Project number	101085626
Project name	Trustworthy artificial intelligence: the European approach
Funding Programme	ERASMUS2027
Project start date	01-10-2022

Deliverable number	1.3
Deliverable name	Teaching materials for the module “Ethical AI” (for Masters in Computer Science)
Work Package number	1
Lead Beneficiary	Lviv Polytechnic National University
Type	DEM — Demonstrator, pilot, prototype R — Document, report
Dissemination level	Public
Due date (in months)	24
Description	The use cases for discovering existing ethical risks in AI technologies, accordingly to a legal framework on AI by collaborative learning group should be created using the software for interactive cooperation. e-format, Ukrainian language
Website link	https://trustai.org.ua/portfolio-item/d3_teaching_materials_ethical_ai/
Author(s)	Anastasiya Doroshenko

Teaching materials for the module “Ethical AI” (for Masters in Computer Science)

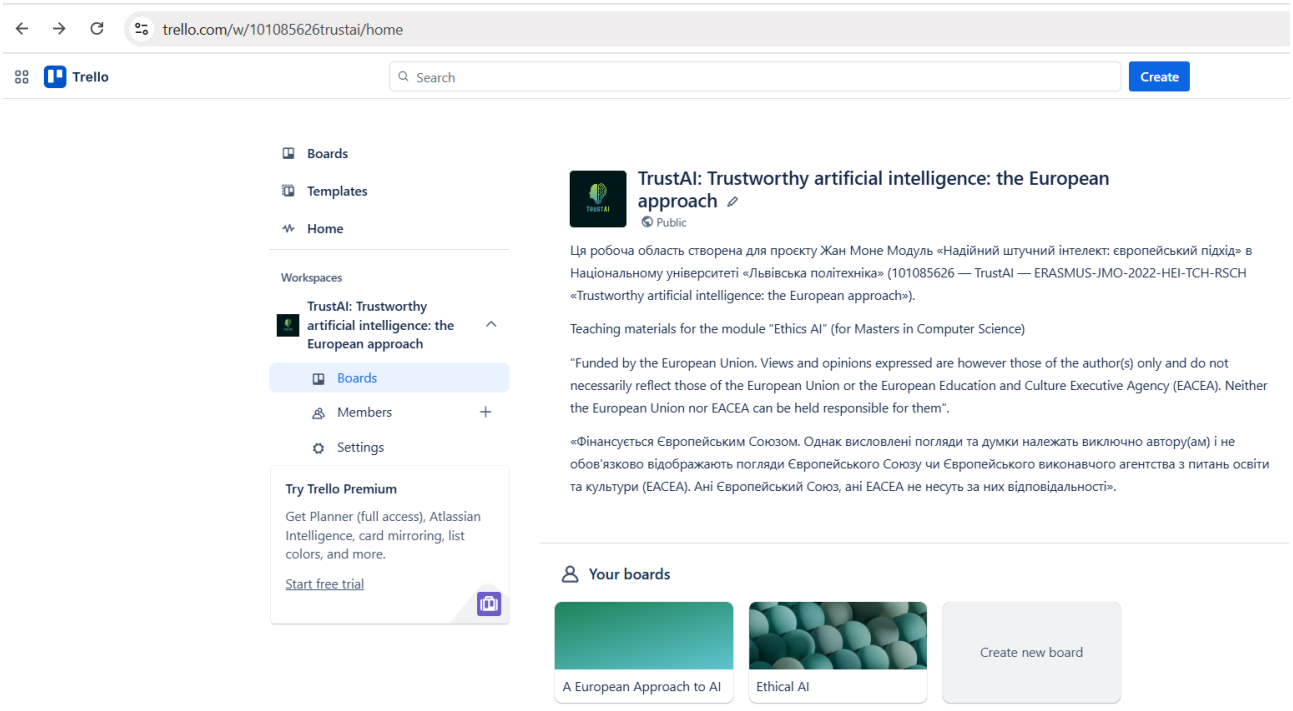
Teaching materials for the module “Ethical AI” (for Masters in Computer Science) consist of 2 parts:

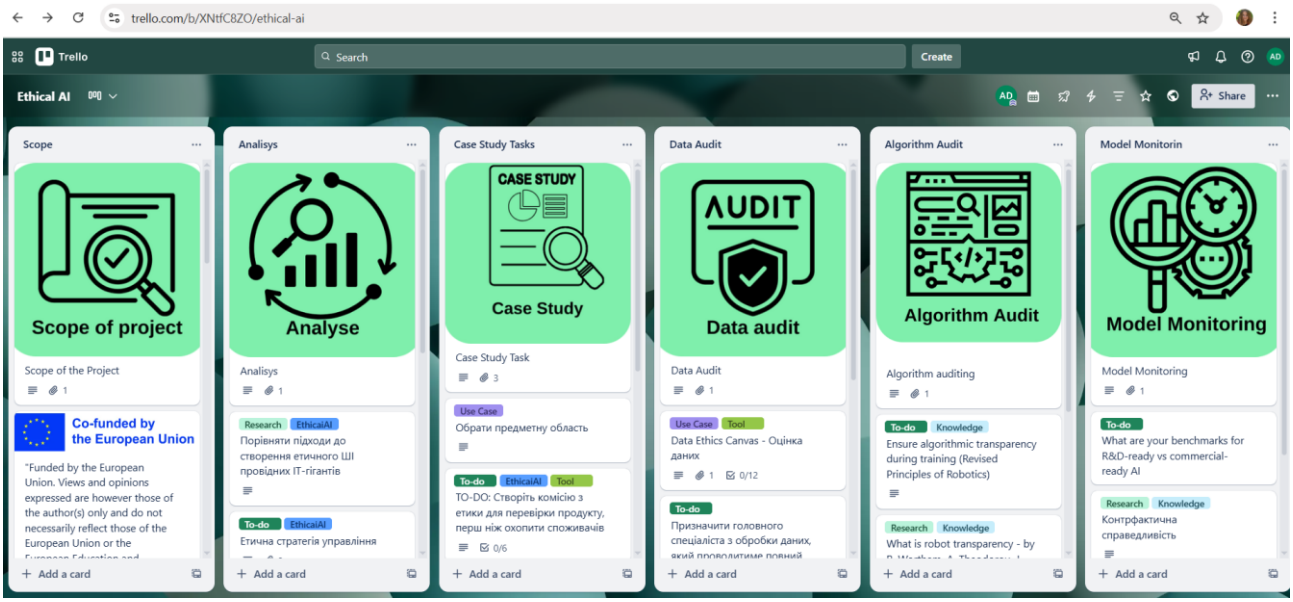
1. The tutorial, which highlights the current problem of assessing the risks of AI Systems and compliance with the requirements of modern European legislation (EU AI Act). The main ethical problems that arise in modern artificial intelligence systems, as well as ways to solve them, are considered. The main principles of ethical AI are described. International standards and global legal acts that determine the requirements for the development of ethical AI systems are considered. For key subject areas - education, medicine and police - options for using artificial intelligence are considered, as well as possible challenges that may arise in them. Recommendations for creating ethical AI in these areas of human life are described.

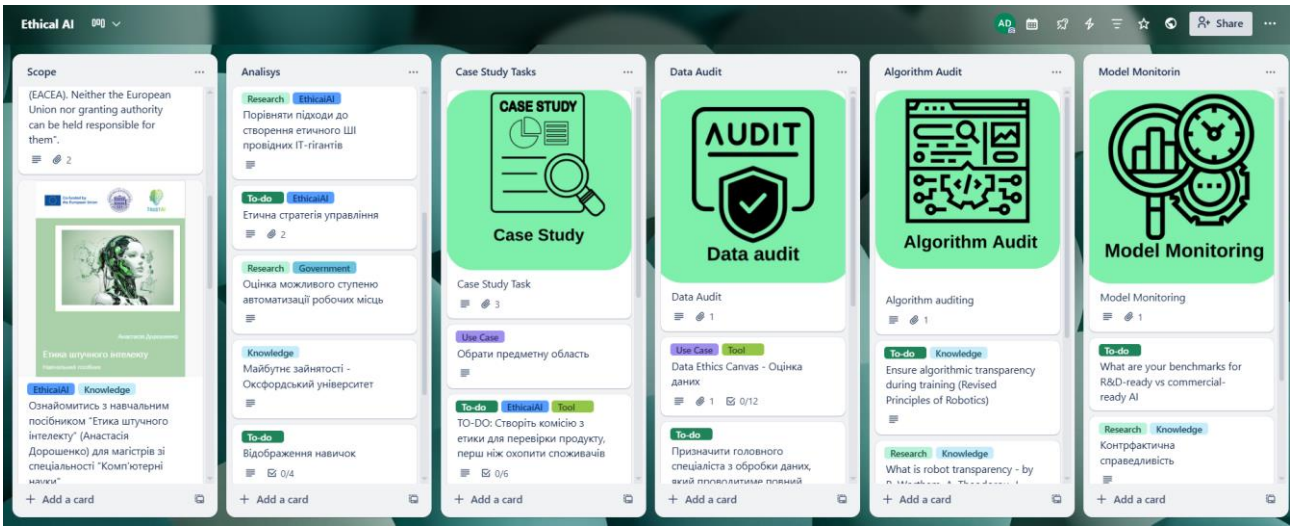
2. The second component of the developed learning materials is an interactive collaboration tool developed using Trello. (https://trello.com/b/XNtfC8ZO/ethical-ai)

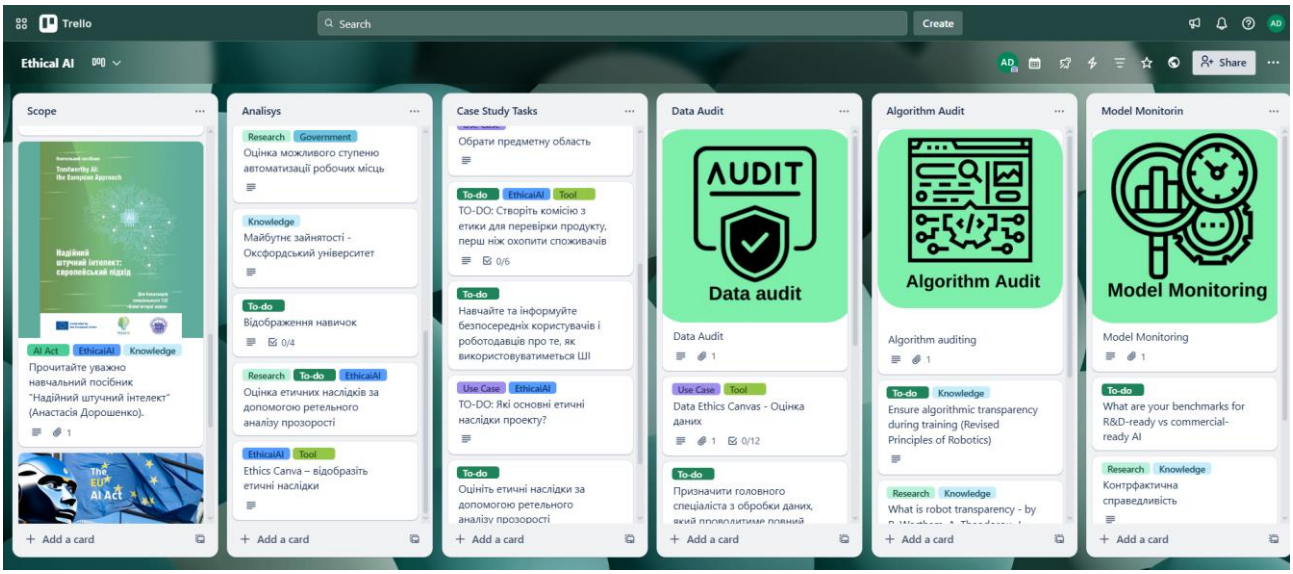
   This tool enables the assessment of an AI system's ethicality, reliability, and fairness. It evaluates key areas such as human oversight, technical robustness, data privacy, fairness, transparency, and resilience to manipulation. Each area is analyzed using specific assessments to ensure the system operates safely, equitably, and in compliance with ethical standards. Also, we can do these assessments for a specific subject area and a specific use case in accordance with the EU legal framework.

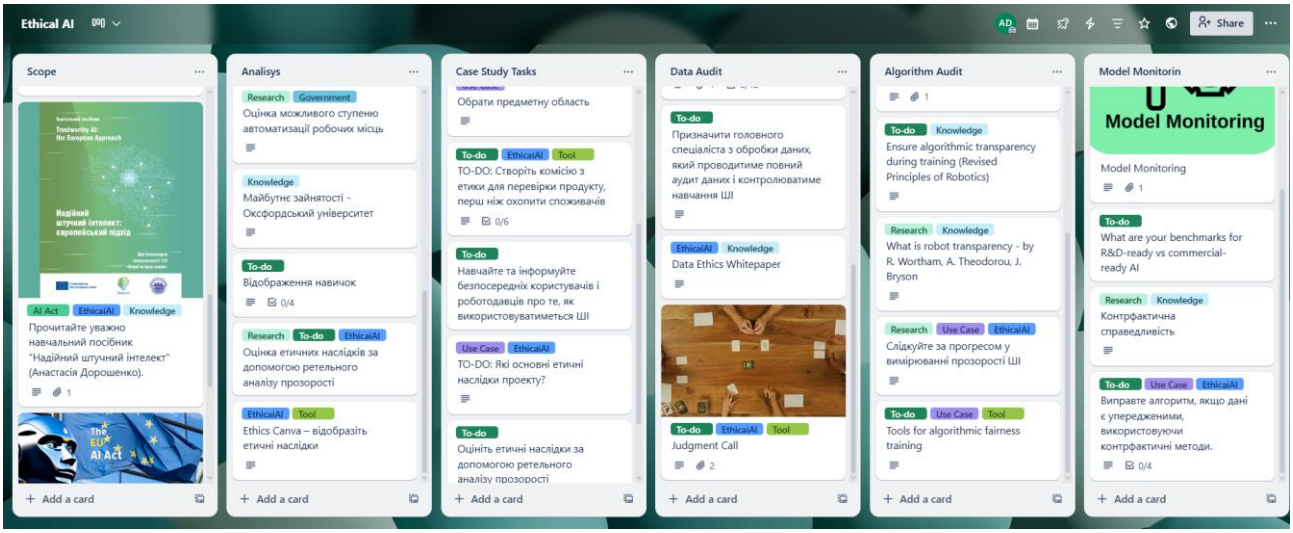
   This tool not only allows you to visualize the risk assessment process for different use cases in accordance with the methodology developed in this course, but also to assign responsibility for each task, track progress and set deadlines.

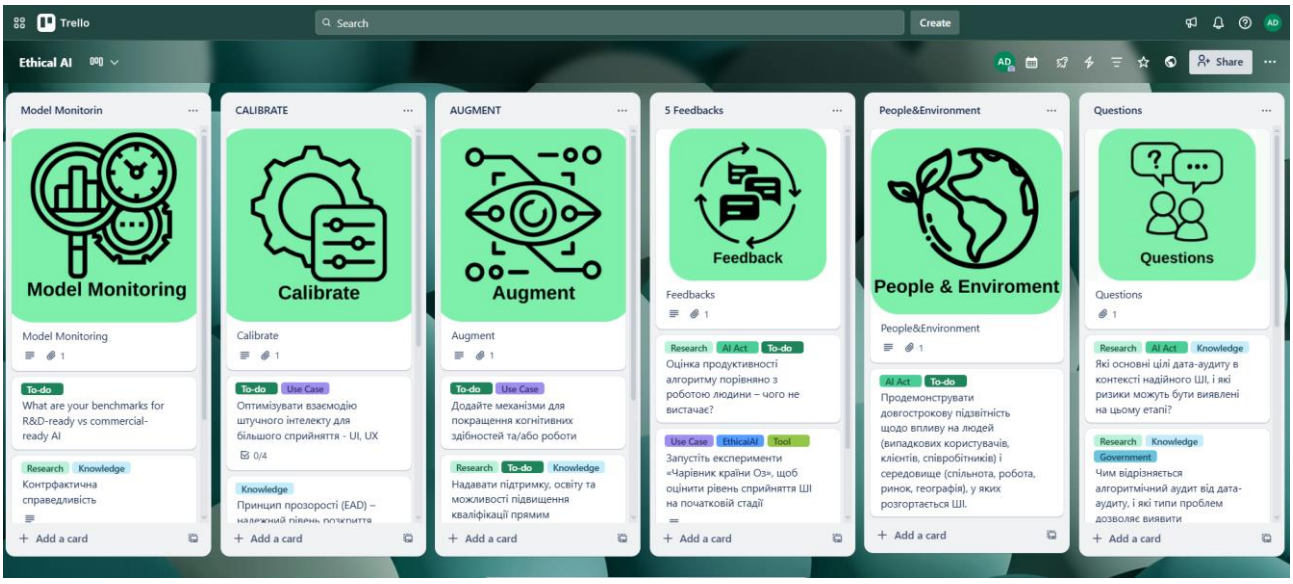


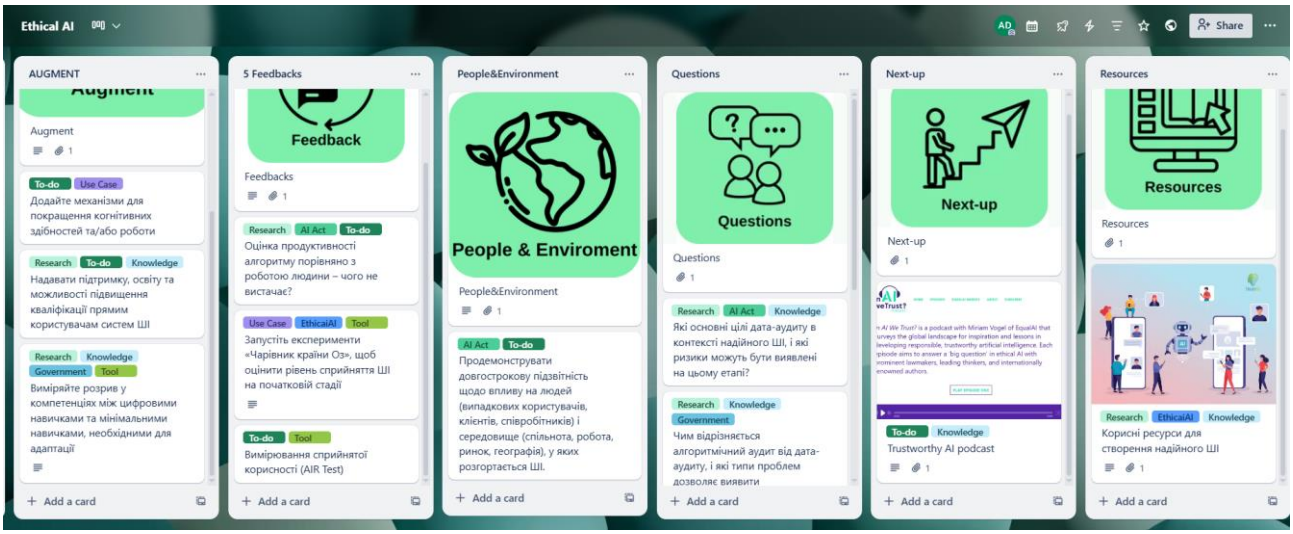

trello.com/b/XNtfC8ZO/ethical-ai
Trello
Search
Create
Ethical AI
Share
Scope
Scope of project
Scope of the Project
Co-funded by the European Union
"Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the
Add a card
Analisys
Analyse
Analisys
Research EthicalAI
Порівняти підходи до створення етичного ШІ провідних IT-гігантів
To-do EthicalAI
Етична стратегія управління
Case Study Tasks
CASE STUDY
Case Study
Case Study Task
Use Case
Обрати предметну область
To-do EthicalAI Tool
TO-DO: Створіть комісію з етики для перевірки продукту, перш ніж охопити споживачів
0/6
Data Audit
AUDIT
Data audit
Data Audit
Use Case Tool
Data Ethics Canvas - Оцінка даних
0/12
To-do
Призначити головного спеціаліста з обробки даних,
Algorithm Audit
Algorithm auditing
To-do Knowledge
Ensure algorithmic transparency during training (Revised Principles of Robotics)
Research Knowledge
What is robot transparency - by
Model Monitorin
Model Monitoring
Model Monitoring
To-do
What are your benchmarks for R&D-ready vs commercial-ready AI
Research Knowledge
Контрфактична справедливість


Ethical AI
Share
Scope
(EACEA). Neither the European Union nor granting authority can be held responsible for them".
EthicalAI Knowledge
Ознайомитись з навчальним посібником "Етика штучного інтелекту" (Анастасія Дорошенко) для магістрів зі спеціальності "Комп'ютерні науки"
Add a card
Analisys
Research EthicalAI
Порівняти підходи до створення етичного ШІ провідних IT-гігантів
To-do EthicalAI
Етична стратегія управління
Research Government
Оцінка можливого ступеню автоматизації робочих місць
Knowledge
Майбутнє зайнятості - Оксфордський університет
To-do
Відображення навичок
0/4
Case Study Tasks
CASE STUDY
Case Study
Case Study Task
Use Case
Обрати предметну область
To-do EthicalAI Tool
TO-DO: Створіть комісію з етики для перевірки продукту, перш ніж охопити споживачів
0/6
Data Audit
AUDIT
Data audit
Data Audit
Use Case Tool
Data Ethics Canvas - Оцінка даних
0/12
To-do
Призначити головного спеціаліста з обробки даних,
Algorithm Audit
Algorithm auditing
To-do Knowledge
Ensure algorithmic transparency during training (Revised Principles of Robotics)
Research Knowledge
What is robot transparency - by
Model Monitorin
Model Monitoring
To-do
What are your benchmarks for R&D-ready vs commercial-ready AI
Research Knowledge
Контрфактична справедливість


Trello
Search
Create
Ethical AI
Share
Scope
AI Act EthicalAI Knowledge
Прочитайте уважно навчальний посібник "Надійний штучний інтелект" (Анастасія Дорошенко).
Add a card
Analisys
Research Government
Оцінка можливого ступеню автоматизації робочих місць
Knowledge
Майбутнє зайнятості - Оксфордський університет
To-do
Відображення навичок
0/4
Research To-do EthicalAI
Оцінка етичних наслідків за допомогою ретельного аналізу прозорості
EthicalAI Tool
Ethics Canva – відобразіть етичні наслідки
Case Study Tasks
Обрати предметну область
To-do EthicalAI Tool
TO-DO: Створіть комісію з етики для перевірки продукту, перш ніж охопити споживачів
0/6
To-do
Навчайте та інформуйте безпосередніх користувачів і роботодавців про те, як використовуватиметься ШІ
Use Case EthicalAI
TO-DO: Які основні етичні наслідки проекту?
To-do
Оцініть етичні наслідки за допомогою ретельного аналізу прозорості
Data Audit
AUDIT
Data audit
Data Audit
Use Case Tool
Data Ethics Canvas - Оцінка даних
0/12
To-do
Призначити головного спеціаліста з обробки даних,
Algorithm Audit
Algorithm auditing
To-do Knowledge
Ensure algorithmic transparency during training (Revised Principles of Robotics)
Research Knowledge
What is robot transparency - by
Model Monitorin
Model Monitoring
To-do
What are your benchmarks for R&D-ready vs commercial-ready AI
Research Knowledge
Контрфактична справедливість

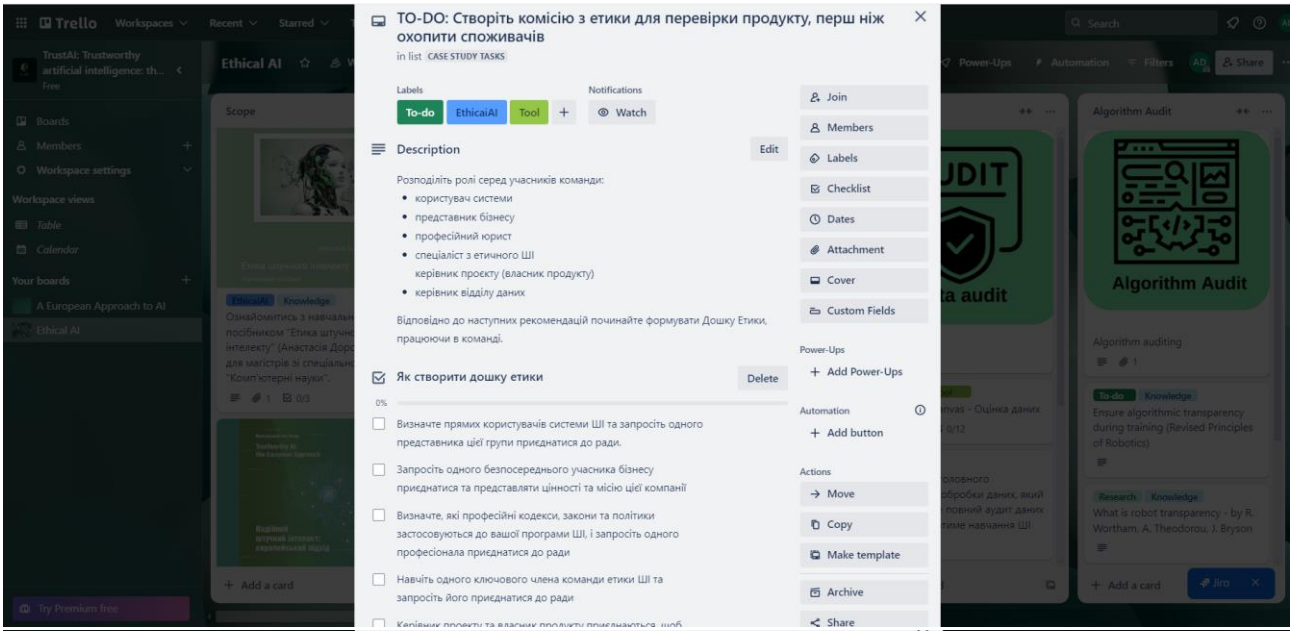


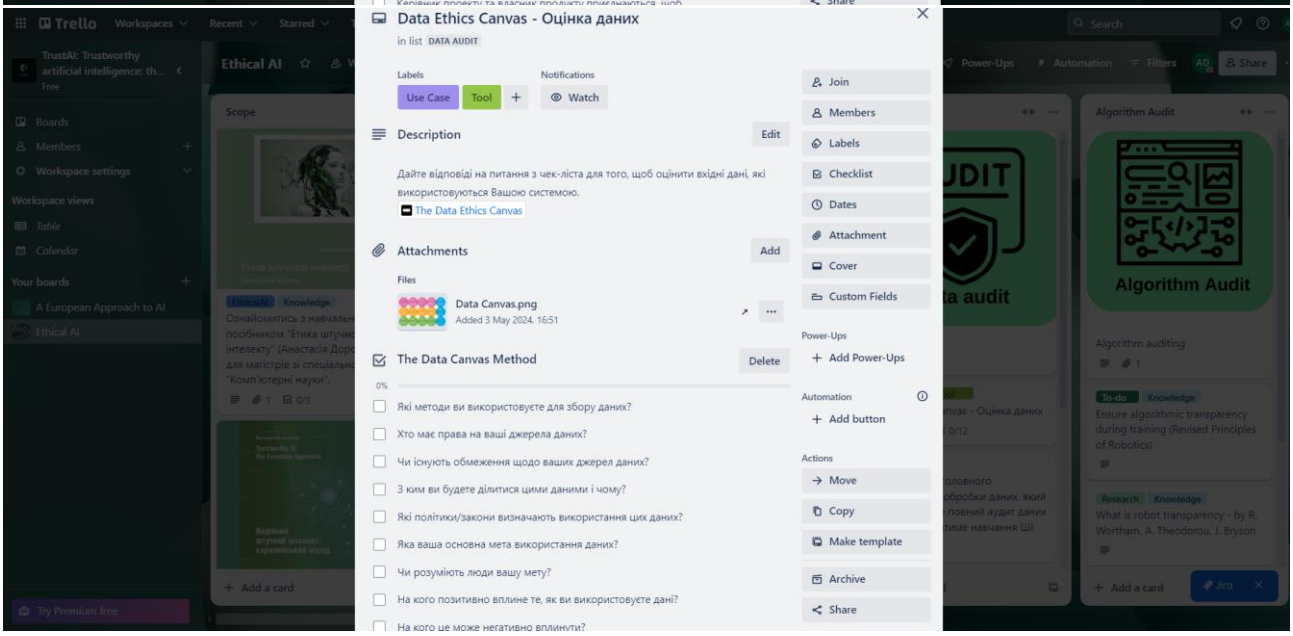


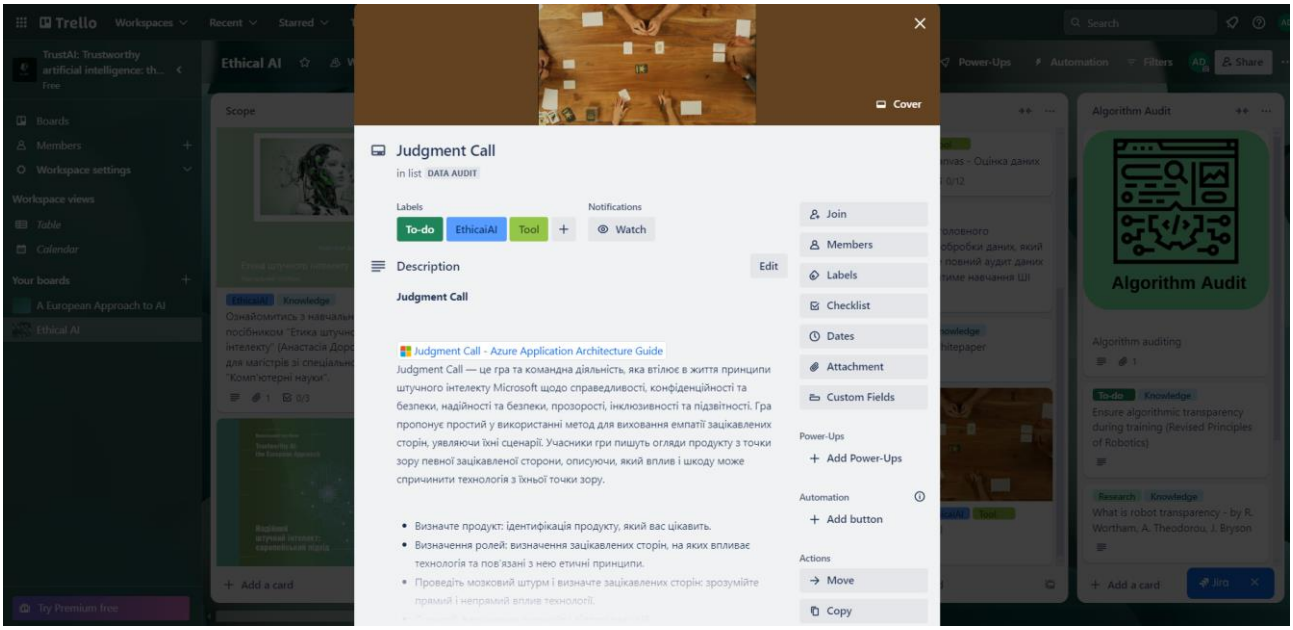


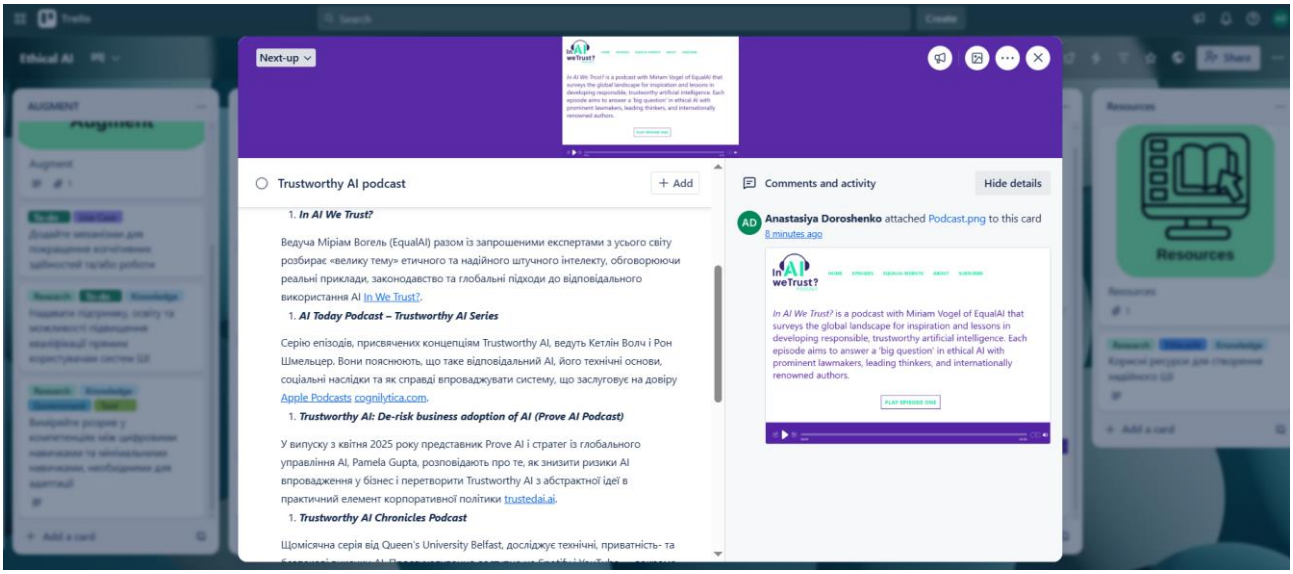
It is also very convenient to be able to store all the necessary work and training materials directly in the project, which greatly facilitates not only the work of the team, but also the teacher's check of the completed assignment.

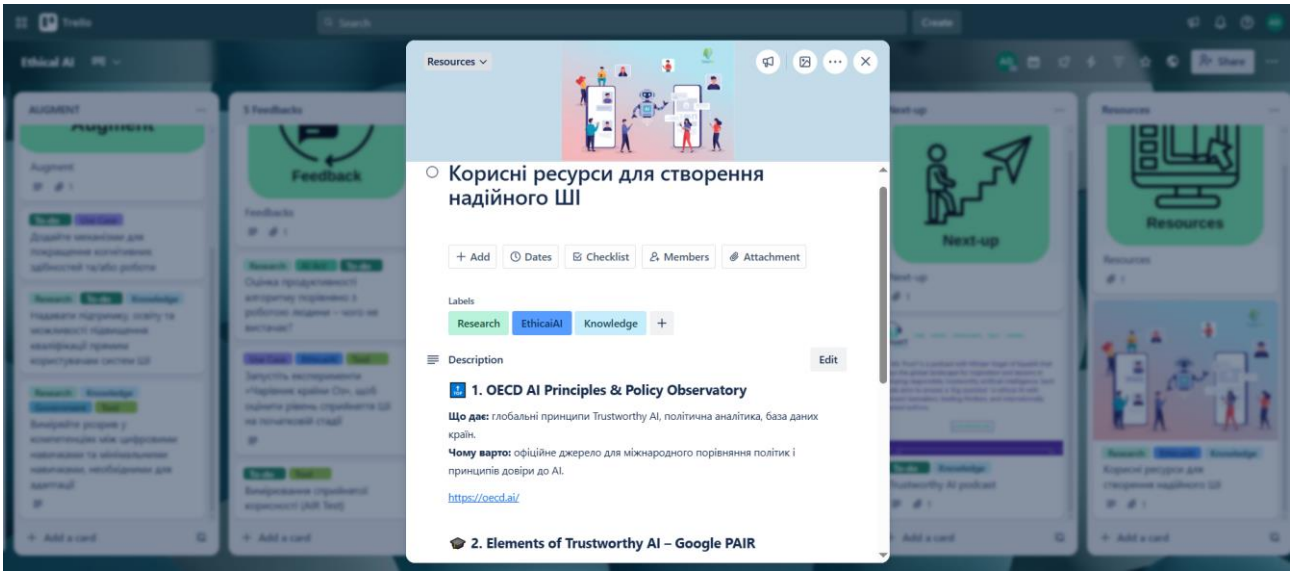
The ability to leave a comment for each step is very useful for the learning process. Accordingly, the teacher can leave a comment or recommendation for each of the completed tasks.


TO-DO: Створіть комісію з етики для перевірки продукту, перш ніж охопити споживачів
in list CASE STUDY TASKS
Labels
To-do
EthicaiAI
Tool
Notifications
Watch
Description
Edit
Розподіліть ролі серед учасників команди:
користувач системи
представник бізнесу
професійний юрист
спеціаліст з етичного ШІ
керівник проєкту (власник продукту)
керівник відділу даних
Відповідно до наступних рекомендацій починайте формувати Дошку Етики, працюючи в команді.
Як створити дошку етики
Delete
0%
Визначте прямих користувачів системи ШІ та запросіть одного представника цієї групи приєднатися до ради.
Запросіть одного безпосереднього учасника бізнесу приєднатися та представляти цінності та місію цієї компанії
Визначте, які професійні кодекси, закони та політики застосовуються до вашої програми ШІ, і запросіть одного професіонала приєднатися до ради
Навчіть одного ключового члена команди етики ШІ та запросіть його приєднатися до ради
Join
Members
Labels
Checklist
Dates
Attachment
Cover
Custom Fields
Power-Ups
Add Power-Ups
Automation
Add button
Actions
Move
Copy
Make template
Archive
Share


Data Ethics Canvas - Оцінка даних
in list DATA AUDIT
Labels
Use Case
Tool
Notifications
Watch
Description
Edit
Дайте відповіді на питання з чек-ліста для того, щоб оцінити вхідні дані, які використовуються Вашою системою.
The Data Ethics Canvas
Attachments
Add
Files
Data Canvas.png
Added 3 May 2024, 16:51
The Data Canvas Method
Delete
0%
Які методи ви використовуєте для збору даних?
Хто має права на ваші джерела даних?
Чи існують обмеження щодо ваших джерел даних?
З ким ви будете ділитися цими даними і чому?
Які політики/закони визначають використання цих даних?
Яка ваша основна мета використання даних?
Чи розуміють люди вашу мету?
На кого позитивно вплине те, як ви використовуєте дані?
На кого це може негативно вплинути?
Join
Members
Labels
Checklist
Dates
Attachment
Cover
Custom Fields
Power-Ups
Add Power-Ups
Automation
Add button
Actions
Move
Copy
Make template
Archive
Share


Cover
Judgment Call
in list DATA AUDIT
Labels
To-do
EthicaiAI
Tool
Notifications
Watch
Description
Edit
Judgment Call
Judgment Call - Azure Application Architecture Guide
Judgment Call — це гра та командна діяльність, яка втілює в життя принципи штучного інтелекту Microsoft щодо справедливості, конфіденційності та безпеки, надійності та безпеки, прозорості, інклюзивності та підзвітності. Гра пропонує простий у використанні метод для виховання емпатії зацікавлених сторін, уявляючи їхні сценарії. Учасники гри пишуть огляди продукту з точки зору певної зацікавленої сторони, описуючи, який вплив і шкоду може спричинити технологія з їхньої точки зору.
Визначте продукт: ідентифікація продукту, який вас цікавить.
Визначення ролей: визначення зацікавлених сторін, на яких впливає технологія та пов'язані з нею етичні принципи.
Проведіть мозковий штурм і визначте зацікавлених сторін: зрозумійте прямий і непрямий вплив технології.
Join
Members
Labels
Checklist
Dates
Attachment
Custom Fields
Power-Ups
Add Power-Ups
Automation
Add button
Actions
Move
Copy


Next-up
Trustworthy AI podcast
+ Add
1. In AI We Trust?
Ведуча Міріам Вогель (EqualAI) разом із запрошеними експертами з усього світу розбирає «велику тему» етичного та надійного штучного інтелекту, обговорюючи реальні приклади, законодавство та глобальні підходи до відповідального використання AI In We Trust?.
1. AI Today Podcast – Trustworthy AI Series
Серію епізодів, присвячених концепціям Trustworthy AI, ведуть Кетлін Волч і Рон Шмельцер. Вони пояснюють, що таке відповідальний AI, його технічні основи, соціальні наслідки та як справді впроваджувати систему, що заслуговує на довіру Apple Podcasts cognilytica.com.
1. Trustworthy AI: De-risk business adoption of AI (Prove AI Podcast)
У випуску з квітня 2025 року представник Prove AI і стратег із глобального управління AI, Pamela Gupta, розповідають про те, як знизити ризики AI впровадження у бізнес і перетворити Trustworthy AI з абстрактної ідеї в практичний елемент корпоративної політики trustedai.ai.
1. Trustworthy AI Chronicles Podcast
Щомісячна серія від Queen's University Belfast, досліджує технічні, приватність- та
Comments and activity
Hide details
Anastasiya Doroshenko attached Podcast.png to this card
8 minutes ago
In AI We Trust? is a podcast with Miriam Vogel of EqualAI that surveys the global landscape for inspiration and lessons in developing responsible, trustworthy artificial intelligence. Each episode aims to answer a 'big question' in ethical AI with prominent lawmakers, leading thinkers, and internationally renowned authors.
Resources


Resources
Корисні ресурси для створення надійного ШІ
+ Add
Dates
Checklist
Members
Attachment
Labels
Research
EthicalAI
Knowledge
Description
Edit
1. OECD AI Principles & Policy Observatory
Що дає: глобальні принципи Trustworthy AI, політична аналітика, база даних країн.
Чому варто: офіційне джерело для міжнародного порівняння політик і принципів довіри до AI.
https://oecd.ai/
2. Elements of Trustworthy AI – Google PAIR
Feedback
Next-up
Resources

Co-funded by the European Union

Анастасія Дорошенко

Етика штучного інтелекту

Навчальний посібник

Етика штучного інтелекту: навчальний посібник для магістрів за спеціальністю «Комп’ютерні науки» / Анастасія Дорошенко. – Національний університет «Львівська політехніка». – Львів, 2024.

Розглянуто основні етичні проблеми, які виникають в сучасних системах штучного інтелекту, а також шляхи їх вирішення. Описано основні принципи етичного ШІ. Розглянуто міжнародні стандарти та світові нормативно-правові акти, які визначають вимоги до розроблення етичних систем ШІ. Для ключових предметних областей – освіти, медицини та поліції - розглянуто варіанти використання штучного інтелекту, а також можливі виклики, які можуть у них виникнути. Описано рекомендації створення етичного ШІ в цих сферах людського життя.

Навчальний посібник написано в межах виконання проекту Жан Моне Модуль «Надійний штучний інтелект: європейський підхід» в Національному університеті «Львівська політехніка» (101085626 — TrustAI — ERASMUS-JMO-2022-HEI-TCH-RSCH «Trustworthy artificial intelligence: the European approach».

"Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them".

«Фінансується Європейським Союзом. Однак висловлені погляди та думки належать виключно автору(ам) і не обов'язково відображають погляди Європейського Союзу чи Європейського виконавчого агентства з питань освіти та культури (EACEA). Ані Європейський Союз, ані EACEA не несуть за них відповідальності».

Table of Contents

1. Навіщо штучному інтелекту етика?

Машина, тупа, нехитра, нездатна розкинути розумом, робить, що накажуть. А тямуща спочатку розмірковує, що вигідніше: вирішити запропоноване завдання чи спробувати від нього відвернутися?

- С. Лем. Футурологічний конгрес

Етика ШІ розглядається у двох основних аспектах: етичні принципи, що лежать в основі прийнятих ШІ рішень, та етична поведінка ШІ у ситуації, що безпосередньо стосується людей. Другий аспект принципово відрізняє етику ШІ від етики інших цифрових технологій.

Впровадження систем ШІ у повсякденне життя пов'язане з безліччю етичних проблем, які вже найближчими роками будуть дедалі серйознішими і складнішими. Першими прикладами стали смертельні наслідки в автокатастрофах із самокерованими автомобілями Tesla (2016) та Uber (2018), протест розробників ШІ в компанії Google проти участі у військових проектах Міноборони США, випадки маніпулювання доступністю інформації, сексизму та расизму в алгоритмах розпізнавання осіб та агресивності з використанням ШІ. Масштабні етичні проблеми виникають під час використання ШІ державними службами контролю за громадянами. Можливі негативні соціальні наслідки застосування алгоритмів у роботі держави широко обговорюються у ЗМІ [1-3], а також в органах влади деяких країн.

Однак етика ШІ принципово відрізняється від етики інших технологій, наприклад, дата-етики. Відносно інших технологій обговорюються загальні питання змін у професійній етиці, етиці застосування, етиці відповіді на соціальні виклики (наприклад, ризик масового безробіття) тощо, тоді як в етиці ШІ є ще зовсім інша, дуже важлива зміна. Етику ШІ технологій від етики інших галузей відрізняє проблема етичної поведінки інтелектуальної системи (ІС) у ситуації, коли її рішення стосується людей.

Принципово важливо, що система ШІ здатна:

- ✓ самостійно приймати рішення щодо людини,
- ✓ аналізувати дані в таких обсягах і з такою швидкістю, як людина робити не в змозі (отже людина не може перевірити вірність рішень).

Відповідно, основна проблема – визначення того, наскільки рішення, які приймає інтелектуальна автономна система (ІАС), відповідають етичним нормам, тобто наскільки вона етична.

Тому ми можемо говорити про два зовсім різні аспекти етики ШІ (Рис. 1.).

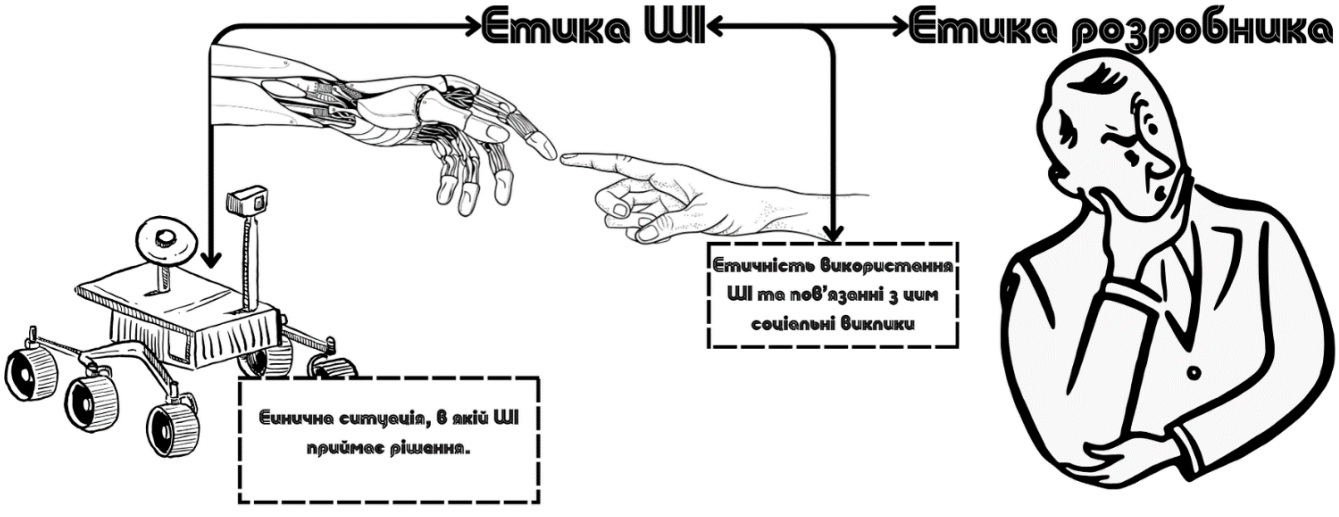


Рис. 1. Два аспекти етики ШІ: етичність рішення та етичність застосування

Перший аспект передбачає, що спочатку програму для ШІ пишуть люди, але надалі система ШІ поводиться вже майже самостійно, може самовдосконалюватись, реструктуризуватись, покращувати свої параметри тощо. Створюючи ІС, яка приймає критично важливі для людини рішення, ми хочемо отримати гарантії , що система керується етичними міркуваннями.

Тому так важливо привнести етичність у саму технологію ШІ. Складність полягає в тому, що моральний вибір — це вибір, який визначається не чітко означеними нормами закону, а приблизно сформульованими правилами, принципами, особистими думками та всім тим, що можна оцінити як «добре» чи «погано». Відповідно, його складно формалізувати та закласти в ШІ.

Другий аспект етики ШІ передбачає аналіз та запобігання етичним колізіям, що виникають у процесі застосування ШІ, до них віднесені: порушення приватності, можлива дискримінація, соціальне розшарування, проблеми працевлаштування тощо. Окремо стоїть тема професійної етики розробників систем ШІ, вона також вимагає розгляду, і в перспективі можливе створення етичних кодексів та рекомендацій для розробників.

В останні кілька років десятки країн включилися до технологічних перегонів з розробки ШІ: до листопада 2019 року було прийнято 34 національні стратегії з ШІ (ще три знаходяться на стадії прийняття). У майбутньому людство не відмовиться від ШІ, швидше за все, ШІ буде все більше, ширше і активніше використовуватися в різних

сферах. Важливо якнайшвидше задати етичні рамки, в яких розвиватиметься ШІ, обмежити можливості його неетичного застосування та спрямувати енергію розробників та ідеї законодавців у русло, що забезпечує максимальну безпеку та вигоду для суспільства.

2. Що таке штучний інтелект?

Величезною проблемою в галузі етики ШІ є відсутність єдиної понятійної бази. Оскільки немає формального чи хоча б конструктивного опису основних положень етики у технічних дослідженнях, моральні аспекти часто обмежуються їх побутовим, інтуїтивним розумінням. Самі фахівці в галузі моральної філософії не завжди чітко уявляють суть досягнень у галузі технологій ШІ. Відповідно, існує розбіжність в поглядах між розробниками, дослідниками та філософами.

Під словосполученням «штучний інтелект» розуміють як комплекс технологій, так і галузь науки.

Штучний інтелект —це комплекс технологічних рішень, що дозволяє імітувати когнітивні функції людини (включаючи самонавчання та пошук рішень без заздалегідь заданого алгоритму), і отримувати при виконанні конкретних завдань результати, які можна порівняти як мінімум з результатами інтелектуальної діяльності. Комплекс технологічних рішень включає інформаційно-комунікаційну інфраструктуру, програмне забезпечення (у тому числі і те, в якому використовуються методи машинного навчання), процеси та сервіси з обробки даних та пошуку рішень.

З іншого боку, **штучний інтелект** — це наука і технологія, куди входять набір засобів, що дозволяють комп'ютеру, виходячи з накопичених знань, давати відповіді на питання і робити з урахуванням цього експертні висновки, тобто отримувати знання, які у нього закладалися розробниками. Область „штучного інтелекту“ є міждисциплінарною і входить до комплексу комп'ютерних наук, а створювані на її основі технології відносяться до інформаційних технологій.

Історія ШІ як науки почалася в 1956 році в США, де на двомісячний семінар у Дартмуті організатори Джон Маккарті, Марвін Мінськ, Клод Шеннон та Натаніель Рочестер запросили визначних американських учених, які вивчають теорію управління, теорію автоматів, нейронних мереж, теорію ігор, та дослідників інтелекту: Артура Самюеля, Аллена Ньюелла, Герберта Саймона, Тренчарда Мура, Рея Соломоноффа та Олівера Селфріджа.

Прийнято виділяти дві групи проблем, які досліджує ШІ як галузь науки:

- **Сильний (інтегральний, універсальний, загальний) ШІ**

Гіпотетичний ШІ, здатний як вирішувати інтелектуальні завдання, так і самостійно ставити цілі, на одному рівні з людським інтелектом або перевершуючи його. Для вирішення проблем сильного ШІ необхідно добре розуміти, як функціонує людський мозок. Нейробіологія накопичила величезну кількість емпіричних знань про анатомію

та фізіологію мозку, молекулярні та генетичні механізми. Проте загальні принципи переробки інформації мозком остаточно не зрозумілі; Відомо лише те, що вони істотно відрізняються від принципів роботи комп'ютера. Тому перспективи реальної розробки сильного ШІ дуже туманні. Ще в 1975 році науковці прогнозували, що створення сильного ШІ відбудеться до 2000 року. Однак і сьогодні, у 2024 найоптимістичніші прогнози появи сильного ШІ – 2040 рік.

- **Слабкий (прикладний) ШІ**

Методи та програмні системи, що вирішують окремі інтелектуальні завдання. Тут успіхи набагато значніші.

Інтелектуальне завдання — це завдання, для вирішення якого людина не має алгоритму. Виконуючи дії за алгоритмом, різні люди завжди отримають той самий результат, причому і хід рішення буде у них однаковим. При розв'язанні інтелектуального завдання люди використовують свої знання, уміння міркувати і кмітливість, як різні індивідууми. Основні успіхи ШІ за 60 років свого існування полягають у формалізації цих інтелектуальних здібностей людини, тобто у розробці методів представлення знань, моделювання міркувань, евристичного пошуку тощо.

Паралельно з методами формалізації інтелектуальних здібностей (символьного, або алгоритмічного, ШІ) розвивався інший напрямок ШІ, заснований на ідеї машинного навчання та технології нейронних мереж. З появою глибокого навчання нейротехнології набули широкого розвитку та стали найбільш тиражованими. Ці інтелектуальні технології поступово розширювали сфери свого застосування, але при цьому породжували завищені очікування.

Тому корисно пам'ятати, що нейротехнології мають деякі важливі обмеження. Вони в основному пов'язані з труднощами оцінки якості навчальної вибірки (зокрема, її повноти), що призводить до проблем з поясненням отриманих результатів і недовіри до них, навіть якщо вони вірні.

Важкі інтелектуальні завдання доводиться вирішувати, наприклад, під час створення безпілотників, особливо автомобілів; в основному саме тому вони ще не виробляються у промисловому обсязі.

Головною проблемою стає оцінка ситуації, можливість відокремити предмет від фону, об'єкт, що рухається від нерухомого, виділити потенційні джерела небезпеки. Вирішити ці проблеми може допомогти машинне навчання на даних та прецедентах. Проблема в тому, що може виникнути ситуація, якої у навчальній вибірці не було.

Деякі види інтелектуальних систем представлені у табл. 1.

Таблиця 1

Види інтелектуальних систем

Вид ІС	Опис та застосування ІС
Інтелектуальні управляючі системи	Управління у виробництві, проектуванні, бортові ІС в авіації
Динамічні робототехнічні системи: роботи, безпілотники	Робот — автономна рухома система, що, керується дистанційно або за допомогою вбудованої програми. Вже є роботи з адаптивною поведінкою, здатні долати перешкоди, «оцінювати» ситуацію, орієнтуватися в просторі
Багатоагентні системи	Колектив роботів із загальною метою, наприклад обстеження території. Управління може йти зовні або зсередини через робота-координатора, але члени системи можуть обходитися і без координатора, взаємодіючи на горизонтальному рівні
Системи підтримки прийняття рішень	Один із напрямків — когнітивні карти. Подібні схеми описують ситуації та зв'язки між ними та показують, як один фактор впливає на інший. З їхньою допомогою можна розрахувати, наприклад, як підвищення податку вплине на доходи бюджету, ціни тощо. Такий інструмент дозволяє простежувати віддалені наслідки рішень, але важливо розуміти, що ІВ не приймає рішення, а лише дає рекомендації. За рішення відповідає людина. Такі системи активно розробляються для різних галузей
Когнітивні дослідження та когнітивне моделювання	Спроба формалізації пізнавальних процесів людини. Ці дослідження спрямовані на те, що називається сильним ШІ
Інженерія знань та онтології	Експертні системи (ЕС) орієнтовані на тиражування досвіду фахівців у галузях, де якість рішень залежить від рівня експертизи та важливий емпіричний досвід фахівців: медицина, юриспруденція, економіка та ін. Багато компаній створюють для внутрішнього користування ЕС за ключовими технологіями

Моделювання міркувань	При моделюванні міркувань використовується логіка, але формалізація класичної логіки – складне завдання. Однією із проблем є формалізація здорового глузду. У міркуваннях людина використовує розпізнавання, досвід, уміння пригадати потрібний прецедент. Моделювання міркувань включає: моделювання міркувань на основі прецедентів аргументації або обмежень, моделювання міркувань з невизначеністю, генерацію та перевірку гіпотез та ін.
Обробка природної мови	Машинний переклад текстів, аналіз текстів: вилучення потрібної інформації, класифікація за змістом, автоматичне реферування, розпізнавання, переклад та генерація мови. Деякі з цих завдань стали успішно вирішуватись за допомогою машинного навчання.

Крім того, елементи ШІ застосовуються в таких сучасних цифрових технологіях, як кіберфізичні системи (розумні будинки, розумні міста), в доповненій реальності тощо.

3. Етичне застосування штучного інтелекту

Впровадженню ШІ та інших цифрових технологій перешкоджає низький рівень довіри громадян до алгоритмів і нових технологій, а також відсутність зрозумілих етичних рамок у застосуванні ШІ. Розглянемо загальну характеристику та найперспективніші підходи до вирішення основних етичних проблем, пов'язаних із застосуванням систем ШІ:

- **відповідальність за етичну/неетичну поведінку ШІ, за прийняття помилкових рішень, збитки через збої тощо;**
- **упередженість алгоритмів (bias);**
- **забезпечення та регулювання прозорості ШІ (пояснювальна компонента);**
- **проблема приватності при застосуванні технологій ШІ ;**
- **надійність технологій ШІ .**

3.1. Відповідальність ШІ

Проблема відповідальності за дії систем ШІ – починаючи від роботи безпілотного автомобіля, постановки медичних діагнозів до гіпотетичних систем прийняття рішень планетарного масштабу – найбільш обговорювана, коли йдеться про застосування ШІ. Проблема відповідальності з'являється в тих сферах довіри, де людині доводиться покладатися на дії системи ШІ, тобто автомобільний транспорт, фармацевтика, медицина, освіта тощо. Також для різних категорій роботів, залежно від ступеня їхньої суспільної небезпеки, контрольованості або здатності до навчання, інститути відповідальності можуть мати окремі нюанси. Деколи взагалі важко відновити фактичні обставини заподіяння шкоди, а проблемна ситуація може отримати різне рішення з погляду конкретної юрисдикції.

Тому існують різні підходи до принципів встановлення відповідальності за дії ШІ, у тому числі:

- **повне звільнення будь-кого** від відповідальності за дії ШІ (за аналогією з обставинами непереборної сили);
- **часткове звільнення від відповідальності** (звільнення конкретної особи від будь-якої відповідальності та одночасна виплата постраждалим компенсації шкоди з різних джерел);
- **відповідальність з вини** , що настає лише залежно від вини конкретного суб'єкта, наприклад виробника, розробника, особи, відповідальної за навчання ШІ, власника, користувача тощо;
- **безвинна відповідальність** (певна особа (швидше за все, виробник) за загальним правилом вважається відповідальною за дії системи ШІ);
- **особиста відповідальність роботів** за умови наділення роботів правосуб'єктністю (правами та обов'язками, статусом електронної особи).

3.2. Ступінь прозорості ШІ

Системи ШІ, здатні самонавчатися, удосконалюватися і розвиватися, стають все складнішими. Сьогодні найбільш складним питанням є те, як саме конкретна система ШІ прийняла рішення, оскільки вона, як правило, робить це за допомогою дуже складного алгоритму. Дії ШІ повинні бути прозорими для широкого кола зацікавлених сторін з низки причин:

- Прозорість важлива для користувачів, оскільки вона формує довіру до системи, надаючи простий спосіб зрозуміти, що та чому робить система.
- Валідація та сертифікація прозорості ІАС є важливими, оскільки вони розкривають процеси, що відбуваються в системі, для проведення перевірки на відповідність ІС чинному законодавству. Наприклад, система автоматизованого ухвалення рішення про видачу кредитів чи рекомендації до вступу в університет повинна розглядати заявки за відкритими та прописаними в законодавстві критеріями. Результат роботи системи повинен збігатися з результатом, досягнутим людиною за тих же розрахунків.
- ІАС має бути прозорою для розслідування аварії, нещасного випадку; так щоб можна було легко простежити, який внутрішній процес спричинив аварію.
- Під час розслідування нещасного випадку прозорість необхідна і адвокатам та іншим експертам.
- Нарешті, революційні технології, наприклад безпілотні автомобілі, мають бути певною мірою прозорі для ширших кіл суспільства, щоб підвищити довіру громадськості до технологій.

Канадська служба з питань імміграції, прийому біженців та громадянства (Immigration, Refugees and Citizenship Canada) з 2014 року розробляє систему для автоматизації міграційної служби. Алгоритмічні системи використовуються на всіх етапах імміграції до Канади, при цьому уряд не розкриває:

- ✓ які критерії використовуються для оцінки мігрантів та біженців;
- ✓ який тип даних буде зібрано та введено в автоматизовану систему;
- ✓ хто матиме доступ до інформації та як вона буде передана іншим відомствам;
- ✓ що держава вважає прийнятною межею похибки цих систем.

Принцип прозорості ШІ, як і ряд інших етичних принципів, закладений в основу EU AI Act. Зазначимо, однак, що прозорість ШІ не може бути абсолютною. Залежно від типу ІС змінюються критерії прозорості, які до них застосовуються. Ступінь прозорості однієї і тієї ж системи може відрізнятися з точки зору функціоналу, цільової аудиторії тощо. Також алгоритм може бути прозорим для одних (наприклад, для розробників

системи скорингу), але непрозорим та незрозумілим для інших (у разі скорингу — для клієнтів банку і навіть його менеджменту, який не має спеціальних знань).

Як і інша інтелектуальна власність, унікальні алгоритми, створені розробниками, не повинні бути повністю розкриті (крім випадків відкритого коду або окремих випадків, обумовлених договором на розробку). Відповідно, пояснювальна компонента ІС повинна працювати таким чином, щоб показувати хід роботи системи, не розкриваючи всієї "механіки" її функціонування. Наявність пояснювальної компоненти – невід'ємна властивість системи ШІ, відома сьогодні як окремий напрям досліджень в сфері ШІ – пояснювальний ШІ (Explainable AI). Якщо пояснюваність відсутня, то відсутня і довіра до такої системи, а отже, її цінність ставлять під сумнів. Наприклад, експертна система може продемонструвати користувачеві весь ланцюжок міркувань, система інтелектуального аналізу даних — видати сформовані нею гіпотези у явному, зрозумілому людині вигляді; система, що доводить теореми, — показати весь ланцюг виведення.

Однак, цілком інша ситуація з одним з найбільш популярних сьогодні методів ШІ – нейронними мережами, які лежать в основі глибокого навчання. Хоча ці методи дають надзвичайно високу точність (до 99%), в них є суттєвий недолік – відсутність пояснювальної компоненти, через що така система є «чорною скринькою». Алгоритми роботи нейронних мереж вкрай складні для інтерпретації, а відповідно, результати їх роботи можуть бути поставлені під сумнів і скасовані людиною. Відсутність розуміння того, як штучний інтелект досягає результатів, є однією з причин низького рівня довіри до сучасних технологій штучного інтелекту і може стати на заваді їх розвитку.

Тому завданням розробника сьогодні є не лише обрати правильний алгоритм та створити модель, яка демонструватиме гарні результати як під час навчання, так і під час використання, але й проаналізувати доступні методи ШІ з точки зору компромісу між їх точністю та прозорістю.

3.3. Упередженість алгоритмів

Проблема упередженості систем ШІ загалом та програм-порадників зокрема — одна з найкритичніших при застосуванні ШІ. Непомітні на перший погляд упередження та припущення, що можуть ховатися в даних, успадковують системі ШІ, навчені на них. Відповідно, це впливає на об'єктивність системи та робить прийняті нею рішення упередженими. У результаті ШІ може мати серйозні зміщення і видавати рекомендації або вчиняти дії, які лише зміцнюють та відтворюють ці упередження. Справедливість алгоритмів — це один із найважливіших напрямів у створенні етичного ШІ.

ШІ може допомогти зменшити упередженість. Як показують дослідження [4], алгоритми здатні допомогти зменшити расову нерівність у системі кримінального правосуддя. Аналогічно автоматизовані системи фінансового страхування можуть бути

корисними для заявників з недооціненою кредитною історією [5]. Нарешті, на відміну від людських, рішення, які приймаються ШІ, в принципі, можуть бути розкриті, досліджені та детально перевірені.

Вивчення кредитної історії при ухваленні рішення про прийом на роботу може зашкодити соціально незахищеним громадянам, хоча наявність зв'язку між якістю кредитної історії та поведінкою на роботі не доведено. У США програма прогнозування злочинів PredPol навчалася на етнічно спотвореній вибірці, тому частіше посилає поліцію на адреси, де мешкають представники етнічних меншин [6]. Навчена на частково вигаданих історіях хвороби програма IBM Watson іноді видає смертельно небезпечні рекомендації щодо лікування раку [7].

Створення справедливого неупередженого ШІ – тема численних дискусій, досліджень, стандартизації тощо. Розглянемо коротко основні можливості зменшення упередженості та дискримінаційності.

Перша – це виявлення та зменшення впливу людських упереджень.

Фахівці лондонської компанії DeepMind запропонували як захист від впливу людських упереджень використовувати метод «гіпотетичної справедливості» (counterfactual fairness, контрфактична справедливість). Щоб сформулювати справедливе і неупереджене судження про громадянина, ШІ формує гіпотетичну ситуацію, в якій даний громадянин має протилежні ознаки: жінка перетворюється на чоловіка, бідний – на багатого, афроамериканець – на білого тощо. Таким чином, реальний статус не впливає на оцінку діянь громадянина. Судження формується у гіпотетичній ситуації. Таке судження вважається вільним від упереджень, отже, справедливим.

Друга можливість — вдосконалення самих систем ШІ, починаючи від способів використання даних і закінчуючи процесами розробки, впровадження та застосування, щоб запобігти закріпленню індивідуальних та суспільних упереджень або виникненню упередженості та пов'язаних із нею проблем. Міждисциплінарне співробітництво може забезпечити розробку та впровадження технічних удосконалень, методів роботи та етичних стандартів [9].

Для того щоб система ШІ була справедливою, необхідно виключити упередженість та дискримінацію у навчальних даних. Як показує практика, навіть за ретельної підготовки даних це не завжди можливо.

Компанія Amazon припинила використовувати систему підбору персоналу після того, як в алгоритмі виявилися похибки, пов'язані з гендерними упередженнями [10]. Алгоритм розпізнавав шаблони слів у резюме, а не необхідні набори навичок для конкретної позиції. В

результаті аналізу роботи системи виявилось, що для навчання системи використовувались переважно резюме білих чоловіків. Відповідно, алгоритм виключав резюме, які такі містили слова, що частіше використовуються жінками. Через це розроблена система мала значне упередження проти жінок під час прийому на роботу.

Упередженість може бути не лише природною, що утворилася випадково через особливості вхідних даних, але і штучною, закладеної навмисно, наприклад у вигляді переваги інтересів деяких третіх осіб. Прикладом такої упередженості може бути невелика навмисна зміна маршруту користувача на карті в навігаційній системі, щоб він проїхав або пройшов повз певну точку, наприклад магазину, який замовив рекламу.

Нарешті алгоритм машинного навчання може виявити статистичні кореляції, які соціально неприйнятні або незаконні. Наприклад, модель іпотечного кредитування виявляє, що в людей похилого віку вища ймовірність не дотриматися графіку платежів, і на цій підставі скорочує обсяг кредитування залежно від віку. Суспільство та правові інститути можуть вважати це незаконною дискримінацією за віком.

У США алгоритмічний аудит програми для прогнозування обсягу необхідної медичної допомоги виявив упереджене ставлення алгоритму до афроамериканців [11]. Незважаючи на єдину методику розрахунку для всіх пацієнтів, алгоритм вважав чорношкірого пацієнта таким, що потребує медичної допомоги менше, ніж білий, навіть якщо у першого більше об'єктивних причин отримати медичну допомогу. У коді не було закладено перевагу білошкірих пацієнтів, і алгоритм працював правильно. Помилковою була початкова ідея розробників, що рівні витрати на медичну допомогу свідчать про однакову потребу в ній, тому алгоритм розраховував рекомендації на підставі витрат пацієнтів на медичну допомогу в минулому. Однак витрати людини на медичні послуги дуже залежать від рівня доходу та соціального стану. Отже, алгоритм закріпив дискримінацію, що існувала ще до його використання: пацієнти, які в минулому отримували менше медичної допомоги через низький рівень доходу, будуть обділені нею і в майбутньому.

Як ми визначаємо і вимірюємо справедливість , щоб мінімізувати упередженість?

Коли ми говоримо про упередженість в системах ШІ та боротьбу із нею, то часто кажемо, що це потрібно для того, аби робота системи та рішення, які нею приймаються, були справедливими. Однак, складність забезпечення справедливості систем ШІ полягає ще й у тому, що саме поняття справедливості потребує окремого визначення та дослідження. Існує трохи більше 20 різних визначень справедливості і навіть вони не є вичерпними. [13]. Як правило, у визначеннях йдеться про індивідуальну справедливість – однакове поводження зі схожими людьми або про групову справедливість [14]. Ймовірно, ніколи не вдасться створити єдине, універсальне визначення справедливості

чи системи показників її виміру. Натомість, швидше за все, знадобляться різні системи показників та стандарти залежно від обставин та варіанту використання [15].

Важливою складовою справедливості систем ШІ є безпосередня участь людини. Хоча статистичні показники справедливості , безумовно, корисні, але вони можуть враховувати нюанси соціальних умов, у яких розгортається система ШІ , і потенційні проблеми, пов'язані, наприклад, зі збором даних [16, 17].

Потрібно відповісти на запитання:

- Де саме і в якій формі необхідне людське судження при розробці та експлуатації ШІ ?
- Хто вирішує, коли система ШІ вже мінімізувала упередженість та придатна для безпечного використання?
- У яких ситуаціях взагалі допустиме повністю автоматизоване прийняття рішень?

Ці питання жоден алгоритм оптимізації не може вирішити самостійно, і їх не можна довірити жодній машині. Вони вимагають людського судження та осмислення з опорою на безліч дисциплін, включаючи гуманітарні науки, особливо соціальні науки, право та етику.

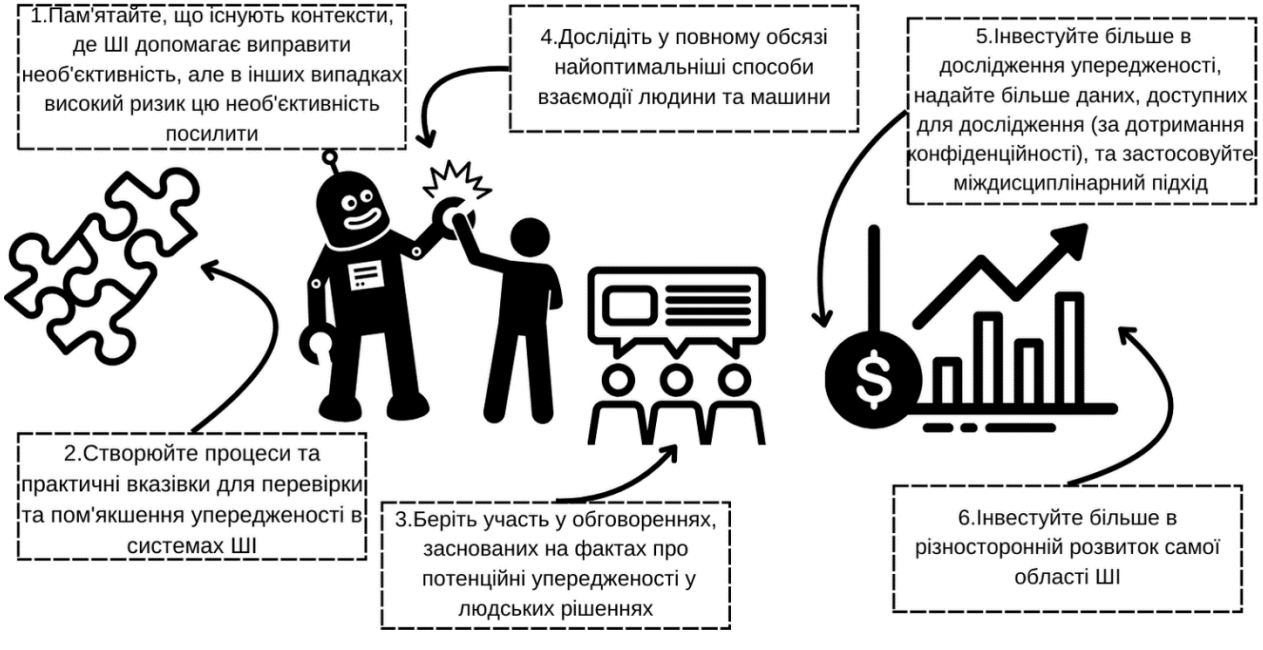


Рис.2. Рекомендації щодо роботи з ШІ для зниження упередженості

Сьогодні різні організації та науковці розробляють різноманітні рекомендації та фреймворки для виявлення та оцінки упередженості. Одні з таких верхньорівневих рекомендацій щодо роботи з ШІ для зниження упередженості розробив Інститут McKinsey [12]. Вони адресовані практичним спеціалістам з ШІ та керівникам (рис. 2).

Упереджене прийняття рішень людьми або машинами не лише призводить до руйнівних наслідків для людей [12], які зазнають дискримінації, а й завдають шкоди всім громадянам та державі загалом, необґрунтовано обмежуючи можливості окремих осіб брати участь та робити свій внесок у розвиток економіки та суспільства. Крім того, мінімізація упередженості в системах ШІ критична, інакше люди не зможуть довіряти цим системам. Останнє необхідно, щоб ШІ зміг реалізувати свій потенціал для держави та для економіки через збільшення продуктивності, а для суспільства завдяки внеску у вирішення нагальних соціальних проблем [18].

3.4. Надійність ШІ

Проблеми безпеки і надійності ШІ , далекі від етики на перший погляд, все ж таки мають безпосереднє відношення до неї і заслуговують на розгляд у кількох аспектах. Перший аспект, суто технічний, стосується надійності та безпеки технічних (програмно-технічних) систем загалом. У звичайних системах існує не менший ризик помилок і технічних збоїв, ніж в інтелектуальних, тому цей аспект не є специфічним для ШІ . Помилка в ПЗ (не інтелектуальному) системи управління ядерним реактором, швидше за все, набагато небезпечніша порівняно з помилками в ШІ для видачі споживчого кредиту.

Другий аспект стосується самої суті роботи ІС. Оскільки ШІ займається вирішенням слабо формалізованих завдань, використовуючи різного роду евристики (складно довести коректність правил, але на практиці вони дають прийнятні результати), правдоподібні міркування та подібні механізми, то найчастіше від систем ШІ і не чекають оптимального, єдино вірного рішення, задовільняючись рішенням "субоптимальним", "розумним", "придатним". Саме для перевірки коректності, контролю потрібна пояснювальна компонента, про яку йшлося вище. Тут проблема надійності перетинається із проблемою прозорості ШІ .

Коли прозора система ШІ починає працювати неправильно, ненадійно, розробники можуть швидко знайти причину помилки. У непрозору систему з «чорною скринькою» можна приховано внести зміни так, що вона досить довго прийматиме рішення по-новому, не викликаючи підозр, і на виявлення помилки (зміни) потрібно набагато більше зусиль і часу. Для вирішення цієї проблеми доцільно організувати моніторинг моделі, який буде аналізувати її роботу після розгортання. Такий моніторинг дасть змогу одразу побачити, якщо поведінка системи відчутно змінилась, якщо погіршилась її точність, а також у випадку появи упередженості в її роботі.

Третім аспектом можна назвати проблему програми-порадника («досвід оператора»), яка безпосередньо стосується етики. Проблема актуальна не тільки для ІС, але саме з початком їхнього застосування вона виходить на новий рівень. Експериментальні дослідження показують, що в умовах невизначеності та дефіциту часу у людей виникає наддовіра до систем ШІ та роботів: люди схильні більше довіряти

системі, ніж собі [19,20]. Довіра до ШІ зростає , якщо програма коментує свої дії [21]. Цілком покладаючись на систему ШІ , людина рідше приймає рішення самостійно і усвідомлено, що сприяє ризику помилок, і втрати кваліфікації. Для власників ІС існує спокуса найняти менш кваліфікованого фахівця та запропонувати йому зарплату скромніше. Тому важливо, щоби на небезпечних виробництвах, у медицині, освіті фахівці зберігали стабільно високий рівень кваліфікації. Не можна знижувати рівень освіти оператора небезпечного виробництва або лікаря, до такого, щоб він тільки знав, як працювати з системою та які кнопки натискати. Навпаки, кожен із них має детально розуміти свою область. Якою б розумною та інтелектуальною не була система ШІ, другий рівень людського контролю потрібен, і відповідних фахівців треба готувати максимально якісно [22].

Стандартизація покликана підвищити надійність систем, поставити мінімальний рівень, перебуваючи нижче якого система не може вважатися надійною. У той же час ця планка не повинна перевищувати можливості сьогоднішніх технологій, щоб розробники не були обмежені в процесі створення нових проривних рішень і не забирали частину технологій з легального ринку.

3.5. Приватність інформації, отриманої за допомогою аналізу метаданих

Алгоритми ШІ здатні отримувати нову персональну інформацію про людей шляхом аналізу великих даних, видобуваючи її з метаданих. Збираючи все більше даних про людину в цифровому профілі, власник алгоритму – компанія, державна організація, органи поліції тощо – може з високим ступенем точності передбачити, що далі відвідуватиме і слухатиме цей користувач, за кого він схильний проголосувати, за допомогою чого можна ним маніпулювати та багато іншого. Розповідаючи про свої смаки, уподобання, місця та сайти, де вони бували, люди не замислюються про те, що колись (можливо, через роки) хтось скористається цією інформацією, щоб зробити статистичні зрізи, вибірки, аналіз тощо).

Вже сьогодні державі та суспільству свідомо чи мимоволі доведеться сформувати своє ставлення до допустимого рівня обробки таких даних. Один із підходів до вирішення проблем приватності, прозорості та справедливості ШІ — це створення стандартів на розробку етично орієнтованих ІС.

В Європейському Союзі ситуацію із регулюванням обробки персональних вдалось значно покращити після того, як у 2016 році було прийнято Загальноєвропейський регламент про захист персональних даних (GDPR, General Data Protection Regulation), який увійшов в дію 25 травня 2018 році. У GDPR чітко прописані досить суворі вимоги до компаній та організацій, які отримують доступ до персональних даних користувачів (контролери) та до компаній чи організацій, які отримують ці дані від контролера для подальшої обробки (процесори). Також визначено як саме необхідно отримувати згоду

користувача на оброблення персональних даних для того, щоб вона вільно наданою та поінформована [33].

Виділено спеціальні типи даних (чутливі дані), збір та обробка яких вимагає додаткових заходів захисту та підвищені вимоги до контролера – це медичні, біометричні, генетичні дані. До чутливих даних, відповідно до GDPR, належать також дані про політичні, релігійні та сексуальні вподобання людини, її приналежність до професійних об'єднань – тобто дані, обробка яких може призвести до дискримінації чи упередження. Саме в GDPR було також окремо приділено увагу користувача про ухвалення рішень щодо нього із використанням засобів автоматизованої обробки (до яких належать і системи ШІ) та вказано, що суб'єкт персональних даних по-перше має право відмовитись від такої автоматизованої обробки, а по-друге – вимагати пояснення щодо того, чому саме це рішення було прийнято на підставі його даних. Це саме те, про що ми говорили вище в контексті пояснюваності і такого напряму, як пояснюваний ШІ.

В Україні захист персональних даних фізичних осіб в Україні регулюється Законом України «Про захист персональних даних» від 01.06.2010 № 2297-VI (Закон № 2297). Згідно з цим Законом обробкою персональних даних є будь-яка дія або сукупність дій, таких як збирання, реєстрація, накопичення, зберігання, адаптування, зміна, поновлення, використання і поширення (розповсюдження, реалізація, передача), знеособлення, знищення персональних даних, у тому числі з використанням інформаційних (автоматизованих) систем [34].

Закон про захист персональних даних регулює такі основні питання обробки персональних даних:

- ✓ вимоги до обробки персональних даних;
- ✓ права суб'єктів персональних даних;
- ✓ підстави для обробки персональних даних;
- ✓ порядок здійснення основних процесів обробки персональних даних (збирання, використання, накопичення та зберігання, поширення, видалення);
- ✓ порядок обробки персональних даних, обробка яких становить особливий ризик для прав і свобод суб'єктів персональних даних тощо.

Проте, в епоху бурхливого розвитку цифрових технологій діюче законодавство про захист персональних даних все більше й більше застаріває. Воно не забезпечує належного захисту суб'єктів персональних даних при обробці персональних даних з використанням інформаційних технологій та не містить достатніх гарантій прав суб'єктів даних у разі витоку персональних даних, що містяться у базах даних володільця персональних даних.

4. Машинна етика

Машинна етика, або етика «поведінки» систем ШІ, викликає низку проблем, передусім розуміння того, в чому полягає етичність рішення. Основна складність полягає у виборі тих самих етичних норм, які закладатимуться у ШІ.

4.1. Підходи до вирішення проблем машинної етики

При розгляді етичних проблем ШІ діапазон тем для міркувань надзвичайно широкий — від небезпеки машин, що «думають» (пригадаємо обговорення проблеми машинних помилок у книзі А. Тюрінга [24]та небезпека для людини у Н. Вінера [25], негативні сценарії можливого розвитку ШІ у Бострема [26]) до цікавих спроб формалізації понять моралі з метою імітувати їх у програмних та технічних системах. Однак реальні досягнення в цьому напрямку виглядають набагато скромнішими, це зумовлено відсутністю потреб і, відповідно, завдань, в яких така формалізація була б необхідна. Сьогодні ситуація змінюється, і питання практичної реалізації етичної компоненти ШІ стають все більш значущими, серед них:

- реалізація машинної етики;
- формалізація етичних понять;
- верифікація та валідація етичної компоненти;
- стандартизація машинної етики;
- стандартизація етичних аспектів ШІ.

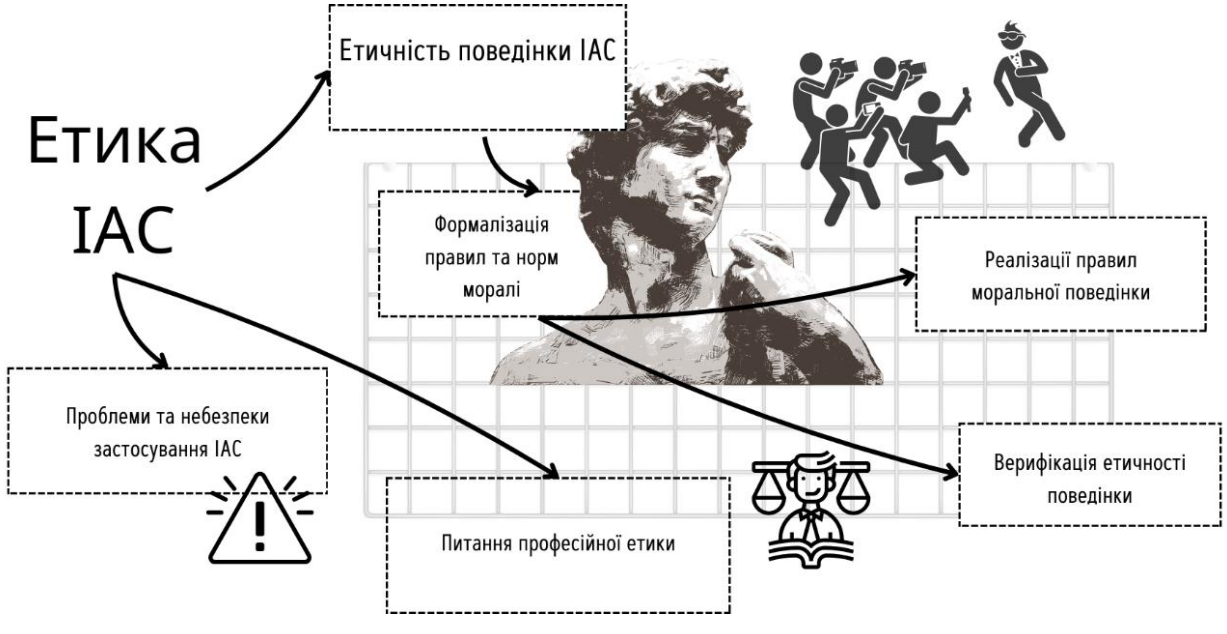


Рис. 3. Машинна етика

Стосовно різних науково-технічних сфер питання етики розуміються в основному з точки зору небезпеки застосування тих чи інших технологій. Проте в області ШІ ситуація зовсім інша, і поряд з проблемами застосування ШІ та професійної етики створення ІС на перший план виходить проблема етичності «поведінки» ІС, так званої машинної етики (рис. 3).

Відповідно, перед вченими, інженерами, юристами та філософами постало завдання виробити підходи до проектування ІАС з етичною компонентою, яка може бути реалізована насамперед в етичних стандартах з проектування ІАС. В етично обумовленому проектуванні предмет дослідження – це ІАС, які роблять вибір того чи іншого значущого, критично важливого для людини чи суспільства дії чи рішення. Нас цікавить ситуація, коли вибір складає основі деяких евристик, заснованих на етичних імперативах.

Евристики *- це правила, які допомагають при пошуку варіантів рішень. Евристики не мають суворого обґрунтування, а основною мотивацією їх використання є лише підвищення ефективності пошуку. Прикладом евристики є правило лівої (правої) руки під час пошуку виходу лабіринті.*

Іншими словами, основне питання – це етичність поведінки самої системи, що викликає цілу низку проблем. Має бути зрозуміло, в чому полягає етичність рішення, на це питання дуже складно отримати однозначну відповідь від моральних філософів. Однак без залучення професійних фахівців з етики розробники етичної компоненти ШІ так і залишаться в полоні своїх побутових уявлень про мораль. Отже, важливо встановити зв'язок між етичними концепціями, поняттями і сутностями і технічними елементами, у яких можуть бути втілені, тобто вирішити питання конструктивних визначень і онтологій .

Онтологія *- це максимально точний, суворий, ясний опис деякого поняття, явища, предметної галузі. Формально онтологія складається з термінів, організованих у таксономію (систему), їх визначень та атрибутів, а також пов'язаних з ними аксіом та правил виведення. Цей термін було введено при розгляді питання взаємодії ІС одна з одною та з людиною.*

З технічної точки зору онтології – це бази знань спеціального типу, які можна читати, розуміти і відчужувати від розробника. Саме вони можуть стати основним інструментом створення етичної компоненти ІС.

Вже досить давно існує групова робототехніка. Замість створення складного пристрою для вирішення комплексного завдання, доцільніше розробити багато простих пристроїв, які, взаємодіючи один з одним, спільно вирішать завдання (патрулювання території, розвідка, будівництво, транспортування вантажів тощо). Під час експериментів виявилося, що ці групи не такі ефективні, як очікувалося. Було

запропоновано гіпотезу: групова робототехніка дасть значні переваги та вийде на новий вищий рівень в ефективності застосування лише в тому випадку, якщо роботи зможуть утворити не просто групу, а соціум з усіма законами соціальної взаємодії: наслідуванням, агресією, домінуванням, спілкуванням тощо.

Управління таким соціумом стає актуальним завданням, оскільки його діяльність має бути спрямована на вирішення прикладних, необхідних людині завдань. У результаті дослідження з'ясувалося, що з різних способів управління соціумом роботів найефективніше управління моральними надбудовами. Для цього передбачаються додаткові правила поведінки робота, цілі їх запровадження – це визначення способів вирішення конфліктів усередині спільноти та забезпечення індивідуальної поведінки, так щоб «індивід» узгоджував свої потреби з потребами соціуму.

Мораль – це більш зручна, легка та безпечна надбудова системи управління, ніж інші можливі (наприклад, управління середовищем). Йдеться про ті моральні імперативи, які можуть визначити характер поведінки агента і соціуму загалом, змушуючи їх поводитися агресивніше або, навпаки, бути більш замкнутими, бути терпимими або нетерпимими до чужинців тощо. Все це визначає характер вирішення прикладних завдань, які стоять перед соціумом штучних агентів.

У цьому плані мораль (етика, моральність) сприймається як механізм врегулювання конфліктних ситуацій між агентами, як спосіб цілепокладання поведінки. У рамках моделювання соціальної поведінки в системах групової робототехніки реалізовано моделі наслідувальної поведінки та соціального навчання. На підставі цього робиться висновок про можливість моделювання такого механізму, як емпатія (чуйність на емоційний стан інших). Емпатія – це зручний механізм визначення мети поведінки робота або ІС. Якщо ІАС здатна визначити емоційний стан контрагента (людини або члена спільноти роботів), то вона може взаємодіяти з людиною за правилами, яка враховує її емоційний стан, і, отже, бути більш етичною. Для роботів, які безпосередньо спілкуються з людиною, ця чуйність може визначити нову якість дружнього «етичного» інтерфейсу.

Швидше за все, перевірка етичності ІАС є найскладнішим питанням. З технічної точки зору найбільш складною є етична верифікація. Вона полягає у комплексі тестів, здатних визначити ступінь етичності ІС. Для з'ясування цього ступеня придатні лише спостереження за реакціями та поведінкою досліджуваної ІАС.

З погляду гіпотетичної процедури етичної верифікації залишається лише провести цикл двоетапних перевірок:

- ✓ пред'явлення чергової неоднозначної ситуації, вирішення якої потребує залучення міркувань моральності (коли ІАС треба здійснювати екстрене гальмування чи жертвувати собою — це проблема морального вибору);
- ✓ Аналіз ланцюжка міркувань ІАС, завдяки якому експертам стане ясно, чому система ухвалила те чи інше рішення.

Другий пункт — це пояснювальна компонента — це те, що відрізняє цю процедуру верифікації від тієї, що пропонується у відомому проекті Moral Machine [27], у якому за «статистичного» підходу до проблем морального вибору ми не зможемо отримати іншого пояснення, чому ІАС зробила саме цей вибір, крім подібного: «...цей вибір здійснено через те, що так робить більшість» або «така була апроксимація при аналізі великої вибірки навчальних прикладів». Крім урахування думки «більшості», можливі інші підходи до визначення того, які етичні норми слід закладати в систему ШІ.

4.2. Алгебра совісті: формалізація етики

Наявність математичного апарату, що дозволяє формалізувати етичні поняття, також потребує уваги. Вже розроблено великий математичний інструментарій, який можна використовуватиме формалізації поняття етики в ШІ. Зокрема, особливий інтерес становлять підходи, що дозволяють оцінювати ті чи інші технології, які використовує ШІ, для відповідності певним вимогам (етичним нормам, критеріям, стандартам тощо). Важливо відзначити, що проблема формалізації етичних норм тісно пов'язана з більш загальним завданням — формалізацією гуманітарного знання.

Таблиця 2

Підходи до формалізації етики в ШІ

Механізм	Опис	Коментар
Булева алгебра	Висловлювання можуть бути лише істинними чи хибними, тобто використовується двійкова логіка	Добре розвинена, є безліч додатків, програмних бібліотек для різних інструментальних засобів тощо. Але не завжди різні етичні проблеми можна суворо розділити на «білі» та «чорні»
Багатозначна логіка	Тип формальної логіки, в якій допускається більше двох істинних значень для висловлювань	Подолання однозначності булевої алгебри. Значна складність реалізації
Нечітка логіка	Узагальнення багатозначної логіки	Подолання однозначності булевої алгебри та складнощів багатозначної логіки.

		Нестійкість щодо вихідних даних (різні методи можуть призводити до різних результатів)
Методи вербального аналізу рішень (ВАР)	Група методів ВАР спирається на досягнення різних наукових дисциплін: когнітивної психології; прикладної математики; теорії організацій тощо.	Поєднання якісної та кількісної інформації, суджень експертів, об'єктивних та суб'єктивних факторів тощо. Пояснення прийнятих рішень даються у термінах предметної сфери, тобто норм етики ШІ. Як недоліки методів ВАР відзначено великі трудові витрати експерта або особи, яка приймає рішення, при роботі в ознаковому просторі великої розмірності

Проблема формалізації етичних норм включає два основні завдання — створення форм уявлень етичних норм (критеріїв, ознак тощо.) і вибір відповідного математичного апарату до роботи з ними: зіставлення, виміру, аналізу тощо.

Не завжди відповідність тим чи іншим нормам можна звести до класичних «так» і «ні». Тому тут актуальним є розгляд та використання різних некласичних логік (наприклад, багатозначних), механізму багатокритеріальної класифікації, ймовірнісних підходів тощо (табл. 2).

4.3. Яку етику закласти у машину?

Механізми, що дозволяють закласти етичні норми в ІС, - це необхідна, але недостатня умова, ті ж норми мають бути у розпорядженні розробників етичної компоненти. У цьому випадку найважливішими, але на даний момент питаннями, що не мають відповіді, виявляються такі:

- Як ми вирішимо, що саме етично для штучної системи в тому чи іншому випадку, а що ні?
- За якими критеріями ми вибиратимемо етичні вчинки для ШІ? Чи буде ця думка більшості людей, чи то думка держави, наприклад правлячої партії, чи то думка особливих людей — моральних філософів?

Проблема вагонетки (Trolley problem) [28] – уявний експеримент в етиці, вперше сформульований в 1967 році англійським філософом Філіппою Фут. Тяжка некерована вагонетка мчить по рейках. На шляху її прямування перебувають п'ятеро людей, прив'язані до рейок божевільним філософом. На щастя, ви можете переключити стрілку, і тоді вагонетка поїде іншим, запасним шляхом. Однак, на запасному шляху знаходиться одна людина, також прив'язана до рейок. Які ваші дії?

Наразі філософська проблема вагонетки стає ілюстрацією до реальних обставин роботи безпілотного транспорту, перед розробниками якого гостро постає проблема: які етичні правила закладати у програму. Досі триває відомий експеримент Массачусетського технологічного інституту «Моральна машина» [27] з метою зібрати відповіді на запитання: якщо є ризик аварії, кого машина повинна буде задавити в тій чи іншій конкретній ситуації? (рис. 4). Система навчається на мільйонах прикладів (близько 40 млн. відповідей) і вибирає правильну, моральну дію за принципом «це рішення більшості».

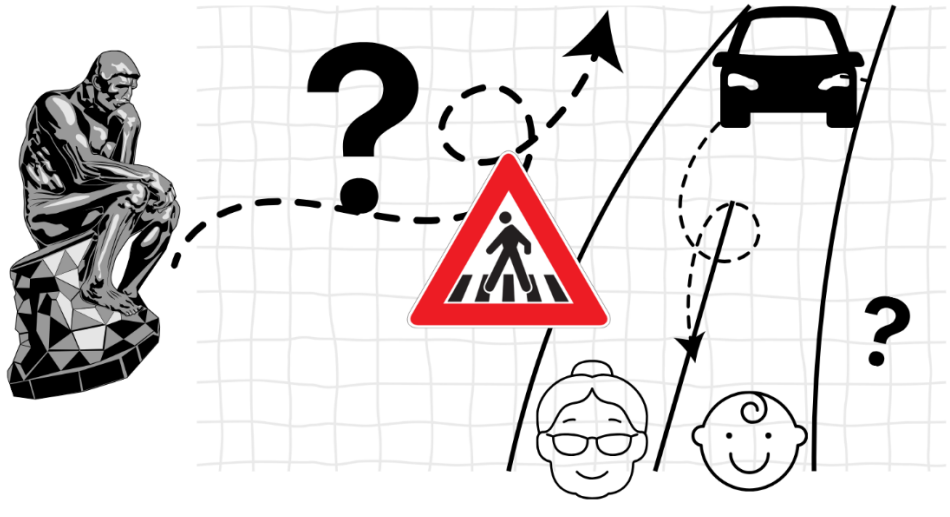


Рис. 4. Проблема вагонетки у застосуванні до безпілотного транспорту

В етиці існує два основні напрямки – утилітаризм та абсолютизм. "Так вирішила більшість, отже, так правильно", - приклад утилітарного підходу, продиктованого міркуваннями матеріальної вигоди, практичності. Абсолютистський підхід з ідеєю «життя людини священне» принципово забороняє жертвувати життям одного в ім'я порятунку навіть кількох людей.

Згідно з моральною філософією, рішення або дія не можуть вважатися етичною лише на підставі того, що так вважає більшість опитаних, незалежно від того, якого підходу вони дотримуються. У сучасному світі технології та цінності можуть змінюватися досить швидко, і в умовах невизначеності цінніше рефлексивне мислення, готовність до діалогу та облік інтересів різних сторін, ніж заздалегідь вшиті на низькому рівні принципи. Ця проблема дуже широко обговорюється в наукових колах, містить багато глибоких філософських проблем, наприклад проблему морального аутсорсингу: чи не віддаємо ми моральні рішення на волю машини, чи допомагає машина нам самим поводитися етично, і якщо так, то якою мірою це допустимо.

Крім того, вибір підходу, який закладатиметься в конкретну систему ШІ, нелегкий для розробників, він вимагає філософського підходу фахівця з етики, який здатний чітко визначити етичні рамки та правила, якими має керуватися ІС, та громадського обговорення.

Культурні особливості етичних норм ШІ. Багато систем ШІ розробляються транснаціональними компаніями, які потім використовують їх відразу в кількох країнах. Вибір конкретних етичних норм для ІС ускладнюється з силу культурних відмінностей, поглядів на моральну і аморальну поведінку. У згаданому експерименті «Моральна машина» брали участь люди з 240 країн, більшість жертвували тваринами на користь людей, злочинцями — на користь законослухняних громадян, а також прагнули врятувати більше життів. Як виявилося, індивідуальні відмінності за статтю, рівнем доходів, релігійністю та політичними поглядами не впливають на вибір, тоді як культурні відмінності завдання виявилися суттєвими для вирішення завдань. Так, у країнах з високим рівнем колективізму учасники експерименту менш схильні керуватися чисельністю врятованих життів і вважають за краще рятувати людей похилого віку ціною життя молодих.

Там, де існує значна різниця в доходах між бідними і багатими, випробувані частіше воліли рятувати життя людей, які мають високий статус в суспільстві, ціною життя людей із нижчим статусом [29].

У країнах Північної Америки та Європи частіше воліли не втручатися у дії машини та жертвувати пішоходами, у країнах Південно-Східної Азії та Близького Сходу частіше вирішували рятувати пішоходів і людей похилого віку, а в країнах Латинської Америки та Південної Африки частіше жертвували літніми на користь молодих, низькостатусними на користь людей високого соціального статусу, а також чоловіками заради жінок.

У зв'язку з культурними відмінностями виникає низка питань:

- Норми якої культури доцільно закладати у ШІ?
- Чи потрібно передбачати роботу ШІ у різних етичних рамках залежно від регіону застосування (етична локалізація)?
- Чи потрібно спочатку різним країнам домовитися про єдиний етичний кодекс (якщо взагалі принципово можна домовитися про це)?

Приклад «Моральна машина» показує, що сьогодні неможливо встановити точні етичні норми навіть у досить вузькій сфері. При цьому аналогічна ситуація виникає і в інших сферах, які, можливо, менш вивчені.

5. Стандарти етичного штучного інтелекту

Розробленням комплекту стандартів етичного ШІ займається організація IEEE. В стандарті встановлюють не жорсткі обмеження, а мінімально допустимий рівень надійності систем ШІ. Згодом вони можуть стати стандартами за замовчуванням для всіх пристроїв з ШІ, що виробляються та експлуатуються в більшості країн світу.

Етика ШІ передбачає, що з етичної та правової точок зору функціонування технічної ІАС буде забезпечене не гірше, ніж якби такою системою управляла людина. Громадські інтереси вимагають, щоб вирішення цих питань не залишалося на розсуд виробників. Необхідно створити етичну систему координат для відповідної техніки, але вкрай важливо не допустити надмірного регулювання: жорсткі етичні стандарти для всіх систем ШІ загальмують розвиток галузі, якщо не зведуть його нанівець. Підходи до створення такої системи можуть бути різними, і загальні рекомендації щодо цих підходів закріплені у низці міжнародних, державних та галузевих документів. Одна з найбільш примітних ініціатив цієї галузі пов'язана зі створенням цілого ряду стандартів, що регулюють етично обумовлене проектування інтелектуальних та автономних систем.

Стандарти розробляються великою міжнародною організацією — Інститутом інженерів у галузі електротехніки та електроніки (Institute of Electrical and Electronics Engineers, IEEE), який об'єднує не лише класичних спеціалістів-електротехніків та електронників, а й представників суміжних областей — спеціалістів з комп'ютерної техніки, програмування та ШІ.

Технічний стандарт *— це основні технічні вимоги, яким повинні відповідати всі пристрої, що розробляються та використовуються в країні, яка ратифікувала цей стандарт. Зазвичай при стандартизації прагнуть зменшити негативні ефекти технологій на людину та максимізувати позитивні.*

Історично склалося, що IEEE – одна з найбільших міжнародних організацій, яка серед багатьох своїх завдань ставить розробку технічних стандартів. Ініціатива зі стандартизації в галузі ШІ та даних з’явилась завдяки прагненню багатьох фахівців з різних галузей науки визначити позицію щодо етики при роботі з системами ШІ.

Перший настановчий документ Ethically Aligned Design описує загалом існуючі у цій сфері питання, що стосуються галузі науки і рекомендує додаткові джерела на тему [30]. У створенні документа брало участь понад 700 експертів з усього світу: не лише програмісти та фахівці з ШІ, а й філософи, культурологи, психологи, які представили погляди на проблему з різних боків.

Стандарти IEEE не обмежують і не забороняють розробку ШІ, програм, математичних методів, появу нових ідей з гуманітарних областей, але задають їм певний напрямок. У більшості областей не використовуються жорсткі заборони, за винятком атомної енергетики та деяких інших. Стандарти IEEE не суперечать тим стандартам, які

зараз обговорюють в ООН [31] або прийняли в Європейському союзі (EU AI Act) [32]. Проекти стандартів IEEE — це не спроба перекрити або звузити бурхливий «потік прогресу», а свідомі зусилля, спрямовані на те, щоб відвести цей потік від села, яке він може затопити.

Крім того, з великою ймовірністю протягом найближчих років етичність систем ШІ стане конкурентною перевагою навіть у ситуації, коли стандарти на етичні системи ще не будуть прийняті. Вже зараз окремі виробники позиціонують своє програмне забезпечення як більш «незалежне», безпечне в плані стеження за користувачами та «етичне». «Етичність», «відповідальність» та «незалежність» ПЗ та ІВ — це тренд наступних років , якщо не десятиліття.

Ухвалення стандартів охорони навколишнього середовища стимулювало колосальний технічний прогрес у сфері енергетики. У генерації енергії викопні види палива замінив газ, з'явилися потужні газові турбіни, що вже працюють на бінарних циклах з вищим коефіцієнтом корисної дії, нові системи рекуперації енергії, нові види акумуляторів і полімерних матеріалів. Почала розвиватися відновлювана енергетика.

Проєкти етичних стандартів IEEE

P7001. Transparency of Autonomous Systems (Прозорість автономних систем)

Стандарт є посібником для оцінки прозорості в процесі розробки ШІ та пропонує механізми для підвищення прозорості (наприклад, обов'язкове захищене зберігання даних датчиків та даних про внутрішній стан аналогічно реєстратору даних польоту або «чорній скриньці»)

P7002. Data Privacy Process (Забезпечення конфіденційності даних)

Стандарт орієнтований на захист приватності громадян та переважно стосується використання персональних даних громадян рекламними мережами за допомогою ІАС. Для стандартизації поки що виділено кілька груп: учасники взаємовідносин «працівник — роботодавець», діти (неповнолітні), студенти

P7003. Algorithmic Bias Considerations (Облік необ'єктивності алгоритму)

Стандарт цілком може стати одним із перших, прийнятих як базові. Він зобов'язує розробників, насамперед систем машинного навчання, як найпопулярнішого зараз напряму, відповідально підходити до навчальних даних, до їх розмітки, тестування та валідації систем

P7004. Standard for Child and Student Data Governance (Управління даними дітей та студентів)

Стандарт повинен буде регулювати роботу алгоритмів із даними дітей та учнів. Захист прав дітей та дитинства – одне з найважливіших завдань соціальної політики будь-якої

держави. З розвитком ШІ технологій збільшується ризик того, що машини в результаті спілкування один з одним прийматимуть рішення на основі вхідних даних, непрозорих для людей, за принципом «чорної скриньки».

6. Штучний інтелект у державі: етика та довіра

Довіра до систем ШІ тісно пов'язана із соціальною довірою: найшвидше ШІ впроваджується у держуправлінні країн із високою довірою до соціальних інститутів. Державне регулювання використання систем ШІ не може покладатися виключно на думку більшості, але й не повинно орієнтуватися виключно на етичні теорії без урахування думки виборців.

Державне управління засноване на збиранні та аналізі величезних обсягів даних, і в цій галузі застосування ШІ має величезний потенціал. Сьогодні використання ШІ в держуправлінні перебуває на стадії просунутих експериментів у багатьох країнах. Насамперед це відбувається за двома основними напрямами: аналітична робота з інформацією та автоматизація рутинних інтелектуальних процесів, яка може призвести до їхньої суттєвої трансформації. Використання ШІ у взаємодії з іншими лініями технологічної трансформації апарату управління може дати такі результати, як гнучка адресна допомога з боку соціальних та комунальних служб, предиктивне надання послуг у охороні здоров'я, реагування у надзвичайних ситуаціях, високотехнологічний ризик-орієнтований нагляд тощо.

Водночас сьогодні світові ЗМІ схильні обговорювати ШІ як сенсацію, підігрівають очікування радикальних змін, страхи та необґрунтовані надії. Деякі дослідники вважають, що розвиток технологій ШІ ще більше збільшить прірву, що поділяє бідних та багатих, але при цьому переоцінюють можливості таких систем. Алгоритми та великі дані, їх можливості у вирішенні соціальних проблем деколи переоцінюються. Тим не менш, готовність використовувати системи ШІ визначається скоріше не їх зрозумілістю для громадян, а ступенем довіри до їх розробників та інших користувачів.

Впровадження алгоритмів, що приймають рішення та передбачають поведінку громадян, згодом призведе до технократичного та бюрократичного управління, коли знизиться відсоток рішень, що приймаються людьми. Багато дослідників попереджають про те, що надмірна залежність від ШІ усуває обіцяний нейтралітет та об'єктивність урядових функцій, створюючи відчуття відсутності контролю у громадян та службовців держсектора.

За даними дослідження «Порівняльний аналіз окремих показників електронного уряду» (Digital Government Benchmarking), проведеного компанією BCG, найшвидшими темпами системи ШІ впроваджуються в державному управлінні в країнах з високою довірою до соціальних інститутів: в Індії, Китаї, Індонезії та ОАЕ . Невипадково використання комп'ютерного зору розпізнавання осіб перехожих на вулицях викликає запеклі публічні дискусії у європейських ЗМІ, тоді як у Китаї впроваджується обов'язкове розпізнавання осіб всім користувачів мобільного зв'язку.

Низька інституційна довіра, характерна для більшості країн, посилює техно-гуманітарний дисбаланс, коли впровадження нових технологій випереджає здатність суспільства домовлятися про правила їх використання. Технологічний оптимізм , у тому числі серед керівників державних установ, виявляється компенсацією соціального песимізму: алгоритми та системи ШІ розглядаються як ліки від корупції, орієнтовані на покращення роботи судів, медичних та освітніх установ.

Психологічні дослідження показують, що довіра до кіберфізичних систем відрізняється від довіри до людей і легко змінюється абсолютною недовірою. Обвал довіри до ШІ та заснованих на ньому робототехнічних систем може спричинити ще більшу кризу довіри до соціальних інститутів – держави, бізнесу та громадських організацій. Крихкість довіри до ІАС ставить перед керівниками державних служб завдання не лише інформувати громадян про те, які програмні рішення використовуються в цифрових державних послугах, а й про те, якими є етичні принципи держслужбовців, які безпосередньо управляють конкретними ІАС.

Існує так званий ефект ШІ: користувачі не знають про те, де використовується ШІ, причому чим частіше застосовується ця технологія, тим менш вона помітна для людини. Так, більшість користувачів не усвідомлюють те, що технології задіяні в пошукових запитах, формуванні стрічки новин і рекомендаціях у соціальних мережах.

Обмеження, що нав'язуються алгоритмами, найчастіше невідомі користувачам, навіть якщо вони описані в документації користувача. Лише третина інтернет-користувачів стверджують, що хоча б колись читали угоди про надання послуг та використання персональних даних. При цьому аналіз цифрових слідів на серверах показує, що насправді угоди читають менше ніж 1% користувачів. Більше того, алгоритми перетворюються на «архітектуру вибору», що підштовхує користувачів до рішень, які мають підвищити якість їхнього життя.

Навіть якщо в основі такого цифрового патерналізму буде захист прав людини, використання алгоритмів, що коригують недосконалість людської природи заради благих цілей, може сприяти зниженню свідомості та рефлексивності суспільства. З цієї точки зору ключові етичні дилеми, що стоять перед керівниками в умовах форсованої цифрової трансформації, зводяться до вибору між швидкістю впровадження суспільно корисних систем ШІ та їх обговоренням з різними зацікавленими сторонами, а також між використанням систем ШІ як альтернативи суб'єктивності людини та їх застосуванням для розвитку критичного мислення користувачів та підвищення усвідомленості рішень, які приймаються громадянами.

Сучасні технології все частіше дозволяють приховати використання роботів у взаємодії з користувачами. Відомо, що довіра до роботів і програм вища у тому випадку, коли надана ними послуга схожа на взаємодію з людиною [36]. Впровадження ШІ в сферу державних послуг може бути пов'язане зі спокусою не інформувати громадян про те, що з ними взаємодіє саме робот. Як показала серія експериментів, люди схильні до

кооперації з ботами, які видають себе за людей. При розумінні, що партнер є роботом, довіра знижується. Більше того, боти, які розкрили себе при взаємодії з людьми, навчаються не чекати від своїх партнерів готовності до співпраці. Ці психологічні закономірності ставлять держслужбовців перед вибором між ефективністю послуги та її прозорістю для користувачів. Особливо складним вибір стає в охороні здоров'я та за надзвичайних ситуацій, коли рекомендації бота можуть врятувати життя.

Щодо проблеми вибору критеріїв етичності того чи іншого рішення, очевидно, що державне регулювання використання систем ШІ не може покладатися виключно на думку більшості. Історії відомі численні приклади того, як сильно громадська думка залежить від психологічних механізмів відчуження моральної відповідальності, колективних страхів та конформізму. Щоб переконатися в цьому, достатньо згадати історію Голокосту, маккартизм, боротьбу за виборчі права жінок тощо. Проте регулятори ринку ШІ не мають можливості орієнтуватися виключно на етичні теорії, не враховуючи думки виборців.

Можливим способом вирішення цієї проблеми може бути створення етичної інфраструктури , що забезпечує обговорення етичних проблем розвитку ШІ з опорою на деліберативні форми демократії, звані асамблеї громадян, громадські групи та міські ради.

У 2018 році в Нью-Йорку було ухвалено закон, спрямований на запобігання дискримінації за допомогою алгоритмів, які використовуються державними службами, він став підставою для створення громадської групи експертів, які проводять аналіз правових та етичних аспектів роботи міських систем автоматизованого прийняття рішень [38]. У листопаді 2019 року в Нью-Йорку з'явилася нова посада – співробітник з питань алгоритмів (Algorithms Management and Policy Officer) в Управлінні справами мера [39]. Його головними обов'язками є боротьба з упередженостями в алгоритмах та підвищення відповідальності за прийняті ними рішення.

Вироблення норм у сфері розвитку ШІ може спиратися на пошук відповідності між етичними теоріями та думкою громадян (рис. 5): наприклад, згідно з практично всіма етичними концепціями та міжнародними опитуваннями, число врятованих життів — важливий критерій для рішень, що приймаються самокерованими автомобілями на дорозі.

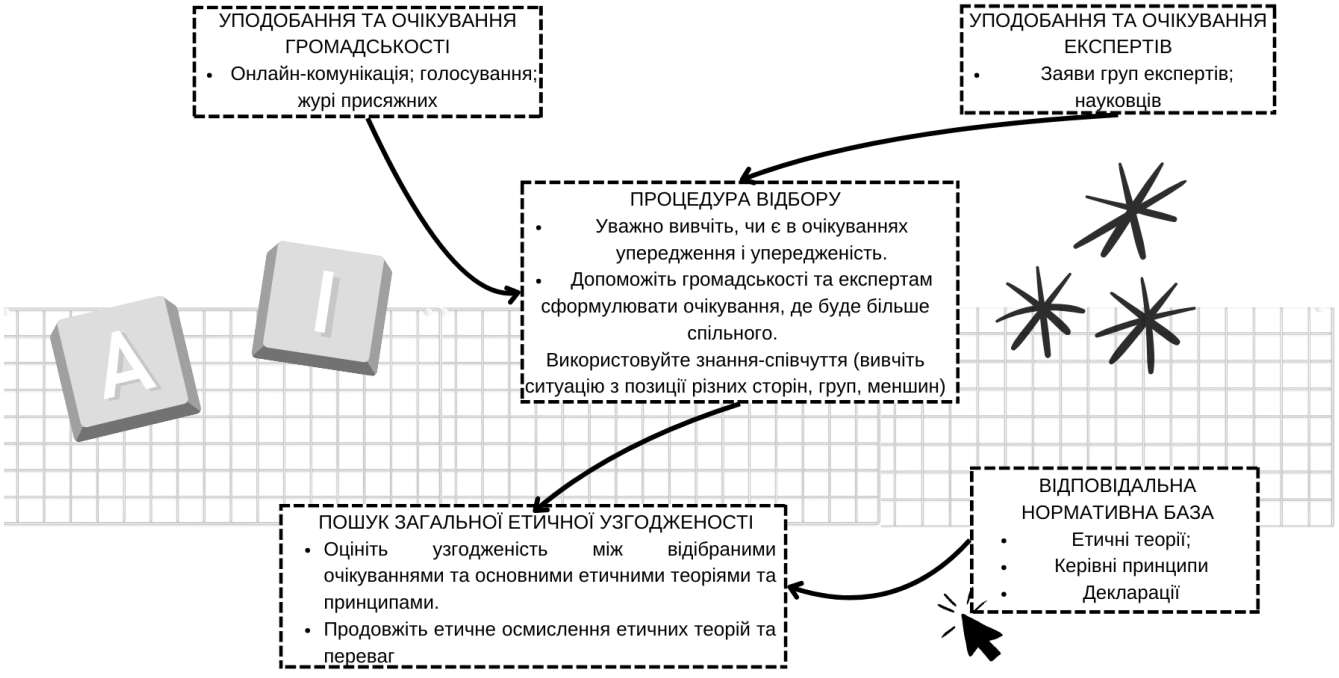


Рис. 5. Етика та суспільство як основа для регулювання ШІ

Безумовно, системи ШІ можуть підвищити увагу держслужбовців до етичних аспектів рішень, які вони ухвалюють щодня. У повсякденному житті ми постійно стикаємося з випадками порушення соціальних норм; якщо ми не бачимо можливості вплинути на ситуацію, то намагаємось скласти з себе моральну відповідальність :

- посилаємося на корпоративні правила чи високі моральні цілі;
- використовуємо евфемізми, описуючи свою поведінку («утримався» замість «промовчав»);
- перекладаємо відповідальність іншим людям («так вирішило керівництво»);
- знелюднюємо тих, хто постраждав від безвідповідальної поведінки («з такими людьми доводиться працювати») або робимо їх винуватцями того, що трапилося («ніхто не намагався ввести його в оману, це був його вибір»).

В основі захисних механізмів лежить прагнення зберегти позитивну самооцінку. Зростаюча цифрова прозорість соціального життя та підтримка прийняття рішень за допомогою систем ШІ можуть послабити переконливість деяких стратегій самовиправдання, а також зробити їх усвідомленішими для нас самих.

Найближчими роками очікується поява програм, заснованих на технологіях ШІ, які допомагатимуть подолати притаманну людям властивість відкладання справ «на потім», компенсувати властиву людям недооцінку майбутнього, візуалізувати сценарії розвитку подій.

У співтовариствах фахівців з аналізу великих даних все більшого визнання набуває перспективність створення таких систем ШІ, які не замінюють людину, а

доповнюють її. Сьогодні стає можливим шукати і залучати до розробки рішень різних експертів як всередині уряду, так і в громадянському суспільстві. Особливо важливими у цій галузі стають системи підтримки колективного інтелекту: ШІ використовується у краудсорсингових спільнотах та публічних дебатах.

Платформа Assembl за допомогою технологій ШІ підтримує обмін думками у муніципальних спільнотах метрополії Великого Парижа. За допомогою технологій ШІ, що надаються компанією Insights. US, Держдепартамент США збирає та аналізує пропозиції щодо вдосконалення процедури видачі паспортів. Платформа POPVOX на основі ШІ дозволяє підвищити ефективність взаємодії виборців із конгресменами у США при розробці нових законодавчих ініціатив. Такі інструменти можуть бути використані і для вироблення правил впровадження систем ШІ в повсякденне життя.

Очевидно, для подолання соціального песимізму належить перетворити ШІ з «протезу» для погано працюючих соціальних інститутів на посередника, що дозволяє людям краще зрозуміти одне одного, розвинути в собі співпереживання і здатність до діалогу для вироблення етично зважених рішень.

7. Оцінка впливу алгоритмів: від технологій до прав людини

В даний час оцінка впливу ШІ та алгоритмічних систем на права людини не включена до процедур оцінки в рамках регулюючого впливу проектів нормативно-правових актів (НПА) та діючих НПА. У перспективі така оцінка може стати обов'язковим етапом розробки ШІ та алгоритмів.

Для визначення етичних та правових наслідків впровадження технологій ШІ формування адекватного правового регулювання необхідно застосовувати методи оцінки впливу, яка є ключовим інструментом «розумного» регулювання (smart regulation) у всьому світі. Найбільш поширеною є оцінка регулюючого впливу (regulatory impact assessment), проте її методологія недостатньо враховує особливості розвитку технологій і пов'язаних з цим викликів.

У сфері розвитку та впровадження нових технологій, у тому числі цифрових, все більшого розвитку набувають спеціалізовані види оцінки, які інтегруються в державне управління та нормотворчий процес.

7.1. Технологічна оцінка ШІ

Процес формування підходів до оцінки впливу технологій на розвиток суспільних відносин, економіки, державного управління та, навпаки, до оцінки впливу регулювання на розвиток технологій активно йде з 1970-х років у межах інституту технологічної оцінки (technology assessment). Її здійснюють спеціальні підрозділи апаратів парламентів або окремі дослідницькі організації, які об'єднані в Європейську мережу парламентської оцінки технологій (European Parliamentary Technology Assessment , EPTA Network). У 2017—2019 роках зазначені підрозділи та організації опублікували такі доповіді на тему ШІ:

Доповідь Інституту Ратенау (RATH, Нідерланди) «Права людини в епоху роботів: проблеми, пов'язані з використанням робототехніки, штучного інтелекту, віртуальної та доповненої реальності » (жовтень 2017 року) [39].

Доповідь містить низку рекомендацій Раді Європи, спрямованих на захист персональних даних, повагу до сімейного життя, гідність особистості, свободу вираження поглядів, та орієнтує на розробку окремої Конвенції про захист прав людини в епоху роботів, а також етичних кодексів та створення комітетів з етики цифрових технологій. та ШІ.

Доповідь Офісу інформації науки та техніки Конгресу (Oficina de Información Científica y Tecnológica para el Congreso de la Unión, Мексика) « Штучний інтелект » (березень 2018 року) [40].

У доповіді особливо наголошено на викликах для системи зайнятості, при цьому підкреслюються перспективи економічного зростання та створення робочих місць, що вимагають високої кваліфікації працівників.

Доповідь Центру науки, технологій та інжинірингу Рахункової Палати (US Government Accountability Office, США) « Штучний інтелект: нові можливості, проблеми та наслідки » (травень 2018 року) [41].

У звіті наголошено на необхідності розробки та прийняття відповідних етичних норм використання ШІ.

Доповідь Офісу з оцінки науки та технологій Парламенту (Розпізнання осіб) (липень 2019 року) [42].

Доповідь орієнтує органи державної влади на розробку законодавчого регулювання, яке забезпечить повагу до основних свобод, суверенітет країни та розвиток етичного ШІ.

7.2. Оцінка впливу на права людини

Оцінку впливу на права людини (Human rights impact assessment) як самостійний інститут вперше відзначено у Керівних засадах підприємницької діяльності в аспекті прав людини ООН, які були схвалені Радою з прав людини у червні 2011 року. Дані Керівні принципи рекомендують бізнесу впровадити оцінку впливу на права людини на всі відповідні внутрішні бізнес-функції та процеси.

Оцінку впливу на права людини слід виконувати
- ✓ на початку реалізації нового виду діяльності;
- ✓ до здійснення серйозних змін у діяльності (наприклад, до виходу ринку, початку збуту продукції, зміни стратегії тощо. буд.);
- ✓ у відповідь на зміну умов;
- ✓ періодично.

У межах Ради Європи оцінка на права людини отримала навіть більшого розвитку, ніж у ООН. Рекомендація CM/Rec (2016) Комітету Міністрів Ради Європи з прав людини та бізнесу наказує проводити цю оцінку не лише самим компаніям, а й державам — членам Ради Європи при здійсненні законодавчого регулювання та інших заходів [43].

У січні 2017 року Консультативний комітет Конвенції Ради Європи із захисту прав фізичних осіб при автоматизованій обробці персональних даних (T-PD) ухвалив Керівні принципи щодо захисту фізичних осіб щодо обробки персональних даних у світі великих

даних . Пріоритетом оголошено етичне використання даних, яке не повинно суперечити етичним цінностям відповідної спільноти, включаючи захист прав людини. Як особливий захід для впровадження оцінки впливу на права людини Керівні принципи передбачають створення спеціальних комітетів з етики всіма операторами персональних даних. У січні 2019 року той же Комітет ухвалив Керівні принципи штучного інтелекту та захисту даних. Розробникам, виробникам та постачальникам послуг ШІ наказано проводити оцінку впливу на права людини.

У травні 2019 року комісар Ради Європи з прав людини опублікував рекомендацію « Штучний інтелект: 10 кроків для захисту прав людини ». Держави-члени Ради Європи повинні створити правову базу для того, щоб вона встановлювала процедури проведення державними органами оцінки впливу на права людини (Human Rights Impact Assessments) систем ШІ, які придбані, розроблені та/або розгорнуті цими органами. Процедури оцінки впливу на права людини повинні бути впроваджені та введені в дію аналогічно до інших форм оцінки впливу, що проводяться державними органами.

7.3. Оцінка впливу алгоритмічних систем: методика та підходи до регулювання

Особлива увага до впливу цифрових технологій, у тому числі ШІ, на права людини призвела до ідеї сформувати спеціальну оцінку впливу алгоритмічних систем (Algorithmic impact assessment) [44]. Алгоритмічна оцінка характеризує ризики для прав людини, етичних та соціальних наслідків дії алгоритмічних систем. При цьому існують різні моделі такої оцінки.

Зокрема, GDPR виконує оцінку впливу на те, як захищаються дані (Data Protection Impact Assessment), тоді як Рада Європи просуває концепцію оцінки впливу на права людини загалом. Відмінності моделей алгоритмічної оцінки також пов'язані з визначенням суб'єкта, що її проводить.

У межах оцінки на захист даних алгоритмічну оцінку проводить сам оператор (контролер) даних, у межах оцінки на права людини — зовнішня третя сторона чи окремий орган з акредитації. Відкритим поки що залишається питання про обов'язковість чи добровільність такої оцінки. Методологія алгоритмічної оцінки повинна забезпечувати реальний захист прав людини і бути зручною для фірм, що її проводять, та інших організацій, щоб не стати додатковим бюрократичним тягарем.

В даний час Комітет Міністрів Ради Європи розробляє Рекомендації щодо впливу на права людини алгоритмічних систем . Цей проект передбачає, що оцінку впливу на права людини проводять як органи державної влади, так і бізнес. Особливістю цього документа є виділення алгоритмічних систем із високими ризиками для прав людини. Оцінка впливу таких систем має включати оцінку можливих трансформацій існуючих

соціальних, інституційних чи управлінських структур та чіткі рекомендації, як запобігти або пом'якшити високі ризики для прав людини.

8. Регулювання штучного інтелекту у світовій практиці

8.1. Національні документи стратегічного розвитку

За різними оцінками сьогодні національні стратегії в тому чи іншому вигляді є більш ніж у тридцяти країнах, включаючи Китай, Корею, Канаду, США, Велику Британію, Францію, Україну [45-48]. У документах стратегічного розвитку, як правило, міститься опис підходів до розвитку технологій ШІ, зокрема:

- існуючий рівень розвитку технологій ШІ у світі, ключові галузі їх впровадження;
- очікування щодо розвитку технологій у короткотерміновій, середньотерміновій та довготерміновій перспективах;
- ключові етапи, завдання та цілі розвитку технологій ШІ в конкретній країні;
- основні проблеми та складності розвитку технологій ШІ;
- план основних заходів, вкладених у розвиток технологій загалом;
- плани фінансової підтримки галузі;
- основні етичні проблеми та питання;
- цільовий стан розвитку технологій.

До таких національних документів належать :

- ✓ Національна стратегія розвитку ШІ France IA (Франція, 2018);
- ✓ Загальноканадська стратегія штучного інтелекту (Канада, 2018);
- ✓ План розвитку технологій штучного інтелекту нового покоління (Китай, 2017);
- ✓ Національна стратегія зі штучного інтелекту (Данія, 2019).

8.2. Закони та підзаконні акти

Закон про ШІ (EU AI Act), який був офіційно прийнятий парламентом під час його пленарної сесії 13 березня 2024 року (було оголошено на сесії у квітні 2024 року), набув чинності в серпні 2024 року і став першим у світі законом, що регулює використання ШІ. Він був розроблений на основі стратегії ЄС щодо розвитку штучного інтелекту і декларує людино-орієнтований підхід, зосереджений на повазі європейських цінностей і прав людини. Він містить рамки для оцінки ризику будь-якого продукту, послуги або системи ШІ [32].

В інших країнах світу теж відбувається розроблення подібних нормативно-правових документів, зокрема ще у 2008 році в Південній Кореї було прийнято акт , формально присвячений робототехніці, а саме специфічного різновиду — роботам, оснащеним ШІ.

У лютому 2017 року Парламент ЄС прийняв резолюцію 2015/2103 (INL) Civil Law Rules on Robotics. Документ стосується насамперед робототехніки, але за змістом та логікою очевидно, що маються на увазі і технології ШІ. У резолюції намічено підходи до регулювання відповідальності за шкоду, запропоновано створити європейську систему реєстрації «розумних» роботів. Особливу популярність резолюція набула завдяки ідеї наділити роботів з ШІ статусом електронних осіб. В інших країнах є акти, присвячені конкретним різновидам систем ШІ, що застосовуються в автоматизованих автомобілях, охороні здоров'я, при реалізації концептів «розумного» міста, у фінансовій сфері. Перші приклади регулювання безпосередньо технологій ШІ відносяться до медицини та державного управління. Також слід звернути увагу на інші акти:

- Зміни до Закону про дорожній рух Німеччини для використання високоавтоматизованих автомобілів (2017);
- Посібник з випробувань автоматизованих транспортних засобів (Австралія, 2017);
- Пробне зведення правил для випробування автономних транспортних засобів на території Китаю (2018);
- Резолюція щодо заборони застосування автономних смертельних систем озброєння (Бельгія, 2018);
- Директива про автоматизоване ухвалення рішень для федеральних установ (Канада, 2019).

8.3. Дослідження етики ШІ

Найчастіше розробці національних стратегій передує велика дослідницька робота, яку виконують різні експертні групи чи наукові установи. Підходи до регулювання ШІ розглядаються у Рекомендаціях з безпілотних автомобілів , розроблених Комісією з етики при Міністерстві транспорту та цифрової інфраструктури Німеччини (2017). У Рекомендаціях запропоновано принципи застосування безпілотних технологій у галузі транспорту. На наднаціональному рівні значущі доповіді експертів Всесвітньої комісії з етики наукових знань та технологій ЮНЕСКО (доповіді про етику робототехніки (2017) , про етику ШІ (2019)), а також такі приклади, як Біла книга зі стандартизації ШІ (Китай, 2018) ; доповідь Палати лордів « Алгоритми у громадському бізнесі та прийнятті рішень» та Зведений звіт Комітету зі штучного інтелекту Палати лордів (Великобританія, 2018).

8.4. Етичні документи в галузі ШІ

Довгий час розробка етичних документів за правилами застосування ШІ була ключовим трендом їх регулювання. До кінця 2019 року по всьому світу прийнято не менше сотні різних актів, посібників, принципів та кодексів, присвячених етиці ШІ. У більшості з них наведено кілька ключових принципів : безпека, дотримання конфіденційності , недискримінація, контрольованість тощо. Серед найбільш відомих з

них можна згадати Монреальську декларацію про відповідальний розвиток штучного інтелекту (2017) [50] та Посібник з етики для надійного ШІ Спеціальної групи експертів високого рівня (HLEG) Ради Європи (2018) [51].

8.5. Стандарти та доктринальні джерела

Міжнародною Організацією зі стандартизації (ISO) створено спеціальний технічний комітет із штучного інтелекту (SC № 42), для розроблення стандартів, присвячених ШІ та великим даним ISO/IEC, що регулюють роботу з великими даними , а також стандарти ISO по роботі з ШІ , включаючи обмеження упередженості ШІ тощо.

До інших прикладів стандартизації належать:

- Глобальна ініціатива етики автономних та інтелектуальних систем (IEEE, 2016) ;
- Рекомендовані практики управління якістю наборів даних для медичного штучного інтелекту (IEEE);
- Проект плану федеральної участі у розробці технічних стандартів ШІ та пов'язаних з ними інструментів (NIST, США, 2019);
- Доповідь « Розробка стандартів для штучного інтелекту: Чути голос Австралії » (Австралія, 2019).

Також вже кілька десятків років досить послідовно формується «робоправо», або «право роботів», як самостійна предметна сфера дослідження. Розглядаються насамперед проблеми відповідальності, правосуб'єктності, забезпечення алгоритмічної прозорості, контрольованості систем ШІ, проблеми авторського та патентного права та багато інших.

8.6. Міжнародні акти з етики ШІ

Проводячи дослідження у сфері регулювання ШІ , автори часто дійшли висновків необхідність вироблення міжнародних правил взаємодії людей у зв'язку з розвитком систем ШІ, зокрема етики застосування ШІ. Ця ідея отримала значну підтримку на різних рівнях, зокрема Доповідь з етики ШІ (2019) Всесвітньої комісії з етики наукових знань та технологій ЮНЕСКО присвячена рекомендаціям щодо структури та змісту такого можливого міжнародного документа. ЮНЕСКО може доповнити численні керівні принципи та вказівки з етики, які в даний час розробляються державними органами, компаніями та громадськими організаціями, міждисциплінарним, універсальним та цілісним підходом до розвитку ШІ на благо людства та в ім'я миру та сталого розвитку.

Сьогодні лідером у визначенні підходів до регулювання ШІ є Рада Європи. Зусиллями різних експертних комітетів цієї організації розроблено [52]:

- Європейська етична хартія Ради Європи з використання ШІ в судових системах
- Посібник із захисту даних під час використання ШІ
- Декларація Комітету Міністрів про маніпулятивні можливості алгоритмів

- рекомендації Комісара СЄ з прав людини « 10 кроків для захисту прав людини під час використання ШІ » тощо.
- Для створення повноцінного нормативного документа у сфері ШІ у 2019 році Радою Європи було сформовано Спеціальний комітет з ШІ . До інших документів у цій сфері належать:
- Принципи І та рекомендації щодо національної політики Експертної групи зі штучного інтелекту (OECP, 2019);
- Проект рекомендацій Ради міністрів щодо впливу алгоритмів на права людини (Рада Європи, 2019).

Всі ці напрацювання та проведені дослідження, суспільні та експертні обговорення необхідності регулювання систем ШІ сприяли тому, що після довгих обговорень в Європейському Союзі було прийнято AI Act, який узагальнив всі актуальні практики та підходи до розроблення та використання систем надійного та етичного ШІ.

Use Case 1. Етика цифрових технологій освіти

Розвиток цифрових технологій докорінним чином змінює весь освітній процес. Особливо відчутно це стало під час пандемії COVID-19, коли через вимушений перехід більшості людства до дистанційної освіти освоювати новітні технології довелось практично всім: учням, вчителям, батькам. Це, своєю чергою, призвело до надзвичайно стрімкого розвитку цифрових освітніх технологій: доповнена та віртуальна реальність, системи штучного інтелекту допомагали зробити освітні системи ще більш ефективними та захопливими. Впровадження цифрових технологій не лише трансформує процес навчання, але й формує нові вимоги до результатів освіти, до компетенцій та навичок, необхідних для життя в новому, цифровому, світі. Ці зміни супроводжуються специфічними етичними проблемами.

В етиці освіти виділяються три основні напрямки :

- **етика як предмет, що викладають;**
- **етика як принципи, що закладені в основу системи освіти;**
- **професійна етика викладачів.**

1. Чи потрібно навчати розробників етиці?

Етика як обов'язковий предмет вивчається, зазвичай, на філософських факультетах. Студентам деяких спеціальностей викладають професійну етику, наприклад, деонтологію в медичних вузах або етику роботи психолога. Для інших спеціальностей етика залишається розділом філософії і не має прикладного значення. Таким чином, у традиційній системі професійного навчання майбутні інженери, фахівці з цифрових технологій, державні службовці не отримують професійної підготовки у сфері етики [53].

Чи має це залишатися незмінним у епоху, коли майже у всі сфери діяльності впроваджуються цифрові технології, на які впливають традиційні етичні норми (наприклад, відеоспостереження, збір персональних даних, скоринг тощо.)? Ситуація з дотриманням етичних норм у цифрову епоху стає критичною і викликає реакцію у відповідь: з'являються етичні кодекси та стандарти, етику конкретних професій починають регулювати на законодавчому рівні. Люди поступово усвідомлюють, що етичні питання слід вирішувати не після того, як технологія вже розроблена, а ще на етапі проектування цифрових рішень (X-by-Desigh).

Швидше за все, прикладну етику технологій слід вводити до освіти як частину навчальної програми. Можливі два варіанти позитивного розвитку ситуації:

- Концепція професійної етики розробляється на державному рівні, закріплюється в нормативних актах та закладається в освітні програми у вигляді стандарту, якого треба обов'язково дотримуватися.
- Навчальні програми містять загальні знання та різні трактування цифрової етики, студентам слід мати уявлення про неї, але дотримуватись чи ні конкретних етичних принципів у професійній діяльності — їх особистий вибір.

Не виключений і третій, негативний варіант — етика так і залишиться частиною філософії та не сприйматиметься як необхідна професійна компетенція.

Важливо, щоб етична основа цифрових рішень була зрозуміла не лише проектувальникам, а й громадянам, які користуватимуться цими рішеннями. Основи етики цифрових технологій бажано викладати вже у школі. Тоді школяр розумітиме як ризики, можливі у цифровому середовищі, і свої права щодо технологій (наприклад, права на приватність та захист своїх персональних даних), так і етичні норми різних професій. Поки що рано говорити про сформовані підходи до викладання етики, але ці питання потрібно обговорювати на різних рівнях [54].

2. Етичні проблеми цифрової освіти

Коли говорять про етику освіти, зокрема вищої, як правило, мають на увазі:

- базові етичні засади, що лежать в основі системи освіти;
- практичні етичні принципи, сформульовані у будь-яких документах, наприклад, у етичному кодексі вузу;
- професійну етику педагога.

Усі названі типи етики розвиваються давно і незалежно від цифрового прогресу. Проте впровадження у навчання цифрових технологій, їх використання учасниками навчального процесу за стінами вишу породжують нові проблеми.

Етичні кодекси є в багатьох університетах і регулюють норми поведінки студентів та викладачів, їх взаємовідносини під час навчального процесу та за його межами. "Об'єктом" застосування етики є такі проблеми, як хабарі викладачам, гендерна нерівність, фаворитизм, плагіат і т.д. До етичних засад відноситься, наприклад, доступність освіти, етичні цінності, які безпосередньо транслюються в навчальному процесі.

Однак, впровадження цифрових технологій в освіту позначається на традиційних етичних проблемах освіти та привносить нові, специфічні, пов'язані з дистанційними технологіями, з обробкою персональних даних учнів, з використанням відомостей про успішність тощо.

Нові етичні проблеми породжуються різними новими трендами освіти. Сьогодні найбільш значущими є персоналізація та адаптивний підхід, дистанційні технології, предиктивна аналітика (рис. 6).

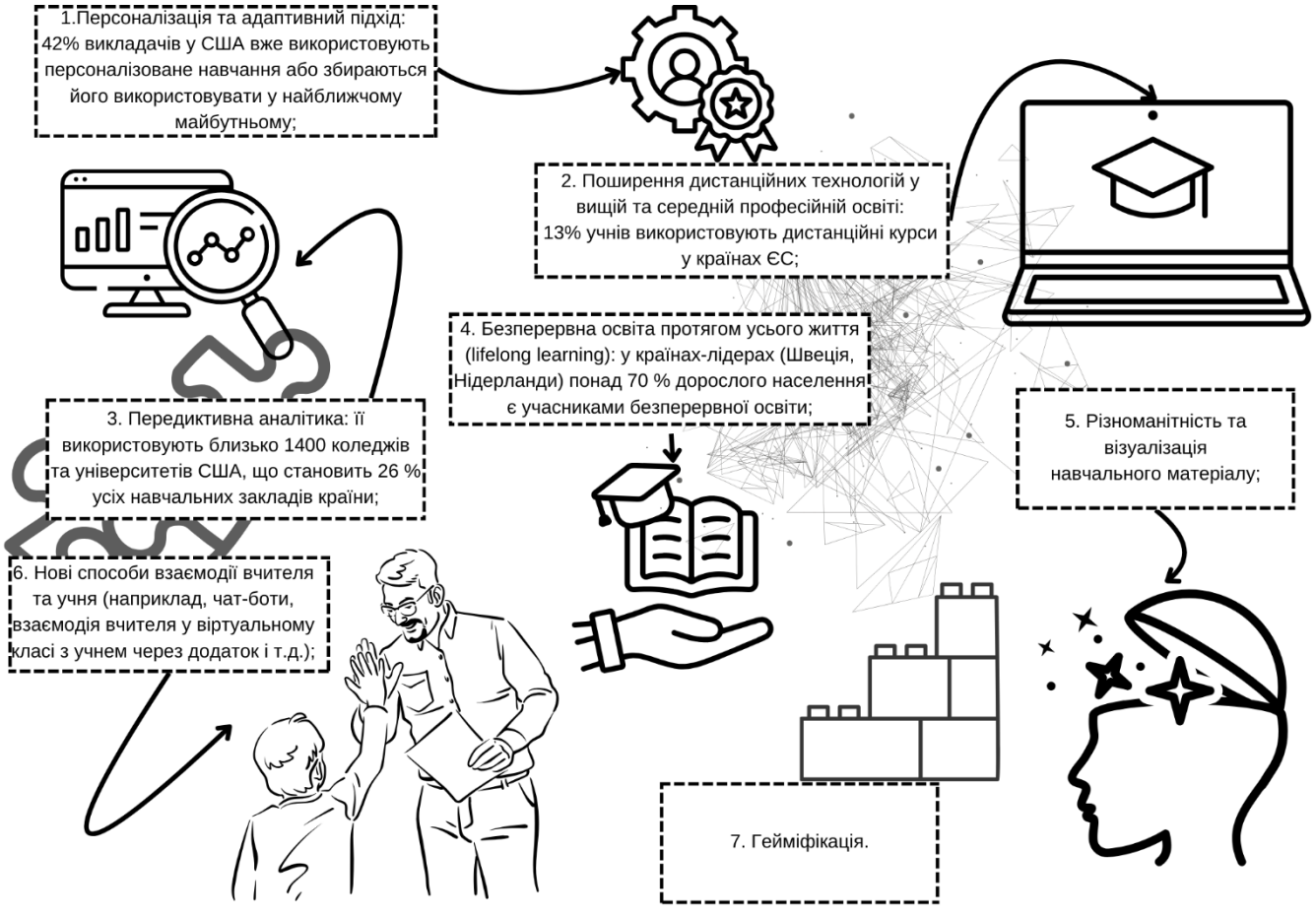


Рис. 6. Глобальні тренди розвитку цифрових технологій в освіті

Поруч із очевидними перевагами ці тренди можуть бути джерелами певних етичних проблем. Розглянемо лише ті з них, реалізація яких пов’язана із використанням технологій ШІ.

Персоналізація та адаптивний підхід при використанні цифрових технологій дозволяють проаналізувати психологічні особливості учня та запропонувати той навчальний матеріал і в тому форматі, який оптимально підходить йому. Це не тільки в рази підвищує якість освіти, а й знижує ризик дискримінації, що існує в сучасній системі очної освіти, коли школярі та студенти зі здібностями, які не досягають або суттєво перевищують середній рівень, опиняються у гіршому становищі порівняно з рештою. Адаптивний підхід має стати наступним кроком у розвитку як дистанційної, так і очної освіти.

Перехід на індивідуальні освітні траєкторії – це поки що завдання майбутнього. Зараз найефективніше змішане навчання, комбінація офлайн- та онлайн-елементів. Наприклад, підхід «перевернутий клас» передбачає, що учень самостійно опрацьовує

матеріал перед уроком: читає, дивиться відео, проходить тести, а на урок приходить для того, щоб вчитель пояснив йому те, що залишилося незрозуміло. Викладач постає як унікальний носій знань, який забезпечує індивідуальний підхід до кожного.

Ще однією етичною проблемою є збір та обробка даних учнів, щоб створити індивідуальний профіль навчання. Проблема не є унікальною для сфери освіти, а має місце у всіх галузях, де використовуються цифрові технології. У міжнародному праві та в практиці дані дітей і студентів традиційно підлягають особливому захисту, оскільки вони розглядаються як уразлива група (аж до того, що для їхнього захисту створюють окремі етичні стандарти). Саме з такими даними зазвичай працюють навчальні заклади. Щоб пропонувати найбільш ефективні рішення щодо адаптації освіти та створення персоналізованого курсу навчання, потрібно збирати дані про учнів та певним чином їх використовувати.

У середній школі № 11 міста Ханчжоу (Китай) встановлено систему «Розумне» око для розпізнавання осіб учнів. Три камери над дошкою спостерігають за класом. Комп'ютер розрізняє на обличчі сім різних емоційних виразів, наприклад, щасливе, сумне, розчароване, перелякане. Якщо вираз обличчя учня змінилося, система може оцінити це як ослаблення уваги та негайно надсилає повідомлення вчителю. Ступінь концентрації учня впливає на його рейтинг.

На чаші ваг тут, з одного боку, повнота відомостей про школяра чи студента, що дозволяє забезпечити зміст і формат навчання, що максимально підходять саме йому, а з іншого — накопичення чутливих даних про учня в навчальному закладі. Йдеться про когнітивні здібності, ставлення до навчання, соціальні зв'язки, результати аналізу його успішності, аж до прогнозів про його можливу кар'єру. Батьки учня та/або його законні представники мають право не ділитися цими даними зі школою чи вузом, але в цьому випадку персоналізоване навчання неможливе.

Гіпотетично можна вдатися до наступного варіанту: зберігається неадаптований усереднений варіант програми, і учні, які не готові надати всі дані про себе, навчаються по ньому, а ті, хто, навпаки, зацікавлений у максимально ефективному навчанні, погоджуються на збирання та обробку всіх даних. У будь-якому випадку важлива можливість відмовитися від збору даних та поінформованість про всі аспекти збору даних, включаючи наслідки такої відмови.

Варто зазначити, що за вимогами EU AI Act застосування такого типу системи для користувачів з ЄС буде заборонене, оскільки вимагає біометричної ідентифікації людини та обробки її обличчя в режимі реального часу [32].

Предиктивна аналітика освіти пов'язана як із створенням персоналізованих освітніх траєкторій, і з рекомендаціями того, яка спеціальність і кар'єра більше підійде тій чи іншій людині. Довгострокові прогнози, як і будь-які рекомендації щодо вибору

професії, пов'язані з певними етичними проблемами. З одного боку, складнощі виникають через те, що між навчанням та перевіркою його ефективності минає багато часу. Наприклад, у молодшого школяра можуть виявити наявність чи нестачу якихось здібностей та вибрати для нього відповідну освітню траєкторію. З роками ці здібності можуть розвинутися або згаснути, та й вимоги до професії зміняться, але дитина вже отримала більш «вузьку» або профільну освіту, ніж могла б в іншій ситуації.

З іншого боку, тут виникає відома проблема "nudging" , коли алгоритми та цифрові технології підштовхують користувача до певної поведінки, обмежуючи його свободу волі. Щодо навчання це посилюється тим, що прогноз може вплинути на самооцінку та мотивацію школяра чи студента, погіршити розвиток вольових якостей: учень не сам вирішує, що він хоче і що він вивчатиме, а змушений дотримуватися вказівок аналітичної системи.

Також існуючим сервісом є оцінка цифрового сліду студента — його успішності, поведінки та участі у громадському житті. Система ШІ пропонує керівництву вищого навчального закладу рекомендації: відрахувати студента з низькою успішністю або направити на додаткові курси.

Крім того, якщо з метою предиктивної аналітики збираються дані про всі досягнення учня, ці дані можуть потрапити до роботодавця, який використовує їх при пошуку співробітника на вакансію.

Етична проблема несанкціонованого використання ПД у сфері професійної освіти може набувати зловісного відтінку «торгівлі» фахівцями. Наприклад, деякі власники освітніх курсів з масовою інтерактивною участю, із застосуванням технологій електронного навчання та відкритим доступом через інтернет (англ. massive open online course) готові продавати відомості про осіб, зарахованих на курс, і розглядають цю практику як спосіб отримання додаткового доходу.

Предиктивна аналітика має схильність навішувати ярлики. Якщо людина погано навчалася в школі, це не означає, що вона недостойна здобути вищу освіту або хорошу роботу. Згодом людина може змінитися, і цифровий слід, що тягнеться з дитинства, — це сумнівна цінність. Тестуючи здібності людини, система освіти має допомагати удосконалювати їх, вирівнювати баланс, давати ще більше можливостей, а не обмежувати розвиток [55].

Тому, пригадуючи розглянуті нами раніше етичні проблеми використання ШІ, варто наголосити на тому, що:

- ✓ Для персоналізованої освітньої системи з адаптивним підходом особливо важливо приділяти увагу якості навчальної вибірки для того, щоб уникнути упередженості;

- ✓ Важливо постійно донавчати систему ШІ для того, щоб її рекомендації відповідали сучасному стану розвитку науки та потребам професійного середовища. Адже якщо учні чи студенти використовуватимуть застарілу версію системи, ми можемо зіштовхнутись із ситуацією, яка була яскраво описана у фантастичному оповіданні Рея Бредбері «Професія», коли діти, що отримували знання за допомогою новіших версій навчальних чипів, мали суттєві переваги на ринку праці;
- ✓ Критично важливою є прозорість алгоритму рекомендацій та етичність розробників такої системи. Адже існуватиме імовірність, що рекомендаційний алгоритм рекомендуватиме набувати тих знань та навичок, які сьогодні потрібні, наприклад, бізнесу чи державі, а не тих, які були б цікаві студенту.

Крім зазначених трендів, в сучасній освіті є загальні етичні проблеми, характерні і для інших областей і відрізняються гостротою, оскільки йдеться про дітей та юнацтво. Це, насамперед, проблеми, пов'язані зі збором даних, втратою приватності, відеоспостереженням тощо. Вже накопичено приклади неетичного збору даних у школах та вишах. Висунуто низку ініціатив, пов'язаних зі збором даних, вони перебувають на стадії тестування і також можуть викликати згодом питання етики, як, наприклад, перспективи використання системи розпізнавання осіб під час НМТ чи ЗНО.

Таким чином, цифрова економіка ставить перед освітою складні етичні проблеми, пов'язані з використанням цифрових технологій у процесі навчання, а також їх впливом на розвиток здібностей і кар'єру людини. Більшість цих проблем поки що не має однозначного рішення. Очевидно, з одного боку, що впровадження проривних технологій в інерційну за своєю природою систему освіти має проводитися з великою обережністю, а з іншого — ці зміни є необхідними і неминучими [56].

Use Case 2. Етика цифрових технологій у поліції

Порівняно з медициною та сферою послуг застосування цифрових технологій у судовій та правоохоронній діяльності менш масштабне, але проблема етики не менш актуальна. Використання сучасних технологій правоохоронними структурами і дає нові можливості щодо забезпечення безпеки громадян, та таїть у собі загрозу правам і свободам людини.

Впровадження цифрових технологій у поліцейську діяльність багато в чому зумовлене необхідністю протистояти злочинному світу, який дедалі краще озброюється. Кримінальний світ активно використовує останні досягнення четвертої промислової революції: технології блокчейну, дрони, ШІ тощо, а наслідки злочинів стають все більш масштабними та тяжкими, що необхідно враховувати у роботі поліції. До сучасних поліцейських структур висувається нова вимога — бути відкритими та прозорими для громадян. Застосування цифрових технологій дозволяє поліції ефективно виконувати свої прямі функції, а також підвищити довіру громадян і знизити ймовірність корупції [57].

Щоб ефективно боротися і з традиційною злочинністю, і з кіберзлочинністю, правоохоронним органам не можна відставати, а краще йти на крок попереду правопорушників у використанні технічних досягнень. Перевагою поліції буде насамперед збирання та обробка величезного обсягу даних, багато з яких накопичуються вже зараз, але не обробляються і не використовуються. При цьому на перший план виходить предиктивна поліцейська діяльність (predictive policing).

Предиктивна поліцейська діяльність *- стратегія передбачення та запобігання ризику злочинів, що базується на ШІ та аналізі великих даних.*

Основні етичні проблеми предиктивної поліцейської діяльності пов'язані не так з інформаційними, як із соціальними технологіями, зокрема з методами збору даних та прийняттям рішень про долю громадян, для яких ШІ передбачив високу ймовірність скоєння злочину.

1. Цифрові технології у роботі поліції

У світі поліцейські служби та судові органи активно розробляють та використовують окремі елементи цифровізації та цілі системи ШІ. За даними Інтерполу та Європолу, більш ніж у 70 країнах поліцейські на практиці використовують ті чи інші види предиктивної аналітики.

Крім вузькоспрямованих систем ШІ для поліцейської роботи створюються універсальні комплексні системи, які допомагають при розслідуванні злочинів та запобіганні ним.

На замовлення Євросоюзу міжнародна команда вчених розробила і в 2013 році запустила систему ePOOLICE (early Pursuit against Organized crime using environmental scanning, the Law and IntelligenCE systems). Система сканує сторінки сайтів, електронну листування, поліцейську інформацію з метою знайти ознаки діяльності організованої злочинності та оцінює ризик скоєння злочину. Для аналізу використовується відео, текстовий контент, фінансові дані, інформація із соціальних мереж, чатів тощо

2. Етичні проблеми цифрової поліції

Етичні проблеми використання цифрових технологій у поліцейській діяльності обумовлені як недосконалістю самих технологій, так і людським фактором. Як і в будь-якій іншій області, однією з ключових проблем є упередженість алгоритмів .

У соціології існує думка , що всі бази даних, відомості для яких зібрані в процесі конкретних дій поліцейських, наприклад оглядів на вулицях, містять упередження. Крім того, кримінальна статистика не відображає реальний рівень злочинів, а лише вказує, про яку кількість злочинів стало відомо державі, і представляє лише низку соціальних характеристик конкретної спільноти (стратифікацію, інтенсивність та близькість взаємодій тощо). Опора на зібрані таким чином дані може призводити до неправильних прогнозів, зловживань щодо меншин та груп з низьким соціальним статусом. Особливо небезпечно те, що такі прогнози будуть легітимізовані, оскільки вважається, що технології об'єктивні, точні і не схильні до впливу людського фактора.

У 2014 році спеціалісти Массачусетського технологічного інституту (США) розробили програму COMPAS з елементами ІІ. Програма призначалася для того, щоб допомогти суддям ухвалити рішення про ув'язнення або звільнення підсудного під заставу. Програма успішно працювала та отримувала позитивні відгуки доти, доки не було помічено, що система апріорі зменшує шанси на звільнення для латиноамериканців, які перебувають у країні нелегально, та афроамериканців із низьким доходом. Проведений глибокий аналіз системи підтвердив повну статистичну обґрунтованість її передбачень. Справді, латиноамериканці та афроамериканці частіше порушували правила визволення під заставу. Але американське суспільство не могло погодитись з таким висновком. В результаті експлуатацію системи було припинено.

Використання цифрових технологій у поліцейській діяльності виявляє існування стигматизації.

Стигматизація *- навішування соціальних ярликів, ув'язування будь-якої якості (як правило, негативного) з окремою людиною або безліччю людей, хоча цей зв'язок відсутній або не доведений. Стигматизація є складовою багатьох стереотипів.*

Використання систем персонального скорингу, алгоритмів ШІ для передбачення скоєння злочину окремою особою, застосування систем корпоративної безпеки для відстеження нетипового та «небезпечного» співробітника можуть призвести до того, що його заздалегідь стигматизуватимуть. Це стане додатковим фактором ризику і збільшить ймовірність скоєння злочину, якого в інших обставинах людина могла і не вчиняти.

2014 року у Фергюсоні (США) поліцейський застрелив афроамериканця Майкла Брауна. З погляду громадськості, інцидент трапився лише тому, що вбитий був чорношкірим і справа відбувалася у неблагополучному районі. На думку прихильників громадянських свобод, системи, що приймають рішення на основі територіальних даних (щільність населення, соціальний склад, розташування барів, церков, шкіл, транспортних вузлів тощо) створюють ризик упередженості щодо неблагополучних районів. Вони стверджують, що прогностична поліцейська діяльність може лише загострити відносини між поліцейськими та афроамериканськими громадами.

Несанкціоноване використання персональних даних як етична проблема є найбільш очевидним і викликає суспільні дискусії, особливо в тих випадках, коли поліція у своїх розслідуваннях використовує відомості про громадян, отримані різними способами. Найчастіше йдеться про відеоспостереження та стеження. Повсюдна установка відеокамер (на вулиці та в офісах) дозволила створювати системи ШІ з використанням алгоритмів аналізу відеозображень [58].

Інше джерело персональних даних — профілі у соціальних мережах та інша інформація, яку громадяни самі розміщують в інтернеті. Поліція використовує у роботі спеціально розроблені програмно-аналітичні системи, які здійснюють автоматизоване стеження за соціальними мережами та семантичний аналіз повідомлень.

Іноді з метою поліцейського розслідування використовуються відомості про громадян , які надають комерційні організації: телекомунікаційні компанії (дані білінгових систем), банки, послуги таксі та інші компанії. За рахунок збору великих даних вони знають про громадян часом набагато більше держави.

У 2019 році в Арізоні (США) в ході розслідування вбивства заарештували людину , оскільки в момент вбивства сигнал з його телефону було зафіксовано поблизу місця злочину. Таку інформацію на запит поліції надала компанія Google. Заарештований провів майже тиждень у в'язниці, доки слідчі не знайшли іншого підозрюваного.

Практика запитів до Google про місцезнаходження користувача вперше була використана федеральними агентами в 2016 році в Північній Каролині і з тих пір поширилася по всій країні. База даних Sensorvault зберігає докладні записи про розташування сотень мільйонів пристроїв по всьому світу за останні десять років, інструмент обслуговування бізнес-процесів компанії Google поступово перетворюється на цифрову мережу для правоохоронних органів. По запиту поліції Google вивантажує з Sensorvault інформацію про пристрої, що відповідають заданим параметрам. Ця практика викликає побоювання та критику громадськості та юристів. Орін Керр, професор права в Університеті Південної Каліфорнії, вважає, що конфіденційність невинних людей, які потрапляють у поле зору поліції завдяки цифровим технологіям, — це нова правова проблема, яку слід вирішувати.

Крім названих вище етичних проблем відзначимо кілька особливо актуальних питань, які потребують подальшого громадського обговорення та вивчення.

Вплив впровадження ШІ та інших цифрових технологій на чисельність поліції. Здається, що цифровізація значно скоротить штат поліції, як це відбувається, наприклад, у банківських структурах. Однак завдяки розширенню та ускладненню системи відеоспостереження поліція отримуватиме суттєво більше інформації про протиправні дії, яким треба давати кримінально-процесуальну оцінку. Потрібно збільшити штат оперативних працівників, слідчих, експертів. Вже зараз потрібно підготувати багато аналітиків великих даних.

Різні етичні рекомендації для застосування ШІ в поліції та судовій системі. У більшості нещодавно прийнятих документів з питань використання ШІ наводяться ті самі рекомендації та обмеження для поліцейських та суддів. Однак насправді в поліції та в судовій системі слід по-різному застосовувати етичні принципи. Принцип прозорості щодо судових баз даних не викликає заперечень. Для поліцейських даних, отриманих під час оперативно-розшукових заходів, є серйозні обмеження. Ступінь прозорості сильно залежить від рівня секретності та можливих негативних наслідків для джерел інформації, потерпілих, підозрюваних та обвинувачених.

Критерії, які поліція використовує при зборі інформації про громадян за допомогою інструментів ШІ. Важливим є консенсус між поліцією та громадянським суспільством щодо цілей використання цих даних. Якщо йдеться про запобігання та розкриття тероризму, корупції, інших злочинів, то такого консенсусу можна досягти. Зібрана інформація використовуватиметься для прогнозування злочинних дій на основі відповідних кримінологічних та криміналістичних критеріїв. Якщо ж інформація стане

підставою для обмеження прав і свобод громадян, для побудови рейтингів соціального кредиту, то поліція та громадянське суспільство навряд чи дійдуть згоди.

3. Підходи до вирішення етичних проблем

Щоб мінімізувати етичні проблеми, які виникають під час використання поліцією цифрових технологій, пропонується кілька підходів:

- забезпечити прозорість ШІ ;
- використовувати можливості цифрових технологій для виявлення упереджень та дискримінації ;
- враховувати інтереси всіх зацікавлених сторін під час використання цифрових технологій, зокрема інтереси держави, громадянського суспільства та пересічних громадян;
- забезпечити нормативне регулювання використання цифрових технологій у поліцейській та судовій діяльності.

Європейський союз демонструє приклад найбільш збалансованого сьогодні врахування інтересів усіх сторін. Глобальний інноваційний центр Інтерполу (Interpol Global Complex for Innovation) та Міжрегіональний науково-дослідний інститут ООН з питань злочинності та правосуддя (United Nations Interregional Crime and Justice Research Institute) називають знаходження справедливого балансу між забезпеченням безпеки, з одного боку, та захистом приватного життя та конфіденційності - з іншого, ключовою етичною проблемою.

Поліцейським органам слід використовувати цифрові технології під контролем державних органів та громадських організацій. Кожен громадянин повинен мати можливість контролювати зберігання та використання своїх персональних даних, і ця умова в даний час не дотримується повною мірою в жодній країні світу частково через недостатнє нормативне регулювання та нерівномірний доступ до технологій різних груп населення. Для етичного використання ШІ правоохоронними органами належить створити систему державного та громадського нагляду. Її завдання — попередити перевищення повноважень співробітниками поліції та слідчими та забезпечити верифікацію даних, які використовуються для навчання алгоритмів ШІ [59].

У межах нормативного регулювання цифрових технологій у поліцейській та судовій діяльності національні та міжнародні органи розробили низку документів. Основним документом подібного роду можна назвати Європейську етичну хартію використання штучного інтелекту в судовій та правоохоронній системах Європейської комісії з ефективності правосуддя Європейської комісії з ефективності правосуддя. Хартія містить значні положення у сфері регулювання використання ШІ та великих даних, підтримує використання ШІ для підвищення ефективності та якості правоохоронної роботи. Цілеспрямовано наголошено на необхідності дотримуватися

прав особи, викладених у Європейській конвенції про захист прав людини та основних свобод, Конвенції про захист фізичних осіб при автоматизованій обробці персональних даних та інших нормативних актах. Хартія проголошує п'ять принципів використання ШІ у судовій та правоохоронній системах:

✓ **Повага до фундаментальних прав особистості**

При створенні та використанні технологій та інструментів ШІ слід переконатися, що вони не порушують основних прав особистості.

✓ **Неприпустимість дискримінації**

Необхідно блокувати можливість будь-якої дискримінації окремих груп та соціальних верств, яка може з'явитися внаслідок застосування статистичних методів при обробці великих даних.

✓ **Забезпечення якості та безпеки алгоритмів**

При використанні великих даних слід перевіряти їх джерела, структуру та зміст; необхідно використовувати математичні моделі, розроблені на міждисциплінарній основі, враховувати як прямі статистичні кореляції, а й соціальні, культурні, економічні та інші чинники.

✓ **Прозорість систем ШІ**

ШІ допустимо використовувати у судовій та правоохоронній системі тільки в тому випадку, якщо забезпечено прозорість вихідних великих даних.

✓ **Забезпечення якості та безпеки алгоритмів**

Особи, пов'язані зі слідством та судочинством, завжди повинні розуміти, на чому ґрунтуються висновки ШІ, які пропонуються як автоматизована експертна думка.

В умовах цифровізації поліція використовує все більш складні технології, які можуть нести потенційну загрозу інформаційній безпеці громадян та нормам суспільної моралі. Для того, щоб запобігти формуванню негативних тенденцій, суспільство має стежити, чи є діяльність поліції максимально відкритою та підконтрольною. Етичні проблеми, пов'язані з роботою поліції, можуть бути вирішені за допомогою вдосконалення технологій, відповідного регулювання та громадського контролю.

Use Case 3. Етика цифрової медицини

При переході до цифрової медицини питання етики стали ключовими і значною мірою визначають швидкість технологічного прогресу у цій сфері. Використання великих даних та технологій ШІ дає можливість підняти на новий рівень діагностику, лікування та систему профілактики захворювань. Однак гостро стоїть питання, якою мірою можна використовувати дані про здоров'я громадян для навчання ШІ: обмежене використання уповільнює розвиток технологій ШІ, а необмежене загрожує дискримінацією та порушенням прав і свобод особистості.

1. Біомедичні дані та великі дані в цифровій медицині

По всьому світу медичні дані збиралися століттями, але застосування інформаційних технологій для збору, зберігання та аналізу дозволило вивести роботу з ними на новий рівень. Існують різні дані, що стосуються фізичного стану та здоров'я людини. Частина даних власне виробляється внаслідок надання медичної допомоги, а частина даних може бути отримана поза зв'язком із цими двома процесами. Вони мають різний правовий статус, різні можливості доступу до них третіх осіб, але використання тих і інших викликає етичні питання.

Біомедичні дані є відомостями, що становлять лікарську таємницю. Дане медичне, правове, соціально-етичне поняття є забороною медичному працівнику повідомляти третім особам інформацію про стан здоров'я пацієнта. Лікарська таємниця – один із найважливіших принципів у професійній медичній етиці, крім того, лікарська таємниця захищена законом [60].

У цифрову епоху медичні документи та інші відомості, що становлять лікарську таємницю, не є єдиним джерелом даних про фізичний стан та здоров'я людини. Соціальні мережі, історія пошукових запитів, дані про пересування та відвідування лікувальних установ, покупки теж стають джерелами даних, які потенційно можуть бути використані, наприклад, при оцінці ризиків у страхуванні або прийомі на роботу, але на такі дані не поширюється закон про лікарську таємницю. Виникає проблема не тільки ефективного, а й етичного обігу та використання медичних даних.

Джерела даних у медицині у найширшому розумінні:

- електронні медичні картки;
- мобільні програми для охорони здоров'я;
- датчики та пристрої моніторингу;
- дані лабораторних досліджень; рентгенівські знімки;
- дані, одержані в ході наукових досліджень за участю груп пацієнтів;
- дані щодо купівлі ліків та інших засобів медичної допомоги пацієнтами;
- дані соцмереж, пошукових запитів тощо.

Використання ШІ, побудованого на аналізі медичних даних, дозволяє змінити життя мільйонів пацієнтів: якісно покращити діагностику, персоналізувати лікування, докорінно змінити прийняття лікарських рішень, розширити можливості раннього виявлення та профілактики захворювань. Дилема полягає в тому, що для найбільш ефективної роботи ШІ потрібні максимально повні дані про пацієнтів як безпосередньо медичні, так і соціальні [61].

Сьогодні не завжди можливо перетворити медичну інформацію на дані , придатні для подальшої роботи. Медична інформація та дані зараз розрізнені, містяться в ізольованих сховищах та несумісних системах та форматах, багато що існує тільки на папері або на плівці, і майже все підлягає законодавчому захисту. Перелічені фактори ускладнюють обмін, обробку та інтерпретацію навіть у США та Великій Британії, де досягнуто високого технічного рівня медичної статистики.

Для того, щоб використовувати медичні дані у наукових дослідженнях, потрібні:

- ✓ якісні набори даних;
- ✓ налагоджений зв'язок між системами;
- ✓ уніфікація даних;
- ✓ вироблення етичних норм щодо використання персональних даних.

Перш ніж використовувати великі дані, потрібно вирішити головне етичне питання: як забезпечити пацієнтові приватність? Технічно можливим є виявлення людей з конкретними особливостями, навіть якщо самі індивіди явно не вказали їх. При цьому важливо відзначити, що великі дані не вимагають однозначної вказівки на належність до тієї чи іншої групи, вони дозволяють виявляти її ознаки автоматично (наприклад, прийом певних ліків може вказувати на ВІЛ-статус). Якщо буде відкритий доступ до чутливих відомостей про стан фізичного, психічного здоров'я, схильність до суїциду тощо, це може призвести до дискримінації при прийомі на роботу, нерівності при отриманні медстрахування тощо.

Дослідники використовують ШІ для прогнозування проблем зі здоров'ям (хвороби серця , інсульт , зниження когнітивних функцій , ризик самогубства). Одна із світових соціальних мереж запровадила алгоритм, який робить висновки про суїцидальні наміри користувачів на основі постів (наприклад, таких фраз, як «З тобою все гаразд?», у поєднанні з «Прощавай» та «Будь ласка, не роби цього»).

Цифрові сліди, які залишає людина, накопичуються протягом усього життя. Це дає нові можливості для розвитку різних технологій, але водночас підвищує вразливість приватного життя, ставлячи під загрозу таємницю особистого життя. Реальність сьогодні така, що мобільний оператор знає про звернення людини по медичну допомогу більше, ніж система охорони здоров'я. Дані, що становлять медичну таємницю, все

більше затребувані третіми сторонами і, наприклад, викликають інтерес не тільки у комерційних структур, а й у державних органів.

У Китаї єдине сховище медичних даних - медична хмара – дозволяє об'єднувати дані та полегшує роботу з ними. Права на збір, обмін та використання даних, на управління хмарою та відповідні обов'язки законодавчо регулюються та передаються на аутсорсинг як державним, так і приватним компаніям. Завдяки такій практиці дослідникам та розробникам ШІ стають доступні великі обсяги даних, що сприяє розвитку технологій.

Захист конфіденційності при зборі особистих даних в Україні організовано в такий спосіб, що серйозно ускладнює використання біомедичних великих даних. Власником своїх біомедичних чутливих персональних даних є пацієнт. Людина підписує дозвіл на обробку персональних даних (інформована згода) при зверненні до будь-якої організації, де йде збір даних. Обробляти їх може тільки та організація, якій він це дозволив, і тільки для тих цілей, які вказані у письмовому дозволі або передбачені законом. Створення ШІ не є медичною послугою, тому займатися подібною діяльністю у лікувальних закладах не можна. Розвиток технологій неможливий без роботи з великими масивами даних, що поєднують дані з локальних джерел, тому варто очікувати, що регламент роботи з великими даними зміниться. Важливо регламентувати умови доступу до біомедичних даних, не порушивши права пацієнта. Розвиток цифрової медицини ставить перед державою принципову етичну дилему: зберегти за пацієнтом право на безмежне володіння даними чи точково змінювати законодавство, дозволивши обробку деперсоніфікованих даних без запиту згоди пацієнта, але зрештою на користь особистості та суспільства [62].

Якщо розглядати анонімізовані (деперсоніфіковані) дані, то заради цілей цифрової економіки дозволено лише збирання і контрольований оборот таких даних всередині державної системи охорони здоров'я. Використовувати дані можна лише для певних потреб, головна з яких – надання медичної допомоги. Створення цифрових сервісів, дослідження в галузі машинного навчання та інші приклади роботи з даними не є медичними послугами, а відповідно, їх збір та обробка персональних з цією метою не відповідатимуть цілям організації, в даному випадку лікувального закладу.

Варіанти вирішення проблеми:

- проводити ретроспективні наукові дослідження та машинне навчання у тій організації, де зібрані дані (дані не вийдуть за межі організації, закон не буде порушено);
- використовувати у зазначених цілях деперсоніфіковані дані.

Говорячи про вирішення етичних питань під час збирання медичних даних, слід зазначити актуальну проблему достовірності даних. Адже існують випадки, коли з тих чи інших причин відбуваються маніпуляції зі статистикою в медицині (табл. 3).

Таблиця 3

Можливі маніпуляції з медстатистикою у сфері охороні здоров'я

Ціль	Спосіб маніпуляції
Забезпечити певні показники (наприклад, зниження смертності від будь-яких причин тощо)	Лікарі вказують як причину смерті супутню патологію або виписують термінального хворого, щоб він помер не в стаціонарі, оскільки смерть у стаціонарі зіпсує статистику
Перестрахування на випадок кримінального переслідування	Неточність, неповнота та недостовірність записів у медкарті
Запобігання штрафам з боку страхової компанії	Неповнота та неточність даних, недостатній або надмірний обсяг лікування, неправильна тактика лікування

У системі охорони здоров'я, де багато чого — від фінансування клінік до зарплати медиків — залежить від показників, статистика перестає бути лише інструментом для спостереження за лікувально-діагностичним процесом, вона перетворюється на інструмент впливу, часом його істотно трансформуючи. Тому виникає питання про те, наскільки можна довіряти таким навчальним даним для системи ШІ.

2. Моделі розвитку та способи регулювання цифрової медицини

Правила використання великих медичних даних залежать від пріоритетів при вирішенні етичних питань та концепції регулювання останніх з боку держави та суспільства (табл. 4). У кожній країні є орган нагляду, який стежить за дотриманням інтересів пацієнтів і регулює ринок медичних послуг. В Україні це Міністерство охорони здоров’я, у США - Управління з контролю за якістю харчових продуктів, медикаментів та косметичних засобів (Food and Drug Administration).

Таблиця 4

Моделі розвитку цифрової медицини

Характеристики	Модель розвитку медицини		
	пацієнтська	технологічна	державна
Країна (регіон)	Європа, лідер - Великобританія	США	Китай
Пріоритети	захист прав та свобод пацієнта	розвиток технологій, інтереси бізнесу	інтереси держави

У Великій Британії після тривалого обговорення всіма зацікавленими сторонами було сформульовано «Кодекс правил щодо використання технологій, заснованих на даних» . За задумом творців, цей кодекс має стати частиною загальної цифрової національної стратегії та допомогти створити середовище, яке підтримує інноваційні технології, що використовують дані, забезпечити безпеку, конкурентоспроможність, дотримання етичних та правових норм.

10 принципів використання технологій, що базуються на даних:

1. Розуміння користувачів, їх потреб, вивчення умов, у яких використовуватиметься алгоритм.
2. Визначення результату та того, як технологія сприятиме його досягненню.
3. Використання даних згідно з принципами та відповідно до цілей, для яких вони були зібрані.
4. Облік етичної складової, прозорість та відповідальність у питаннях використання даних.
5. Використання відкритих стандартів.
6. Прозорість алгоритмів.
7. Визначення, який тип алгоритму розробляється або розгортається, етичний аналіз використовуваних даних, перевірка їхньої ефективності та того, як вони будуть інтегровані в систему охорони здоров'я.
8. Відкриття даних про ефективність передбачуваного використання та співвідношення ціни та якості.
9. Забезпечення безпеки даних як невід'ємна частина розробки.
10. Визначення комерційної стратегії структури чи організації.

У Китаї збирання та зберігання великих даних організовано таким чином, що їх можна використовувати для проведення масштабних досліджень. Медичні дані поєднуються з інформацією із соцмереж, географічним розташуванням, економічними та екологічними даними. Інтеграція джерел інформації до наборів даних, які можна аналізувати, є ключем до використання великих даних. Крім того, великі дані генеруються та обробляються з такою швидкістю, яка дозволяє оперативно використовувати їх для профілактики та лікування захворювань.

Основними проблемами при цифровізації медицини стають:

- забезпечення рівних прав для бізнесу під час розвитку цифрової медицини;
- обмеження корупції у сфері медичних послуг;
- захист даних та приватності пацієнтів.

Можливий шляхом вирішення цієї проблеми може бути підхід, при якому враховується, що пацієнт має безумовне право на захист своїх персональних даних, але при цьому анонімізовані дані належать державі.

Для реалізації такого підходу необхідне:

- створення законодавчої бази;
- створення масиву анонімізованих даних;
- призначення державного оператора до роботи з даними;
- забезпечення контрольованого доступу дослідників до даних.

Для розвитку систем ШІ в галузі медицини доступ до якісних медичних даних є життєво важливим. При зміні законодавства у сфері захисту ПД важливо звернути увагу на баланс між забезпеченням приватності пацієнта та розвитком технологій.

Висновки

Технології штучного інтелекту розвиваються надзвичайно швидко та допомагають людству у вирішенні найрізноманітніших задач: діагностичні медичні та промислові системи, розумні міста, кожна з компонент яких використовує штучний інтелект, смартфони, порохотяги та беспілотні автомобілі – штучний інтелект всюди. Системи комп'ютерного зору, машинного навчання, генеративний штучний інтелект показують результати вражаючої точності.

Однак, такий прогрес та поширення систем ШІ, який безперечно орієнтований на те, щоб полегшити та покращити життя людей, вирішити економічні та екологічні проблеми, зробити суспільство інклюзивнішим та вирішити багато інших важливих проблем, все частіше супроводжується гучними скандалами в пресі та науковій спільноті. Чи то голосовий помічник, який, виявляється, 24 години на день прослуховує та записує те, що відбувається навколо, скорингові системи в Нідерландах, які руйнують життя людей, несправедливо звинувативши їх в махінаціях із податками, упереджені рекрутингові системи, які не рекомендують брати на роботу людей, які проживають в «неблагонадійних» районах міста, не зважаючи на їхні особистісні чи професійні якості. Кожен з таких випадків доводить, що сьогодні основним критерієм для оцінки систем ШІ має бути не лише точність або швидкодія, а етичність та справедливість.

Над цим вже давно замислюються як науковці, так і влада багатьох країн. І чим більш активно розвивають технології та методи ШІ, тим більш актуальною стає необхідність його регулювання, формулювання принципів, дотримуючись яких ми зможемо бути впевнені, що системи ШІ, які ми розробляємо чи які ми використовуємо – надійні, відповідальні, неупереджені та заслуговують на нашу довіру.

У цьому навчальному посібнику детально розглянуто те, які проблеми з погляду етичності можуть виникати в системах ШІ, а також детально описано різноманітні підходи, завдяки яким ці ризики можна зменшити чи загалом позбавитись їх. Окремо приділено увагу етичності систем ШІ, які використовуються в таких чутливих сферах життя людей, як освіта, медицина та правоохоронна діяльність.

Список використаної літератури

1. Simonite T. AI experts want to end «black box» algorithms in government // Wired Business.
2. Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland Uk. ACM, New York, NY, USA. https://doi.org/10.1145/3290605.3300233
3. Mike Elgan. The case against teaching kids to be polite to Alexa. https://www.fastcompany.com/40588020/the-caseagainst-teaching-kids-to-be-polite-to-alexa, 2018. Accessed: January 22, 2023.
4. Kleinberg J., Lakkaraju H., Leskovec J.et al. Human decisions and machine predictions // The Quarterly Journal of Economics. 2018. Vol. 133, no 1. P. 237–293.
5. Gates S. W., Perry V. G., Zorn P. M. Automated underwriting in mortgage lending: Good news for the underserved? // Housing Policy Debate. 2002. Vol. 13, no 2. P. 369–391.
6. Lum K., Isaac W. To predict and serve? // Significance. 2016. Vol. 13, no 5. P. 14–19.
7. Ross C., Swetlitz I. IBM's Watson supercomputer recommended «unsafe and incorrect» cancer treatments, internal documents show // STAT+
8. Chiappa S. Path-Specific Counterfactual Fairness // [Silvya Chiappa.]
9. Tackling bias in artificial intelligence (and in humans) // McKinsey.
10. Vincent J. Amazon Reportedly Scraps Internal AI Recruiting Tool That Was Biased against Women // The Verge
11. Strickland E. Racial Bias Found in Algorithms That Determine Health Care for Millions of Patients // IEEE Spectrum.
12. Silberg J., Manyika J. Tackling bias in artificial intelligence (and in humans) // McKinsey Global Institute.
13. Narayanan A. Tutorial: 21 fairness definitions and their politics / FAT. [S.l.,] 2018.
14. Zemel R., Wu Y., Swersky K. et al., Learning Fair Representations //Proceedings of the 30th International Conference on Machine Learning.2013. Vol. 28, no 3. P. 325–333.
15. Chiappa S., Isaac W.S.A causal Bayesian networks viewpoint on fairness // Privacy and Identity Management. Fairness, Accountability, and Transparency in the Age of Big Data // IFIP Advances in Information and Communication Technology. 2019. Vol. 547. P. 3–20.
16. Green B., Hu L. The myth in the methodology: Towards a recontextualization of fairness in machine learning // 35th International Conference on Machine Learning. Stockholm, 2018;
17. Richardson R., Schultz J., Crawford K. Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice // New York University Law Review Online. 2019. March.
18. Manyika J.,Bughin J.The promise and challenge of the age of artificial intelligence / McKinsey Global Institute // McKinsey.
19. Robinette P., Howard A., Wagner A.R. Conceptualizing overtrust in robots: Why do people trust a robot that previously failed? // Autonomy and Artificial Intelligence: A Threat or Savior? Cham: Springer International Publishing, 2017. P. 129–155. DOI:10.1007/978-3-319-59719-5_6;
20. Wagner A.R., Borenstein J., Howard A. Overtrust in the Robotic Age // Communications of the ACM. 2018. Vol. 61, no 9. P. 22–24. DOI:10.1145/3241365.
21. Pynadath D. V., Barnes M., Wang N. et al. Transparency Communication for Machine Learning in Human-Automation Interaction // Human and Machine Learning. Visible, Explainable, Trustworthy and Transparent / Ed. by J. Zhou, F. Chen. Cham: Springer, 2018. P. 75–90. DOI: 10.1007/978-3-319-90403-0_5.
22. GDPR
23. Закон України про захист персон
24. Turing A.M. Computing machinery and Intelligence // Mind. 1950. Vol.. 54, no 236. P. 433–460.

25. Wiener N. Some Moral and Technical Consequences of Automation // Science. 1960. Vol. 131, no 3410. P. 1355–1358.
26. Bostrom N., Yudkowsky E. The Ethics of Artificial Intelligence // Cambridge Handb. Artif. Intell. 2011. P. 1–20.
27. https://www.moralmachine.net/
28. Foot P. The Problem of Abortion and the Doctrine of the Double Effect in Virtues and Vices. Oxford: Basil Blackwell, 1978.
29. Awad E., Dsouza S., Kim R. et al. The Moral Machine experiment // Nature. 2018. Vol. 563. P. 59–64. DOI:10.1038/s41586-018-0637-6.
30. "Ethically Aligned Design - A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems," in *Ethically Aligned Design - A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems* , vol., no., pp.1-294, 31 March 2019.
31. UK AI Council. (2021). AI Roadmap. Available online: https://www.aicouncil.org.uk/ai-roadmap (accessed on 30 May 2024).
32 European Commission. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 24 July 2024 on a European approach for Artificial Intelligence (AI Act). Official Journal of the European Union. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ%3AL_202401689 (accessed on 27 July 2024)
33. Регламент Європейського Парламенту і Ради (ЄС) 2016/679 про захист фізичних осіб у зв'язку з опрацюванням персональних даних і про вільний рух таких даних, та про скасування Директиви 95/46/ЄС (Загальний регламент про захист даних) від 27 квітня 2016 року. URL: https://zakon. rada.gov.ua/laws/show/984_008-16#Text (дата звернення: 23.05.2021 р.)
34. Про захист персональних даних : Закон України від 01.06.2010 р. №2297- VI / Верховна Рада України. Офіційний вісник України від 09.07.2010 р. Офіц. вид. 2010. № 49. С. 199, стаття 1604, код акта 51762/2010.
35. Understanding society: the power and perils of data // Ipsos.
36. Calhoun C. S., Bobko Ph., Gallimore J. J. et al. Linking precursors of interpersonal trust to human-automation trust: An expanded typology and exploratory experiment // Journal of Trust Research. 2019. Vol. 9, no 1. P. 28–46. DOI: 10.1080/21515581.2019.157973
37. NYC Automated Decision Systems Task Force Report. [S.l.,] 2017.
38. Mayor de Blasio Signs Executive Order to Establish Algorithms Management and Policy Officer // NYC.
39. van Est R., Kool L. Human rights in the robot age: challenges arising from the use of robotics, artificial intelligence, and virtual and augmented reality. // Rathenau Instituut.
40. Pérez Orozco B., Rentería Rodríguez M. E.INCyTU Note 12 Artificial Intelligence // Foro Consultivo.
41. Technology Assessment: Artificial Intelligence: Emerging Opportunities, Challenges, and Implications // U.S. Government Accountability Office.
42. Baichere D. Facial recognition // Assemblee Nationale.
43. Recommendation CM/Rec (2016) 3 to member States on human rights and business // Council of Europe.
44. Responsibility and AI // Council of Europe.
45. The Global AI Strategy Landscape // Holon IQ.
46. National and international AI Strategies // Future of Lise Institute.
47. Building an AI World: Report on National and Regional AI Strategies // CIFAR.
48. Dutton T., Barron B., Boskovic G. Building an AI World Report on National and Regional AI Strategies // CIFAR.
49. Asilomar AI Principles // Future of Life.

50. Montréal Declaration: Responsible AI // Université de Montréal.
51. Ethics guidelines for trustworthy AI // European Commission.
52. Report of COMEST on robotics ethics // COMEST.
53. European Commission, 2020c. Proposal for a regulation laying down harmonized rules on Artificial Intelligence. URL: https://eur-lex.europa.eu/legal-content/EN/ TXT/?uri=CELEX:52021PC0206. (Accessed 04 November 2023).
54. European Commission, 2019. Communication–Building trust in human centric Artificial Intelligence. URL: https://ec.europa.eu/digital-single-market/en/news/ communication-building-trust-human-centric-artificial-intelligence. (Accessed 20 November 2023).
55. European Commission, 2020a. Ethics guidelines for trustworthy AI. URL: htt ps://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf. (Accessed 17 November 2023).
56. European Commission, Horizon 2020 programme - guidance–How to complete your ethics self-assessment. URL: https://ec.europa.eu/research/participants/ data/ref/h2020/grants_manual/hi/ethics/h2020_hi_ethics-self-assess_en.pdf. (Accessed 17 November 2023).
57. European Commission, 2020b. Horizon Europe strategic plan 2021-2024. URL: https://ec.europa.eu/commission/presscorner/detail/en/ip_21_1122, doi: 10.2777/083753. (Accessed 19 November 2023)
58. European Commission. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 24 July 2024 on a European approach for Artificial Intelligence (AI Act). *Official Journal of the European Union*. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ%3AL_202401689 (accessed on 27 July 2024)
59. International Organization for Standardization. (2015). ISO 9001:2015 - Quality management systems - Requirements. Available online: https://www.iso.org/standard/62085.html (accessed on 27 July 2024).
60. U.S. White House. (2020). Guidance for Regulation of Artificial Intelligence Applications. Available online: [URL] (accessed on 27 July 2024).
61. IEEE. (2019). Ethically Aligned Design. [Online]. Available: https://ethicsinaction.ieee.org
62. UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence. Available online: https://unesdoc.unesco.org/ark:/48223/pf0000379981 (accessed on 30 July 2024).