у зрозумілій формі: тип аномалії, ключові параметри, що вплинули на класифікацію, ступінь впевненості та опис прийнятої дії [70]. Для високочутливих систем доцільно впроваджувати багаторівневі механізми пояснень: спрощені індикатори для чергового персоналу, розширені пояснення для аналітиків, детальні технічні журнали для аудиторів. Політика пояснюваності повинна містити вимоги до формату журналів, частоти їх оновлення, структури збереження та процедур доступу.

У межах рекомендацій необхідно також розробити політику контролю автономності системи. Ця політика визначає, які дії система AI може виконувати самостійно, а для яких потрібне погодження оператора або іншого контролюючого суб’єкта. Важливим є встановлення чітких порогів автономних рішень, що враховують критичність керованих процесів. У разі електроенергетичних, транспортних чи телекомунікаційних систем неприпустимо дозволяти автоматичне блокування функціональних вузлів без верифікації оператором, навіть якщо модель демонструє високу точність. Політика має передбачати процедури для трьох режимів роботи: повністю

автономного, змішаного (автономність із верифікацією) та режиму лише рекомендацій. Усі зміни режимів повинні фіксуватися в журналах та супроводжуватися обґрунтуванням.

Окремим напрямом є розроблення внутрішніх процедур оцінювання ризиків, пов'язаних з використанням AI у критичних системах. Організація повинна розробити матрицю ризиків, що включає технологічні, етичні, операційні та правові ризики. До технологічних належать некоректні класифікації, нестабільність під навантаженням, схильність до хибних спрацювань; до етичних - непрозорість рішень, неконтрольована автономність, потенційні упередження; до операційних - ризики простоїв, несанкціоновані втручання; до правових - невідповідність нормативним вимогам, відсутність належного журналювання [71]. Рекомендації передбачають регулярне оновлення матриці ризиків після кожного значного оновлення системи, зміни архітектури або появи нових даних.

Організаціям слід формулювати політики тестування та валідації AI-рішень. У рамках таких політик передбачаються процедури моделювання сценаріїв інцидентів, тестування граничних випадків, симуляції аномалій у цифрових двійниках та регулярні перевірки стабільності моделей. Тестування має включати контрольні набори даних, що містять рідкісні та високоризикові події, моделювання якісно нових атак, які модель ще не зустрічала, а також оцінку реакції на часткову відмову каналів зв'язку або обладнання. Політика тестування встановлює порогові значення допустимих показників: максимальний рівень хибнопозитивних і хибнонегативних рішень, час реагування, відсоток відхилень у пограничних сценаріях. Усі результати тестування повинні зберігатися у централізованих сховищах і підлягати перевірці під час внутрішніх або зовнішніх аудитів.

Створення політики щодо оновлення та модифікації моделей штучного інтелекту є також важливим. Оскільки моделі можуть змінювати поведінку після додаткового навчання або заміни архітектури, організації повинні встановити процедури контролю за версіями, перевіркою оновлень,

регламентами відкату до попередніх версій у разі некоректної роботи. Рекомендовано впроваджувати механізм «перевірки у пісочниці», де оновлена модель тестується в ізольованому середовищі перед розгортанням у виробничих умовах. Політика також повинна включати вимоги до документування змін: зазначення причин модифікації, очікуваного впливу, результатів тестування та оцінки ризиків, що дозволяє забезпечити прозорість процесу оновлення.

Суттєву роль відіграє політика управління інцидентами, що стосуються роботи AI. Організації повинні мати процедури виявлення та документування випадків, коли модель діяла некоректно, порушила регламент, дала непояснюване рішення або стала джерелом потенційної загрози. Політика має визначати терміни фіксації інциденту, відповідальних осіб, порядок проведення аналізу відхилень, процедури швидкого усунення наслідків та заходи з недопущення повторення аналогічних помилок. У межах цієї політики доцільно передбачати автоматичні механізми зупинки або обмеження роботи моделі при фіксації ознак некоректності.

Для забезпечення всебічного контролю необхідна політика зовнішнього та внутрішнього аудиту AI. Зовнішній аудит має на меті підтвердити дотримання етичних та безпекових вимог незалежними експертами, тоді як внутрішній аудит забезпечує регулярну оцінку дотримання організаційних процедур [72]. Політика повинна регламентувати частоту аудитів, вимоги до тестових матеріалів, правила перевірки журналів та протоколів, критерії оцінювання моделі. Організація має передбачити механізми усунення недоліків, виявлених під час аудиту, та контроль за впровадженням коригувальних заходів.

Окремим аспектом є політика підвищення кваліфікації персоналу. Оскільки ефективність роботи AI у кіберзахисті залежить не лише від моделі, а й від компетентності операторів, організація повинна забезпечити системне навчання співробітників, включаючи тренінги щодо пояснюваності рішень, правильного тлумачення вихідних даних, виявлення помилкових спрацювань,

аналізу журналів та використання інструментів аудиту. Політика має включати регулярні атестації персоналу, проведення симуляцій аварійних ситуацій та оновлення навчальних програм відповідно до змін у технологіях.

Отже, організаціям рекомендовано розробити політику взаємодії між AI-системами та іншими компонентами інфраструктури. Це включає визначення протоколів обміну даними, вимог до форматів ідентифікаторів подій, регламентів відмовостійкості, а також правил інтеграції з існуючими системами моніторингу. Політика має забезпечити уніфікацію каналів зв'язку та мінімізацію можливостей для ін'єкцій фальшивих даних. Важливим елементом є розроблення механізмів контролю сумісності під час оновлення окремих модулів системи, що запобігає потенційним конфліктам між компонентами, які можуть призвести до збоїв або обмеження функціональності (рис. 3.5).


Рис. 3.5. Модель інтеграції AI з іншими компонентами критичної інфраструктури

**Висновки до розділу 3**

У межах розділу було послідовно визначено, обґрунтовано та практично продемонстровано методику впровадження етичних підходів у розробку й експлуатацію AI-систем кіберзахисту критичних об'єктів, що охоплює формування критеріїв етичності й ефективності, їх валідацію на прикладному кейсі та розробку організаційних рекомендацій для сталого впровадження таких підходів у практику. Сукупність отриманих результатів дозволяє

стверджувати, що етичність і технічна надійність неможливо розглядати окремо у контексті критичної інфраструктури: лише системний, методологічно узгоджений підхід забезпечує контрольованість автономних рішень, зниження ймовірності небажаної поведінки та відповідність нормативним, технологічним і суспільним вимогам до безпеки.

Результати опрацювання критеріїв оцінювання AI-рішень довели, що етичність може бути формалізована у вигляді вимірюваних параметрів, які дозволяють об’єктивно оцінювати вплив системи на конфіденційність, цілісність, доступність, пояснюваність, недискримінаційність та відповідність регуляторним вимогам. Таке формалізоване представлення забезпечує не лише підвищення прозорості процесу розробки, а й можливість інтеграції етичних стандартів у процедури аудиту та технічного контролю. Встановлено, що баланс між точністю, стабільністю, операційною стійкістю та відповідністю етичним вимогам може бути досягнутий лише завдяки системному поєднанню технічних і організаційних заходів.

Практичний аналіз кейсу застосування AI у захисті критичної енергетичної системи продемонстрував, як визначені критерії функціонують у реальному середовищі, де на якість рішень впливають нерівномірність даних, високі вимоги до сталості роботи, а також потенційні ризики, пов’язані з автономними діями AI. Дослідження показало, що навіть високоточні моделі можуть створювати етичні виклики, зокрема через неповноту журналювання рішень, неоднакову поведінку в різних технічних сегментах або недостатню адаптивність під час пікових навантажень. Це підтвердило необхідність впровадження механізмів багаторівневих пояснень, розширених процедур аудиту, прозорості моделей і регулярного доопрацювання наборів даних. Прикладне тестування у середовищі цифрового двійника також засвідчило ефективність методики сценарного моделювання, яка дозволяє виявляти приховані технічні та етичні недоліки ще до реального розгортання системи.

Надані рекомендації для організацій підтвердили, що жодна технічна методика не може бути повною без належного інституційного забезпечення.

Для ефективного впровадження етичного AI у сфері кіберзахисту необхідні комплексні політики, які формалізують принципи відповідального розвитку, встановлюють процедури контролю, визначають зони відповідальності та забезпечують циклічне покращення систем. Найбільш критичними елементами є: встановлення етичних вимог на ранніх етапах життєвого циклу; прозорість алгоритмічних рішень; створення незалежних механізмів аудиту та моніторингу; ідентифікація ризиків автономності; забезпечення кадрової підготовки; інтеграція вимог до кіберстійкості у всі процеси розробки та експлуатації. Доведено, що лише у поєднанні з корпоративними регламентами технічні критерії перетворюються з рекомендаційного набору у реальні інструменти контролю та запобігання технологічним і етичним загрозам.

Загалом результати розділу демонструють, що впровадження етичних підходів у системи кіберзахисту критичних об’єктів можливе лише за умов поєднання технічних, організаційних та нормативних механізмів. Етичність AI у критичному середовищі не є абстрактним принципом, а виступає елементом безпеки, що безпосередньо впливає на надійність і контрольованість технологічних процесів. Представлена методика забезпечує можливість системного оцінювання рішень, формує базу для структурованого аудиту, створює передумови для підвищення довіри до AI-засобів і мінімізує ризики, пов’язані з використанням автономних технологій у високоризикових сферах. Отримані результати підтверджують, що етичне впровадження AI не є додатковою опцією, а становить органічну складову безпечної цифрової трансформації критичної інфраструктури.

# ВИСНОВКИ

У ході виконання дипломної роботи було здійснено комплексне теоретичне, нормативне, практичне та методологічне дослідження етичних аспектів застосування технологій штучного інтелекту у сфері кіберзахисту критичних інформаційних систем, що дало змогу сформувати цілісне бачення концептуальної природи етичних викликів, визначити стандартизовані підходи до їх мінімізації та надати практичні рекомендації для організацій, що здійснюють розвиток і експлуатацію AI-рішень у високоризикових інфраструктурах.

1. На основі аналізу встановлено, що впровадження штучного інтелекту в критичні системи не може розглядатися виключно як технологічне оновлення: воно потребує системного перегляду парадигми управління кіберзахистом, посилення нормативних механізмів, переосмислення принципів безпеки та формування нової моделі відповідальності, яка передбачає узгодженість між операторами, розробниками, регуляторами та суспільством.

Дослідження ролі критичних інформаційних інфраструктур показало, що їх функціонування забезпечує сталість національної економіки, безпеку громадян та оперативність державного управління. З огляду на системну залежність держави і суспільства від сталої роботи енергетичних, транспортних, фінансових, комунікаційних та інших життєво важливих систем, будь-яке зниження рівня їх стійкості може мати суттєві економічні, соціальні та безпекові наслідки. Це визначає особливу відповідальність щодо впровадження AI саме в цій сфері, оскільки навіть незначна помилка або хибне спрацювання автономної системи здатні спричинити масштабні порушення технологічних процесів.

Аналіз сучасних загроз підтвердив, що критична інфраструктура є постійною ціллю для державних та недержавних суб'єктів, що використовують методи складних тривалих атак, приховані маніпуляції

телеметрією, компрометацію каналів керування та зловживання доступами. На цьому фоні штучний інтелект виступає необхідним інструментом підвищення рівня ситуаційної обізнаності, стабільності систем та швидкості реагування на кіберінциденти.

Вивчення можливостей AI у кіберзахисті продемонструвало, що технології машинного навчання та глибинного аналізу даних здатні забезпечувати переваги у виявленні аномалій, поведінковому аналізі, прогнозуванні загроз, автоматизації рішень та формуванні адаптивних механізмів оборони. AI-системи уможливлюють обробку великих обсягів телеметрії, багатовимірний аналіз, виявлення слабко виражених індикаторів компрометації та побудову моделей, що враховують контекст функціонування об’єкта. Проте така автономність створює також нові ризики, пов’язані з непрозорістю рішень, неконтрольованістю поведінки, дисбалансами у даних, потенційними можливостями прихованих маніпуляцій та непередбачуваними наслідками у разі аномальних умов експлуатації. Саме тому у роботі було проаналізовано нормативно-правові та етичні рамки застосування AI у сфері кіберзахисту. Встановлено, що міжнародні підходи (ЄС, США) зосереджуються на принципах прозорості, підзвітності, справедливості, недискримінаційності та безпеки, а також на ризик-орієнтованій моделі регулювання. Україна перебуває на шляху адаптації таких підходів, що відкриває можливості для гармонізації стандартів із міжнародними практиками.

2. У другому розділі було визначено основні етичні проблеми впровадження AI у контексті кіберзахисту критичних систем. Виявлено, що найсуттєвішими викликами є недостатня пояснюваність, яка ускладнює аудит і встановлення відповідальності; можливі прояви алгоритмічної дискримінації; неочевидність механізмів прийняття рішень; високий ризик зловживань у разі отримання доступу до моделі або даних; а також імовірність виникнення «небажаної автономності» системи. Окрему увагу приділено ризикам використання AI з боку зловмисників, які можуть застосовувати генеративні моделі для автоматизації атак, створення фішингових кампаній, маніпуляції

індикаторами, обходу механізмів виявлення або підробки аномальних сигналів. Аналіз міжнародних підходів до етичного регулювання підтвердив, що провідні країни світу надають перевагу розбудові системи багаторівневого контролю, включно з незалежними аудитами, сертифікаційними механізмами, обов'язковим документуванням рішень та впровадженням інструментів оцінювання ризиків на всіх етапах життєвого циклу AI.

3. У третьому розділі було розроблено і запропоновано методику впровадження етичних підходів у розробку та експлуатацію AI-систем кіберзахисту критичних об’єктів. Методика включає формування критеріїв оцінювання моделей, практичну перевірку таких критеріїв на прикладі цифрового двійника енергетичної підстанції та рекомендації щодо інституційного впровадження етичного AI. Критерії охоплюють пояснюваність, стійкість, точність, збалансованість даних, відповідність нормативам, недискримінаційність, операційну надійність та контроль автономності. Завдяки багаторівневій структурі критеріїв було сформовано основу для цілісного аудиту моделей. Практичне застосування методики на прикладному кейсі показало, що системи AI можуть мати приховані слабкі місця, які не виявляються у звичайних тестах, але стають очевидними під час сценарного моделювання. Зокрема було встановлено, що нерівномірність навчального набору призводить до різної чутливості моделі в окремих сегментах енергетичного обладнання, а неповнота журналювання рішень унеможливлює повноцінну форензіку інцидентів. Дослідження підтвердило, що правильне застосування критеріїв етичності дозволяє системно підвищувати надійність AI-рішень.

4. Запропоновані рекомендації для організацій продемонстрували, що впровадження етичного AI є не лише технічним, а передусім управлінським завданням. Політики організацій повинні визначати правила використання даних, стандарти прозорості, процедури аудиту, механізми перевірки стійкості моделей, процеси документування та перегляду рішень, а також план реагування на інциденти, пов'язані з роботою AI. Необхідною умовою є

впровадження спеціалізованих ролей, таких як AI-аудитор, офіцер з етики AI, інженер із валідації моделей, а також формування міждисциплінарних груп, що поєднують технічні, юридичні та організаційні компетенції. Таким чином, створюється інституційна система, здатна забезпечити безпечний життєвий цикл AI у критичних середовищах.

Отримані результати дозволяють сформулювати узагальнений висновок: впровадження етичних підходів до застосування AI у кіберзахисті критичних систем є ключовою умовою забезпечення національної та технологічної стійкості. Штучний інтелект підвищує рівень адаптивності захисних механізмів, але вимагає суворих правил контролю, прозорості та відповідальності, що мінімізують ризики автономного прийняття рішень і потенційних негативних наслідків. Методика, розроблена в межах роботи, створює основу для уніфікованого оцінювання моделей, їх практичної апробації та інтеграції етичних стандартів у всі етапи їх розробки та експлуатації. Подальший розвиток регуляторного поля, удосконалення процедур аудиту, адаптація організаційних структур і підготовка фахівців є необхідними передумовами безпечного використання AI у критичних інформаційних інфраструктурах. Застосування таких підходів відкриває можливості для підвищення довіри до технологій, зниження рівня системних ризиків і забезпечення сталого та етично відповідального розвитку цифрової держави.

# СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Ethics in artificial intelligence: an approach to cybersecurity / A. López González et al. *Inteligencia artificial*. 2024. Vol. 27, no. 73. P. 38–54. URL: https://doi.org/10.4114/intartif.vol27iss73pp38-54 (date of access: 24.09.2025).

2. Goel P. K. Ethical considerations in implementing artificial intelligence in cybersecurity. *Redefining security with cyber AI*. 2024. P. 72–91. URL: https://doi.org/10.4018/979-8-3693-6517-5.ch005 (date of access: 26.09.2025).

3. Kulothungan V. Securing the AI frontier: urgent ethical and regulatory imperatives for ai-driven cybersecurity. *2024 IEEE international conference on big data (bigdata)*, Washington, DC, USA, 15–18 December 2024. 2024. P. 5602–5609. URL: https://doi.org/10.1109/bigdata62323.2024.10826010 (date of access: 26.09.2025).

4. National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0)**.** *NIST*. Gaithersburg, MD : NIST, 2023. 84 с. URL: https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf (date of access: 28.09.2025).

5. National Institute of Standards and Technology. AI Risk Management Framework Supplementary materials. *NIST*. Gaithersburg, MD : NIST, 2024. URL: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf (date of access: 28.09.2025).

6. UNESCO. Recommendation on the Ethics of Artificial Intelligence. *UNESCO*. Paris : UNESCO, 2021. URL: https://unesdoc.unesco.org/ark:/48223/pf0000380455 (date of access: 28.09.2025).

7. IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design. First edition. *IEEE*. Piscataway, NJ : IEEE, 2019. URL: https://engagestandards.ieee.org/rs/211-FYL-955/images/EAD1e.pdf (date of access: 30.09.2025).

8. IEEE Standards Association. Ethically Aligned Design — Version 2 (Request for Input). *IEEE SA*. 2024. URL: https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf (date of access: 30.09.2025).

9. European Commission. Proposal for a Regulation laying down harmonised rules on Artificial Intelligence (AI Act). *European Commission. COM(2021)206 final*. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206 (date of access: 30.09.2025).

10. European Union Agency for Cybersecurity (ENISA). Artificial Intelligence – Cybersecurity challenges. *ENISA*, 2020. URL: https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges (date of access: 03.10.2025).

11. ENISA. Multilayer Framework for Good Cybersecurity Practices for AI. *ENISA*, 2023. URL: https://www.enisa.europa.eu/sites/default/files/publications/Multilayer%20Framework%20for%20Good%20Cybersecurity%20Practices%20for%20AI.pdf (date of access: 03.10.2025).

12. ISO/IEC JTC 1/SC 42. ISO/IEC TR 24028:2020 – Information technology – Artificial intelligence – Overview of trustworthiness in artificial intelligence. Geneva : ISO/IEC, 2020. URL: https://www.iso.org/standard/77624.html (date of access: 03.10.2025).

13. ISO/IEC. ISO/IEC 27001:2013 (updated 2022) – Information security management systems – Requirements. Geneva : ISO/IEC. — URL: https://www.iso.org/standard/54534.html (date of access: 13.10.2025).

14. U.S. Department of Homeland Security. Framework for Using AI in Critical Infrastructure. DHS, 2024. URL: https://apnews.com/article/ (date of access: 13.10.2025).

15. Goldilock. NATO-backed report. Agentic malware and AI-driven cyber threats. Axios, 2025. URL: https://www.axios.com/2025/01/07/goldilock-agentic-malware-2027-doomsday (date of access: 13.10.2025).

16. Henschke A. Cybersecurity, Critical Infrastructure, and Ethics 4TU.Research, 2022. URL: https://www.4tu.nl/ethics/downloads/default/files/henschke-cybesecurity-critical-infrastructure-and-ethics.pdf (date of access: 17.10.2025).

17. Bostrom N., Yudkowsky E. The ethics of artificial intelligence Cambridge Handbook of Artificial Intelligence. Cambridge : Cambridge Univ. Press, 2014. C. 316–334. (date of access: 17.10.2025).

18. Floridi L., Cowls J., Beltrametti M. et al. AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines. 2018. Vol. 28, №4. P. 689–707. (date of access: 19.10.2025).

19. Mittelstadt B.D. Principles alone cannot guarantee ethical AI. Nature Machine Intelligence. 2019. Vol. 1. P. 501–507. (date of access: 19.10.2025).

20. Hecht B., et al. Ethics in cyber-defence systems: balancing automation and human oversight. Journal of Cybersecurity Practice. 2021. Vol. 2, №3. (date of access: 19.10.2025).

21. Goodfellow I., McDaniel P., Papernot N. Making machine learning robust against adversarial inputs. Communications of the ACM. 2018. Vol. 61, №7. P. 56–66. (date of access: 20.10.2025).

22. Szegedy C., Zaremba W., Sutskever I. Intriguing properties of neural networks. ICLR Workshop Paper, 2014. (date of access: 20.10.2025).

23. Sokol K., Flach P. Explainability and interpretability in security-focused AI systems. IEEE Security & Privacy. 2020. Vol. 18, №3. P. 25–33. (date of access: 21.10.2025).

24. Ribeiro M.T., Singh S., Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. KDD, 2016. P. 1135–1144. (date of access: 21.10.2025).

25. Doshi-Velez F., Kim B. Towards a rigorous science of interpretable machine learning. 2017. URL: https://arxiv.org/abs/1702.08608 (date of access: 23.10.2025).

26. Cummings M.L. Automation and accountability in autonomous cyber-defence. Ethics Inf Technol. 2017. Vol. 19, №2. P. 137–149. (date of access: 23.10.2025).

27. Brundage M., Avin S., Clark J. et al. The malicious use of artificial intelligence: forecasting, prevention, and mitigation. Cambridge : Future of Humanity Institute OpenAI, 2018. URL: https://maliciousaireport.com/ (date of access: 23.10.2025).

28. UK National Cyber Security Centre (NCSC). Securing AI systems — guidance. London : NCSC, 2021. URL: https://www.ncsc.gov.uk/collection/ai (date of access: 23.10.2025).

29. European Commission. High-Level Expert Group on AI — Ethics Guidelines for Trustworthy AI. Brussels : EC, 2019. URL: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (date of access: 27.10.2025).

30. O'Neil C. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. New York : Crown, 2016. 272 c. (date of access: 27.10.2025).

31. Taddeo M., Floridi L. How AI can be a force for good. Science. 2018. Vol. 361. P. 751–752. (date of access: 27.10.2025).

32. Raji I.D., Smart A., White R.N. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. FAT Conference, 2020. (date of access: 27.10.2025).

33. Kuner C., Marelli M. Regulating AI in cybersecurity: data protection and human rights implications. International Data Privacy Law. 2021. Vol. 11, №2. (date of access: 27.10.2025).

34. Abadi M., et al. Differential privacy for deep learning with only a few lines of code. ICML Workshop, 2016. (date of access: 27.10.2025).

35. Finlayson S.G., et al. Adversarial attacks against medical deep learning systems. Science. 2019. Vol. 363, №6433. P. 1287–1290. (date of access: 27.10.2025).

36. Carroll J.M., et al. Human oversight in autonomous cyber-defense: ethical and operational considerations. ACM Transactions on Cyber-Physical Systems, 2022. (date of access: 27.10.2025).

37. Chertoff M., Simon T. AI and national security: balancing innovation and risk. Journal of Strategic Security. 2020. Vol. 13, №4. (date of access: 27.10.2025).

38. Horvitz E., et al. Principles of mixed-initiative user interfaces. AAAI, 1999. (date of access: 04.11.2025).

39. Bellovin S.M., et al. Risks of machine learning in cyber operations. Communications of the ACM. 2020. Vol. 63, №7. P. 46–54. (date of access: 04.11.2025).

40. Kroll J.A., et al. Accountable algorithms. University of Pennsylvania Law Review. 2017. Vol. 165. P. 633–705. (date of access: 04.11.2025).

41. National Academies of Sciences. AI and the future of cybersecurity: opportunities and risks. Washington, DC : NAS, 2020. (date of access: 04.11.2025).

42. Woolley S., et al. AI governance for critical infrastructure. Atlantic Council, 2021. URL: https://www.atlanticcouncil.org/ (date of access: 04.11.2025).

43. Brundage M., et al. The malicious use of AI in cybersecurity: scenarios and policy responses. Policy brief, 2018. (date of access: 04.11.2025).

44. Srinivasan S., et al. Secure and private ML pipelines for critical systems. IEEE Transactions on Dependable and Secure Computing. 2022. (date of access: 04.11.2025).

45. Russell S., Norvig P. Artificial Intelligence: A Modern Approach. 4-е вид. London : Pearson, 2021. (date of access: 12.11.2025).

46. Kshetri N. Big data’s role in expanding access to critical infrastructure protection. *Journal of Cybersecurity*. 2019. (date of access: 12.11.2025).

47. Miller T. Explainable AI: a guide for policymakers. Brookings Institution, 2019. URL: https://www.brookings.edu/ (date of access: 12.11.2025).

48. Wachter S., Mittelstadt B., Russell C. Why fairness cannot be automated: bridging the gap between EU non-discrimination law and AI. Law, Innovation and Technology. 2021. (date of access: 12.11.2025).

49. OECD. Recommendation of the Council on AI OECD, 2019. URL: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 (date of access: 12.11.2025).

50. European Parliamentary Research Service. The AI Act: implications for critical systems and cybersecurity - EPRS Briefing, 2024. (date of access: 12.11.2025).

51. ENISA. AI Threat Landscape. AI-specific cybersecurity threats - ENISA report series (2020–2024). URL: https://www.enisa.europa.eu/ (date of access: 12.11.2025).

52. Raji I.D., et al. AI Now 2018 Report - Recommendations for civil society and policy. AI Now Institute, 2018. URL: https://ainowinstitute.org/AI_Now_2018_Report.pdf (date of access: 12.11.2025).

53. Mogensen P., et al. Ethical governance of AI in the energy sector (critical infrastructure use-case). Energy Policy. 2022. (date of access: 15.11.2025).

54. Papernot N., McDaniel P., Goodfellow I. Practical black-box attacks against machine learning. ASIACCS, 2016. (date of access: 15.11.2025).

55. Shneiderman B. Bridging the gap between ethics and practice: guidelines for human-centered AI. ACM Interactions, 2020. (date of access: 15.11.2025).

56. Taddeo M., Floridi L. Ethical governance of cybersecurity systems powered by AI. Ethics and Information Technology. 2019. Vol. 21. P. 29–41. (date of access: 15.11.2025).

57. Carter E.R., Cole J. Regulatory and Ethical Challenges of AI Deployment in Critical Cybersecurity Infrastructure ResearchGate, 2025. URL: https://www.researchgate.net/publication/394535286_Regulatory_and_Ethical_Challenges_of_AI_Deployment_in_Critical_Cybersecurity_Infrastructure (date of access: 15.11.2025).

58. Андрух А. Cyber security of critical infrastructure. CIMS FTI, 2024. URL: https://cims.fti.dp.ua/j/article/view/210 (date of access: 19.11.2025).

59. Lata M., Kumar V. Cyber security techniques in cloud environment: comparative analysis of public, private and hybrid cloud. *Edpacs*. 2025. P. 1–21. URL: https://doi.org/10.1080/07366981.2025.2449743 (date of access: 19.11.2025).

60. Li T., Yan L. SIEM based on big data analysis. *Cloud computing and security*. Cham, 2017. P. 167–175. URL: https://doi.org/10.1007/978-3-319-68505-2_15 (date of access: 19.11.2025).

61. Network attack prediction method based on threat intelligence / J. Wang et al. *Cloud computing and security*. Cham, 2018. P. 151–160. URL: https://doi.org/10.1007/978-3-030-00012-7_14 (date of access: 19.11.2025).

62. Network security analysis for cloud computing environment / L. Xie et al. *International journal of modeling, simulation, and scientific computing*. 2022. URL: https://doi.org/10.1142/s1793962322500544 (date of access: 24.11.2025).

63. Security issues and research challenges in cloud computing / B. B. Jagadale et al. *International journal of engineering and computer science*. 2024. Vol. 13, no. 05. P. 26158–26171. URL: https://doi.org/10.18535/ijecs/v13i05.4820 (date of access: 24.11.2025).

64. Security threat analysis in cloud computing / M. Iqbal Fadillah et al. *JATI (jurnal mahasiswa teknik informatika)*. 2024. Vol. 9, no. 1. P. 992–998. URL: https://doi.org/10.36040/jati.v9i1.12528 (date of access: 24.11.2025).

65. Singh U., Tiwari A., Sharma S. Data security in cloud computing. *International journal of engineering applied sciences and technology*. 2020. Vol. 04, no. 10. P. 170–173. URL: https://doi.org/10.33564/ijeast.2020.v04i10.033 (date of access: 24.11.2025).

66. Surveying and analyzing security, privacy and trust issues in cloud computing environments / D. Sun et al. *Procedia engineering*. 2011. Vol. 15. P. 2852–2856. URL: https://doi.org/10.1016/j.proeng.2011.08.537 (date of access: 24.11.2025).

67. Artificial intelligence in cybersecurity: protecting national infrastructure: A USA review / Adebunmi Okechukwu Adewusi et al. *World journal of advanced research and reviews*. 2024. Vol. 21, no. 1. P. 2263–2275. URL: https://doi.org/10.30574/wjarr.2024.21.1.0313 (date of access: 24.11.2025).

68. Cybersecurity and artificial intelligence (AI) / C. Rios-Campos et al. *South florida journal of development*. 2024. Vol. 5, no. 8. P. e4276. URL: https://doi.org/10.46932/sfjdv5n8-021 (date of access: 27.11.2025).

69. Ethical and strategic implications of AI in cybersecurity / N. Pardhi et al. *2025 IEEE international conference on computer, electronics, electrical engineering & their applications (IC2E3)*, Srinagar Garhwal, India, 15–16 May 2025. 2025. P. 1–6. URL: https://doi.org/10.1109/ic2e365635.2025.11166708 (date of access: 27.11.2025).

70. Ethical Challenges in AI-Driven Cybersecurity Decision-Making / Emmanuel Cadet et al. *International journal of scientific research in computer science, engineering and information technology*. 2024. Vol. 10, no. 3. P. 1031–1064. URL: https://doi.org/10.32628/cseit25113577 (date of access: 27.11.2025).

71. Marzoog Al-Mukhtar W. N. AI in cybersecurity: transformative approaches to safeguarding information technology systems. *Turkish journal of computer and mathematics education (TURCOMAT)*. 2024. Vol. 15, no. 3. P. 391–412. URL: https://doi.org/10.61841/turcomat.v15i3.14945 (date of access: 29.11.2025).

72. Vemuri N., Thaneeru N., Tatikonda V. M. Securing trust: ethical considerations in AI for cybersecurity. *Journal of knowledge learning and science technology ISSN: 2959-6386 (online)*. 2023. Vol. 2, no. 2. P. 167–175. URL: https://doi.org/10.60087/jklst.vol2.n2.p175 (date of access: 29.11.2025).