МІНІСТЕРСТВО ОСВІТИ ТА НАУКИ УКРАЇНИ
КРИВОРІЗЬКИЙ ФАХОВИЙ КОЛЕДЖ
ДЕРЖАВНОГО НЕКОМЕРЦІЙНОГО ПІДПРИЄМСТВА
«ДЕРЖАВНИЙ УНІВЕРСИТЕТ «КИЇВСЬКИЙ АВІАЦІЙНИЙ ІНСТИТУТ»
Циклова комісія *комп'ютерних систем та мереж*
(повна назва циклової комісії)

Допустити до захисту
Голова випускової циклової комісії
комп'ютерних систем та мереж
(повна назва циклової комісії)
(підпис) *Ірина КРАВЧУК*
(ім'я, ПРІЗВИЩЕ)
« 10 » 06 2025 р.

КВАЛІФІКАЦІЙНА РОБОТА
(ПОЯСНЮВАЛЬНА ЗАПИСКА)

ВИПУСКНИКА ОСВІТНЬО-ПРОФЕСІЙНОГО СТУПЕНЯ
ФАХОВИЙ МОЛОДШИЙ БАКАЛАВР

Тема: *«Дослідження питань етичних норм в процесі розвитку штучного Інтелекту»*

Група: *3-011* Спеціальність: *123 «Комп'ютерна інженерія»*

Здобувач освіти (підпис) *Богдан ШАПОВАЛ* (ім'я, ПРІЗВИЩЕ)

Керівник роботи (підпис) *Сергій РУДИЙ* (ім'я, ПРІЗВИЩЕ)

Консультант з оформлення пояснювальної записки (підпис) *Оксана ОСАДЧА* (ім'я, ПРІЗВИЩЕ)

Кривий Ріг 2025 р.

КРИВОРІЗЬКИЙ ФАХОВИЙ КОЛЕДЖ
ДЕРЖАВНОГО НЕКОМЕРЦІЙНОГО ПІДПРИЄМСТВА
«ДЕРЖАВНИЙ УНІВЕРСИТЕТ «КИЇВСЬКИЙ АВІАЦІЙНИЙ ІНСТИТУТ»

Відділення *комп'ютерної та програмної інженерії*
Циклова комісія *комп'ютерних систем та мереж*
Освітньо-професійний ступінь *фаховий молодший бакалавр*
Спеціальність *123 «Комп'ютерна інженерія»*

ЗАТВЕРДЖУЮ
Голова випускової циклової комісії
комп'ютерних систем та мереж
(повна назва циклової комісії)
(підпис) *Ірина КРАВЧУК* (ім'я, ПРІЗВИЩЕ)
« *01* » *03* 2025 р.

ЗАВДАННЯ

НА КВАЛІФІКАЦІЙНУ РОБОТУ ЗДОБУВАЧУ ОСВІТИ

ШАПОВАЛА Богдана Олександровича
(прізвище, ім'я, по батькові)

1. Тема роботи *Дослідження питань етичних норм в процесі розвитку штучного інтелекту*

Керівник роботи *Рудий Сергій Володимирович, викладач вищої категорії*
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

затверджені наказом по коледжу від « *04* » *04* 2025 року № *50-ст*

2. Строк подання здобувачем освіти роботи з _____ по _____

3. Вихідні дані до роботи *Структурований підхід до інтеграції етичних принципів на всіх етапах життєвого циклу ШІ. Підвищення довіри до технологій ШІ та зменшенню соціальних і правових ризиків.*

4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити)

Обґрунтування необхідності етичних норм у розвитку штучного інтелекту

Аналіз відповідальних фреймворків країн-лідерів

Розробка стратегії пом'якшення упередженості штучного інтелекту

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень)

Презентація Microsoft PowerPoint

6. Консультанти розділів роботи (проекту)

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв

7. Дата видачі завдання _____

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів кваліфікаційної роботи	Строк виконання етапів роботи	Примітка
1	*Узгодження технічного завдання з керівником кваліфікаційної роботи*	*04.04.2025-07.04.2025*	*виконано*
2	*Підбір та вивчення науково-технічної літератури за темою кваліфікаційної роботи*	*08.04.2025-14.04.2025*	*виконано*
3	*Обґрунтування необхідності етичних норм у розвитку штучного інтелекту*	*15.04.2025-21.04.2025*	*виконано*
4	*Аналіз відповідальних фреймворків країн-лідерів*	*22.04.2025-28.04.2025*	*виконано*
5	*Розробка стратегії пом'якшення упередженості штучного інтелекту*	*29.04.2025-02.05.2025*	*виконано*
7	*Написання та оформлення пояснювальної записки*	*26.05.2025-30.05.2025*	*виконано*
8	*Попередній захист кваліфікаційної роботи*	*09.06.2025-12.06.2025*	*виконано*
9	*Захист кваліфікаційної роботи*		

Здобувач освіти _____ *Богдан ШАПОВАЛ*
(підпис) (ім'я, ПРІЗВИЩЕ)

Керівник роботи _____ *Сергій РУДИЙ*
(підпис) (ім'я, ПРІЗВИЩЕ)

StrikePlagiarism.com

Дата звіту 6/4/2025
Дата редагування —

Звіт не був оцінений

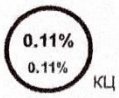
Звіт подібності

метадані

Назва організації
Ukrainian national aviation university
Заголовок
123_Шаповал_2025-КПІ
Автор Науковий керівник / Експерт
ШаповалРудий С.
підрозділ
Криворізький Фаховий коледж

Обсяг знайдених подібностей

Коефіцієнт подібності визначає, який відсоток тексту по відношенню до загального обсягу тексту було знайдено в різних джерелах. Зверніть увагу, що високі значення коефіцієнта не автоматично означають плагіат. Звіт має аналізувати компетентна / уповноважена особа.



25	10724	83494
Довжина фрази для коефіцієнта подібності 2	Кількість слів	Кількість символів

Тривога

У цьому розділі ви знайдете інформацію щодо текстових спотворень. Ці спотворення в тексті можуть говорити про МОЖЛИВІ маніпуляції в тексті. Спотворення в тексті можуть мати навмисний характер, але частіше характер технічних помилок при конвертації документа та його збереженні, тому ми рекомендуємо вам підходити до аналізу цього модуля відповідально. У разі виникнення запитань, просимо звертатися до нашої служби підтримки.

Заміна букв		0
Інтервали		0
Мікропробіли		3
Білі знаки		0
Парафрази (SmartMarks)		13

Подібності за списком джерел

Нижче наведений список джерел. В цьому списку є джерела із різних баз даних. Колір тексту означає в якому джерелі він був знайдений. Ці джерела і значення Коефіцієнту Подібності не відображають прямого плагіату. Необхідно відкрити кожне джерело і проаналізувати зміст і правильність оформлення джерела.

10 найдовших фраз		Колір тексту
ПОРЯДКОВИЙ НОМЕР	НАЗВА ТА АДРЕСА ДЖЕРЕЛА URL (НАЗВА БАЗИ)	КІЛЬКІСТЬ ІДЕНТИЧНИХ СЛІВ (ФРАГМЕНТІВ)
1	https://arxiv.org/pdf/2402.15770	22 0.21 %
2	https://journal.iistr.org/index.php/BST/article/view/631	15 0.14 %
3	https://link.springer.com/article/10.1007/s43681-023-00307-3	14 0.13 %
4	https://link.springer.com/chapter/10.1007/978-981-97-9251-1_14	14 0.13 %
5	https://deepai.org/publication/ai-ethics-principles-in-practice-perspectives-of-designers-and-developers	14 0.13 %

РЕФЕРАТ

Дипломна робота «Дослідження питань етичних норм в процесі розвитку штучного інтелекту» містить _59_ сторінок, _10_ рисунків, _17_ використаних джерел.

ЕТИЧНІ НОРМИ, ШТУЧНИЙ ІНТЕЛЕКТ (ШІ), ПРОЗОРІСТЬ, СПРАВЕДЛИВІСТЬ, КОНФІДЕНЦІЙНІСТЬ, ПІДЗВІТНІСТЬ.

Дипломна робота присвячена обґрунтуванню необхідності етичних норм у розвитку штучного інтелекту (ШІ) з метою забезпечення його безпечного та відповідального використання. У роботі розглянуто ключові етичні аспекти, такі як прозорість алгоритмів, пояснюваність рішень, запобігання упередженості, захист конфіденційності даних та підзвітність ШІ-систем.

Запропоновано системний підхід до інтеграції етичних принципів у розробку ШІ, що включає формулювання та впровадження етичних стандартів, створення механізмів контролю за відповідністю цим стандартам та забезпечення їхнього дотримання на всіх етапах життєвого циклу ШІ. Представлено методи аналізу впливу етичних норм на функціонування ШІ-систем та розглянуто роль міжнародних нормативно-правових актів у регулюванні етичних питань у сфері ШІ.

Результати дослідження підкреслюють необхідність розробки та впровадження етичних регламентів для запобігання потенційним загрозам, пов’язаним із використанням ШІ. Запропоновані підходи сприятимуть підвищенню довіри суспільства до ШІ-технологій, забезпеченню їхньої справедливості та мінімізації ризиків для користувачів.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

ADWIN - Adaptive Windowing

AI - Artificial intelligence

AWS - Autonomous Weapons Systems

CX - Counterfactual Explanations

DLT - Distributed Ledger Technology

FRL - Fair Representation Learning

FTA - Fairness through Awareness

LIME - Local Interpretable Model-agnostic Explanations

SHAP - Shapley Additive Explanations

XAI - Explainable AI

OECP - Організації економічного співробітництва та розвитку ШІ - Штучний інтелект

ВСТУП

Розширення штучного інтелекту (ШІ) у таких сферах, як охорона здоров'я, фінанси та зв'язок, викликало серйозні етичні проблеми, пов'язані з прозорістю, справедливістю та конфіденційністю. Вирішення цих питань є критично важливим для відповідального розвитку та впровадження ШІ. У даній дипломній роботі пропонується провести дослідження, що формує комплексну етичну основу, яка спрямована на мінімізацію упередженості та забезпечення підзвітності технологій ШІ.

Порівняльний аналіз міжнародної політики штучного інтелекту в різних регіонах, зокрема в Європейському Союзі, США та Китаї, здійснюється за допомогою аналітичних інструментів, таких як діаграми Венна та декартові графіки. Використання цих інструментів дозволяє систематично оцінювати етичні принципи, які визначають розвиток ШІ в різних юрисдикціях. Результати дослідження виявляють значні відмінності у пріоритетах глобальних регіонів щодо прозорості, чесності та конфіденційності, що ускладнює створення єдиних етичних стандартів.

Останніми роками в етиці штучного інтелекту відбувся значний прогрес, що включає дискусії про прозорість, справедливість та конфіденційність у розробці ШІ.

Дослідження у [1] висвітлюють проблему упередженості у великих мовних моделях, що може сприяти збереженню соціальної нерівності. Аналогічно, у [2] розглядають відповідність базових моделей проекту Закону ЄС про ШІ, наголошуючи на необхідності підзвітності. Крім того, у [3] були запропоновані нові стратегії даних для інтеграції етичних принципів у розвиток ШІ.

Ключові регуляторні рамки, такі як Закон ЄС про штучний інтелект [4] та стандарти NIST (2024), підкреслюють важливість розробки прозорих та справедливих систем ШІ, які враховують конфіденційність даних та суспільні цінності. Це дослідження спирається на сучасні досягнення, пропонуючи порівняльний аналіз стратегічних підходів до етичного розвитку ШІ у світовому контексті.

Політика етики ШІ значно залежить від культурного, соціально-політичного та економічного контексту. Європейський Союз приділяє особливу увагу захисту прав особи та прозорості, США зосереджуються на інноваціях і ринкових механізмах, а Китай інтегрує ШІ у державну політику, наголошуючи на соціальній гармонії та безпеці [5]. Ці розбіжності ускладнюють розробку універсальних етичних стандартів та вимагають ретельного аналізу.

У дипломній роботі передбачається аналіз спільних етичних принципів, таких як справедливість, прозорість та конфіденційність, оцінюючи їхню реалізацію у різних країнах. Використовуючи інструменти порівняння, такі як діаграми Венна та декартові графіки, передбачається демонстрація як відмінності, так і можливі точки зближення політик. Основною метою дипломної роботи є не лише виявлення розбіжностей, а й пошук шляхів гармонізації етичного регулювання ШІ на глобальному рівні. В роботі пропонується стратегія пом’якшення упередженості, етичного моніторингу та створення міждисциплінарних команд розробників.

Мета та завдання дипломної роботи має три основні цілі. По-перше, прагнення дослідити етичні аспекти, властиві розробці штучного інтелекту. Це включає детальний аналіз фундаментальних етичних принципів прозорості,

справедливості та конфіденційності в системах ШІ, а також розуміння їх взаємозв'язку та значення окремо.

По-друге, критичний аналіз різних глобальних нормативно-правових рамок щодо ШІ, з особливим акцентом на політики Європейського Союзу, США та Китаю. Мета полягає в тому, щоб виявити їхні подібності та відмінності, а також зрозуміти їхнє значення для міжнародного управління ШІ.

По-третє, робота має на меті узагальнити підходи та практики, спрямовані на виявлення та пом'якшення упередженості в системах ШІ, забезпечуючи їхню справедливість і надійність.

Таким чином, у даній дипломній роботі буде розглядатися нагальна потреба в інтеграції етичних принципів у розробку та використання ШІ. Це дозволить не лише

створювати технологічно досконалі рішення, а й забезпечувати їх відповідність суспільним цінностям та міжнародним стандартам. Глобальна співпраця у сфері етики ШІ стане ключовим фактором у забезпеченні сталого та відповідального розвитку цієї технології.

РОЗДІЛ 1
ОБҐРУНТУВАННЯ НЕОБХІДНОСТІ ЕТИЧНИХ НОРМ У РОЗВИТКУ ШТУЧНОГО ІНТЕЛЕКТУ

1.1 Необхідність етичних міркувань в сфері штучного інтелекту

Поява штучного інтелекту (ШІ) ознаменувала нову епоху в технологічній еволюції, суттєво вплинувши на різні сектори, зокрема охорону здоров'я, фінанси, транспорт і комунікації [7]. Ця безпрецедентна інтеграція ШІ в суспільну структуру вимагає невідкладної розробки надійних етичних рамок. Вони повинні охоплювати складнощі, властиві технологіям ШІ, такі як конфіденційність даних, непрозорість алгоритмів, справедливість у процесі прийняття рішень і ширший суспільний вплив.

Етичні міркування щодо ШІ виходять за межі академічних дискусій, маючи значні наслідки у реальному світі. Серед ключових питань — конфіденційність даних і необхідність отримання усвідомленої згоди, оскільки персональна інформація часто використовується для роботи алгоритмів ШІ. Не менш важливими є прозорість і пояснюваність цих алгоритмів, що є критично необхідними для підтримки довіри суспільства, особливо у високоризикових сферах, таких як юридичні рішення або медична діагностика. Крім того, важливим етичним викликом є забезпечення справедливості та уникнення вбудованих упереджень у системах ШІ, оскільки вони мають тенденцію відтворювати та закріплювати існуючі суспільні нерівності.

Потреба в етично обґрунтованому ШІ продиктована не лише необхідністю запобігання шкоді та забезпечення справедливості, а й стратегічною метою сприяння сталим, соціально корисним та загальноприйнятим інноваціям.

1.2. Етичні аспекти розробки штучного інтелекту

Етичні аспекти відіграють найважливішу роль у розробці ШІ, оскільки гарантують безпеку, справедливість і прозорість цих систем. Діаграма Венна (рис.1.1) ілюструє взаємозв’язок трьох ключових аспектів: прозорості, справедливості та конфіденційності.

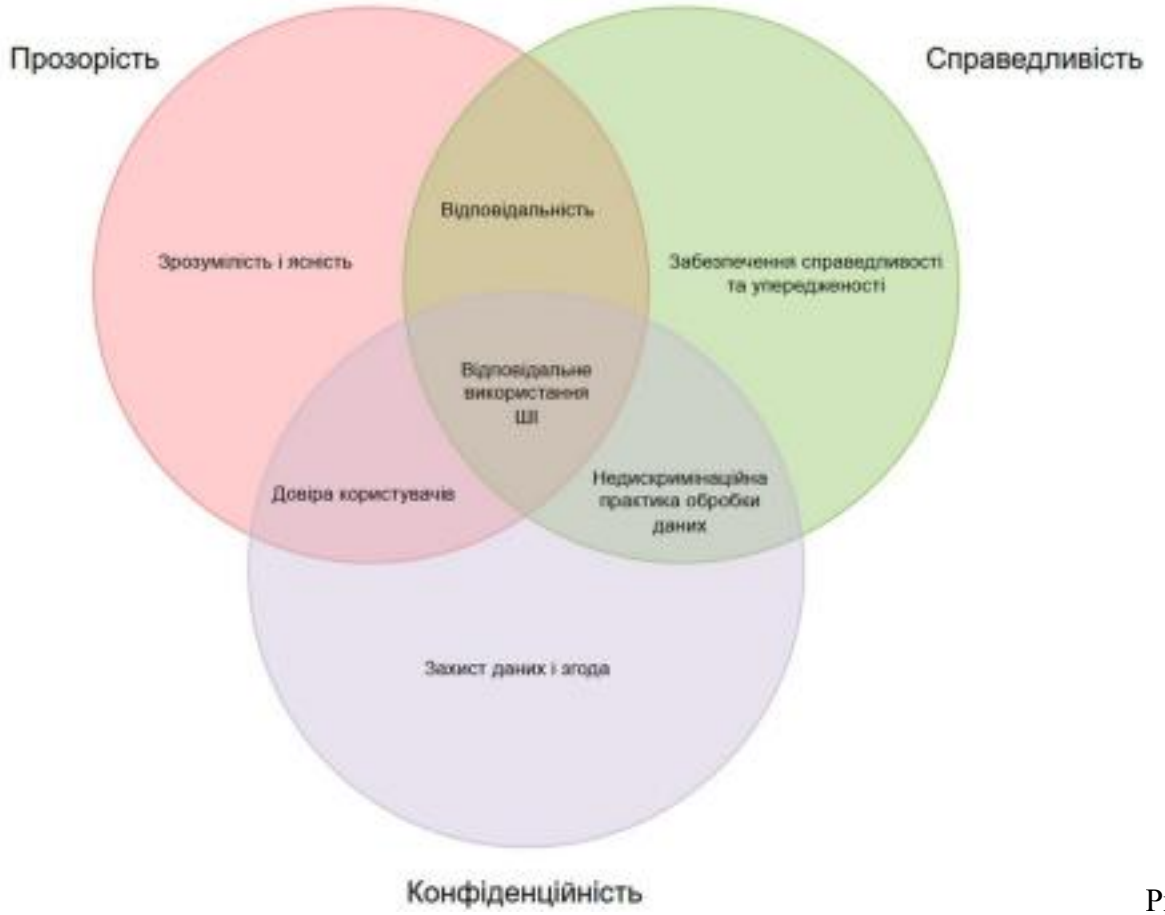


Рисунок 1.1 - Взаємопов’язані концепції Основ для розробки систем штучного інтелекту, які втілюють прозорість, пояснюваність і зрозумілість

Прозорість, пояснюваність та зрозумілість означають, що системи ШІ мають бути зрозумілими для користувачів. Справедливість вимагає, щоб ШІ-системи розроблялися без упередженості та дискримінації. Конфіденційність, або «Захист

даних і згода», зосереджується на захисті персональних даних і необхідності отримання згоди на їх використання.

Ці аспекти взаємопов’язані, а їхні перетини в діаграмі Венна показують області їхнього перекриття. Наприклад, перетин прозорості та справедливості називається підзвітністю, що підкреслює важливість прозорого й справедливого ухвалення рішень ШІ. Перетин прозорості та конфіденційності, відомий як довіра користувачів, наголошує на необхідності прозорості у використанні та захисті даних [3]. Перетин справедливості та конфіденційності, відомий як недискримінаційні практики використання даних, підкреслює необхідність узгодження

конфіденційності зі справедливістю, щоб уникнути дискримінації.

Центральний перетин діаграми Венна представляє ідеал відповідального використання ШІ, який збалансовує прозорість, справедливість і конфіденційність. Для цього використовується блок-схема, що надає покрокове керівництво щодо впровадження етичних принципів у розробку ШІ (рис. 2.1).

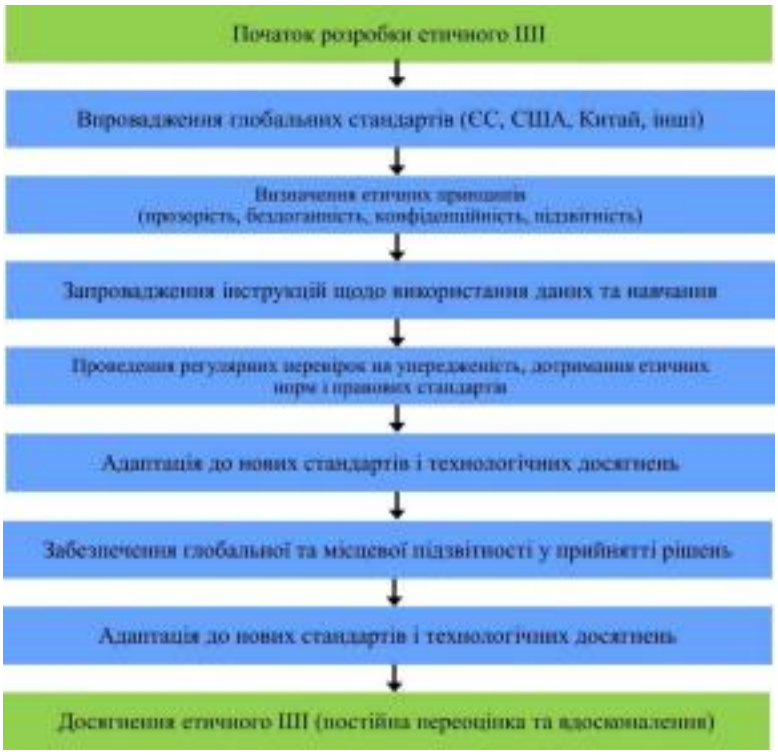


Рисунок 1.2 - Основа для розробки систем ШІ, які втілюють етичні принципи та забезпечують безпеку та довіру усіх користувачів

Процес починається з прихильності до етичної розробки ШІ та визначення ключових етичних принципів: прозорості, справедливості та конфіденційності. Далі впроваджуються рекомендації щодо використання даних і навчання ШІ, що відповідають цим принципам. Регулярно проводяться аудити, щоб виявляти та виправляти упередження, а також забезпечувати відповідність етичним стандартам.

Також необхідно встановити чіткі межі відповідальності та підзвітності при ухваленні рішень ШІ. Постійне вдосконалення на основі зворотного зв’язку від користувачів і зацікавлених сторін має стати невід’ємною частиною процесу. Мета полягає у створенні ШІ-систем, що повністю відповідають етичним принципам та гарантують безпеку й довіру усіх користувачів.

У процесі розробки штучного інтелекту необхідно приділяти першочергову увагу прозорості, пояснюваності та зрозумілості, щоб забезпечити його етичну розробку та впровадження:

- прозорість означає доступність інформації про роботу ШІ-систем для користувачів і зацікавлених сторін;
- пояснюваність тісно пов’язана з прозорістю і стосується здатності ШІ бути зрозумілим та інтерпретованим людиною, бажано простою, не технічною мовою;
- зрозумілість гарантує, що цілі та результати роботи ШІ комунікуються чітко та зрозуміло.

Ці елементи відіграють вирішальну роль у формуванні та підтримці довіри користувачів, забезпечуючи передбачуваність роботи ШІ. Крім того, прозорість і пояснюваність мають критичне значення для підзвітності, оскільки дозволяють притягати до відповідальності розробників та користувачів ШІ за наслідки його використання [4].

Забезпечення справедливості в системах штучного інтелекту передбачає створення алгоритмів, які ухвалюють рішення без упередженості чи необ'єктивності [1]. Це вимагає свідомих зусиль у проєктуванні ШІ, щоб він не відтворював наявні

упередження та не створював нові [3]. Проте досягнення справедливості в ШІ є складним завданням, оскільки системи часто навчаються на реальних даних, які можуть містити вроджені упередження.

Перетин справедливості, конфіденційності та підзвітності є складним, але необхідним аспектом етичного ШІ. Забезпечення справедливості часто вимагає ретельної роботи з чутливими даними, зберігаючи при цьому прозорість і відповідальність у процесах ухвалення рішень. Такий баланс є ключовим у зменшенні упередженості та забезпеченні рівноправності у функціонуванні ШІ-систем.

Конфіденційність і захист даних є критично важливими етичними аспектами,

які необхідно враховувати під час розробки ШІ. Це передбачає захист персональної та чутливої інформації від несанкціонованого доступу та відповідальне використання даних. Важливу роль у формуванні етичних стандартів ШІ відіграють нормативні акти та стандарти, такі як «Загальний регламент із захисту даних» (GDPR) у Європейському Союзі (GDPR 2018; ICO 2018), що встановлюють суворі вимоги до обробки даних.

Конфіденційність, справедливість і довіра користувачів тісно взаємопов'язані. Захист конфіденційності відіграє ключову роль у формуванні довіри, що є необхідною умовою для прийняття та успіху ШІ-систем. Крім того, належне поводження з даними є критично важливим для забезпечення справедливості, оскільки зловживання даними може призвести до упереджених результатів.

1.3 Стратегії впровадження етичних принципів у ШІ

Включення етичних міркувань у розробку ШІ вимагає системного підходу. «Дорожня карта» для цього, передбачає спочатку зобов'язання дотримуватися етичних принципів. Далі слід впровадити рекомендації щодо використання даних та навчання ШІ, щоб узгодити їх з цими принципами. Регулярні аудити необхідні для виявлення та виправлення упереджень, щоб забезпечити відповідність етичним стандартам.

Встановлення чітких ліній відповідальності та підзвітності в рішеннях, що приймаються на основі ШІ, також є критично важливим.

Щоб ефективно інтегрувати етичні принципи у розробку ШІ, необхідний систематичний підхід. Покроковий алгоритм передбачає:

- формулювання та прийняття етичних принципів, таких як прозорість, справедливість і конфіденційність;
- впровадження чітких правил використання даних та навчання ШІ, щоб забезпечити дотримання цих принципів;

- регулярні аудити для виявлення та усунення упередженості, а також контролю відповідності етичним стандартам;
- визначення відповідальності та підзвітності в ухваленні рішень на основі ШІ;
- безперервне вдосконалення ШІ на основі зворотного зв’язку від користувачів і зацікавлених сторін.

Кінцева мета — створення ШІ-систем, які втілюють етичні принципи та забезпечують безпеку та добробут усіх користувачів.

1.4 Висновок до першого розділу

Розвиток штучного інтелекту (ШІ) супроводжується значними етичними викликами, що вимагають комплексного підходу до їх вирішення. У першому розділі роботи було обґрунтовано необхідність запровадження етичних норм у сфері ШІ, визначено ключові аспекти етичності у розробці цих систем та розглянуто стратегії інтеграції етичних принципів у штучний інтелект.

Одним із ключових висновків є те, що етичні міркування у сфері ШІ виходять далеко за межі академічних дискусій і безпосередньо впливають на суспільство. Зокрема, конфіденційність даних, прозорість алгоритмів, справедливість у прийнятті рішень і підзвітність є основними чинниками, які формують довіру користувачів до

ШІ. Ці аспекти мають бути враховані на всіх етапах розробки та впровадження технологій ШІ.

Прозорість та пояснюваність алгоритмів є фундаментальними принципами, що дозволяють користувачам розуміти, як функціонує ШІ і на яких засадах ухвалюються рішення. Це, своєю чергою, підвищує рівень довіри та зменшує ризики необґрунтованих або дискримінаційних рішень. Крім того, справедливість у розробці ШІ передбачає створення алгоритмів, які не містять упередженості, а також здатні забезпечити рівноправне ставлення до всіх користувачів незалежно від їхніх

соціальних, демографічних чи інших характеристик.

Ще одним важливим аспектом є захист персональних даних і забезпечення конфіденційності. Використання великих обсягів даних у навчанні ШІ накладає серйозні обмеження на розробників, оскільки необхідно дотримуватися чинних нормативних актів і стандартів. Наприклад, Загальний регламент захисту даних (GDPR) вимагає отримання усвідомленої згоди на обробку персональних даних, що є необхідною умовою етичного використання ШІ.

Етичні аспекти ШІ тісно взаємопов'язані та утворюють складну систему взаємозалежностей. Наприклад, баланс між прозорістю та конфіденційністю потребує ретельного підходу, щоб з одного боку забезпечити відкритість рішень ШІ, а з іншого — гарантувати захист даних користувачів. Подібний баланс необхідний і між справедливістю та підзвітністю: відповідальність за роботу ШІ має бути чітко визначена, щоб уникнути правової невизначеності.

Для успішного впровадження етичних норм у розробку та використання ШІ необхідний системний підхід. Це включає формулювання чітких правил та стандартів, регулярні аудити та моніторинг функціонування ШІ, а також створення механізмів зворотного зв'язку з користувачами. Важливим фактором є міждисциплінарний підхід, який дозволяє залучати до розробки етичних принципів спеціалістів із різних сфер, включаючи інженерію, право, філософію та соціальні науки.

Отже, інтеграція етичних норм у розвиток штучного інтелекту є критично важливим завданням, яке визначає не лише технічну досконалість цих систем, а й їхню відповідність суспільним цінностям. Етична розробка ШІ сприятиме підвищенню рівня довіри до технологій, зменшенню ризиків упередженості та зловживань, а також забезпечить сталість та безпеку їхнього використання. Впровадження прозорих, справедливих та підзвітних алгоритмів стане важливим кроком до створення ШІ, що працює на благо всього суспільства.

РОЗДІЛ 2
АНАЛІЗ ВІДПОВІДАЛЬНИХ ФРЕЙМВОРКІВ КРАЇН-ЛІДЕРІВ

У цьому розділі розглядаються фреймворки ШІ Європейського Союзу, США та Китаю. Використовуючи детальну блок-схему, планується проаналізувати, як ці фреймворки впливають на всі етапи життєвого циклу ШІ-проєкту – від ініціації до построзгортання.

Крім того, передбачається аналіз за допомогою діаграми Венна, щоб порівняти: - акт про ШІ ЄС (EU AI Act) [2];

- принципи ШІ США (US AI Principles) [8];

- етичні настанови Китаю щодо ШІ (China AI Ethics Guidelines) [5]. Цей аналіз висвітлить унікальні особливості кожного фреймворку та спільні елементи, такі як дотримання етичних стандартів, захист конфіденційності та справедливість. Також буде продемонстровано, як ці фреймворки відображають різні підходи до розробки та регулювання ШІ.

2.1 Аналіз практичних реалізацій

Хоча основний фокус цієї дипломної роботи – теоретичний аналіз, необхідно враховувати зростаючу потребу в емпіричних дослідженнях для підтвердження концепцій етики ШІ. Тому ми розглянемо два ключові кейс-стаді, які наочно демонструють застосування етичних фреймворків у реальних умовах. Ці дані були отримані з глибоких аналізів нещодавніх впроваджень ШІ в двох секторах: охорона здоров'я і фінансовий сектор, які є критичними сферами, де етичні міркування є надзвичайно важливими.

Перший випадок дослідження розглядає впровадження систем діагностики на основі ШІ в європейській системі охороні здоров'я, зокрема зосереджуючи увагу на використанні ШІ IBM Watson Health у діагностиці онкології. Провівши структуровані

інтерв'ю з медичними працівниками та переглянувши звіти про дотримання вимог, спостерігалося, як суворі вимоги до прозорості та підзвітності, передбачені Законом ЄС про ШІ [4], вплинули на дизайн системи ШІ та її операційну прозорість. Це емпіричне свідчення демонструє, як системи ШІ були модифіковані для відповідності регуляторним стандартам, зокрема щодо пояснювальності діагностичних рекомендацій, наданих медичним працівникам і пацієнтам.

Другий приклад передбачає емпіричну оцінку систем виявлення шахрайства штучним інтелектом у секторі фінансових послуг у Сполучених Штатах. Завдяки безпосередній взаємодії з галузевими експертами та аналізу процесів внутрішнього аудиту було досліджено, як система виявлення шахрайства на основі штучного інтелекту JP Morgan узгоджується з гнучкими, орієнтованими на інновації принципами NIST AI Risk Management Framework [8, 9]. Дослідження показує, як ці структури дозволяють адаптувати стратегії управління ризиками, надаючи компаніям автономію для адаптації своїх етичних стандартів, зберігаючи при цьому баланс між інноваціями та підзвітністю.

Ці тематичні дослідження надають емпіричні докази на підтримку теоретичного та політичного аналізу. Вони ілюструють, як глобальні рамки етики штучного інтелекту є не просто абстрактними концепціями, а робочими рекомендаціями, які мають відчутний вплив на проектування, розгортання та відповідність стандартам штучного інтелекту. Ці реальні приклади, можна використати як надійну основу для розуміння динаміки управління ШІ в різних галузях.

2.2 Аналіз блок-схеми розробки та розгортання

Створення та впровадження систем штучного інтелекту є складним процесом, який вимагає відповідності міжнародним нормам для забезпечення етичних і відповідальних результатів. У даному пункті представлено вичерпну блок-схему, яка

об’єднує фреймворки ШІ з Європейського Союзу, Сполучених Штатів і Китаю, відображаючи їх застосування протягом життєвого циклу проекту ШІ. Блок-схема базується на Законі ЄС про штучний інтелект [4], Принципах штучного інтелекту США [9] і етичних принципах штучного інтелекту Китаю [10]. Він починається з ініціювання проекту ШІ, де визначаються цілі та збираються відповідні дані. Рекомендації Закону ЄС про ШІ щодо етичного використання даних і прозорості є важливими на цьому етапі. Керівні принципи штучного інтелекту США вступають у дію, коли проект просувається до обробки даних і розробки моделі штучного інтелекту, наголошуючи на інноваціях, справедливості та підзвітності. Під час етапу валідації та тестування модель має відповідати встановленим цілям і відповідати етичним стандартам у цих рамках. На етапі розгортання інтегруються принципи етики штучного інтелекту Китаю, які віддають пріоритет соціальній гармонії, національній безпеці та глобальній співпраці. Після розгортання безперервний моніторинг і технічне обслуговування є важливими для того, щоб система ШІ функціонувала належним чином і відповідала етичним стандартам. Цей етап має вирішальне значення для включення відгуків і розуміння, що дозволяє вносити ітераційні вдосконалення на основі реальних показників і впливу. На рис. 2.1 показано, наскільки глобальні структури ШІ необхідні та застосовні на різних етапах розробки та розгортання ШІ. Ця інтеграція забезпечує цілісний підхід до відповідальних практик ШІ, які відповідають світовим стандартам і етичним міркуванням.

На рис. 2.2 можна виділити етапи розробки та розгортання систем ШІ. Вони включають найновіші системи штучного інтелекту в усьому світі, в тому числі з ЄС, США і Китаю, і висвітлюють критичні моменти, де ці рамки перетинаються. Потік процесу починається з початку проекту ШІ та переходить до визначення цілей і збору відповідних даних. На цій стадії можна розглянути вказівки Закону про штучний інтелект ЄС.

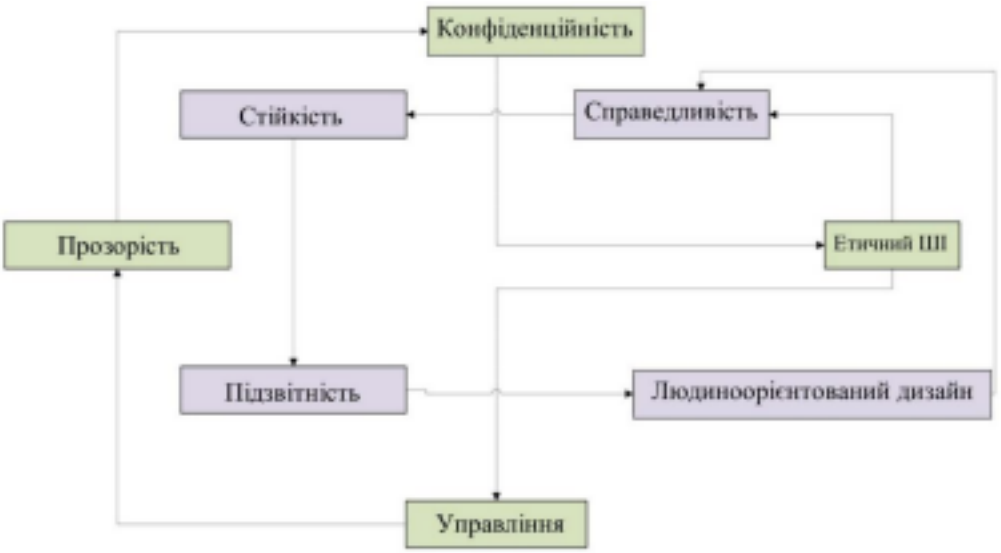


Рисунок 2.1 - Ключові елементи, знайдені в глобальних структурах ШІ

Потім зібрані дані обробляються, і модель штучного інтелекту розробляється відповідно до принципів, викладених у Рекомендаціях щодо штучного інтелекту США.

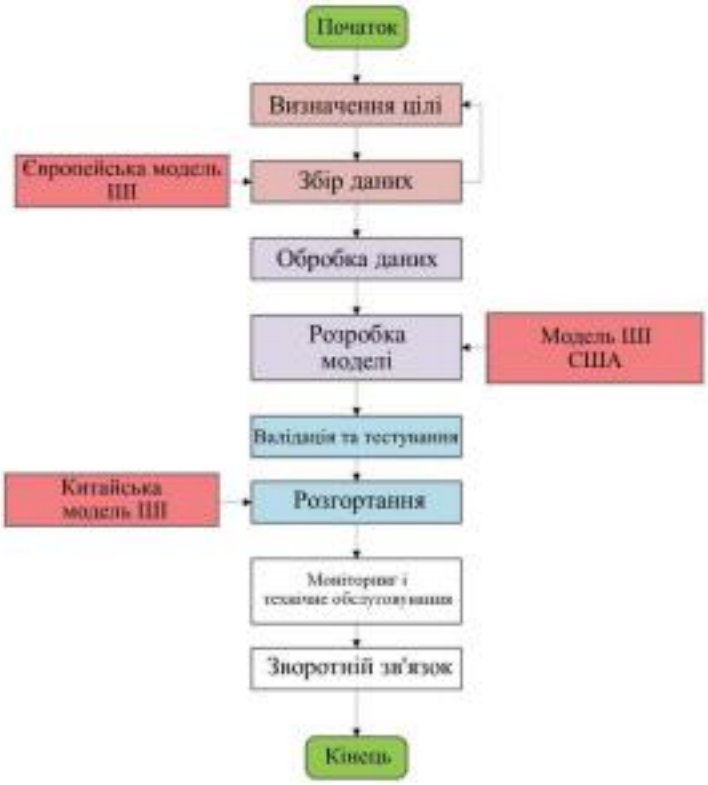


Рисунок 2.2 - Прикладне проектування відповідальних та етичних практик ШІ шляхом інтеграції глобальних структур штучного інтелекту на різних етапах розробки та розгортання ШІ

Потім модель перевіряється на відповідність поставленим цілям. Розгортання

системи штучного інтелекту в реальному середовищі є наступним кроком, і тут беруть участь міркування з етичних принципів китайського штучного інтелекту. Постійний моніторинг і підтримка системи ШІ після розгортання є важливими. Блок-схема також включає цикл зворотного зв'язку, який передбачає перегляд цілей і процесів на основі зворотного зв'язку та нових ідей. Після завершення всіх кроків цикл проекту ШІ завершується. Блок-схема на рис. 2.2 забезпечує відповідальне й етичне застосування штучного інтелекту шляхом інтеграції глобальних структур ШІ на різних етапах розробки та розгортання.

Діаграма Венна на рис. 2.3, представляє рамки ШІ з ЄС, США та Китаю.

Рисунок 2.3 - Різні глобальні інфраструктури ШІ, що збігаються в одних областях, але зберігають унікальні характеристики в інших

Кожне коло на діаграмі представляє структуру штучного інтелекту в іншому

регіоні: Закон ЄС щодо штучного інтелекту, принципи штучного інтелекту США та етичні норми штучного інтелекту в Китаї. Перекриття між колами вказують на сфери спільного фокусування або принципи, спільні для цих структур. Окремі розділи висвітлюють унікальні аспекти кожної структури. Ця діаграма Венна наочно демонструє, як різні глобальні структури штучного інтелекту перетинаються в одних областях, зберігаючи унікальні характеристики в інших. Вона відображає різноманітні підходи до управління ШІ та етики в цих регіонах.

Діаграма всебічно аналізує рамки ШІ з Європейського Союзу, Сполучених Штатів і Китаю. Він підкреслює сфери співпраці та розбіжності між трьома регіонами та демонструє складність структур штучного інтелекту в різних регіонах.

Закон Європейського Союзу про ШІ зосереджується на нагляді з боку людини, недискримінації та дотриманні нормативних вимог. Принципи штучного інтелекту США наголошують на заохоченні інновацій, громадській довірі та відкритому співробітництві. Етика штучного інтелекту Китаю надає перевагу соціальній гармонії, національній безпеці та глобальній співпраці. Ці унікальні елементи відображають індивідуальні пріоритети та культурні перспективи кожного регіону.

ЄС і США поділяють цінності прозорості та етичних стандартів. ЄС і Китай наголошують на захисті конфіденційності та етичних стандартах. США та Китай знаходять спільну мову в етичних стандартах і зосередженні на справедливості. Ці спільні області між двома структурами вказують на потенціал для глобальної співпраці та гармонізації етики та правил ШІ.

Центральне перекриття на діаграмі Венна висвітлює сфери, де ЄС, США та Китай мають спільні принципи, такі як етичні стандарти та захист конфіденційності. Ці сфери пропонують можливості для глобальної співпраці та гармонізації етики та правил ШІ.

Кожен регіон має унікальні напрямки, які відображають його пріоритети та культурні перспективи. Ці різні підходи можуть призвести до різних підходів у

розробці та управлінні ШІ. Діаграма Венна демонструє потенціал для співпраці та розходження в глобальному ландшафті ШІ.

Аналіз діаграми Венна на рис. 2.4 показує складність структур штучного інтелекту в різних регіонах і потенціал для співпраці та розходжень. У ньому підкреслюється необхідність глобальної співпраці та гармонізації етики та правил штучного інтелекту, щоб забезпечити відповідність розвитку та управління ШІ етичним і суспільним цінностям.

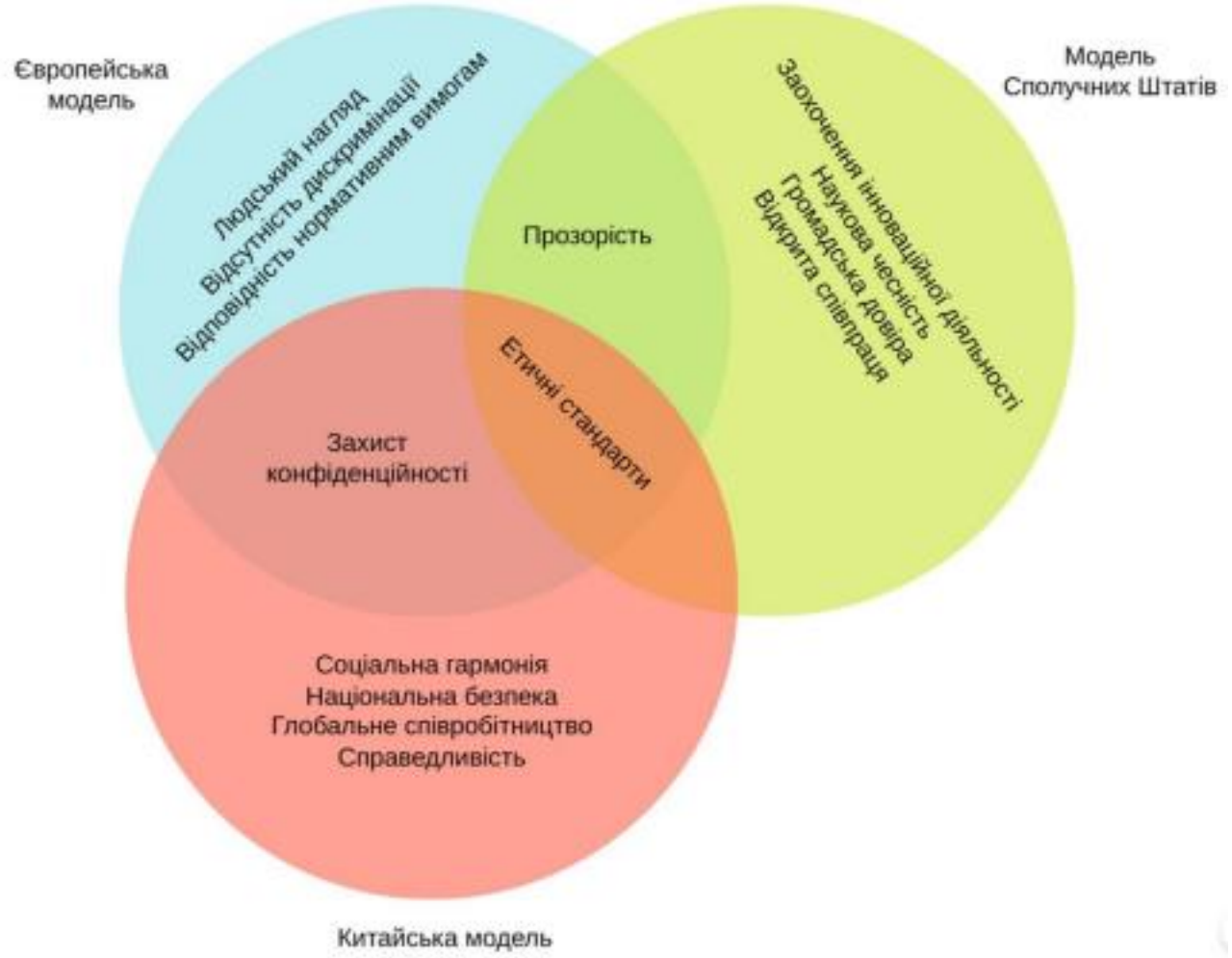


Рисунок 2.4 - Складність структур штучного інтелекту в різних регіонах і потенціал для співпраці та розбіжності

Діаграма Венна на рис. 2.4 показує Закон ЄС про штучний інтелект (світло блакитним кольором), принципи штучного інтелекту США (світло-зеленим), а рамки етики штучного інтелекту Китаю – кораловим.

2.3 Аналіз застосування глобальних фреймворків ШІ у реальному світі

Щоб забезпечити більш чітке розуміння того, як на практиці функціонують фреймворки штучного інтелекту, важливо вивчити реальне застосування цих фреймворків у різних регіонах. Відповідний приклад можна знайти у застосуванні Закону Європейського Союзу про ШІ у сфері охорони здоров'я. Використання діагностичних інструментів на основі штучного інтелекту, таких як Watson Health від IBM, було предметом ретельної перевірки відповідно до суворих правил ЄС щодо прозорості та можливості пояснення. Відповідно до Закону про штучний інтелект, компанії, які розгортають такі системи в секторах високого ризику, зобов'язані надати детальну документацію щодо алгоритмів, які використовуються, а також механізми пояснення, які дозволяють медичним працівникам і пацієнтам розуміти рішення, керовані ШІ. Ця нормативна вимога призвела до модифікації моделей штучного інтелекту для забезпечення відповідності, зокрема шляхом надання більш прозорих процесів прийняття рішень, які можуть бути перевірені регуляторними органами охорони здоров'я.

Навпаки, у розгортанні штучного інтелекту в секторі фінансових послуг можна спостерігати більш орієнтований на інновації підхід Сполучених Штатів, який втілює структура управління ризиками штучного інтелекту NIST. Наприклад, система виявлення шахрайства на основі штучного інтелекту JP Morgan працює в рамках системи, яка наголошує на зменшенні ризиків за допомогою найкращих практик і галузевих стандартів, а не жорсткого регуляторного контролю. Структура NIST заохочує компанії розробляти внутрішню політику, адаптовану до їхніх операційних ризиків, що забезпечує більшу гнучкість у впровадженні ШІ. У результаті JP Morgan розробив власні методи постійного моніторингу та аудиту моделей штучного інтелекту, щоб переконатися, що вони залишаються ефективними, збалансовуючи потребу в інноваціях з етичними міркуваннями.

Управління ШІ в Китаї, яке надає пріоритет державному контролю та суспільній гармонії, можна побачити у використанні урядом систем розпізнавання облич для громадської безпеки. Розгортання таких систем, що регулюється Положеннями Китаю про адміністрування інформаційних послуг Інтернету

глибокого синтезу, Законом Китайської Народної Республіки (КНР) про захист особистої інформації (2022 р.), ілюструє, як держава використовує штучний інтелект у рамках, що надають пріоритет національній безпеці. У цьому випадку технології штучного інтелекту використовуються для моніторингу громадських місць, але їхні етичні наслідки, зокрема щодо конфіденційності, розглядаються в рамках моделі управління, яка суттєво відрізняється від моделі західних демократій. Акцент китайського уряду на штучному інтелекті як інструменті суспільної стабільності підкреслює унікальне застосування їхньої структури на практиці.

Ці приклади висвітлюють різноманітні підходи до структур штучного інтелекту в різних регіонах і секторах, демонструючи, як глобальні політики штучного інтелекту формуються контекстними факторами та застосовуються в практичних сценаріях із високим рівнем впливу. Вивчаючи такі випадки з реального світу, ми можемо краще зрозуміти сильні сторони й обмеження цих структур, а також проблеми, пов’язані з гармонізацією етики ШІ в глобальному масштабі.

2.4 Аналіз рекомендацій щодо застосувань глобальних фреймворків ШІ

2.4.1 Технічні рішення для впровадження етичних принципів у системи ШІ

Впровадження таких етичних принципів, як справедливість, прозорість і підзвітність, у системи штучного інтелекту потребує складних алгоритмічних підходів, які забезпечують досягнення цих цілей без шкоди для продуктивності системи. Алгоритмічну справедливість можна вирішити за допомогою таких методів,

як диференціальна справедливість і навчання справедливого представлення. Наприклад, такі алгоритми, як модель навчання справедливому представленню (Fair Representation Learning, FRL), спрямовані на пом’якшення упередженості шляхом перетворення необроблених даних у приховане представлення, незмінне щодо

чутливих атрибутів, таких як раса чи стать, без втрати важливої передбачуваної значущості. Метод FRL застосовує змагальне навчання, щоб гарантувати, що модель не може легко вивести чутливі атрибути, таким чином зменшуючи упередження, зберігаючи точність. Це може бути особливо корисним у таких секторах, як фінанси, де історичні упередження в наборах даних кредитного рейтингу часто призводять до несправедливих результатів. Включення цих обмежень справедливості на етапі навчання моделі гарантує, що система штучного інтелекту не поширюватиме дискримінаційні моделі в даних.

Прозорість покращується завдяки використанню пояснюваних методів штучного інтелекту (explainable AI, XAI). Одним із поширених підходів є впровадження локальних інтерпретованих модельно-незалежних пояснень (Local Interpretable Model-agnostic Explanations, LIME), які надають користувачам інтерпретовані наближення складних моделей, що дозволяє кінцевим користувачам і аудиторам зрозуміти й оцінити окремі прогнози. LIME працює, спотворюючи вхідні дані та спостерігаючи за тим, як зміни впливають на прогнози, таким чином створюючи простіші моделі, які можна інтерпретувати локально навколо конкретних випадків. Цей метод особливо цінний у галузях високого ризику, таких як охорона здоров’я, де розуміння обґрунтування діагнозів, керованих штучним інтелектом, має вирішальне значення для зміцнення довіри та підзвітності. Наприклад, LIME було ефективно застосовано в медичній візуалізації, щоб пояснити, як системи штучного інтелекту ідентифікують ділянки пухлини, забезпечуючи прозорість як для клініцистів, так і для пацієнтів.

2.4.2 Технічні рішення для впровадження етичних принципів у законодавстві

Щоб створити більш надійне законодавство щодо штучного інтелекту, яке б відповідало на виклики реального світу, необхідний збір імовірнісних даних. Важливе рішення полягає в моделюванні на основі даних, яке використовує реальні ймовірнісні розподіли результатів ШІ в різних областях. Ці симуляції можуть використовувати байєсівські моделі висновків для аналізу ймовірності етичних

помилок, таких як упереджені рішення або порушення прозорості, за різних нормативних сценаріїв. Наприклад, байєсівські моделі можуть оцінити ймовірність упереджених результатів у системах схвалення позик на основі різноманітних нормативних обмежень, дозволяючи посадовим особам кількісно оцінювати компроміси між суворим регулюванням та інноваціями. Включаючи такі ймовірнісні оцінки, посадові особи можуть розробити законодавство, засноване на емпіричних даних, гарантуючи, що етичні рекомендації є практичними та застосовними в різних секторах.

Більше того, збір кількісних даних із розгортань штучного інтелекту в реальному світі може використовувати такі методи, як диференційована конфіденційність, щоб захистити конфіденційну інформацію, зберігаючи при цьому значущу інформацію. Наприклад, у сфері охорони здоров'я збір великомасштабних даних пацієнтів за допомогою діагностичних інструментів штучного інтелекту, зберігаючи конфіденційність пацієнтів, можна досягти за допомогою диференціальних алгоритмів конфіденційності, які вносять шум у набори даних, гарантуючи, що окремі записи неможливо буде повторно ідентифікувати. Це дозволяє регулюючим органам збирати точні статистичні дані про продуктивність системи штучного інтелекту, наприклад точність прогнозів і частоту помилок, не порушуючи законів про конфіденційність. Потім ці точки даних реального світу можна використовувати для точного налаштування законодавчих рамок, щоб вони відображали практичне використання ШІ та відповідали стандартам конфіденційності.

2.4.3 Алгоритмічні рішення для забезпечення чесного та етичного ШІ для кінцевих користувачів

Щоб переконатися, що системи штучного інтелекту сприймаються кінцевими користувачами як чесні та етичні, у життєвий цикл розробки можна інтегрувати кілька алгоритмічних підходів. Одним із багатообіцяючих методів є використання обмежень справедливості в оптимізації моделі, таких як зрівняні шанси та демографічний паритет. Алгоритм Equalized Odds [11] гарантує, що система

штучного інтелекту має однакові істинно позитивні та хибно позитивні показники для різних демографічних груп, гарантуючи, що жодна група не отримає непропорційної вигоди чи не постраждає від рішень системи. Ця техніка була успішно реалізована в судових системах, де моделі штучного інтелекту використовуються для рекомендацій щодо звільнення під заставу та винесення вироку, зменшуючи расову невідповідність, яка зазвичай спостерігається в попередніх моделях.

Алгоритми навчання з урахуванням справедливості також можна вбудувати в конвеєри машинного навчання для моніторингу та коригування упередженості під час процесу навчання. Наприклад, структура справедливості через обізнаність (Fairness through Awareness, FTA) коригує межі рішень у моделях, щоб забезпечити однакове ставлення до подібних осіб, тим самим зменшуючи несправедливе упередження. Цей алгоритм обчислює відстані в просторі, чутливому до справедливості, і гарантує, що люди, які знаходяться поблизу в цьому просторі, отримують подібні прогнози. Це було застосовано в алгоритмах найму, щоб гарантувати справедливе ставлення до кандидатів із подібною кваліфікацією, незалежно від демографічних ознак.

Крім того, залучення кінцевих користувачів до систем ШІ можна покращити за допомогою інтерактивних механізмів прозорості. Наприклад, контрфактичні пояснення (counterfactual explanations, CX) можна використовувати, щоб надати користувачам практичне уявлення про те, як можуть змінитися рішення, якщо змінити певні вхідні дані. Наприклад, у системах кредитного скорингу, CX може інформувати користувача про те, що його позику було відхилено через низький кредитний рейтинг,

і запропонувати конкретні кроки, такі як зменшення боргу за кредитною карткою, які призведуть до схвалення. Надаючи користувачам чітку, практичну інформацію, ці системи не тільки збільшують довіру, але й дають користувачам змогу більш значуще залучатися до рішень, керованих ШІ.

2.4.4 Алгоритмічна підзвітність і постійний моніторинг

Щоб підтримувати постійну справедливість і етичні стандарти, необхідний постійний моніторинг систем ШІ. Цього можна досягти за допомогою алгоритмічних структур аудиту, які регулярно оцінюють системи ШІ на дотримання етичних принципів після розгортання. Інструменти пост-спеціального аудиту справедливості, такі як AI Fairness 360 (AIF360), надають набір інструментів з відкритим кодом, який вимірює та пом'якшує упередженість у розгорнутих моделях. Ці інструменти можна інтегрувати в процеси управління штучним інтелектом, гарантуючи, що моделі залишатимуться справедливими та неупередженими, коли вони стикаються з новими даними в реальному середовищі. AIF360 оцінює справедливість за допомогою багатьох показників, таких як різнорідний вплив і статистичний паритет, і забезпечує постійне повторне калібрування моделей для підтримки етичної ефективності.

Включення алгоритмічних систем звітності з циклами зворотного зв'язку в реальному часі, забезпечує швидке виявлення та пом'якшення упереджень, викликаних змінами в розподілі даних (дрейф даних). Такі методи, як алгоритми виявлення дрейфу, включно з ADWIN (Adaptive Windowing), безперервно відстежують продуктивність моделей ШІ та запускають повторне навчання, коли виявляються значні відхилення від очікуваної поведінки. Автоматизуючи виявлення етичних порушень і перекалібруючи моделі у відповідь, ці системи гарантують, що ШІ з часом залишається ефективним і етичним.

2.5 Висновок до другого розділу

У цьому розділі використовується аналіз діаграми Венна для порівняння систем штучного інтелекту в Європейському Союзі (ЄС), Сполучених Штатах (США) і Китаї. Діаграма представляє кожну структуру, підкреслюючи їхні унікальні особливості та області перекриття, розкриваючи як потенціал співпраці, так і різні підходи.

Закон ЄС про штучний інтелект надає пріоритет нагляду з боку людини, недискримінації та суворому дотриманню нормативних вимог. Це відображає наголос ЄС на захисті прав громадян в епоху цифрових технологій. Принципи штучного інтелекту США віддають перевагу сприянню інноваціям, забезпеченню суспільної довіри та сприянню відкритому співробітництву. Це відображає наголос США на розробці ШІ, орієнтованій на ринок та інновації. З іншого боку, китайська етика AI наголошує на важливості соціальної гармонії, національної безпеки та глобальної співпраці. Це відображає підхід Китаю до збалансування технологічного прогресу з соціальною стабільністю та державною безпекою.

Перетини цих рамок на діаграмі Венна підкреслюють спільні принципи та потенційні сфери міжнародного співробітництва. Наприклад, ЄС і США наголошують на етичних стандартах і прозорості. ЄС і Китай поділяють спільну увагу до захисту конфіденційності та етичного використання ШІ. США та Китай сходяться на заохоченні етичних стандартів і справедливості в ШІ.

У центральному перетині діаграми Венна, де перетинаються всі три основи, лежить спільне зобов’язання дотримуватися етичних стандартів, захисту конфіденційності та забезпечення справедливості. Ця спільна основа пропонує можливості для глобальної співпраці та гармонізації етики та правил ШІ.

Проте відмінні аспекти кожної структури відображають різні пріоритети та культурні перспективи кожного регіону. Ці відмінності можуть призвести до різних підходів до розвитку ШІ та управління в усьому світі. Таким чином, діаграма Венна

підкреслює потенціал для співпраці та підкреслює необхідність розуміння та поваги до різноманітних точок зору в глобальному ландшафті ШІ.

РОЗДІЛ 3
РОЗРОБКА СТРАТЕГІЇ ПОМ’ЯКШЕННЯ УПЕРЕДЖЕНОСТІ

ШІ 3.1 Розширений аналіз політик ШІ країн світу

Цей розділ розширюється до ретельного аналізу рамок політики етики штучного інтелекту в усьому світі. Він охоплює значні регіони, такі як Європейський Союз, Сполучені Штати, Китай, Канада, Японія, Індія та Австралія. На рис. 3.1 представлено дані у вигляді декартових графіків, які порівнюють ці рамки за чотирма ключовими етичними параметрами: прозорість, підзвітність, справедливість і конфіденційність.

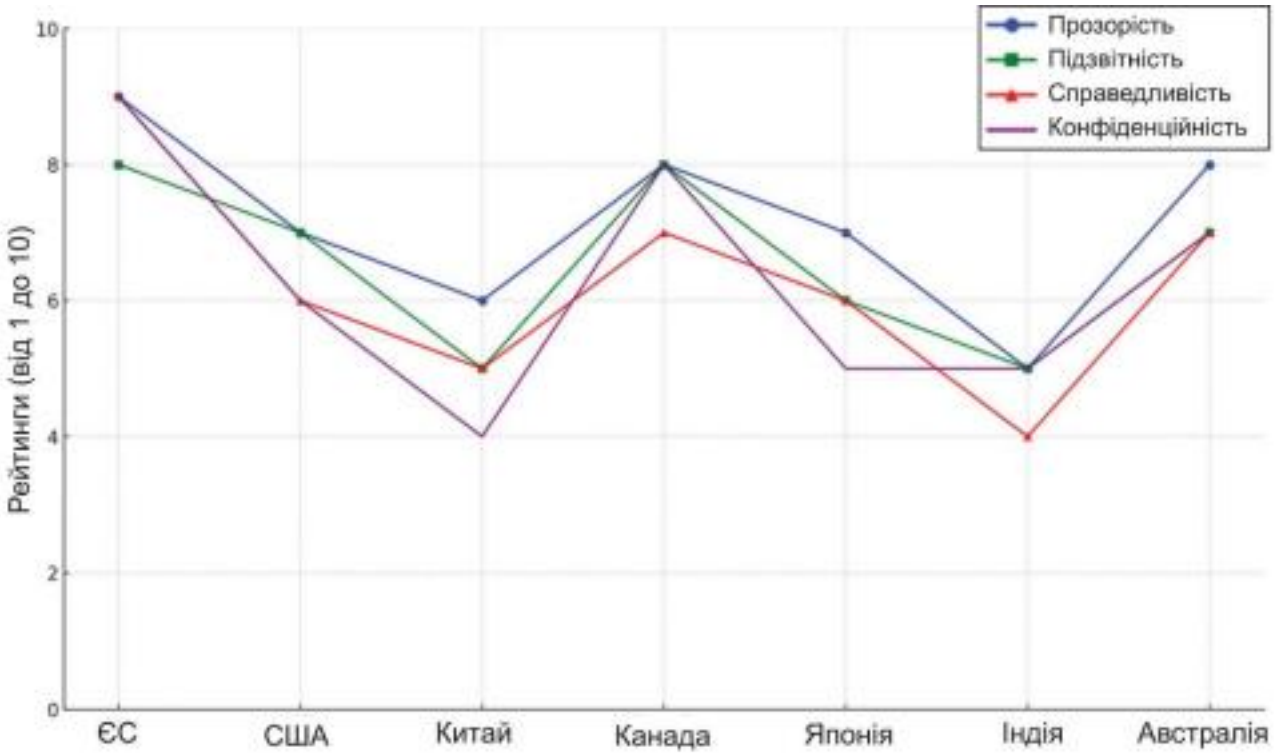


Рисунок 3.1 - Декартовий графік порівняння основ етичної політики в різних країнах

Кожен фреймворк детально оцінювався, що забезпечує глибоке розуміння того, як країни віддають пріоритет цим параметрам у своїй політиці ШІ. На цьому графіку висвітлюються унікальні пріоритети та напрямки діяльності кожної структури, а також

наголошується на різноманітності та спільному в глобальних підходах до етики ШІ. Ці знання можуть використати для визначення областей для вдосконалення та розробки всеосяжних, етичних політик ШІ. Крім того, цей розділ розширює цей аналіз, включаючи додаткові критичні аспекти, такі як людський нагляд і національна безпека, розширюючи сферу, щоб охопити ширші соціально-політичні

наслідки технології ШІ.

3.2 Визначення критеріїв для оцінки основ політики етики штучного інтелекту в різних країнах

На рис. 3.2 аналізуються глобальні тенденції та представлено дорожню карту для гармонізації етики штучного інтелекту, виступаючи за постійний міжнародний діалог і співпрацю для сприяння відповідальній та етичній глобальній екосистемі штучного інтелекту.

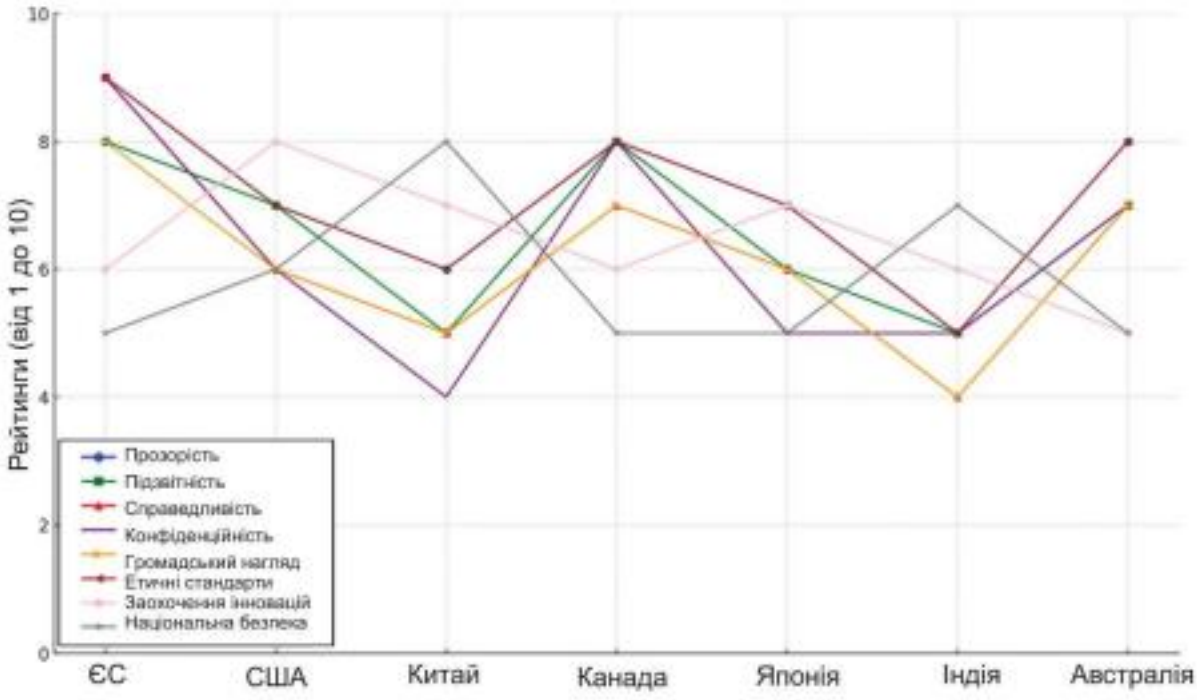


Рисунок 3.2 - Декартовий графік глобальної політики щодо етики штучного інтелекту: інструмент для розуміння пріоритетів різних країн щодо етики ШІ

Рейтинги основ політики етики штучного інтелекту в кожній країні ґрунтувалися на детальному аналізі публічних документів, офіційних документів щодо політики, нормативних вказівок та наукової літератури. Оцінки за кожним виміром, прозорість, підзвітність, справедливість, конфіденційність, людський нагляд, етичні стандарти, заохочення інновацій і національна безпека, були оцінені відповідно до таких конкретних критеріїв:

3.2.1 Прозорість (шкала 1–10):

Прозорість оцінювалася за ступенем, до якого система кожної країни передбачає відкритість щодо розробки, впровадження та процесів прийняття рішень систем ШІ:

- 9-10: Країни з чіткими, законодавчо закріпленими вимогами щодо прозорості систем штучного інтелекту, включаючи мандати щодо пояснюваності та публічної підзвітності (наприклад, Європейський Союз);

- 6-8: Країни, які заохочують прозорість, але не передбачають її як юридичну вимогу для всіх програм ШІ (наприклад, Сполучені Штати);

- 4-5: Країни, де прозорість згадується в політиці, але з невеликою кількістю практичних механізмів забезпечення (наприклад, Китай);

- 1-3: Мінімальна або відсутність офіційної уваги до прозорості в структурі ШІ. **3.2.2 Відповідальність (Шкала 1–10)**

Підзвітність вимірює надійність правових і регуляторних механізмів, які покладають на розробників, компанії та уряди відповідальність за результати систем ШІ:

- 9-10: Країни з чітко визначеними рамками відповідальності, які покладають чітку відповідальність на розробників або операторів ШІ (наприклад, Закон ЄС про ШІ);

- 6-8: Країни, де підзвітність заохочується через добровільні рамки відповідності, але не має обов'язкового дотримання (наприклад, США з NIST AI Risk Management Framework);

- 4-5: Країни з нечіткими заходами підзвітності, які часто здійснюються на розсуд приватних організацій або не мають централізованого регулювання (наприклад, Японія);

- 1-3: Країни, де системи підзвітності не існують або все ще перебувають на ранніх стадіях розробки.

3.2.3 Справедливість (Шкала 1–10)

Справедливість оцінювалася на основі того, наскільки добре політична структура країни бореться з упередженнями в алгоритмах ШІ та забезпечує справедливі результати для демографічних груп.

- 9-10: Країни з чіткими вимогами справедливості для систем штучного інтелекту, обов'язковими перевірками справедливості та методами пом'якшення упередженості (наприклад, Канада, ЄС);

- 6-8: Країни, які заохочують справедливість, але з менш суворими або необов'язковими практиками аудиту (наприклад, США);

- 4-5: Країни, де справедливість є бажаною метою з обмеженою практичною реалізацією (наприклад, Індія);

- 1-3: Мінімальна або відсутність уваги до справедливості в регулюванні ШІ.

3.2.3 Конфіденційність (Шкала 1–10)

Конфіденційність оцінювалася за тим, як система кожної країни захищає дані користувачів у контексті ШІ та наскільки вона відповідає світовим стандартам, таким як GDPR:

- 9-10: Країни з жорсткими правилами конфіденційності, включаючи чіткі правила використання даних AI (наприклад, ЄС, Канада);

- 6-8: Країни із загальними законами про конфіденційність даних, але обмеженими вказівками щодо ШІ (наприклад, США, Японія);

- 4-5: Країни з правилами конфіденційності, які застосовуються непослідовно або недостатньо розвинені щодо ШІ (наприклад, Індія, Австралія);

- 1-3: Країни з мінімальною увагою до конфіденційності в контексті ШІ або де закони про захист даних не виконуються ефективно (наприклад, Китай). **3.2.4 Громадський контроль (Шкала 1–10)**

Цей критерій оцінював роль людського нагляду в процесі прийняття рішень ШІ, особливо в секторах високого ризику, таких як охорона здоров'я чи автономні транспортні засоби:

- 9-10: Країни, які передбачають обов'язковий нагляд з боку людини за

рішеннями щодо штучного інтелекту з високим рівнем ризику, забезпечуючи втручання людини в критичні сфери (наприклад, Закон ЄС про штучний інтелект);

- 6-8: Країни, які рекомендують, але законодавчо не запроваджують людський контроль (наприклад, США);

- 4-5: Країни, де згадується людський нагляд, але механізми примусового виконання нечіткі або відсутні (наприклад, Японія, Індія);

- 1-3: Майже або зовсім не наголошується на людському нагляді в рамках політики ШІ.

3.2.5 Етичні стандарти (шкала від 1 до 10)

Етичні стандарти оцінювалися на основі того, наскільки політика країни щодо штучного інтелекту відповідає глобальним етичним рамкам (таким як рекомендації ЮНЕСКО щодо етики штучного інтелекту) і сприяє етичному розвитку штучного інтелекту:

- 9-10: Країни з чітко визначеною міжнародно узгодженою етичною основою для ШІ (наприклад, ЄС, Канада);

- 6-8: Країни з етичними рекомендаціями щодо штучного інтелекту, але обмежені за сферою застосування або правозастосуванням (наприклад, Японія, США);

- 4-5: Країни, які згадують етичний штучний інтелект, але не мають узгодженої, доступної для виконання системи (наприклад, Індія);

- 1-3: Мінімальна або відсутність офіційної уваги до етики ШІ в державній політиці.

3.2.6 Заохочення інновацій (шкала 1–10)

Цей критерій оцінював баланс між просуванням інновацій ШІ та дотриманням етичних принципів. Країни, які заохочують інновації, зберігаючи міцні етичні рамки, отримали вищі бали:

- 9-10: Країни з політикою, орієнтованою на інновації, які підтримують дослідження та розробки штучного інтелекту, дотримуючись етичних принципів

(наприклад, США, Канада);

- 6-8: Країни з сильною інноваційною політикою, але менш суворим етичним правозастосуванням (наприклад, Японія);

- 4-5: Країни, де просуваються інновації, але ціною етичних стандартів (наприклад, Китай);

- 1-3: Країни, де інновації в штучному інтелекті пригнічуються через надмірне регулювання або брак ресурсів (наприклад, мінімальний фокус). **3.2.7 Національна безпека (1–10 шкала)**

Національну безпеку оцінювали на основі того, як країни включають штучний інтелект у свої стратегії національної безпеки, включаючи оборону, кібербезпеку та спостереження:

- 9-10: Країни, де ШІ відіграє значну роль у системах національної безпеки, мають чітку політику щодо військового ШІ, спостереження та кіберзахисту (наприклад, Китай, США);

- 6-8: Країни, які включають ШІ в політику національної безпеки, але з меншою кількістю чітких правил щодо його використання в обороні (наприклад, Австралія, Японія);

- 4-5: країни, у яких згадується штучний інтелект у контексті національної безпеки, але не мають конкретної політики (наприклад, Індія);

- 1-3: Мінімальна увага до штучного інтелекту для національної безпеки або політик, які все ще знаходяться на ранній стадії розробки (наприклад, ЄС).

3.3 Обґрунтування вибору критеріїв і балів

Вибрані критерії для оцінки основ політики етики штучного інтелекту, прозорості, підзвітності, справедливості, конфіденційності, контролю з боку людини, етичних стандартів, заохочення інновацій і національної безпеки були ретельно відібрані, щоб відобразити основні параметри, важливі для етично надійних,

соціально корисних і технологічно відповідальних систем ШІ. Ці аспекти добре закріплені в політичному дискурсі та науковій літературі як стовпи етики та управління штучним інтелектом, гарантуючи, що розробка штучного інтелекту узгоджується з суспільними цінностями та пом'якшує потенційну шкоду.

3.3.1 Прозорість

Прозорість є «перечепою» етики штучного інтелекту, що підкреслюється як в академічних, так і в регуляторних дискусіях [12]. Прозорі системи ШІ дозволяють зацікавленим сторонам зрозуміти, як приймаються рішення, і забезпечують підзвітність. Вибір прозорості як критерію підтверджується нормативними документами, такими як GDPR Європейського Союзу та Закон про штучний інтелект, які висувають чіткі вимоги до систем ШІ, щоб вони були зрозумілими та відкритими для громадського контролю. Дослідження показали, що брак прозорості є однією з головних причин недовіри громадськості до систем ШІ [13]. Таким чином, країни з чіткими юридичними повноваженнями щодо прозорості отримали вищі бали, тоді як країни з добровільними або нечіткими інструкціями щодо прозорості отримали нижчі. **3.3.2 Відповідальність**

Підзвітність гарантує, що розробники, оператори та користувачі ШІ несуть відповідальність за результати, створені системами ШІ. Цей критерій виправдовується визнанням у літературі того, що без чітких структур підзвітності стає важко

вирішувати проблеми збоїв або шкоди, спричинені системами ШІ [14]. Закон ЄС про штучний інтелект містить комплексні положення, які передбачають юридичну відповідальність, забезпечуючи надійну модель підзвітності. Країни, такі як Сполучені Штати, з добровільним дотриманням через такі структури, як NIST AI Risk Management Framework, отримали проміжні бали через відсутність можливості виконання. Необхідність підзвітності також є основною темою в академічній літературі, особливо в контексті складних систем штучного інтелекту, де в розробці та розгортанні беруть участь численні зацікавлені сторони [12, 15].

3.3.3 Справедливість

Справедливість у системах штучного інтелекту усуває проблеми щодо упередженості та дискримінації, які є добре задокументованими проблемами в програмах штучного інтелекту [16]. Країни з явними вимогами до справедливості у своїй політиці штучного інтелекту, такі як Європейський Союз і Канада, отримали вищі бали, оскільки їхні рамки зобов'язують перевірку справедливості та практики пом'якшення упередженості. Література про справедливість у штучному інтелекті часто вказує на обмеження алгоритмічних систем для забезпечення справедливих результатів для демографічних груп без спеціального регуляторного втручання. Країни з мінімальними або невиконавчими положеннями про справедливість, такі як Індія та Китай, отримали нижчі бали через відсутність надійних механізмів пом'якшення упередженості.

3.3.4 Конфіденційність

Конфіденційність є критичною проблемою в ШІ, особливо в системах, які покладаються на величезну кількість особистих даних. GDPR в ЄС встановлює високий глобальний стандарт захисту даних і конфіденційності, виправдовуючи високу оцінку для ЄС у цьому вимірі. Навпаки, такі країни, як Сполучені Штати, де правила конфіденційності, такі як HIPAA, є специфічними для домену та не є універсальними для систем ШІ, отримали нижчі бали. Конфіденційність як критерій ґрунтується на принципі, згідно з яким етичний штучний інтелект повинен захищати

права людей контролювати свої дані, що є поширеним у науковій та політичній літературі [14].

3.3.5 Людський нагляд

Людський нагляд за прийняттям рішень у сфері штучного інтелекту має важливе значення для запобігання надмірній залежності від автоматизованих систем, особливо в таких критичних секторах, як охорона здоров'я та правоохоронні органи (Європейський парламент 2023). Закон ЄС про штучний інтелект знову лідує, передбачаючи обов'язковий людський нагляд за додатками штучного інтелекту з

високим ризиком. Слід зосередитися на важливості збереження людського судження під час прийняття рішень за допомогою штучного інтелекту, особливо у випадках, пов'язаних із моральними чи правовими наслідками. Країни, які рекомендують, але не передбачають обов'язковий нагляд з боку людини, отримали нижчі бали, оскільки добровільний нагляд часто не дає результатів у реальних програмах, особливо там, де оперативна ефективність має пріоритет над втручанням людини.

3.3.6 Етичні стандарти

Етичні стандарти все частіше вважаються життєво важливими для узгодження розвитку ШІ з людськими цінностями. Рекомендація ЮНЕСКО щодо етики штучного інтелекту та подібні ініціативи Організації економічного співробітництва та розвитку (ОЕСР) надали плани етичного штучного інтелекту, зосереджуючись на таких принципах, як благодійність, нешкідливість, автономія та справедливість. Такі країни, як Канада та ЄС, які прийняли комплексні етичні принципи, отримали високі бали. Ці стандарти мають вирішальне значення для того, щоб ШІ працював у межах моральних і правових норм. Навпаки, країни, в яких відсутні конкретні етичні рамки для штучного інтелекту, такі як Індія та Китай, отримали нижчі бали, що відображає недостатній розвиток етичних міркувань у їхній політиці щодо штучного інтелекту.

3.3.7 Заохочення інноваційної діяльності

Баланс між заохоченням інновацій штучного інтелекту та дотриманням етичних стандартів є ключовим питанням для відповідних посадових осіб. Такі

країни, як Сполучені Штати, отримали високі оцінки в цьому вимірі завдяки своїй орієнтованій на інновації політиці, як-от NIST AI Risk Management Framework, яка сприяє галузевим рішенням і сприяє створенню сприятливого середовища для досліджень і розробок штучного інтелекту. У літературі підтримується думка про те, що інновації процвітають, коли є гнучкість і мінімальні регулятивні витрати, але із застереженням, що не можна нехтувати етичними перешкодами. Країни, які надмірно зосереджені на регулюванні, що потенційно придушують інновації, або в яких бракує достатніх стимулів для досліджень ШІ, отримали нижчі бали.

3.3.8 Національна безпека

Міркування національної безпеки, зокрема щодо розробки автономних систем зброї (Autonomous Weapons Systems, AWS) і кібербезпеки, посиленої штучним інтелектом, стають критично важливим компонентом політики штучного інтелекту. Такі країни, як США та Китай, де штучний інтелект відіграє значну роль у національних оборонних стратегіях, отримали високі бали. Література про автономні системи підкреслює важливість регулювання штучного інтелекту для запобігання небажаним наслідкам у військових застосуваннях. Країни, які ще не інтегрували штучний інтелект у свої стратегії національної безпеки або мають недостатньо розвинене управління штучним інтелектом у цій сфері, наприклад ЄС, отримали нижчі бали.

Оцінки для кожного параметра були отримані з комбінації таких джерел: - Політичні документи та нормативні акти: ключові нормативні рамки, такі як Закон ЄС про штучний інтелект, Адміністрація Байдена-Гарріса оголошує про нові дії для сприяння відповідальним інноваціям ШІ, які захищають права та безпеку американців (Білий дім, 2023 р.), GDPR (GDPR 2018; ICO 2018), NIST AI Risk Management Framework (NIST 2024a, 2024c; Tabassi 2023) і національні стратегії штучного інтелекту були безпосередньо проаналізовані, щоб оцінити силу та комплексність механізмів управління штучним інтелектом у кожній країні;

- Академічна література: використовувалися основоположні тексти про етику штучного інтелекту, справедливість, прозорість і підзвітність, щоб визначити базові очікування щодо того, що є передовою практикою в кожному вимірі [12, 16, 17];

- Реальні практичні приклади: приклади впровадження штучного інтелекту в різних секторах, включаючи охорону здоров’я, фінанси та національну безпеку, були досліджені для контекстуалізації практичного впливу кожної політики.

Критерії були обрані для забезпечення комплексної оцінки підходу кожної країни до етики ШІ, зосереджуючись як на нормативній жорсткості, так і на практичному застосуванні етичних принципів. Кожна оцінка відображає ступінь

вирішення ключових проблем, пов'язаних з управлінням штучним інтелектом, у системі країни, забезпечуючи збалансовану оцінку, що ґрунтується як на аналізі політики, так і на академічних висновках.

3.4 Порівняльний аналіз з використанням декартових графіків

Як результат роботи у попередніх кроках дипломної роботи, надається детальний порівняльний аналіз принципів політики етики ШІ з глобальної точки зору. Аналіз охоплює такі основні регіони, як Європейський Союз, Сполучені Штати, Китай, Канада, Японія, Індія та Австралія. Дослідження розглядає ці рамки на основі ключових етичних аспектів, включаючи прозорість, підзвітність, справедливість і конфіденційність, використовуючи серію декартових графіків.

Кожен фреймворк оцінюється за шкалою від 1 до 10 за цими параметрами (які були визначені у 3.2.1 – 3.2.7). Формат декартового графіка допомагає візуалізувати, як система кожної країни відповідає цим критичним сферам. Наприклад, система ЄС надає пріоритет конфіденційності та підзвітності, що відображено в його високих оцінках у цих сферах. З іншого боку, система США може отримати вищу оцінку щодо прозорості через її фокус на відкритих даних та інноваціях. Рамки Китаю, які наголошують на соціальній гармонії та національній безпеці, можуть мати різні сильні

та слабкі сторони. Цей порівняльний аналіз дає змогу зрозуміти пріоритети та сфери діяльності різних країн, а також підкреслює різноманітність і спільність у підходах до етики штучного інтелекту в усьому світі. Графічне представлення допомагає зрозуміти складність кожної структури, пропонуючи чітке уявлення про те, як країни орієнтуються в етичному ландшафті розвитку ШІ. Декартовий графік на рис. 3.1 порівнює рамки політики етики ШІ в різних країнах, таких як ЄС, США, Китай, Канада, Японія, Індія та Австралія. Він оцінює їх за прозорістю, підзвітністю, чесністю та конфіденційністю. На рис. 3.1 представлено огляд того, як країни

визначають пріоритетність різних елементів етики штучного інтелекту в рамках своєї політики. Для рейтингу використовуються чотири аспекти: прозорість, підзвітність, справедливість і конфіденційність. Кожен аспект оцінюється за шкалою від 1 до 10. Блакитна лінія означає прозорість, яка відображає те, як політика відкрито повідомляється та реалізується. Зелена лінія вказує на підзвітність, вимірюючи ступінь відповідальності розробників і користувачів штучного інтелекту за свої системи згідно з рамками. Червона лінія означає справедливість, вимірюючи ступінь, до якого політики забезпечують справедливі та неупереджені системи ШІ. Нарешті, фіолетова лінія показує важливість конфіденційності користувачів і захисту даних у політиках. Графік пропонує цінну інформацію про глобальний ландшафт етики ШІ. Він підкреслює схожість і відмінності в національних підходах до регулювання штучного інтелекту, які можуть допомогти в розробці майбутньої політики. Оцінки кожного аспекту для кожної країни можуть допомогти визначити сфери, які необхідно покращити у їхній політиці.

Нарис. 3.2 представлена розширена матриця оцінювання, яка включає додаткові виміри, які все більше стосуються етики ШІ. Ці аспекти включають роль людського нагляду та національної безпеки, що відображає ширші соціально політичні наслідки технології ШІ. Включно з людським наглядом підкреслюється необхідність людського втручання та судження в системах штучного інтелекту, принцип, який сильно підтримується рамками ЄС. І навпаки, національна безпека має

вирішальне значення в таких країнах, як Китай і США, де участь ШІ в обороні та розвідці є ключовим питанням. Цей комплексний аналіз підкреслює глобальні тенденції в рамках політики щодо штучного інтелекту та потенціал для гармонізації етики штучного інтелекту. За допомогою сформованого графіку досліджується можливість зближення принципів і стандартів, незважаючи на різноманітні культурні, політичні та соціальні контексти кожного регіону. Хоча повна одноманітність може бути недосяжною, потенціал для міжнародної співпраці та досягнення консенсусу щодо основних принципів є значним. Така гармонізація може сприяти встановленню загальноприйнятих норм і стандартів, забезпечуючи

відповідність розвитку штучного інтелекту етичним і суспільним цінностям. Аналіз дає змогу зробити висновок, що потрібен постійний діалог і співпраця між країнами для розвитку відповідальної та етичної глобальної екосистеми ШІ. Декартовий графік на рис. 3.2 містить лінії, що представляють концепції, натхненні попередніми діаграмами Венна, які досліджували проблеми, пов'язані з управлінням ШІ. Лінії на графіку представляють різні аспекти етики ШІ, які були визначені як необхідні. Блакитна лінія символізує прозорість, яка означає відкритість і ясність у комунікації політики ШІ. Зелена лінія позначає підзвітність, що означає ступінь відповідальності за розробку та використання ШІ. Червона лінія означає справедливість, яка вказує на важливість забезпечення неупереджених систем ШІ. Фіолетова лінія символізує конфіденційність, що підкреслює важливість конфіденційності та захисту даних користувачів. Помаранчева лінія символізує людський нагляд, що означає участь і нагляд людей у процесах ШІ. Коричнева лінія позначає етичні стандарти, що означає дотримання етичних принципів штучного інтелекту. Рожева лінія означає заохочення інновацій, що відображає підтримку інноваційного розвитку ШІ. Сіра лінія представляє національну безпеку, підкреслюючи роль ШІ в національній безпеці. Ці рядки дають повне уявлення про те, як різні країни розглядають численні аспекти етики ШІ у своїй політиці. Вони відображають ширший спектр міркувань у глобальному дискурсі щодо управління ШІ. Розглядаючи різні аспекти етики штучного інтелекту, можна

створювати справедливі, прозорі та підзвітні політики, одночасно заохочуючи інновації та захищаючи конфіденційність і дані користувачів. Це важливо для зміцнення довіри до штучного інтелекту та забезпечення його використання на благо суспільства.

3.5 Розробка стратегії зниження упередженості в ШІ для підвищення етичності

Упередженість у системах штучного інтелекту є надзвичайно важливою, оскільки вона може значно вплинути на справедливість і ефективність цих технологій. Були розглянуті різні стратегії, спрямовані на пом’якшення упередженості, представлені у вигляді діаграми (рис. 3.3), яка демонструє взаємозв’язок і колективну важливість цих стратегій.

Діаграма сформована при використанні ряду взаємопов’язаних методів, щоб продемонструвати, як вони доповнюють і підсилюють один одного. Основні стратегії включають забезпечення різноманітності даних і використання наборів даних, що представляють усі відповідні демографічні показники, щоб запобігти упередженим моделям ШІ. Регулярні аудити мають вирішальне значення для виявлення та усунення упереджень, які можуть виникнути з часом. Крім того, діаграма наголошує на важливості використання інструментів виявлення зміщень, які використовують спеціалізовані алгоритми для виявлення та усунення зміщень у системах ШІ. Діаграма на рис.3.3 підкреслює важливість спільного впровадження цих стратегій, вказуючи, що найефективніший підхід до пом’якшення упередженості включає багатогранні зусилля.

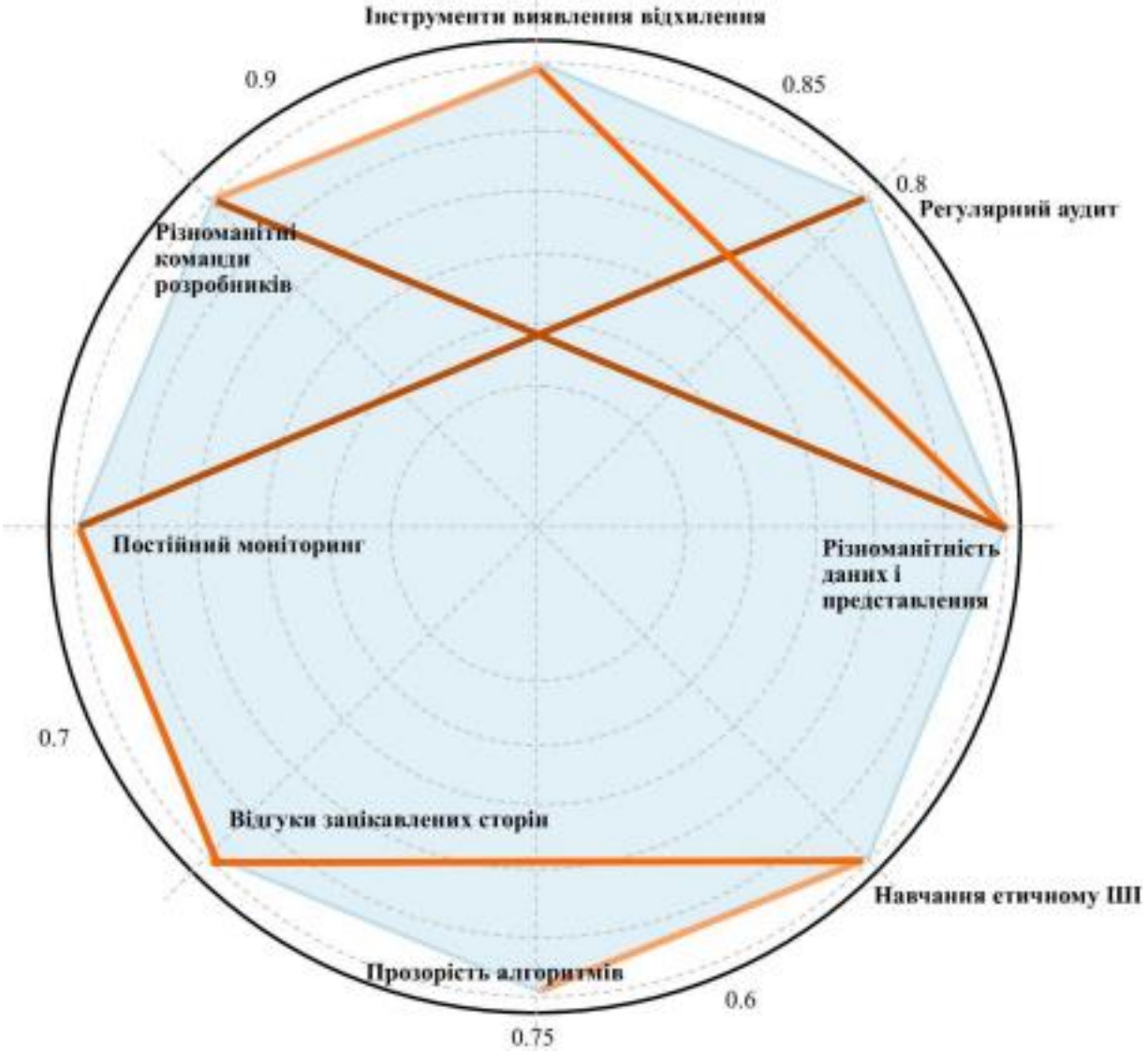


Рисунок 3.3 - Значення колективної реалізації стратегій штучного інтелекту: зв’язки та ваги в стратегіях для пом’якшення упередженості в Al (позначено кольором за силою)

Це включає в себе технічні рішення, організаційні та процедурні заходи для забезпечення того, щоб системи штучного інтелекту розроблялися та експлуатувалися таким чином, щоб мінімізувати упередженість і сприяти справедливості.

Діаграма на рис 3.3 показує зв’язки та присвоєні ваги між різними стратегіями пом’якшення упередженості в ШІ. Відтінок ліній відповідає міцності з'єднання, а ваги позначені на схемі. Стратегії розташовані по колу, щоб підкреслити їх взаємозв’язок і

важливість. Застосовуючи їх спільно, можна значно зменшити ризик упередженості в системах ШІ, що призведе до більш справедливих і надійних рішень ШІ. Однією з ключових стратегій є забезпечення різноманітності та представлення даних. Це передбачає використання різноманітних даних, що представляють усі відповідні демографічні показники, щоб уникнути упереджених моделей у системах ШІ.

Іншою важливою стратегією є інструменти виявлення упередженості. Ці інструменти використовують спеціалізоване програмне забезпечення для виявлення упереджень в алгоритмах ШІ, які потім можна виправити.

Різноманітні команди розробників також можуть допомогти мінімізувати несвідомі упередження при проектуванні та розробці систем ШІ. Проведення етичних тренінгів зі штучного інтелекту для спеціалістів з ШІ є ще одним способом пом'якшити упередженість у сфері штучного інтелекту. Цей тренінг навчає фахівців зі штучного інтелекту етичним міркуванням і уникає упередженості.

Алгоритмічна прозорість також має вирішальне значення для зменшення упередженості в ШІ. Зробивши роботу алгоритмів штучного інтелекту прозорою, упередження можна виявити та виправити швидше.

Залучення різних зацікавлених сторін, у тому числі з непрофільних груп, для надання зворотного зв'язку щодо систем штучного інтелекту, також може значно зменшити упередженість.

Нарешті, постійний моніторинг систем штучного інтелекту є важливим для швидкого виявлення та усунення будь-яких упереджень, які можуть виникнути з часом. Спільна реалізація цих стратегій може значно зменшити ризик упередженості в системах ШІ що, веде до більш справедливих і надійних рішень.

Блок схема розробленої стратегії зменшення упередженості в системах штучного інтелекту наведене у вигляді блок-схеми (рис. 3.4). Ця схема надає контекст та демонструє взаємозв'язки між стратегіями, що забезпечує комплексний підхід до подолання проблеми упередженості.

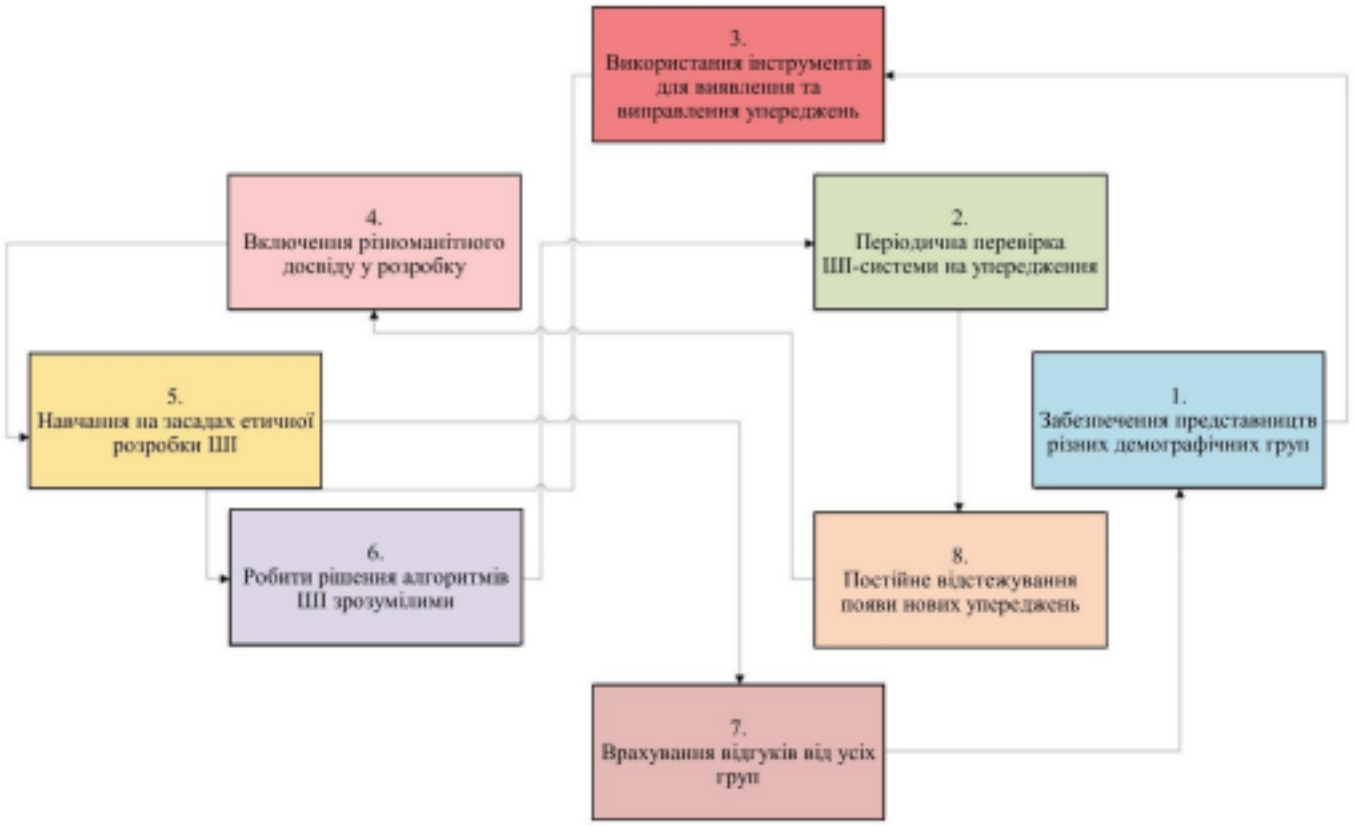


Рисунок 3.4 - Стратегії пом’якшення упередженості ШІ

Однією з ключових стратегій є забезпечення представництва різних демографічних груп у даних, що використовуються для навчання моделей ШІ. Це тісно пов’язано із застосуванням інструментів виявлення упереджень, які допомагають ідентифікувати та усувати потенційні викривлення у моделях. Широкий діапазон даних сприяє формуванню більш збалансованих та точних алгоритмів.

Регулярний аудит ШІ-систем є ще одним важливим аспектом зменшення упередженості. Періодичні перевірки дозволяють відстежувати появу нових викривлень, що може бути наслідком змін у вихідних даних або адаптації системи до нових сценаріїв використання. Постійний моніторинг допомагає зберігати неупередженість системи та запобігати появі небажаних тенденцій у її роботі.

Використання спеціалізованих інструментів для виявлення та корекції упередженості відіграє значну роль у підвищенні точності та прозорості ШІ. Такі інструменти сприяють виявленню потенційних загроз та роблять алгоритми більш

зрозумілими для аналізу. Це, своєю чергою, полегшує виявлення та виправлення

проблем ще на етапі розробки або тестування системи.

Включення представників із різноманітним досвідом у команди розробників є ще однією важливою стратегією. Робота над створенням ШІ за участі спеціалістів із різних сфер дозволяє уникнути однобічного підходу та значно зменшує ймовірність виникнення несвідомих упереджень. Це безпосередньо пов'язано із етичним підходом до навчання, який передбачає врахування різноманітних точок зору при розробці ШІ
систем.

Освіта та навчання спеціалістів на засадах етичної розробки є критично важливим елементом подолання упередженості в ШІ. Це навчання має включати теми, пов'язані з етичною відповідальністю, важливістю збалансованого набору даних та методами виявлення викривлень. Взаємозв'язок між етичним навчанням та врахуванням зворотного зв'язку від усіх зацікавлених сторін допомагає сформувати комплексне бачення впливу ШІ на суспільство.

Прозорість алгоритмів та зрозумілість їхніх рішень є ще одним важливим напрямом боротьби з упередженістю. Чітка документація та зрозумілий принцип роботи алгоритмів полегшують аудит та оцінку справедливості їхніх рішень. Ця стратегія тісно пов'язана з проведенням регулярних перевірок, які сприяють підтримці високого рівня прозорості.

Важливим аспектом є також врахування відгуків від усіх груп користувачів, зокрема тих, які традиційно недостатньо представлені. Це дозволяє системам ШІ адаптуватися до реальних потреб та зменшувати ризики виникнення викривлень у їхніх рішеннях. Постійне спостереження за можливими новими упередженнями та швидке реагування на них сприяє вдосконаленню алгоритмів та їхньої етичної коректності.

Застосування всіх цих стратегій у комплексі дозволяє створювати більш етичні та справедливі системи штучного інтелекту, які враховують інтереси всіх користувачів. Інтеграція етичного навчання, розмаїття у командах розробників та

технічних механізмів контролю сприяє формуванню системи, що відповідає сучасним етичним стандартам та мінімізує ризики упередженості у ШІ.

3.5 Висновок до третього розділу

У цьому розділі було здійснено комплексний аналіз міжнародних політик щодо етичного використання штучного інтелекту, визначено критерії оцінки основ етичної політики в різних країнах, а також розроблено стратегію пом'якшення упередженості в ШІ.

Дослідження показало, що існує значна різниця між підходами різних країн до етичного регулювання ШІ. Основні етичні параметри, такі як прозорість, підзвітність, справедливість і конфіденційність, мають різні рівні пріоритетності в кожній державі. Це підкреслює необхідність глобальної гармонізації етичних стандартів та постійного міжнародного діалогу.

Було обґрунтовано вибір ключових критеріїв для оцінки політики етики ШІ, серед яких: прозорість алгоритмів, підзвітність розробників, дотримання конфіденційності, людський нагляд, етичні стандарти, заохочення інновацій та національна безпека. Оцінка цих критеріїв дозволяє не лише класифікувати існуючі політики, а й визначити напрями їх удосконалення.

Запропонована стратегія зменшення упередженості в ШІ передбачає комплексний підхід, що включає як технічні, так і організаційні рішення. Основними складовими цієї стратегії є:

- використання різноманітних та репрезентативних наборів даних для навчання ШІ;
- проведення регулярних аудитів систем ШІ для виявлення та усунення упередженості;
- розробка та застосування інструментів виявлення упередженості в алгоритмах;

- формування команд розробників із різноманітним професійним і соціальним досвідом;
- впровадження етичних тренінгів для спеціалістів у сфері ШІ;
- забезпечення алгоритмічної прозорості та зрозумілості прийнятих рішень;
- врахування зворотного зв'язку від широкого кола зацікавлених сторін; - постійний моніторинг систем ШІ для оперативного виявлення нових форм упередженості.

Блок-схема запропонованої стратегії демонструє взаємозв'язок між цими елементами та наголошує на важливості комплексного підходу. Спільне впровадження перелічених заходів дозволить значно зменшити ризик упередженості в ШІ, що сприятиме створенню більш справедливих, надійних і соціально відповідальних систем штучного інтелекту.

ВИСНОВКИ ДО ДИПЛОМНОЇ РОБОТИ

Запровадження справедливості, прозорості та підзвітності в системах штучного інтелекту, хоча і має вирішальне значення для забезпечення етичних стандартів, створює значний фінансовий тягар і ускладнює роботу, особливо в секторах, де швидкі інновації є конкуренційною необхідністю. Однією з основних економічних витрат є збільшення складності розробки систем ШІ, які дотримуються етичних принципів. Впровадження алгоритмів навчання з урахуванням справедливості, таких як демографічний паритет або вирівняні шанси, потребує додаткових обчислювальних ресурсів і ретельного тестування на етапі навчання. Ці обмеження справедливості є не просто доповненнями, але вимагають фундаментального перегляду алгоритмічного дизайну, особливо у випадках, коли оптимізація продуктивності суперечить справедливості. Наприклад, у сфері фінансових послуг для забезпечення того, щоб алгоритми схвалення позик не виявляли упередженості, може знадобитися перепідготовка моделей з різними

наборами даних і застосування обмежень справедливості протягом усього циклу розробки.

Цей розширений процес розробки вимагає вищих витрат на оплату праці, потребує більших інвестицій у інфраструктуру та часто призводить до більш тривалих часових рамок для досягнення нормативної відповідності. Крім того, методи збереження конфіденційності, такі як диференційована конфіденційність і інтегроване навчання, додають додаткової складності. Інтегроване навчання, яке дає змогу навчати моделі в розподілених наборах даних без централізації конфіденційних даних, потребує складнішої архітектури системи та безпечних каналів зв'язку, що збільшує як вартість, так і технічну складність впровадження. В оперативному плані вплив строгих етичних принципів відчувається через необхідність постійного дотримання вимог і постійного моніторингу систем ШІ. Етичні рамки, такі як Закон Європейського Союзу про штучний інтелект, передбачають, що програми штучного інтелекту з

високим ступенем ризику, особливо в таких сферах, як охорона здоров'я та кримінальне правосуддя, піддаються постійному аудиту, щоб забезпечити дотримання етичних стандартів після розгортання. Ці операційні витрати посилюються необхідністю інтегрувати інструменти моніторингу справедливості в реальному часі, такі як AI Fairness 360, які перевіряють дрейф упереджень або аномалії прийняття рішень, коли системи ШІ стикаються з новими даними. Ці інструменти вимагають постійних обчислювальних ресурсів, підтримки інфраструктури та персоналу, присвяченого аудиту та повторному калібруванню моделі. Для таких галузей, як фінансові послуги, де системи штучного інтелекту розгортаються в середовищах реального часу, як-от високочастотна торгівля, підтримання справедливості та відповідності додає рівні складності в робочий процес. Ця постійна потреба у повторному калібруванні також може призвести до простою, під час якого системи повинні бути переоцінені та оновлені, що призводить до затримок у процесах прийняття рішень і потенційних порушень безперервності бізнесу. Фінансовий вплив цих етичних вимог також впливає на інноваційні цикли та

швидкість виходу на ринок.

У висококонкурентних секторах, таких як автономне водіння або діагностика на основі штучного інтелекту, час виходу на ринок часто має вирішальне значення для отримання переваги першого. Компанії, які інвестують значні кошти в дотримання етичних норм, як-от у прозорість моделей, перевірку справедливості та пояснення, можуть зіткнутися із затримками у виведенні продуктів на ринок. Наприклад, вимога щодо інтеграції механізмів пояснюваності, таких як додаткові пояснення Шеплі (SHAP, Shapley Additive Explanations) або LIME, у моделі AI часто вимагає додаткових етапів розробки та тестування. Це подовжує загальний графік проекту та може поставити компанії в невигідне конкурентне становище порівняно з тими, хто надає перевагу швидкому розгортанню, а не етичному контролю. Затримка впливає не лише на короткостроковий дохід, але й на довгострокове стратегічне позиціонування, особливо в галузях, де технологічне лідерство є ключовим для збереження частки ринку.

Окрім витрат на розробку та операційних витрат, відповідність законодавству та регуляторний ризик є важливими фінансовими факторами для компаній, які дотримуються суворих етичних принципів. Нормативно-правові рамки, як-от GDPR і майбутній Акт ЄС про штучний інтелект, передбачають суворі покарання за невиконання, а штрафи можуть сягати до 4% від загального доходу компанії за порушення вимог щодо конфіденційності та прозорості даних. Щоб пом’якшити ці ризики, компаніям часто доводиться інвестувати значні кошти в команди юристів, зовнішні аудити та інфраструктуру відповідності. Це створює додатковий рівень витрат, оскільки компанії повинні виділяти ресурси не лише на початкову розробку, а й на постійне управління відповідністю. Циклічний характер відповідності, коли системи необхідно постійно оновлювати, перевіряти та повторно сертифікувати для відповідності стандартам, що розвиваються, створює довгострокові фінансові зобов’язання, які виходять далеко за межі початкового впровадження систем ШІ. Незважаючи на ці витрати, нові технології пропонують потенційні рішення, які можуть пом’якшити деякі фінансові та операційні навантаження, пов’язані з етичним

ШІ. Системи автоматичного машинного навчання (AutoML) дедалі більше здатні включати перевірку справедливості та прозорості в свої розробки, зменшуючи потребу в ручному втручанні та, таким чином, знижуючи витрати на робочу силу. Крім того, технології розподіленого реєстру (DLT, Distributed Ledger Technology), такі як блокчейн, можуть допомогти відстежувати рішення штучного інтелекту прозорим і незмінним способом, тим самим спрощуючи аудити після розгортання та знижуючи витрати на підтримку етичних стандартів. Тим не менш, хоча ці технології пропонують певне полегшення, вони мають свої власні технічні проблеми та інфраструктурні витрати, які потребують додаткових інвестицій та досвіду для ефективного впровадження. Запровадження строгих етичних принципів у розробці ШІ значно впливає на економічні та операційні аспекти проектів ШІ. Незважаючи на те, що ці рекомендації мають вирішальне значення для забезпечення справедливості, прозорості та підзвітності, вони спричиняють значні витрати на кожному етапі

життєвого циклу штучного інтелекту, від розробки до моніторингу та дотримання вимог після розгортання. Збалансування цих етичних зобов'язань із потребою в інноваціях і конкурентоспроможності на ринку залишається проблемою, особливо для компаній, які працюють у високодинамічних і конкурентоспроможних секторах. Ландшафт управління штучним інтелектом, що розвивається, у поєднанні з новими економічними технологіями матиме вирішальне значення для визначення того, як компанії орієнтуються на фінансові та операційні наслідки етичної розробки ШІ.

У даній дипломній роботі розглядалися етичні вимоги, пов'язані з штучним інтелектом, зокрема принципи прозорості, справедливості та конфіденційності, у контексті його переважного впливу на такі сектори, як охорона здоров'я, фінанси та комунікація. Розгортання технологій штучного інтелекту в цих сферах несе з собою глибокі етичні проблеми, які вимагають міцної та інклюзивної системи для захисту індивідуальних прав і інтересів суспільства. Завдяки порівняльному аналізу міжнародної політики щодо штучного інтелекту в Європейському Союзі, Сполучених Штатах і Китаї, проведені в цій роботі дослідження прояснили суперечливі етичні пріоритети, які формують управління штучним інтелектом у всьому світі. Ця робота

роз’яснює етичні принципи конфіденційності, прозорості та справедливості, розглядаючи регіональні проблеми та взаємозалежності.

Розрізняючи, як ці принципи діють незалежно, але інтерактивно в різних структурах, в даній роботі пропонується вдосконалена концептуальна основа, необхідна для глобального управління. Основним внеском є запропонований набір критеріїв інтеграції (сумісність, нормативна згуртованість, адаптованість до культури та прозорість процесу). Ці критерії забезпечують структуровану основу для узгодження етичних принципів у різних міжнародних системах, підтримуючи транскордонну сумісність штучного інтелекту, дотримуючись при цьому регіональних цінностей і нормативних підходів. Графічні представлення представляють окремі та пов’язані взаємозалежності та конфлікти між фреймворками, уникаючи надмірного спрощення та підвищуючи аналітичну ясність. Це дозволило проводити аналіз

візуальним методом, для розуміння етичних стосунків у глобальному управлінні ШІ. Аналіз показує помітні відмінності в тому, як різні регіони збалансовують вимоги інновацій з етичними принципами конфіденційності, справедливості та підзвітності. У той час як певні юрисдикції, такі як Європейський Союз, наголошують на суворому регуляторному нагляді та захисті даних, інші, зокрема Сполучені Штати, застосовують більш гнучкий підхід, орієнтований на інновації. Ці розбіжності підкреслюють складність, пов’язану з прагненням до гармонізованого глобального стандарту етичного управління ШІ. Тим не менш, це дослідження сформулювало кілька стратегічних втручань для пом’якшення алгоритмічної упередженості, включаючи розгортання алгоритмів з урахуванням справедливості, регулярні аудити та залучення різноманітних команд розробників. Ці втручання мають важливе значення для розвитку справедливих і надійних систем ШІ. Крім того, проведена робота підкреслила критичну потребу в постійній міжнародній співпраці та діалозі для подолання прогалин у глобальних структурах етики ШІ.

Стає все більш очевидним, що жодна окрема юрисдикція не може повною мірою вирішити багатогранні етичні виклики, які створює ШІ окремо. Натомість

шлях вперед вимагає узгоджених спільних зусиль, які використовують спільні принципи, поважаючи регіональні відмінності в регулятивних і культурних пріоритетах. Це дослідження висуває аргумент про те, що етичні міркування повинні бути вбудовані на кожному етапі життєвого циклу розробки ШІ, від початку до розгортання та далі.

Розроблена стратегія пом'якшення упередженості спрямована на те, щоб проінформувати регуляторів і розробників ШІ, заохочуючи пошук систем штучного інтелекту, які є не лише інноваційними та технологічно передовими, але й відповідають найвищим етичним стандартам. Оскільки штучний інтелект продовжує розвиватися та використовувати свій трансформаційний потенціал, потреба в пильності, адаптивності та транскордонній співпраці залишається першорядною для забезпечення того, щоб ці технології служили загальному благу, сприяючи справедливості, підзвітності та довірі до їх застосування.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. *Bender, E. M.*, T. Gebru, A. McMillan-Major, and S. Shmitchell. 2021. Про небезпеку стохастичних папуг: чи можуть мовні моделі бути занадто великими? FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–23. doi: 10.1145/ 3442188.3445922.
2. *Bommasani, R., K.* Klyman, D. Zhang, and P. Liang. 2023. Чи дотримуються постачальники базової моделі проект закону ЄС про штучний інтелект? Center for Research on Foundation Models (CRFM): Stanford Center for Research on Foundation Models.
3. *Aldoseri, A., K.* N. Al-Khalifa, and A. M. Hamouda. 2023. Переосмислення стратегії даних та інтеграції для штучного інтелекту: Concepts, opportunities, and challenges. Applied Sciences 13 (12):7082. doi: 10.3390/APP13127082.
4. *European Parliament.* 2023. AI act: A step closer to the first rules on artificial intelligence |news| European Parliament. Інтернет-посилання. Метод доступу: https://www.europarl.europa.eu/news/en/press-room/20230505IPR84904/ai-act a-step-closer-to-the-first-rules-on-artificial-intelligence.
5. *Roberts, H., J.* Cowls, J. Morley, M. Taddeo, V. Wang, and L. Floridi. 2021. Китайський підхід до штучного інтелекту: аналіз політики, етики та

регулювання. AI & Society 36 (1):59–77. doi: 10.1007/s00146-020-00992-2.
6. *UNESCO. 2023*. Recommendation on the ethics of artificial intelligence | UNESCO. Інтернет-посилання. Метод доступу: https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence.
7. *Hosny, A., C.* Parmar, J. Quackenbush, L. H. Schwartz, and H. J. W. L. Aerts.
2018. Штучний інтелект в радіології. (8):500. doi: 10.1038/S41568-018-0016-5.
8. *Tabassi, E.* 2023. Структура управління ризиками ШІ | NIST. doi:
10.6028/NIST.AI.100-1.
9. *NIST. 2023a*. Структура управління ризиками ШІ | NIST. National Institute of Standards and Technology. Інтернет-посилання. Метод доступу: https://www.nist.gov/itl/ai-risk-management-framework
10. *Положення* про адміністрування інформаційних послуг Інтернету глибокого синтезу, Закон Китайської Народної Республіки про захист особистої інформації (PRC). 2022.
11. *Інтернет-посилання*. Метод доступу: https://mlu-explain.github.io/equality-of odds/

12. *Floridi, L., J.* Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, et al. 2018. AI4People — етична основа для доброго суспільства ШІ: можливості, ризики, принципи та рекомендації. Minds and Machines 28 (4):689–707. doi: 10.1007/s11023-018-9482-5.
13. *Wachter, S., B.* Mittelstadt, and C. Russell. 2023. Упередженість у сфері охорони здоров'я небезпечна. Але також і алгоритми справедливості |WIRED. Інтернет-посилання. Метод доступу: https://www.wired.com/story/bias statistics-artificial-intelligence-healthcare/
14. *Mittelstadt, B. 2019*. Самі по собі принципи не можуть гарантувати етичний ШІ. Nature Machine Intelligence 1 (11):501–07. doi: 10.1038/s42256-019-0114-4
15. *Turilli, M.*, and L. Floridi. 2009. Етика інформаційної прозорості. Ethics and Information Technology 11 (2):105–12. doi: 10.1007/s10676-009-9187-9 16. *Binns, R.* 2018. Справедливість у машинному навчанні: уроки політичної філософії. Proceedings of Machine Learning Research, vol. 81, 149–59, PMLR. Інтернет-посилання. Метод доступу: https://proceedings.mlr.press/v81/binns18a.html.
17. *Jobin, A.*, M. Ienca, and E. Vayena. 2019. Глобальний ландшафт етичних принципів ШІ. Nature Machine Intelligence 2019 1 (9):389–99. doi: 10.1038/s42256-019-0088-2.

КРИВОРІЗЬКИЙ ФАХОВИЙ КОЛЕДЖ
ДЕРЖАВНОГО НЕКОМЕРЦІЙНОГО ПІДПРИЄМСТВА
«ДЕРЖАВНИЙ УНІВЕРСИТЕТ «КИЇВСЬКИЙ АВІАЦІЙНИЙ ІНСТИТУТ»

РЕЦЕНЗІЯ
на кваліфікаційну роботу

випускника спеціальності: 123 «Комп'ютерна інженерія»
відділення: комп'ютерної та програмної інженерії
циклова комісія: комп'ютерних систем та мереж
Богдана ШАПОВАЛА
(ім'я, прізвище)

Кваліфікаційна робота присвячена вкрай актуальній проблематиці — забезпеченню етичності у розвитку та застосуванні технологій штучного інтелекту. У сучасному світі, де ШІ дедалі більше впливає на прийняття рішень у сферах економіки, медицини, правосуддя та соціального життя, питання прозорості, пояснюваності, відповідальності та захисту прав користувачів набувають особливої ваги. Автором роботи глибоко проаналізовано ключові етичні виклики, пов'язані з алгоритмічною упередженістю, браком прозорості у процесах прийняття рішень, відсутністю механізмів контролю та недостатнім захистом персональних даних.

У роботі запропоновано системний підхід до інтеграції етичних принципів у процес створення та експлуатації ШІ-систем. Обґрунтовано важливість впровадження етичних стандартів на всіх етапах життєвого циклу ШІ — від розробки моделей до їхнього впровадження у реальні практики. Особливу увагу приділено аналізу ролі міжнародних нормативно-правових актів та ініціатив, які формують глобальну етичну рамку для ШІ — таких як документи ЮНЕСКО, рекомендації ЄС та позиції провідних технологічних компаній.

Кваліфікаційна робота відзначається цілісною структурою, логічною послідовністю викладу матеріалу та ґрунтовністю аналітичного підходу. Автор продемонстрував здатність критично осмислювати складні міждисциплінарні питання, поєднуючи технічні знання зі знаннями в сфері етики, права та соціальних наук. Матеріал викладено грамотно, з дотриманням академічного стилю, а висновки підкріплені аргументованими джерелами.

Загалом робота має високу теоретичну та практичну цінність, може бути використана як основа для подальших досліджень або впровадження етичних регламентів у політику розробки ШІ в організаціях. Враховуючи актуальність теми, якість виконання та рівень аналітичного мислення здобувача, кваліфікаційна робота заслуговує на оцінку «добре».

Рецензент *викладач*
(науковий ступінь, посада)

«____»_____ 2025 р. _____ Тетяна РУБАН
(підпис) (ім'я, прізвище)

З рецензією ознайомлений _____ Богдан ШАПОВАЛ
(підпис) (ім'я, прізвище)