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

Deliverable number	1.2
Deliverable name	Teaching materials for the course “Trustworthy artificial intelligence: the European approach” (for Bachelors of Computer Science)
Work Package number	1
Lead Beneficiary	Lviv Polytechnic National University
Type	R — Document, report
Dissemination level	Public
Due date (in months)	24
Description	The Course Notes will consist of approx. 80 pages, A4 format, font Times New Roman, 12 p. The teaching materials will provide students with detailed, up-to-date information on the course topics and will complement all learning activities during the courses, the workshops and the seminars. Furthermore, the public availability for unlimited download makes possible the dissemination of the course notes to the general public interested in the aforementioned topics, thus broadening the impact of this deliverable beyond its primary target. E-book, Ukrainian language
Website link	https://trustai.org.ua/portfolio-item/d2_teaching_materials/
Author(s)	Anastasiya Doroshenko

Навчальний посібник

Trustworthy AI: the European Approach

Надійний штучний інтелект: європейський підхід

Для бакалаврів
спеціальності 122
«Комп'ютерні науки»

Co-funded by the European Union

TRUSTAI

Національний університет Львівська політехніка 1816

Надійний штучний інтелект: європейський підхід: навчальний посібник для бакалаврів за спеціальністю «Комп'ютерні науки» /Анастасія Дорошенко. – Національний університет «Львівська політехніка». – Львів, 2024.

Висвітлено сучасний стан розвитку систем штучного інтелекту, описано основні принципи та передумови виникнення надійного штучного інтелекту. Розглянуто ризико-орієнтований підхід до створення систем ШІ та стратегії зменшення ризиків для користувачів, проаналізовано етичні принципи впровадження систем ШІ та важливість міждисциплінарної співпраці для забезпечення ефективності ШІ. Розглянуто законодавче регулювання використання систем ШІ із дотриманням вимог сучасного Європейського законодавства із регулювання використання ШІ (AI Act), а також здійснено порівняння з українським підходом.

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

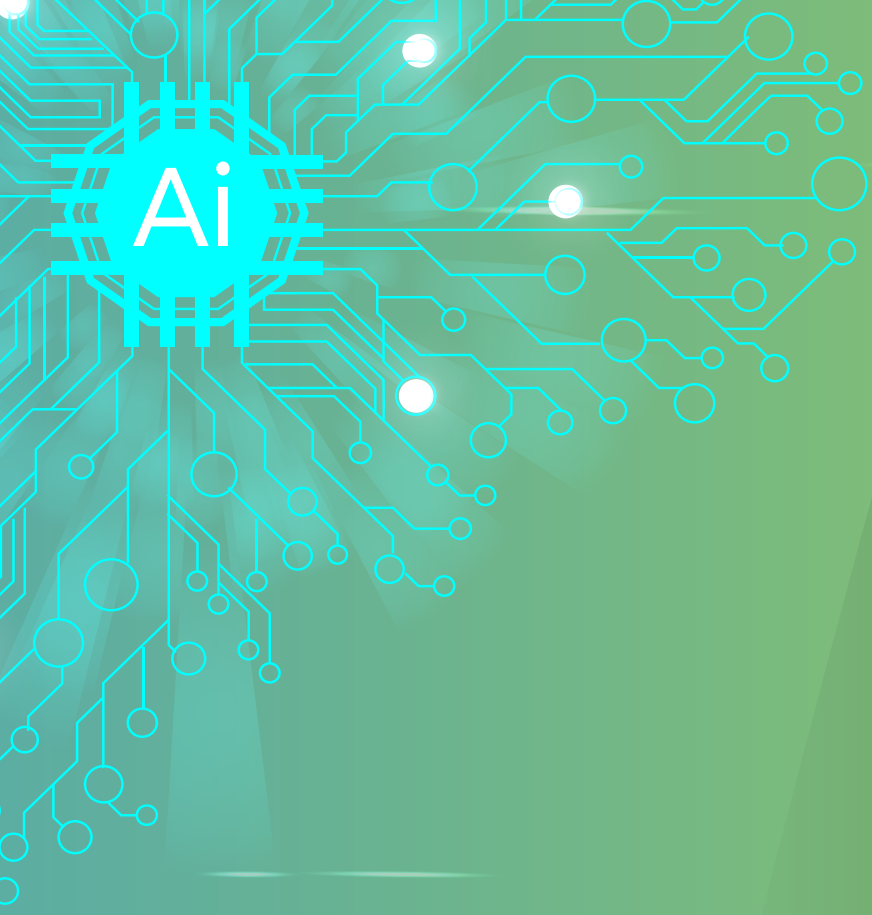
Фінансується Європейським Союзом. Однак висловлені погляди та думки належать виключно автору(ам) і не обов'язково відображають погляди Європейського Союзу чи Європейського виконавчого агентства з питань освіти та культури (EACEA). Ані Європейський Союз, ані EACEA не несуть за них відповідальності.

Co-funded by the European Union

TRUSTAI

Національний університет Львівська політехніка 1816 Lvivska Politekhnika National University

ЗМІСТ

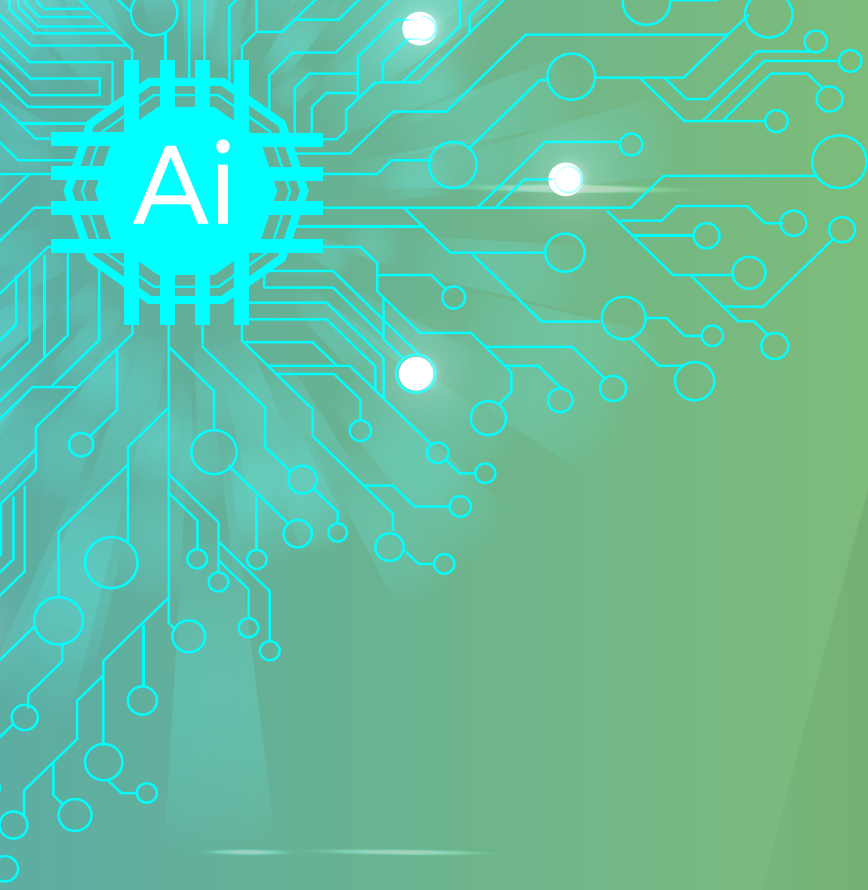


ВСТУП

Значний прогрес у застосуванні штучного інтелекту (ШІ) для прийняття рішень у сфері охорони здоров'я, медичної діагностики та в інших сферах водночас викликав занепокоєння щодо справедливості та упередженості систем ШІ. Це особливо критично в таких сферах, як охорона здоров'я, працевлаштування, кримінальне правосуддя, оцінка кредитоспроможності та все частіше в генеративних моделях ШІ, які створюють синтетичні дані. Такі системи можуть призвести до несправедливих результатів і закріпити існуючу нерівність, зокрема генеративні упередження, які впливають на представлення окремих осіб у синтетичних даних. У навчальному посібнику розглянемо такі виклики для справедливого, надійного та відповідального використання ШІ, як упередженість – розглянемо її джерела, вплив і стратегії пом'якшення, етичність, прозорість та законність розроблення та використання систем ШІ.

Серед джерел упередженості для систем ШІ можна виокремити такі, як упередженість даних, алгоритмів і людських рішень. Зміщення (bias) на кожному з цих етапів потенційно здатне підсилити існуючі проблеми упередження ШІ, коли моделі можуть відтворювати та посилювати суспільні стереотипи. Важливо розуміти вплив на суспільство упереджених систем штучного інтелекту та ризики, пов'язані з їх використанням, зосереджуючись на збереженні нерівності та зміцненні шкідливих стереотипів, особливо тому, що генеративний штучний інтелект стає все більш поширеним у створенні контенту, який впливає на сприйняття громадськості. Розглянемо різні стратегії зменшення ризиків використання систем ШІ для користувачів, обговоримо етичні принципи впровадження систем ШІ та проаналізуємо важливість міждисциплінарної співпраці для забезпечення ефективності.

Варто зазначити, що для усунення упередженості в штучному інтелекті необхідно використовувати цілісний підхід, що включає різноманітні та репрезентативні набори даних, підвищену прозорість і підзвітність у системах штучного інтелекту, а також дослідження альтернативних парадигм штучного інтелекту, які надають пріоритет справедливості та етичним аспектам. Дотримання цих принципів є запорукою розроблення справедливих і неупереджених систем штучного інтелекту.



1

СУЧАСНИЙ СТАН РОЗВИТКУ СИСТЕМ ШІ

Перш ніж розпочати розмову про надійний ШІ, необхідно визначити, що таке ШІ, які завдання він може вирішувати сьогодні та якими методами.

Штучний інтелект — це галузь інформатики, яка займається створенням систем, здатних виконувати завдання, що потребують людського інтелекту, такі як розпізнавання образів, навчання, міркування, прийняття рішень та розуміння природної мови.

Штучний інтелект (ШІ) належить до комп'ютерних систем, здатних виконувати складні завдання, які історично могла виконувати лише людина, наприклад міркувати, приймати рішення або вирішувати проблеми.

Відповідно до більш формалізованого визначення, запропонованого стандартом ISO/IEC 22989:2022, «Штучний інтелект – це технічна та наукова галузь, присвячена розробленій системі, яка генерує такі результати, як контент, прогнози, рекомендації чи рішення для заданого набору цілей, визначених людиною» [1].

Хоча це визначення штучного інтелекту є точним з технічного погляду, його важко зрозуміти середньостатистичній людині, далекій від наукових досліджень. Як зрозуміти в реальному житті, що ми маємо справу із штучним інтелектом?

Сьогодні термін «штучний інтелект» описує широкий спектр технологій, які забезпечують роботу багатьох послуг і товарів, якими ми користуємося щодня: від додатків, які рекомендують серіали чи відео у соціальних мережах, до чат-ботів, які надають підтримку клієнтам у режимі реального часу. Але чи справді все це є штучним інтелектом, яким більшість з нас його уявляє? А якщо ні, то чому ми так часто вживаємо цей термін?

Необхідно розуміти, що штучний інтелект – це лише практичний інструмент, а не панацея. Він працює настільки добре, наскільки правильно підібрані, навчені та налаштовані алгоритми та методи машинного навчання, що керують його діями. ШІ може справді добре виконувати конкретне завдання, але для цього потрібна велика кількість даних для навчання та велика кількість повторень. ШІ вчиться аналізувати великі обсяги даних, розпізнавати закономірності та робити прогнози чи рішення на основі цих даних, постійно покращуючи свою точність з часом.

Однак, сьогодні можливості штучного інтелекту вийшли за межі простої обробки даних і включають розробку машин, здатних навчатися, міркувати та вирішувати проблеми. Машинне навчання стало настільки «компетентним», що генерує все: від програмного коду до зображень, статей, відео та музики. Це наступний рівень ШІ, так званий генеративний ШІ, який відрізняється від традиційного ШІ своїми можливостями та застосуванням. Тоді як традиційні системи штучного інтелекту в основному використовуються для аналізу даних і прогнозування, генеративний штучний інтелект йде ще далі, створюючи нові дані, схожі на навчальні дані.

Все це є черговим етапом еволюції штучного інтелекту, основні віхи якої відображено на рис. 1.

Еволюція Штучного Інтелекту

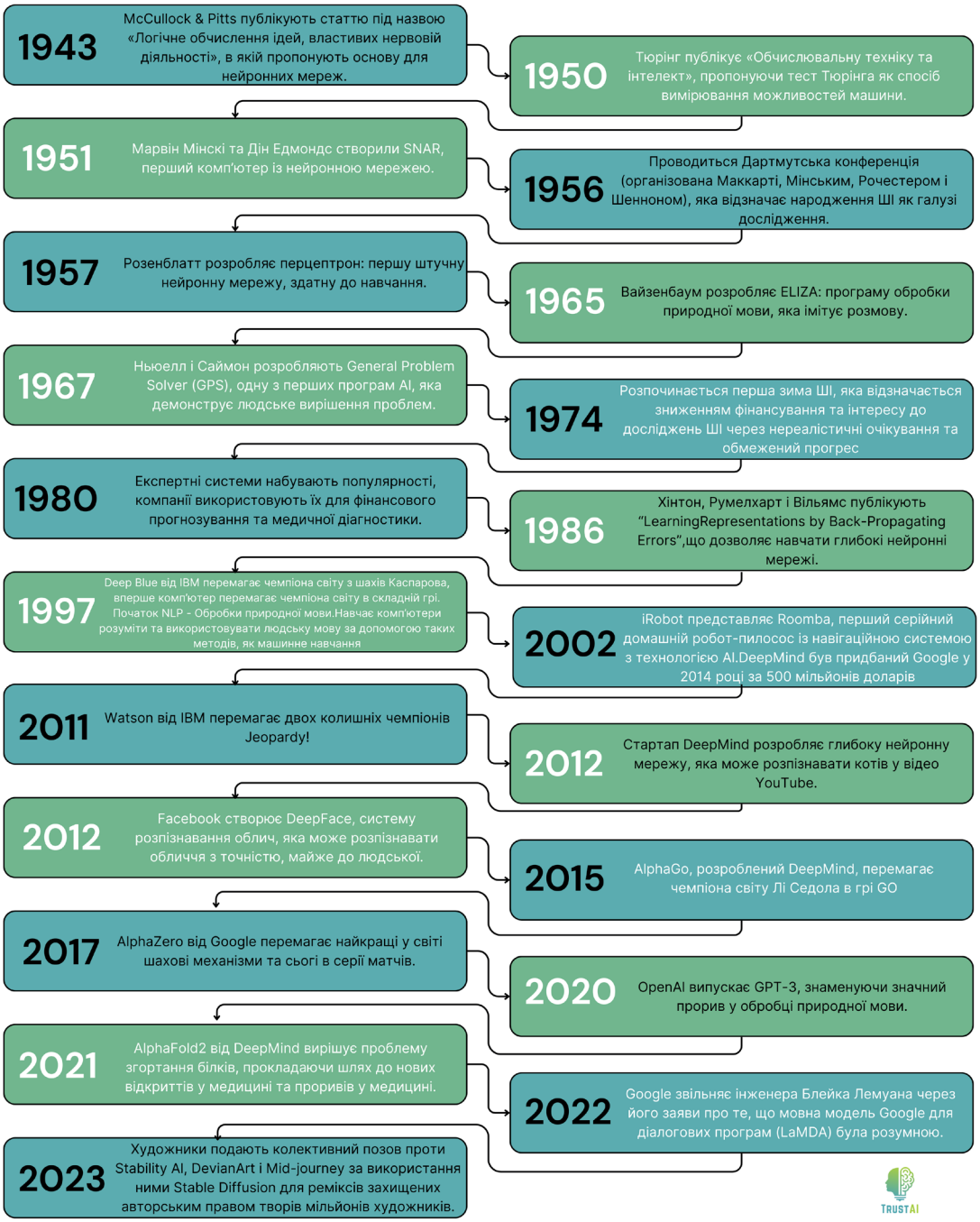


Рис.1.1. Основні віхи еволюції штучного інтелекту

1.1. Еволюція ШІ: від слабкого до супер ШІ

Ранні ітерації додатків штучного інтелекту, з якими ми найчастіше взаємодіємо сьогодні, були побудовані на традиційних моделях машинного навчання. Ці моделі покладаються на алгоритми навчання, розроблені та підтримувані спеціалістами з обробки даних. Інакше кажучи, традиційні моделі машинного навчання потребують втручання людини для обробки нової інформації та виконання будь-яких нових завдань, які виходять за рамки початкового навчання [2].

Наприклад, у 2011 році Apple додала Siri до своєї iOS. Ця рання версія Siri була навчена розуміти набір дуже конкретних заяв і запитів. Щоб розширити базу знань і функціональність Siri, знадобилося втручання людини.

Проте можливості штучного інтелекту постійно розвиваються після прориву в розробки штучних нейронних мереж у 2012 році, які дозволяють машинам брати участь у навчанні з підкріпленням і моделювати, подібно до того, як людський мозок обробляє інформацію [3].

На відміну від базових моделей машинного навчання, моделі глибокого навчання дозволяють додаткам штучного інтелекту навчитися виконувати нові завдання, які потребують людського інтелекту, брати участь у новій поведінці та приймати рішення без втручання людини [4]. У результаті глибоке навчання забезпечило автоматизацію завдань, створення вмісту, прогнозне обслуговування та інші можливості в різних галузях.

Завдяки глибокому навчанню та іншим досягненням сфера штучного інтелекту залишається в постійному та швидкому стані зміни. Наше колективне розуміння реалізованого ШІ та теоретичного ШІ продовжує змінюватися, тобто категорії ШІ та термінологія ШІ можуть відрізнятися від одного джерела до іншого. Однак типи штучного інтелекту можна значною мірою зрозуміти, вивчивши дві охоплюючі категорії: можливості ШІ та функції ШІ.

1.2. Три типи ШІ на основі можливостей

Вузький ШІ

Вузький штучний інтелект (Narrow AI), також відомий як слабкий ШІ (Weak AI), є єдиним типом ШІ, який існує сьогодні. Будь-яка інша форма ШІ є теоретичною. Його можна навчити виконувати конкретне вузьке завдання, часто набагато швидше та краще, ніж із цим може впоратись людський розум. Однак він не може працювати за межами визначеного завдання. Натомість він націлений на одну підмножину когнітивних здібностей і вдосконалює цей спектр. Siri, Alexa від Amazon і IBM Watson® є прикладами вузького штучного інтелекту. Навіть ChatGPT від OpenAI вважається формою

вузького штучного інтелекту, оскільки він обмежений одним завданням текстового чату.

Загальний ШІ (Сильний ШІ)

Загальний штучний інтелект (AGI, Artificial General Intelligence), також відомий як «Сильний ШІ» (Strong AI), сьогодні є не більш ніж теоретичною концепцією. AGI може використовувати попередні знання та навички для виконання нових завдань у іншому контексті без потреби людей навчати основні моделі. Ця здатність дозволяє AGI навчатися та виконувати будь-яке інтелектуальне завдання, яке під силу людині.

Супер ШІ

Супер ШІ зазвичай називають штучним суперінтелектом і, як і AGI, він є суто теоретичним. Якщо його коли-небудь буде реалізовано, штучний суперінтелект думатиме, міркуватиме, навчатиметься, виноситиме судження та матиме пізнавальні здібності, які перевершуватимуть людські.
Програми, що володіють можливостями Super AI, розвинуться далі, ніж розуміння людських почуттів і досвіду, щоб відчувати емоції, мати потреби, власні переконання та бажання.

Розглянемо більш детально кожен з цих типів штучного інтелекту.

Вузький ШІ

Штучний інтелект (ШІ) охоплює різноманітний спектр можливостей, які можна загалом класифікувати на дві категорії: слабкий ШІ та сильний ШІ. Слабкий штучний інтелект, який часто називають вузьким штучним інтелектом (Artificial Narrow Intelligence, ANI), втілює в собі системи, ретельно розроблені для виконання конкретних завдань із чітко визначеними параметрами. Ці системи працюють в обмеженому колі знань і не мають можливості для загального інтелекту. Думайте про них як про спеціалістів, навчених ефективно виконувати певні функції.

Незважаючи на свою назву, слабкий штучний інтелект не є слабким; це потужний центр, що стоїть за багатьма програмами штучного інтелекту, з якими ми взаємодіємо щодня. Ми бачимо приклади вузького ШІ навколо нас. Від блискавичних реакцій Siri та Alexa до майстерності IBM Watson у обробці даних і безперебійної навігації безпілотних автомобілів, ANI сприяє видатним інноваціям, які формують наш світ [5-7].

Розглянемо деякі інші приклади додатків, що використовують вузький ШІ, які характеризуються спеціалізованими алгоритмами, розробленими для конкретних завдань:

- **Розумні помічники**

  Цифрові голосові помічники, які часто називають найкращим прикладом слабкого штучного інтелекту, використовують обробку природної мови для конкретних завдань, таких як налаштування будильників, відповіді на запитання та керування пристроями розумного будинку [5].

- **Чат-боти**

  Якщо ви коли-небудь спілкувалися в чаті з вашим улюбленим електронним магазином, швидше за все, ви спілкуєтеся з ШІ. Багато платформ обслуговування клієнтів використовують алгоритми ANI, щоб відповідати на поширені запити, залишаючи людям можливість виконувати завдання вищого рівня [7].

- **Механізми рекомендацій**

  Ви коли-небудь замислювалися, як Netflix завжди знає, який фільм ви хочете переглянути, або як Amazon чи Розетка передбачають вашу наступну покупку? Ці платформи використовують ANI, щоб аналізувати ваші звички перегляду чи купівлі разом зі звичками схожих користувачів, щоб надавати персоналізовані пропозиції [8].

- **Програми для навігації**

  Як дістатися від А до Б, не заблукавши? Навігаційна програма, як-от Google Maps, — це програмна програма, яка використовує ANI, призначену для надання вказівок у реальному часі користувачам під час навігації з одного місця в інше [9].

- **Фільтри електронної пошти для спаму**

  Комп'ютер використовує штучний вузький інтелект, щоб дізнатися, які повідомлення ймовірно є спамом, а потім перенаправляє їх із папки «Вхідні» до папки зі спамом [10].

- **Функції автовиправлення**

  Коли ваш смартфон виправляє ваші помилки під час написання, ви відчуваєте силу слабкого штучного інтелекту, який працює у вашому повсякденному житті [11]. Завдяки використанню алгоритмів і даних користувача ці функції інтелектуального введення тексту забезпечують більш плавну та ефективну композицію тексту на різних пристроях, а функція T9 передбачає, яке слово ми плануємо написати та надає нам підказки.

- **Автопілоти в автомобілях**

  Сьогодні це вже не фантастика, що автомобіль може рухатись без водія. Автопілоти, які використовують алгоритми для керування транспортом на основі даних від сенсорів, але не можуть самостійно приймати рішення в ситуаціях, що виходять за межі навчання [12-13].

Кожна з цих програм демонструє силу ANI для виконання чітко визначених завдань шляхом аналізу великих наборів даних і дотримання спеціалізованих алгоритмів. Отже, наступного разу, коли ви дивуватиметесь можливостям штучного інтелекту, пам'ятайте, що саме слабкий штучний інтелект забезпечує ці чудові інновації, формуючи наш світ у спосіб, який ми колись вважали неможливим.

Сильний ШІ

Сильний ШІ (Strong AI), або загальний ШІ (Artificial General Intelligence, AGI) — це гіпотетична концепція системи, яка може мати інтелект, аналогічний людському, тобто вона здатна виконувати будь-які завдання, що вимагають інтелекту, в різних доменах, включаючи ті, до яких вона не була спеціально навчена.

Таким чином, концепція сильного штучного інтелекту спрямована на розробленя систем, здатних вирішувати широкий спектр завдань із рівнем кваліфікації, який задовольняє людські стандарти. На відміну від вузьких аналогів ШІ, системи сильного ШІ прагнуть володіти формою загального інтелекту, що дозволяє їм адаптуватися, навчатися та застосовувати знання в різних сферах [14]. По суті, мета полягає в тому, щоб створити штучні об'єкти, наділені когнітивними здібностями, подібними до людських, здатні брати участь в інтелектуальних зусиллях, що охоплюють різні сфери.

Хоча сильний штучний інтелект є більш теоретичним, і сьогодні немає практичних прикладів, в яких вже використано системи сильного ШІ, це не означає, що дослідники штучного інтелекту не зайняті вивченням його потенційних розробок.

Поява генеративного ШІ та стрімкий ріст його популярності призвели до чергового сплеску хвилювань щодо того, що ШІ приведе людство до антиутопічного майбутнього. Видатні лідери думок, від діячів Кремнієвої долини до легендарних вчених, попереджають, що якщо штучний інтелект перетвориться на сильний штучний інтелект, який здатний так само навчатися та виконувати інтелектуальні завдання, як і люди, — цивілізація опиниться під серйозною загрозою [15]. І хоча подібні статті активно супроводжуються тривожними ілюстраціями із зображенням роботів-гуманоїдів більшість авторитетних дослідників та науковців схиляються до того, що питання створення реальних систем сильного ШІ навряд чи буде вирішене в найближчі десятиріччя, незважаючи на значні успіхи та прогрес у створенні систем слабкого ШІ.

Які можливості перетворили б ШІ на AGI?

Щоб зрозуміти складність досягнення справжнього людського інтелекту, варто розглянути деякі здібності, якими справжньому AGI потрібно буде оволодіти.

- **Сенсорне сприйняття**

  У той час як глибоке навчання забезпечило значний прогрес у комп'ютерному зорі, системи штучного інтелекту далекі від розвитку людських можливостей сенсорного сприйняття. Наприклад, системи, навчені за допомогою глибокого навчання, все ще мають погану узгодженість кольорів: системи безпілотних автомобілів були обдурені маленькими шматочками чорної стрічки або наклейками на червоному знаку зупинки [16]. Для будь-якої людини почервоніння знака «стоп» все ще є цілком очевидним, але система, заснована на глибокому навчанні, вводить в оману, вважаючи, що знак «стоп» — це щось інше. Сучасні системи комп'ютерного зору також найчастіше не здатні отримувати глибину та тривимірну інформацію зі статичних зображень.

  Люди також можуть визначати просторові характеристики навколишнього середовища за звуком, навіть якщо слухають монофонічний телефонний канал. Ми можемо зрозуміти фоновий шум і сформувати уявну картину того, де хтось знаходиться, коли розмовляємо з ним по телефону (на вулиці, на фоні звуків автомобілів, що рухаються). Системи штучного інтелекту ще не в змозі відтворити чітко це людське сприйняття.

- **Дрібна моторика**

  Будь-яка людина може легко дістати з кишені в'язку ключів. Мало хто з нас дозволив би роботам-маніпуляторам або гуманоїдним рукам виконати це завдання за нас. Дослідники в цій галузі працюють над цією проблемою. Недавня демонстрація показала, як навчання з підкріпленням може навчити руку робота складати кубик Рубіка. Хоча Клод Шеннон створив робота для складання кубика десятиліття тому, ця демонстрація ілюструє спритність програмування пальців робота на одній руці для маніпулювання складним об'єктом.

- **Розуміння природної мови**

  Люди записують і передають навички та знання за допомогою книг, статей, публікацій у блогах, а останнім часом і відео з інструкціями. ШІ повинен мати можливість використовувати ці джерела інформації з повним розумінням. Люди пишуть із неявним припущенням про загальні знання читача, і величезна кількість інформації припускається та не вимовляється. Якщо ШІ не матиме цієї основи здорового глузду, він не зможе працювати в реальному світі [17].

Професори Нью-Йоркського університету Гарі Маркус і Ернест Девіс описують цю вимогу детальніше у книзі «Перезавантаження штучного інтелекту», вказуючи, що ці здорові знання важливі навіть для найпростіших завдань, які виконували системи ШІ. Як зазначає Дуглас Хофстадтер, той факт, що безкоштовні сервіси машинного перекладу стали доволі точними завдяки глибокому навчанню, не означає, що штучний інтелект близький до справжнього розуміння прочитаного, оскільки він не володіє розумінням контексту в кількох реченнях, з чим легко справляються навіть діти раннього віку [18]. Різноманітні звіти про те, як штучний інтелект успішно склав вступні іспити або успішно здав тести з природничих наук у восьмому класі, є кількома прикладами того, як вузьке рішення штучного інтелекту можна легко сплутати з людським інтелектом.

- **Вирішення проблеми**

  У будь-якій програмі загального призначення робот (або механізм штучного інтелекту, що живе в хмарі) повинен буде вміти діагностувати проблеми, а потім вирішувати їх. Домашній робот мав би розпізнати, що лампочка перегоріла, і або замінити лампочку, або повідомити ремонтника. Для виконання цих завдань робот потребує певного аспекту здорового глузду, описаного вище, або здатності запускати симуляції для визначення можливостей, правдоподібності та ймовірностей. Сьогодні жодна відома система не має ані такого здорового глузду, ані такої можливості моделювання загального призначення.

- **Навігація**

  GPS у поєднанні з такими можливостями, як одночасна локалізація та картографування (SLAM), досяг значного прогресу в цій галузі [19]. Проте проектування дій через уявні фізичні простори не є набагато вищим порівняно з поточними можливостями відеоігор. Потрібні роки роботи, щоб створити надійні системи, які можна використовувати без допомоги людини. Сучасні академічні демонстрації не наблизилися до досягнення цієї мети.

- **Творчість**

  Коментатори, які побоюються суперінтелекту, припускають, що як тільки штучний інтелект досягне людського рівня, він швидко вдосконалиться за допомогою процесу початкового завантаження, щоб досягти рівня інтелекту, який набагато перевищує рівень інтелекту будь-якої людини. Але щоб досягти цього самовдосконалення, системам ШІ доведеться переписати свій власний код. Цей рівень самоаналізу вимагатиме від системи

штучного інтелекту розуміння величезної кількості коду, створеного людьми, і визначення нових методів його вдосконалення [20]. Машини продемонстрували здатність малювати картини та складати музику, але для творчості людського рівня потрібен подальший прогрес.

- **Соціальна та емоційна активність**

  Щоб роботи та ШІ були успішними в нашому світі, люди повинні хотіти з ними взаємодіяти, а не боятися їх. Роботу потрібно буде розуміти людей, інтерпретуючи вираз обличчя або зміни в тоні, які виявляють приховані емоції. Деякі обмежені додатки вже використовуються, наприклад, системи, розгорнуті в контакт-центрах, які можуть виявляти, коли голоси клієнтів звучать сердито або стурбовано, і надавати їм допомогу швидше. Але враховуючи власні труднощі людей з правильною інтерпретацією емоцій, а також проблеми зі сприйняттям, які обговорювалися вище, штучний інтелект, здатний до співпереживання, здається віддаленою перспективою [20].

Основні характеристики сильного ШІ:

- **Здатен розуміти та вирішувати широкий спектр завдань так само, як це робить людина.**
- **Має самосвідомість і здатність до міркування, планування та вирішення проблем.**
- **Може вчитися та адаптуватися до нових ситуацій без необхідності додаткового навчання на конкретному завданні.**
- **Може застосовувати свої знання в одній галузі для вирішення проблем в іншій (узагальнення знань).**

Сильний ШІ в теорії:

- **Сьогодні сильного ШІ не існує. Всі сучасні системи — це варіанти слабкого ШІ, навіть найбільш передові моделі, такі як GPT, які можуть справлятися з великою кількістю завдань, але лише в межах певного контексту.**
- **Сильний ШІ вимагає розроблення нових моделей, які можуть відтворювати людське мислення в його повноті, включаючи саморефлексію, емоції, інтуїцію та розуміння контексту.**

В якості можливих прикладів сильного ШІ (якщо він буде створений), можна навести робота, який може самостійно вивчати нові науки та вирішувати завдання на

рівні або краще за людину або систему, здатну вести складну розмову, приймати етичні рішення та розуміти метафори чи абстрактні поняття так, як це робить людина.

3 Штучний суперінтелект (ASI)

У теорії виокремлюють ще третій тип ШІ, який гіпотетично може з'явитись після появи сильного ШІ – це штучний суперінтелект (рис. 2).

Штучний суперінтелект (Artificial superintelligence, ASI) — це тип ШІ, який перевершує людський інтелект і може виконувати будь-які завдання краще, ніж людина. Системи ASI не тільки розуміють людські почуття та досвід, але також можуть викликати власні емоції, переконання та бажання, подібні до людських.

Хоча існування ASI все ще є гіпотетичним, очікується, що здатність таких систем приймати рішення та вирішувати проблеми буде набагато кращою, ніж людські. Система ASI зможе думати, розв'язувати головоломки, робити судження та приймати рішення самостійно [21].

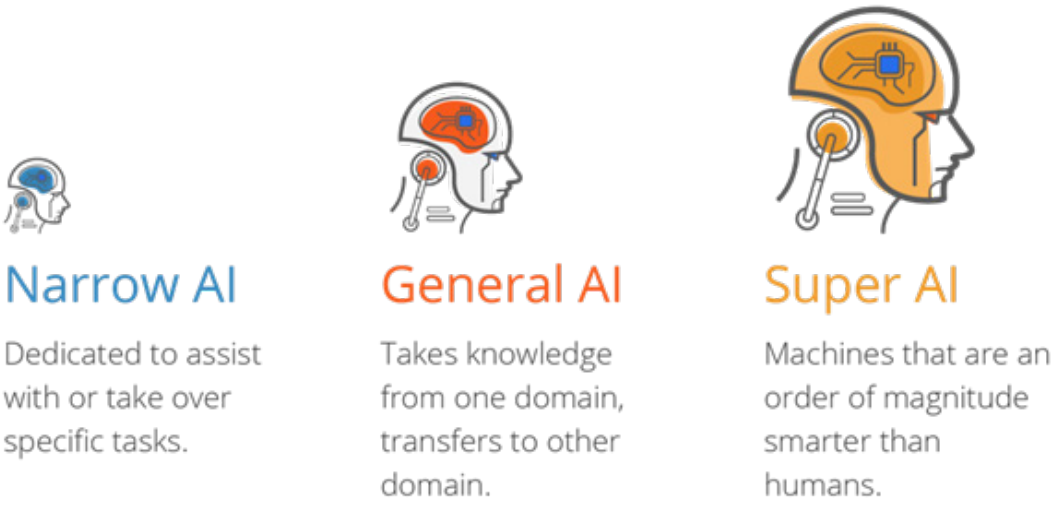


Рис.1.2. Три типи ШІ на основі можливостей

Таблиця 1

Порівняння різних типів ШІ

Категорія	Вузький ШІ	Сильний ШІ	Супер ШІ
Визначення	Вузький штучний інтелект зосереджений на конкретному, окремому або цілеспрямованому завданні та не має функціональних можливостей саморозширення для вирішення незнайомих проблем.	Сильний ШІ може виконувати широкий спектр завдань, міркувати, навчатися та покращувати когнітивні здібності, порівнянні з людськими.	Штучний супер інтелект демонструє інтелект, який перевищує людські можливості.
Призначення	Вузький штучний інтелект запрограмований на роботу в межах набору попередньо визначених функцій для усунення або вирішення конкретної проблеми.	Сильний ШІ матиме власний розум і зможе виконувати будь-які завдання, які може передбачити його «розум».	Штучний супер інтелект перевершить людський інтелект, щоб виконати будь-яке завдання краще, ніж його людські аналоги.
Модель ШІ	Вузький штучний інтелект використовує запрограмовані моделі фіксованої області.	Сильний ШІ самонавчається та обґрунтовує робоче середовище.	Штучний супер інтелект самонавчається та розвивається з власною свідомістю.
Самосвідомість	Вузькому штучному інтелекту бракує самосвідомості, штучної свідомості чи когнітивних здібностей.	Сильний штучний інтелект вважатиметься справді передовим, розумним і повністю самосвідомим, що означає, що він матиме здоровий глузд, креативність і здатність виражати емоції.	Супер ШІ моделюватиме людські міркування та досвід, щоб розвинути власне емоційне розуміння, переконання та бажання.

Категорія	Вузький ШІ	Сильний ШІ	Супер ШІ
Обробка даних	Вузький штучний інтелект класифікує дані за допомогою машинного навчання, обробки природної мови, штучних нейронних мереж і глибокого навчання.	Сильний ШІ використовує кластеризацію та асоціацію за допомогою передових версій машинного навчання, глибокого навчання, НЛП і штучних нейронних мереж.	Супер ШІ може використовувати людський мозок як модель для виявлення поведінкового інтелекту та розуміння й інтерпретації людських емоцій і переживань.
Передача знань	Вузький ШІ не сприяє передаванню знань в інші області чи завдання.	Сильний ШІ використовує передачу знань у нові області та завдання.	Супер ШІ незмінно здійснюватиме передачу знань між завданнями та доменами.
Наслідки	Вузький ШІ перевершує людей у конкретних повторюваних завданнях, таких як водіння, медична діагностика та фінансові консультації.	Сильний ШІ конкурує з людьми в усіх починаннях: від отримання університетського диплому до лікування невідкладних медичних випадків.	Супер ШІ перевершує людей у досягненні суспільних цілей і сприянні дослідженню космосу, але також загрожує самому існуванню людської раси.
Етап ШІ	Сьогоднішній ШІ	Штучний інтелект майбутнього – за різними оцінками від 2030 до 2060 року	Незабаром після AGI

Однак, виникає питання – як визначити, до якого типу систем ШІ належить розроблена система, та й взагалі – чи є вона системою ШІ? Це питання виникало перед дослідниками і людством ще з часів створення перших систем ШІ у 1940-50 х. рр. Для відповіді на це питання британський математик і вчений Алан Тюрінг, засновник ШІ, у 1950 році запропонував експеримент, який мав на меті визначити, чи може машина демонструвати ознаки інтелекту, еквівалентні або невідмінні від людського. Цей експеримент відомий за назвою Тест Тюрінга і був розроблений для відповіді на питання: «Чи можуть машини мислити?» [22]. Оскільки «мислення» — це складне і суб'єктивне поняття, Тюрінг запропонував простий критерій для оцінки інтелекту машин: якщо машина може вести розмову, і співрозмовники не можуть відрізнити її від людини, то її можна вважати інтелектуальною.

Тест полягає в тому, що людина (експерт) спілкується одночасно з двома суб'єктами — одним із них є людина, а іншим — машина. Спілкування відбувається через текстовий інтерфейс, так що експерт не може бачити або чути учасників, лише читати їхні відповіді на свої питання.

Механіка тесту:

1. **Три учасники:** людина-експерт, машина, і ще одна людина.
2. **Експерт задає питання** обом учасникам (машині та людині) через текстовий канал, де ні голос, ні вигляд не впливають на оцінку.
3. **Задача машини:** імітувати поведінку людини настільки добре, щоб експерт не зміг відрізнити її від справжньої людини.
4. **Задача експерта:** визначити, який із двох співрозмовників є машиною, а хто — людиною.

Якщо експерт не може чітко відрізнити машину від людини або помилково ідентифікує машину як людину, тоді вважається, що машина пройшла тест Тюрінга.

Тест Тюрінга мав на меті запропонувати практичний підхід до питання штучного інтелекту: не фокусуючись на тому, що таке свідомість або мислення, Тюрінг запропонував оцінювати лише поведінкові прояви. Якщо машина демонструє поведінку, подібну до людської, то можна припустити, що вона «мислить» [22].

Попри широку популярність та застосовуваність тесту Тюрінга деякі дослідники визнали недоліки запропонованого підходу. Зокрема такі, як обмеженість тесту, яка полягає в тому, що машина може пройти тест, не володіючи справжнім інтелектом або свідомістю. Вона може імітувати поведінку людини без реального розуміння контексту або значення. Також серед слабких місць тесту Тюрінга вказують на його спрямованість

на мовні навички, адже тест оцінює лише здатність машини вести розмову, але не охоплює багато інших аспектів інтелекту, таких як здатність до навчання, розв'язання задач або маніпуляція об'єктами в реальному світі [23].

Експериментом, що мав довести недосконаліть тесту Тюрінга, стала «Китайська кімната» Джона Серла: Філософ Джон Серл стверджував, що машина може «проходити» тест Тюрінга, просто слідуючи правилам для створення відповідей, не маючи справжнього розуміння того, що вона говорить [24]. Його аргумент полягав у тому, що правильні відповіді не свідчать про справжній інтелект або розуміння.

Сьогодні тест Тюрінга залишається класичною концепцією та широко використовується для визначення «інтелекту» машин. Однак, вчені розглядають й інші критерії для оцінки штучного інтелекту, такі як здатність до навчання, адаптація, робота з фізичними об'єктами та пояснюваність рішень.

Тест Тюрінга залишається важливою точкою відліку в дискусіях про штучний інтелект, але сучасний ШІ розвивається в напрямку більш комплексних і практичних перевірок інтелекту [23].

Зокрема Роджер Брукс, відомий дослідник у галузі робототехніки та штучного інтелекту, запропонував альтернативу класичному тесту Тюрінга, яка зосереджується на фізичній взаємодії штучного інтелекту з реальним світом. Брукс вважав, що штучний інтелект не можна оцінювати лише на основі здатності вести бесіду або імітувати людське мислення в текстовій формі. Замість цього він стверджував, що справжній інтелект повинен демонструвати здатність до взаємодії з фізичним світом, орієнтуватися в ньому та вирішувати завдання в реальних умовах.

Брукс запропонував такий критерій: штучний інтелект повинен бути втілений у фізичному тілі (роботі) і здатен діяти в складному, динамічному середовищі так, як це робить людина або жива істота. Він наголошував на тому, що важливою частиною інтелекту є можливість орієнтуватися в просторі, розпізнавати об'єкти, маніпулювати ними, а також адаптуватися до змін. Наприклад, робот, здатний пройти через кімнату, уникнути перешкод, взяти об'єкт, перенести його і використовувати для вирішення проблеми — це приклад втілення інтелекту за критеріями Брукса.

Брукс називав цей підхід «інтелектом, заснованим на поведінці» (behavior-based AI). Це підхід, за яким інтелект формується через взаємодію з реальним світом, а не через виконання абстрактних або текстових завдань. Тобто, Брукс віддавав перевагу практичному, фізично втіленому інтелекту, а не суто когнітивним або мовним перевіркам, як у тесті Тюрінга.

Роджер Брукс у своїй критиці традиційних підходів до штучного інтелекту запропонував чотири способи вимірювання прогресу в розвитку ШІ, що зосереджуються на практичних аспектах, пов'язаних із взаємодією з фізичним світом. Ці способи наголошують на важливості втілення ШІ в роботах та системах, що можуть ефективно функціонувати в реальному середовищі, а не лише у віртуальному або текстовому контексті.

1. Мобільність (Mobility)

Брукс вважав, що здатність ШІ (робота) до переміщення в фізичному світі є ключовим фактором для вимірювання інтелекту. Робот має бути здатним до орієнтування

в просторі, розпізнавання перешкод і вибору оптимальних шляхів для пересування – наприклад, автономні роботи, що можуть переміщатися в складних середовищах (будівельні роботи або роботи-прибиральники).

2. Маніпуляція (Manipulation)

Здатність системи ШІ взаємодіяти з фізичними об'єктами, маніпулювати ними та використовувати їх для досягнення цілей також є критерієм прогресу. Зокрема, це включає здатність брати об'єкти, аналізувати їхні характеристики (вагу, форму, текстуру) та використовувати їх для виконання складних завдань. Наприклад, роботи, які здатні виконувати завдання на складальних лініях або в хірургії.

3. Соціальна взаємодія (Social Interaction)

ШІ повинен бути здатним до комунікації та соціальної взаємодії з людьми або іншими роботами. Це включає розуміння соціальних сигналів, мови, жестів і реагування на них відповідним чином. Така взаємодія є важливою для систем, які працюють у команді або обслуговують людей. Наприклад, роботи-асистенти, що можуть допомагати літнім людям або працювати в сфері обслуговування.

4. Навчання та адаптація (Learning and Adaptation)

Здатність системи набувати нових навичок і адаптуватися до змін у середовищі — ще один важливий спосіб вимірювання прогресу. ШІ повинен не тільки виконувати заздалегідь запрограмовані завдання, а й вчитися на основі нових даних та досвіду. Це передбачає здатність робота змінювати свою поведінку, коли він стикається з новими умовами або непередбаченими перешкодами. Наприклад, робот, який може вчитися новим рухам або коригувати свою стратегію під час гри в складну гру, таку як футбол.

1.3. Чотири типи ШІ на основі функцій

Штучний інтелект охоплює широкий спектр можливостей, кожна з яких виконує різні функції та цілі. Розуміння чотирьох типів ШІ проливає світло на еволюцію машинного інтелекту:

Перші дві функціональні категорії ШІ характерні для вузького ШІ – це ШІ реактивної машини та ШІ з обмеженою пам'яттю:

1. ШІ реактивної машини

Реактивні машини — це системи штучного інтелекту без пам'яті та призначені для виконання дуже специфічних завдань. Оскільки вони не можуть згадати попередні

результати чи рішення, вони працюють лише з наявними на даний момент даними. Реактивний штучний інтелект походить від статистичної математики та може аналізувати величезні обсяги даних для отримання, здавалося б, інтелектуального результату.

Приклади ШІ реактивної машини

- **IBM Deep Blue:** шаховий суперкомп'ютер IBM переміг шахового гросмейстера Гаррі Каспарова наприкінці 1990-х років, аналізуючи фігури на дошці та передбачаючи ймовірні результати кожного ходу.
- **Механізм рекомендацій Netflix.** Рекомендації Netflix щодо перегляду базуються на моделях, які обробляють набори даних, зібрані з історії переглядів, щоб надати клієнтам вміст, який їм найбільше сподобається.

2. ШІ з обмеженою пам'яттю

На відміну від реактивного машинного штучного інтелекту, ця форма штучного інтелекту може згадувати минулі події та результати та контролювати конкретні об'єкти чи ситуації в режимі реального часу. Штучний інтелект з обмеженою пам'яттю може використовувати дані минулого та поточного моментів, щоб прийняти рішення про напрямок дій, який, швидше за все, допоможе досягти бажаного результату.

Однак, хоча штучний інтелект з обмеженою пам'яттю може використовувати минулі дані протягом певного періоду часу, він не може зберігати ці дані в бібліотеці минулого досвіду для використання протягом тривалого періоду. Оскільки ШІ з обмеженою пам'яттю з часом навчається на більшій кількості даних, він може покращити продуктивність.

Приклади ШІ з обмеженою пам'яттю

- **Генеративний штучний інтелект**

Генеративні інструменти штучного інтелекту, такі як ChatGPT, Bard і DeepAI, покладаються на можливості штучного інтелекту з обмеженою пам'яттю, щоб передбачити наступне слово, фразу чи візуальний елемент у контенті, який він генерує.

- **Віртуальні помічники та чат-боти**

Siri, Alexa, Google Assistant, Cortana та IBM Watson Assistant поєднують обробку природної мови (NLP) та штучний інтелект з обмеженою пам'яттю, щоб розуміти запитання та запити, виконувати відповідні дії та давати відповіді.

- **Автономні автомобілі**

Автономні транспортні засоби використовують ШІ з обмеженою пам'яттю, щоб розуміти навколишній світ у режимі реального часу та приймати зважені рішення щодо того, коли застосовувати швидкість, гальмувати, повертати тощо.

3. Теорія розуму ШІ

Теорія розуму ШІ (Theory of Mind AI) — це функціональний клас ШІ, який підпадає під визначення «Сильний ШІ». Незважаючи на те, що сьогодні це нереалізована форма штучного інтелекту, штучний інтелект із функцією Theory of Mind міг би розуміти думки й емоції інших суб'єктів. Це розуміння може вплинути на те, як ШІ взаємодіє з оточуючими. Теоретично це дозволить штучному інтелекту симулювати людські стосунки. Розроблення штучного інтелекту на основі теорії розуму може революціонізувати широкий спектр сфер, включаючи взаємодію людини з комп'ютером і соціальну робототехніку, забезпечивши більш чуйну та інтуїтивну поведінку машин.

Оскільки Теорія розуму ШІ могла б визначити людські мотиви та міркування, вона персоналізувала б свою взаємодію з людьми на основі їхніх унікальних емоційних потреб і намірів. Theory of Mind AI також міг би зрозуміти та контекстуалізувати художні твори та есе, чого не можуть зробити сучасні генеративні інструменти AI.

Емоційний ШІ — це теорія розумового ШІ, яка зараз розробляється. Дослідники штучного інтелекту сподіваються, що він матиме здатність аналізувати голоси, зображення та інші види даних, щоб розпізнавати, симулювати, контролювати та адекватно реагувати на людей на емоційному рівні. Сьогодні Emotion AI не в змозі зрозуміти людські почуття та реагувати на них.

4. Самосвідомий ШІ

Самосвідомий ШІ (Self-Aware AI) — це свого роду функціональний клас AI для додатків, які володіють суперможливостями AI. Як і теорія розуму штучного інтелекту, самосвідомий ШІ є суто теоретичним. Якщо колись цього буде досягнуто, він матиме здатність розуміти власні внутрішні стани та риси, а також людські емоції та думки. Він також матиме власний набір емоцій, потреб і переконань. Поки що ці типи штучного інтелекту зустрічаються лише у фантастичному світі наукової фантастики, популяризованої культовими фільмами, такими як « Той, що біжить по лезу».

Додаткові можливості та практичне застосування технологій ШІ

Розглянемо детальніше типи та класи, на які поділяються системи ШІ, а також які методи вони використовують (рис.1.3.).

Рис. 1.3. Класифікація систем ШІ

Системи комп'ютерного зору

Системи вузького ШІ з комп'ютерним зором можна навчити інтерпретувати та аналізувати візуальний світ. Це дозволяє інтелектуальним машинам ідентифікувати та класифікувати об'єкти на зображеннях і відео [36].

Застосування комп'ютерного зору передбачає:

- Розпізнавання та класифікацію зображень
- Виявлення об'єктів
- Відстеження об'єктів
- Розпізнавання обличчя
- Пошук зображень на основі вмісту

Комп'ютерний зір має вирішальне значення для сценаріїв використання, які передбачають взаємодію машин зі штучним інтелектом і взаємодію із фізичним світом

навколо них. Приклади включають безпілотні автомобілі та машини, які керують складами та іншими середовищами [38].

Робототехніка

Роботи в промислових умовах можуть використовувати вузький штучний інтелект для виконання рутинних, повторюваних завдань, такі як обробка матеріалів, складання товару та перевірка якості [40]. У сфері охорони здоров'я роботи, оснащені вузьким штучним інтелектом, можуть допомогти хірургам контролювати життєво важливі функції та виявляти потенційні проблеми під час процедур [37].

Агротехніка може займатися автономною обрізкою, переміщенням, проріджуванням, посівом і обприскуванням. А розумні пристрої для дому, такі як iRobot Roomba, можуть орієнтуватися в інтер'єрі будинку за допомогою комп'ютерного зору та використовувати дані, що зберігаються в пам'яті, для розуміння прогресу.

Експертні системи

Експертні системи, оснащені можливостями вузького штучного інтелекту, можна навчити на сформованій експертом базі знань, щоб імітувати процес прийняття рішень людиною та застосовувати досвід для вирішення складних проблем. Ці системи можуть оцінювати величезні масиви даних, щоб виявити тенденції та закономірності для прийняття рішень. Вони також можуть допомогти підприємствам передбачити майбутні події та зрозуміти, чому відбулися минулі події [36].

Машинне та глибоке навчання

Центральними для цих досягнень є машинне та глибоке навчання: два підкласи штучного інтелекту, які стимулюють багато інновацій, які ми бачимо сьогодні. Незважаючи на те, що ці терміни пов'язані, кожен із них має своє окреме значення [37].

- **Контрольоване навчання** (Supervised learning – навчання із вчителем)

Під час контрольованого навчання алгоритм навчається на попередньо визначеному наборі даних, де кожен приклад має вхідні та відповідні вихідні дані, для того, щоб робити прогнози щодо нових, невідомих даних.

- **Неконтрольоване навчання** (Unsupervised learning – навчання без вчителя)

Навчальні дані за такого підходу не мають будь-яких попередньо визначених міток або вихідних даних – алгоритм навчається виявляти приховані залежності, структури чи групи в даних.

- **Навчання з підкріпленням** (Reinforcement learning)

Створюється агент, який може взаємодіяти з навколишнім середовищем і навчатися методом проб і помилок. У процесі навчання та функціонування агент отримує зво-

ротний зв'язок у формі винагород або штрафів під час виконання дій, які дозволяють йому навчатися та покращувати ефективність з часом.

Глибоке навчання (Deep Learning) – це підмножина машинного навчання, зосереджена на навчанні штучних нейронних мереж із кількома рівнями, натхненними структурою та функціями людського мозку, які складаються із взаємопов'язаних вузлів (нейронів), між якими передаються сигнали [37].

Завдяки автоматичному видобуванню взаємозв'язків та залежностей із необроблених даних за допомогою кількох рівнів абстракції ці алгоритми штучного інтелекту чудово справляються з розпізнаванням зображень і мови, обробкою природної мови та багатьма іншими сферами. Глибоке навчання може використовувати великі набори даних із вхідними даними великої розмірності, але через їх складність вимагає значної обчислювальної потужності та тривалого навчання.

Серед найпоширеніших методів глибокого навчання можна виокремити такі:

1. Глибокі нейронні мережі (Deep Neural Networks, DNN) –

моделі з багатьма прихованими шарами, кожен з яких відповідає за виявлення складніших патернів. Такі архітектури, як MLP (Multilayer Perceptron), використовуються для задач класифікації та регресії, однак мають обмеження в роботі зі складними структурованими даними (зображення, мова) [37].

2. Згорткові нейронні мережі (Convolutional Neural Networks, CNN) –

використовуються для обробки даних із просторовою структурою, зокрема зображень та відео. Згорткові нейронні мережі забезпечують більш масштабований підхід до задач класифікації зображень і розпізнавання об'єктів порівняно з іншими ШНМ, використовуючи принципи лінійної алгебри, зокрема множення матриць, для ідентифікації шаблонів у зображенні [38]. Тим не менш, вони можуть бути вимогливими до обчислень, вимагаючи графічних процесорів (GPU) для навчання моделей. До найпоширеніших прикладів застосування належать алгоритми розпізнавання облич, обробка медичних зображень (аналіз рентгенівських знімків) тощо.

3. Рекурентні нейронні мережі (RNN) і LSTM (Long Short-Term Memory) –

архітектури, що використовуються для послідовних даних, таких як часові ряди або природна мова. LSTM вирішує проблему зникнення градієнтів у довгих послідовностях завдяки спеціальній структурі пам'яті [37-39].

Найпоширенішими прикладами застосування є створення моделей для аналізу текстів та машинного перекладу.

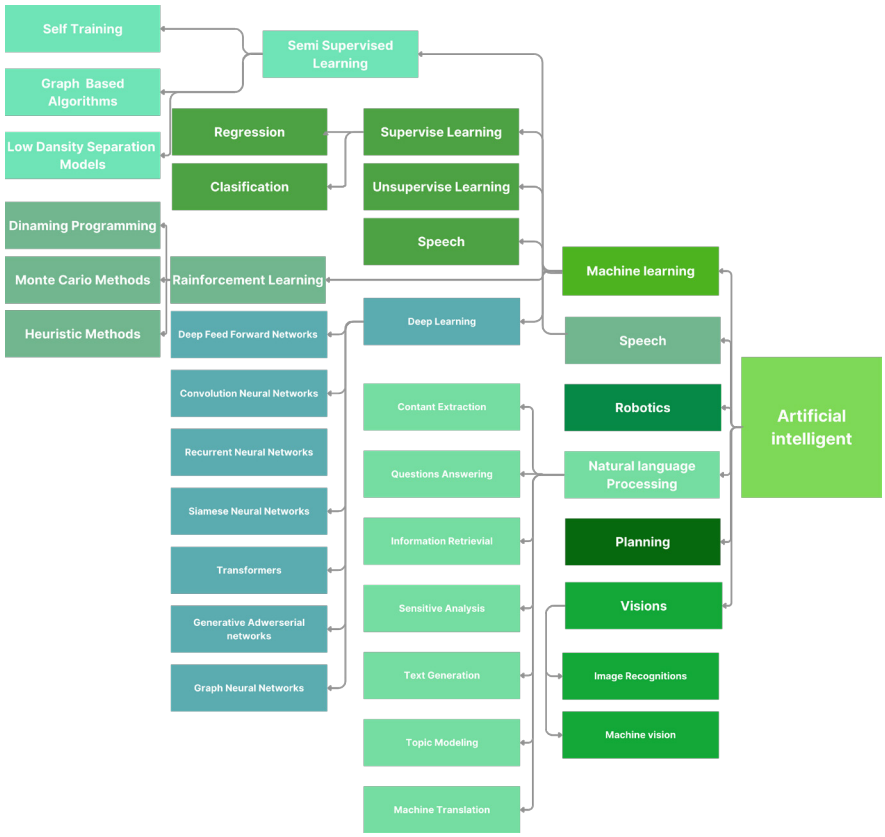


Рис.1.4. Класифікація методів ШІ

Які досягнення могли б прискорити еволюцію систем ШІ?

Зменшення витрат на зберігання інформації за останні два десятиліття призвело до появи концепції «великих даних». Обчислювальний прогрес у графічних процесорах дозволяє унікальним чином застосовувати алгоритм до набагато більших нейронних мереж. Завдяки цим нейронним мережам, навченим на дуже великих наборах даних, науковці досягли надзвичайно великих успіхів у галузі глибокого навчання. Поєднання даних, алгоритмів і обчислювальних досягнень спричинило переломний момент у розвитку систем ШІ. Щоб знайти наступну потенційну точку перелому та стрибка у розвитку штучного інтелекту, корисно знову розглянути ландшафт, використовуючи ці три компоненти.

Основні алгоритмічні досягнення та нові робототехнічні підходи.

Цілком можливо, що знадобляться абсолютно нові підходи, щоб наблизити ШІ до рівня собаки чи дворічної дитини. Одним із прикладів, який вивчають дослідники, є концепція втіленого пізнання. Їхня гіпотеза полягає в тому, що роботам потрібно буде

вчитися в навколишньому середовищі за допомогою безлічі органів чуття, подібно до того, як це роблять люди на ранніх стадіях життя, і що вони повинні будуть відчувати фізичний світ через тіло, схоже на людське, щоб когнітивно розвиватися так само, як і люди. У зв'язку з тим, що фізичний світ вже розроблений навколо людей, цей підхід має переваги. Це позбавляє нас від необхідності переробляти наші фізичні інтерфейси — від дверних ручок до сходів і кнопок ліфта [40].

Весь прогрес у глибокому навчанні досягається завдяки алгоритму зворотного поширення похибки, який дозволяє великим і складним нейронним мережам навчатися на навчальних даних. У 1986 році Хінтон разом із колегами Девідом Румельхартом і Рональдом Вільямсом опублікував «Навчання внутрішнім представленням через помилки зворотного поширення» [41]. Знадобилося ще 26 років, перш ніж збільшення обчислювальної потужності та зростання «великих даних» дозволили використовувати це відкриття в сьогоднішньому масштабі. Незважаючи на те, що безліч дослідників удосконалювали спосіб використання зворотного поширення в глибокому навчанні, жодне з цих удосконалень не трансформувало таким же чином. (Нещодавня робота Хінтона над «капсульними мережами» цілком може бути одним із таких алгоритмічних досягнень, який міг би, серед інших застосувань, подолати обмеження сучасних нейронних мереж у машинному зорі).

Глибоке навчання передбачає стан «чистого аркуша», і весь «інтелект» можна отримати з навчальних даних. Будь-хто, хто хоч раз спостерігав, як народжуються ссавці, зрозуміє, що щось на зразок оленя починає життя з певним рівнем вбудованих знань. Він стає на ноги протягом 10 хвилин, за кілька годин починає годуватись і ходити. Як зазначають Маркус і Девіс у Rebooting AI, «справжній прогрес у ШІ, на нашу думку, розпочнеться з розуміння того, які типи знань і уявлень потрібно вбудувати до навчання, щоб запустити решту». Недавній успіх глибокого навчання, можливо, відвернув увагу дослідників від більш фундаментальної когнітивної роботи, необхідної для досягнення прогресу в AGI [37].

Основні досягнення в обчислювальній техніці.

Застосування графічних процесорів для навчання глибоких нейронних мереж стало критичним кроком, завдяки якому стали можливими великі досягнення за останні кілька років. Графічні процесори дозволили паралельно застосовувати складні обчислення, необхідні для алгоритму зворотного поширення, що дало можливість навчати надзвичайно складні нейронні мережі за кінцевий час. Перш ніж очікувати будь-якого подальшого експоненціального зростання в бік AGI, подібна переломна точка в обчислювальній інфраструктурі повинна бути узгоджена з унікальними алгоритмічними досягненнями [38].

Квантові обчислення зараз розглядають як одне із потенційних обчислювальних досягнень, яке зможе змінити наше суспільство. Але квантові обчислення пропонуються не як заміна сучасним пристроям, а для надскладних статистичних проблем, з якими не можуть впоратися поточні обчислювальні потужності. Більше того, перший реальний доказ того, що квантові комп'ютери можуть впоратися з такими проблемами, з'явився лише наприкінці 2019 року, і лише для суто математичних вправ, які взагалі не

використовувалися в реальному світі [39]. Апаратне та програмне забезпечення для вирішення таких проблем, як ті, які потрібні для розвитку штучного інтелекту, можуть з'явитися не раніше 2035 року або пізніше. Тим не менш, квантові обчислення залишаються однією з найбільш імовірних можливих переломних точок, за якими слід уважно стежити.

Значне зростання обсягу даних із нових джерел.

Розгортання мобільної інфраструктури 5G є одним із технологічних досягнень, які позиціонуються як такі, що забезпечать значне збільшення даних завдяки тому, як технологія може сприяти розвитку Інтернету речей (IoT). Проте проведені дослідження виявили перепони на шляху впровадження 5G, особливо в економічному плані для операторів. Крім того, в опитуванні 2019 року оператори повідомили, що вони не бачать IoT як основну мету для 5G, оскільки наявні можливості IoT, ймовірно, достатні для більшості випадків використання. Як наслідок, малоймовірно, що 5G сам по собі стане основною переломною точкою для збільшення обсягу даних і подальшого покращення навчальних даних. Більшість переваг, можливо, вже з'явилися.

Нові підходи до робототехніки можуть створити нові джерела навчальних даних. Наприклад, розміщення людиноподібних роботів навіть з лише основними функціями серед людей, якщо кількість таких роботів буде значною, дасть змогу отримувати великі набори даних, які імітують наші власні органи чуття, можуть допомогти замкнути цикл зворотного зв'язку навчання, що покращить сучасний рівень. Удосконалені безпілотні автомобілі – один із таких прикладів: дані, зібрані автомобілями, які вже є на ринку, діють як навчальний набір для майбутніх пристроїв із автономним керуванням [41]. Крім того, проводиться багато досліджень щодо взаємодії людини з роботом. Знайшовши початкові варіанти використання людиноподібних роботів, це дослідження могло б значно додати навчальні дані, необхідні для розширення їхніх можливостей.

САМОКОНТРОЛЬ

Запитання для самоконтролю:

1. Що таке штучний інтелект (ШІ) і які основні завдання він може вирішувати сьогодні?
2. Як ШІ визначається згідно зі стандартом ISO/IEC 22989:2022?
3. Які приклади вузького штучного інтелекту (ANI) ми використовуємо в повсякденному житті?
4. Яка різниця між вузьким ШІ, загальним ШІ та супер ШІ?
5. Чим генеративний ШІ відрізняється від традиційного ШІ?
6. Яка роль даних у навчанні систем ШІ?
7. Чому важливо правильно налаштувати та навчити алгоритми ШІ для досягнення найкращих результатів?
8. Як глибоке навчання впливає на здатність ШІ навчатися виконувати нові завдання без втручання людини?
9. Які типи завдань можна автоматизувати за допомогою ШІ?
10. Які ризики та виклики пов'язані з розвитком генеративного ШІ?

Контрольні запитання:

Перший рівень складності

Тестові питання:

1.1. Яка основна відмінність між традиційним машинним навчанням та глибоким навчанням?

a) Глибоке навчання використовує ручне програмування, а машинне навчання - нейронні мережі

b) Глибоке навчання базується на великих нейронних мережах, тоді як машинне навчання використовує простіші алгоритми

c) Машинне навчання не потребує даних для навчання, а глибоке навчання потребує великих обсягів даних

d) Глибоке навчання більше залежить від втручання людини, ніж машинне навчання

1.2. Що таке вузький штучний інтелект (ANI)?

a) Це тип ШІ, що перевищує людський інтелект

b) Це системи, що виконують одне конкретне завдання з високою ефективністю

c) Це загальна концепція, яка дозволяє машинам вирішувати всі типи завдань

d) Це гіпотетична система ШІ, що володіє емоціями та самосвідомістю

1.3. Як штучний інтелект використовується в системах прогнозного обслуговування?

a) ШІ аналізує великі дані для прогнозування поломок обладнання

b) ШІ замінює всіх технічних спеціалістів

c) ШІ лише автоматизує планування робочих графіків

d) ШІ приймає рішення без необхідності у навчанні

1.4. Яка з наведених технологій є прикладом генеративного ШІ?

a) Голосові помічники (Siri, Alexa)

b) Генерація текстів, зображень та музики (ChatGPT, DALL·E)

c) Механізми рекомендацій фільмів та серіалів

d) Системи для фільтрації спаму в електронній пошті

1.5. Що є прикладом автоматизації завдань за допомогою ШІ?

a) Написання дослідницьких статей вручну

b) Автоматичне розпізнавання облич на фотографіях

c) Виконання лише ручних операцій на виробництві

d) Участь людей у прийнятті рішень для кожного процесу

Другий рівень складності

2.1. Як визначається ШІ у контексті сучасних технологій і що він включає?

2.2. Які основні етапи еволюції ШІ ви можете назвати?

2.3. У чому різниця між слабким і сильним штучним інтелектом?

2.4. Як штучні нейронні мережі сприяли розвитку ШІ після 2012 року?

2.5. Наведіть приклади систем вузького штучного інтелекту та опишіть їх функції.

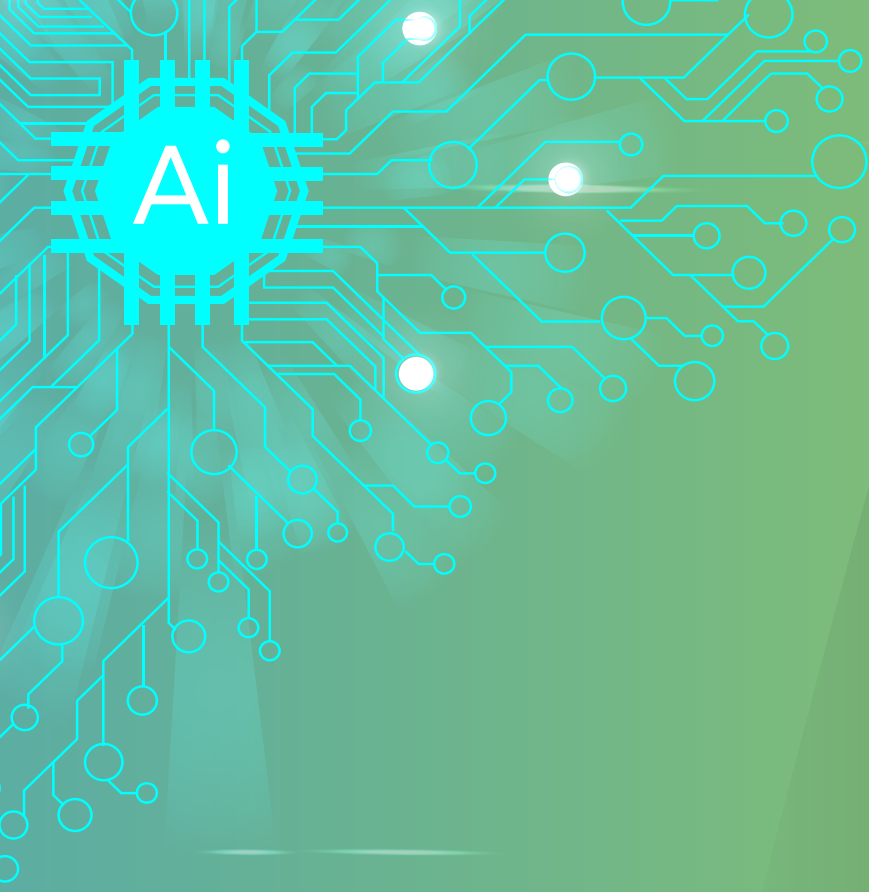
2.6. Як генеративний ШІ використовує дані для створення нового контенту?

2.7. Які характеристики має штучний інтелект, здатний до навчання без втручання людини?

2.8. Які галузі найбільше використовують технології ШІ і як це впливає на їхню діяльність?

2.9. Чому сильний ШІ залишається теоретичним концептом, і які перспективи його розвитку?

2.10. Як ШІ може перевершити людський інтелект у контексті суперінтелекту?



2

ПЕРЕДУМОВИ ВИНИКНЕННЯ НАДІЙНОГО ШТУЧНОГО ІНТЕЛЕКТУ

Зростаюче використання систем штучного інтелекту, а особливо стрімке поширення та повсюдне використання систем генеративного ШІ, посилили дискусії щодо справедливості та упередженості в штучному інтелекті, оскільки потенційні упередження та дискримінація стають більш очевидними, а ризики від них більш значущими. У цьому розділі ми розглянемо джерела можливих ризиків, що здатні порушувати права та свободи користувачів, наслідки, пов'язані зі несправедливістю та упередженістю в ШІ та стратегії їх пом'якшення чи уникнення [42]. Останнім часом дослідження виявляють все більше прикладів упередження щодо певних груп в системах штучного інтелекту, які здійснюють розпізнавання обличчя, застосовуються в процесі найму чи інших скорингових системах. Ці упередження можуть увічнити системну дискримінацію та нерівність із згубним впливом на окремих осіб і громади в таких сферах, як найм, кредитування та кримінальне правосуддя. Відповідно, дослідники та практики розробляють різні стратегії зменшення потенційних ризиків, такі як покращення якості даних чи використання явно справедливих алгоритмів для прозорості використання систем ШІ [43].

Розглянемо джерела і наслідки упередженості в штучному інтелекті, що пов'язані із навчальними даними, алгоритмами, що використовуються та упередження користувачів, а також їхні етичні наслідки.

2.1. Джерела упередженості в ШІ. Визначення упередження в ШІ та його типи

Штучний інтелект має потенціал революціонізувати багато галузей і покращити життя людей незліченною кількістю способів. Однак однією з основних проблем, які виникають у процесі розробки та розгортання систем штучного інтелекту, є наявність упередженості. Під упередженістю матимемо на увазі появу систематичних помилок, які виникають у процесі прийняття рішень, що призводить до несправедливих результатів. У контексті штучного інтелекту упередженість може виникати з різних джерел, включаючи збір даних, розробку алгоритму та людську інтерпретацію [44]. Моделі машинного навчання, які є різновидом систем штучного інтелекту, можуть виявляти та відтворювати шаблони упередженості, наявні в даних, які використовуються для їх навчання, що призводить до несправедливих або дискримінаційних результатів. Важливо виявити та усунути упередженість у ШІ, щоб переконатися, що розроблювана система ШІ є справедливою та є однаково справедливою для всіх користувачів.

Розглянемо більш детально джерела, наслідки та стратегії пом'якшення та уникнення упередженості в ШІ.

2.1.1. Класифікація упередженості в системах ШІ залежно від джерела зміщення (Bias)

Джерела упередженості в ШІ можуть виникати на різних етапах життєвого циклу

машинного навчання, включаючи збір даних, розробку алгоритму та взаємодію з користувачем.

Зміщення даних (Data Bias) виникає, коли дані, які використовуються для навчання моделей машинного навчання, є нерепрезентативними або неповними, що призводить до упереджених результатів. Це може статися, коли дані збираються з упереджених джерел або коли дані неповні, не мають важливої інформації або містять помилки [45].

Алгоритмічне зміщення (Algorithmic Bias), яке виникає, коли алгоритми, які використовуються в моделях машинного навчання, мають властиві зміщення, які відображаються на їхніх результатах. Це може статися, коли алгоритми базуються на упереджених припущеннях або коли вони використовують упереджені критерії для прийняття рішень [46].

Упередженість користувачів (User Bias) виникає, коли люди, які використовують системи штучного інтелекту, свідомо чи несвідомо вносять у систему власні переконання чи упередження. Це може статися, коли користувачі надають упереджені навчальні дані або коли вони взаємодіють із системою у спосіб, який відображає їхні власні упередження [47].

Щоб пом'якшити ці джерела упередженості, були запропоновані різні підходи, включаючи розширення набору даних, алгоритми з урахуванням упередженості та механізми зворотного зв'язку з користувачами. Розширення набору даних передбачає додавання більш різноманітних даних до навчальних наборів даних для підвищення репрезентативності та зменшення упередженості. Алгоритми з урахуванням зміщень включають розробку алгоритмів, які враховують різні типи зміщень і спрямовані на мінімізацію їх впливу на результати системи. Механізми зворотного зв'язку з користувачами передбачають запит відгуків від користувачів, щоб допомогти виявити та виправити упередження в системі [48].

Рис. 2.1. Класифікація типів зміщень у системах ШІ

Таблиця 2 ілюструє приклади різних типів зміщень і серйозні наслідки таких упереджень у системах ШІ, наголошуючи на необхідності ретельної оцінки та стратегій пом'якшення для усунення таких упереджень.

Таблиця 2

Характеристика різних типів упереджень ШІ

Тип зміщення	Опис	Приклади
Зміщення вибірки	Виникає, коли навчальні дані не є репрезентативними для населення, яке вони обслуговують, що призводить до низької продуктивності та упереджених прогнозів для певних груп.	Алгоритм розпізнавання обличчя, навчений переважно на білих особах, погано працює на представниках інших рас.
Алгоритмічний зсув	Результати розробки та реалізації алгоритму можуть визначити пріоритет певних атрибутів і призвести до несправедливих результатів.	Алгоритм, який надає пріоритет віку чи статі, що призводить до несправедливих результатів при прийнятті на роботу.
Упередженість представництва	Вузький ШІ перевершує людей у конкретних повторюваних завданнях, таких як водіння, медична діагностика та фінансові консультації.	Сильний ШІ конкурує з людьми в усіх починаннях: від отримання університетського диплому до лікування невідкладних медичних випадків.
(репрезентативність вибірки)	Трапляється, коли набір даних неточно представляє сукупність, яку він має моделювати, що призводить до неточних прогнозів.	Набір медичних даних, який недостатньо представляє жінок, що призводить до менш точного діагнозу для пацієнтів.
Упередженість підтвердження	Матеріалізується, коли система ШІ використовується для підтвердження існуючих упереджень або переконань її творців або користувачів.	Система штучного інтелекту, яка передбачає успіх кандидатів на роботу на основі упереджень менеджера з найму.

Тип зміщення	Опис	Приклади
Зсув вимірювання	Виникає, коли збір даних або вимірювання систематично надмірно або недостатньо представляють певні групи.	Опитування, яке збирає більше відповідей від міських жителів, що призводить до недостатнього представлення думок мешканців сільської місцевості.
Упередження взаємодії	Виникає, коли система ШІ взаємодіє з людьми упереджено, що призводить до несправедливого ставлення.	Чат-бот, який по-різному реагує на чоловіків і жінок, що призводить до упередженого спілкування.
Генеративне зміщення	Зустрічається в генеративних моделях штучного інтелекту, таких як ті, що використовуються для створення синтетичних даних, зображень або тексту. Генеративне упередження виникає, коли результати моделі непропорційно відображають конкретні атрибути, точки зору або шаблони, наявні в навчальних даних, що призводить до спотворених або незбалансованих представлень у створеному вмісті.	Модель генерації тексту, навчена переважно на літературі західних авторів, може надмірно репрезентувати західні культурні норми та ідіоми, недостатньо або неправильно представляючи інші культури. Подібним чином модель генерації зображень, навчена на наборах даних із обмеженим розмаїттям портретів людей, може мати проблеми з точним представленням широкого діапазону етнічних груп.

2.1.2. Вплив упередженості в ШІ

Швидкий розвиток штучного інтелекту (ШІ) приніс численні переваги, але він також пов'язаний з потенційними ризиками та проблемами. Однією з ключових проблем є негативний вплив упередженості ШІ на окремих людей і суспільство. Упередженість у штучному інтелекті може закріпити й навіть посилити існуючу нерівність, призводячи до дискримінації маргіналізованих груп і обмежуючи їхній доступ до основних послуг. Окрім увічнення гендерних стереотипів і дискримінації, це також може призвести до нових форм дискримінації за кольором шкіри, етнічною приналежністю чи зовнішністю. Щоб переконатися, що системи штучного інтелекту є чесними та справедливими та задовольняють потреби всіх користувачів, вкрай важливо визначити та пом'якшити упередженість ШІ. Крім того, використання упередженого штучного інтелекту має

численні етичні наслідки, включаючи потенційну дискримінацію, відповідальність розробників і політиків, підрив громадської довіри до технологій і обмеження людської волі та автономії. Усунення цих етичних наслідків вимагатиме узгоджених зусиль усіх зацікавлених сторін, і важливо розробити етичні рекомендації та нормативні рамки, які сприятимуть справедливості, прозорості та підзвітності у розробці та використанні систем ШІ.

Були численні приклади упередженості систем ШІ в різних галузях: від охорони здоров'я до кримінального правосуддя. Одним із добре відомих прикладів є система COMPAS, яка використовується в системі кримінального правосуддя Сполучених Штатів і передбачає ймовірність повторного вчинення підсудним злочину. Дослідження, проведене ProPublica, показало, що система була упереджена щодо підсудних афроамериканців, оскільки вони з більшою ймовірністю були визнані групою високого ризику, навіть якщо вони раніше не були судимі. Інше дослідження виявило схожі упередження в подібній системі, що використовується в штаті Вісконсін [25].

У сфері охорони здоров'я виявилося, що система штучного інтелекту, яка використовується для прогнозування рівня смертності пацієнтів, має упередження щодо пацієнтів-афроамериканців. Дослідження [26-28] виявили, що система з більшою ймовірністю призначає афроамериканським пацієнтам оцінки вищого ризику, навіть якщо інші фактори, такі як вік і стан здоров'я залишаються однаковими. Це упередження може призвести до того, що афроамериканським пацієнтам буде відмовлено в доступі до медичної допомоги або вони отримають лікування неналежного рівня.

Іншим прикладом упередженості систем штучного інтелекту є технологія розпізнавання облич, яку використовують правоохоронні органи. Дослідження, проведене Національним інститутом стандартів і технологій (NIST), виявило, що технологія розпізнавання обличчя була значно менш точною для людей із темнішим відтінком шкіри, що призвело до більш високого рівня помилкових спрацьовувань [29]. Це упередження може мати серйозні наслідки, такі як незаконні арешти або засудження.

Нарешті, з розвитком генеративних систем штучного інтелекту (GenAI) зростає ризик шкідливих упереджень [30-32]. Повідомлялося про яскравий випадок упередженості GenAI, коли моделі перетворення тексту в зображення, такі як StableDiffusion, OpenAI DALL-E та Midjourney, демонстрували расові та стереотипні упередження у своїх результатах [33].

Коли їм було запропоновано створити зображення генеральних директорів, ці моделі переважно створювали зображення чоловіків, що відображало гендерні упередження. Це упередження відображає недостатню представленість жінок на посадах генеральних директорів у реальному світі. Крім того, коли їм пропонується створити зображення злочинців або терористів, моделі показують значно більше кольорових людей [34].

Цей інцидент підкреслює ризик того, що генеративний штучний інтелект увічнить суспільні упередження. Моделі GenAI, навчені на зображеннях з Інтернету, ймовірно, страждають від цієї упередженості, оскільки дані віддзеркалюють існуючі розбіжності. Цей приклад підкреслює критичну потребу в різноманітних і збалансованих навчальних наборах даних у розробці ШІ, щоб забезпечити справедливі та репрезентативні результати генеративних моделей [35].

Негативні наслідки упередженості в штучному інтелекті можуть бути значними, впливаючи на окремих людей і суспільство. Дискримінація є ключовою проблемою, коли мова заходить про упереджені системи штучного інтелекту, оскільки вони можуть увічнити та навіть посилити існуючу нерівність. Наприклад, упереджені алгоритми, що використовуються в системі кримінального правосуддя, можуть призвести до несправедливого ставлення до певних груп, особливо до кольорових людей, які, швидше за все, будуть помилково засуджені або отримають суворіші вироки.

Упередженість ШІ також може мати негативний вплив на доступ людини до основних послуг, таких як охорона здоров'я та фінанси. Упереджені алгоритми можуть призвести до недостатнього представництва певних груп, таких як кольорові люди або люди з нижчим соціально-економічним становищем, у системах оцінки кредитоспроможності, ускладнюючи для них доступ до позик чи іпотеки.

Крім того, упередженість у ШІ також може увічнити гендерні стереотипи та дискримінацію. Наприклад, алгоритми розпізнавання облич, навчені на основі даних, які в основному складаються з чоловіків, можуть неточно розпізнавати жіночі обличчя, зберігаючи гендерні упередження в системах безпеки. Коли моделям генеративного штучного інтелекту (GenAI) пропонується створити образи керівників, вони, як правило, зміцнюють стереотипи, зображуючи керівників переважно як чоловіків [23].

Окрім увічнення існуючої нерівності, упередження в ШІ також можуть призвести до нових форм дискримінації, наприклад, дискримінації за кольором шкіри, етнічною приналежністю чи навіть зовнішністю. Ті самі моделі GenAI, які демонструють гендерні упередження, що, можливо, не дивно, також зображують злочинців або терористів як кольорових людей.

Публічне розгортання цих систем може призвести до серйозних наслідків, таких як відмова в обслуговуванні, можливості працевлаштування або навіть незаконні арешти чи засудження. Ризик подвійний: на індивідуальному рівні він впливає на сприйняття людьми самих себе та інших, потенційно впливаючи на їхні можливості та взаємодію.

На суспільному рівні широке використання таких упереджених систем штучного інтелекту може закріпити дискримінаційні наративи та перешкодити зусиллям щодо рівності та інклюзивності. Оскільки штучний інтелект стає все більш інтегрованим у наше повсякденне життя, потенціал такої технології для формування культурних норм і соціальних структур стає більш значним, що робить обов'язковим вирішення цих упереджень на етапах розвитку систем ШІ, щоб пом'якшити їхній шкідливий вплив [43-50].

2.1.3. Обговорення етичних наслідків упередженого ШІ

Використання упередженого ШІ має численні етичні наслідки, які необхідно враховувати. Однією з головних проблем є можливість дискримінації окремих осіб або груп на основі таких факторів, як раса, стать, вік або інвалідність [31-33]. Коли системи штучного інтелекту упереджені, вони можуть увічнити існуючу нерівність і посилити дискримінацію маргіналізованих груп. Особливо це стосується чутливих сфер, таких як охорона здоров'я, де упереджені системи ШІ можуть призвести до нерівного доступу до лікування або завдати шкоди пацієнтам [34].

Іншим етичним питанням є відповідальність розробників, компаній і урядів за те, щоб системи штучного інтелекту розроблялися та використовувалися чесно та прозоро. Якщо система ШІ є упередженою та дає дискримінаційні результати, відповідальність лежить не лише на самій системі, але й на тих, хто її створив і розгорнув [51]. Таким чином, надзвичайно важливо встановити етичні принципи та нормативні рамки, які притягують осіб, відповідальних за розробку та використання систем ШІ, до відповідальності за будь-які дискримінаційні результати.

Крім того, використання упереджених систем штучного інтелекту може підірвати довіру суспільства до технологій, призводячи до зниження рівня впровадження та навіть відмови від нових технологій. Це може мати серйозні економічні та соціальні наслідки, оскільки потенційні переваги штучного інтелекту можуть бути нереалізовані, якщо люди не довірятимуть технології або якщо її розглядатимуть як інструмент дискримінації [38].

Нарешті, важливо розглянути вплив упередженого штучного інтелекту на свободу дій і автономію людини. Коли системи штучного інтелекту упереджені, вони можуть обмежити індивідуальні свободи та зміцнити динаміку суспільної влади [26]. Наприклад, система штучного інтелекту, яка використовується в процесі найму, може непропорційно виключати кандидатів з маргіналізованих груп, обмежуючи їхні можливості доступу до можливостей працевлаштування та сприяння розвитку суспільства [27].

Вирішення етичних наслідків упередженого штучного інтелекту вимагає узгоджених зусиль усіх зацікавлених сторін, включаючи розробників, політиків та суспільства в цілому. Саме тому важливо розробити етичні рекомендації та нормативні рамки, які сприятимуть справедливості, прозорості та підзвітності у розробці та використанні систем ШІ [52]. Крім того, важливо брати участь у критичних дискусіях щодо впливу штучного інтелекту на суспільство та надавати людям можливість відповідально та етично брати участь у формуванні майбутнього штучного інтелекту.

2.2. Стратегії пом'якшення упередженості в ШІ

Дослідники та практики запропонували різні підходи для пом'якшення упередженості в ШІ. Ці підходи включають попередню обробку даних, застосування конкретних типів моделей та відповідальність за прийняття рішення після обробки системою ШІ. Проте кожен підхід має свої обмеження та проблеми, такі як відсутність різноманітних і репрезентативних навчальних даних, складність виявлення та вимірювання різних типів упередженості та потенційні компроміси між справедливістю та точністю. Крім того, існують етичні міркування щодо того, як визначити пріоритетність різних типів упередженості та яким групам віддавати пріоритет у пом'якшенні упередженості.

Незважаючи на ці проблеми, пом'якшення упередженості в штучному інтелекті має важливе значення для створення справедливих і неупереджених систем, які приносять користь усім людям і суспільству. Щоб подолати ці проблеми та забезпечити використання систем штучного інтелекту на благо всіх, необхідні постійні дослідження та розробка підходів до пом'якшення.

2.2.1. Огляд існуючих підходів до пом'якшення зміщення в ШІ

Пом'якшення упередженості в штучному інтелекті є складним і багатогранним завданням. Розглянемо декілька найбільш популярних підходів для вирішення цієї проблеми.

Одним із поширених підходів є попередня обробка даних, які використовуються для навчання моделей ШІ, щоб переконатися, що вони є репрезентативними для всього населення, включаючи історично маргіналізовані групи. Це може включати такі методи, як надмірна вибірка, недостатня вибірка або генерація синтетичних даних [30]. Наприклад, дослідження [45] продемонструвало, що надмірна вибірка темношкірих людей покращує точність алгоритмів розпізнавання обличчя для цієї групи. Попередня обробка даних передбачає виявлення та усунення упереджень у даних перед навчанням моделі. Це можна виконати за допомогою таких методів, як збільшення даних, яке передбачає створення синтетичних точок даних для збільшення репрезентації недостатньо представлених груп, або за допомогою змагального зміщення, яке передбачає навчання моделі стійкості до певних типів упереджень. Документування таких упереджень набору даних і процедур розширення є надзвичайно важливим.

Ще один підхід до пом'якшення упередженості в штучному інтелекті полягає у ретельному виборі моделей, які використовуються для аналізу даних. Дослідники запропонували використовувати методи відбору моделей, які віддають перевагу справедливості, наприклад ті, що базуються на груповій справедливості [47] або індивідуальній справедливості [53]. Наприклад, у дослідженні [54] запропоновано метод вибору класифікаторів, які забезпечують демографічний паритет, забезпечуючи рівномірний розподіл позитивних і негативних результатів між різними демографічними групами. Також окремий підхід полягає у використанні методів відбору моделі, які віддають пріоритет справедливості та пом'якшують упередженість. Це можна зробити за допомогою таких методів, як регуляризація, яка карає моделі за створення дискримінаційних прогнозів, або за допомогою методів ансамблю, які поєднують кілька моделей для зменшення упередженості [55].

Рішення про постобробку є ще одним підходом до пом'якшення упередженості в ШІ. Це передбачає коригування виходу моделей ШІ, щоб усунути упередженість і забезпечити справедливість. Наприклад, дослідники запропонували методи постобробки, які коригують рішення, прийняті моделлю, для досягнення вирівняних шансів, що гарантує рівномірний розподіл хибно-позитивних і хибно-негативних результатів між різними демографічними групами [47].

Хоча ці підходи є перспективними для пом'якшення упередженості в ШІ, вони також мають обмеження та проблеми. Наприклад, попередня обробка даних може займати багато часу і не завжди бути ефективною, особливо якщо дані, які використовуються для навчання моделей, уже є упередженими. Крім того, методи вибору моделі можуть бути обмежені відсутністю консенсусу щодо того, що є справедливістю, а методи постобробки можуть бути складними та вимагати великої кількості додаткових даних. Тому вкрай важливо продовжувати дослідження та розробку нових підходів для пом'якшення упередженості в ШІ [56].

У сфері генеративного штучного інтелекту боротися з упередженням ще складніше, оскільки це вимагає цілісної стратегії, яка починається з попередньої обробки даних для забезпечення різноманітності та репрезентативності. Вона передбачає зби-

рання та включення до навчальної вибірки максимально різноманітних джерел даних, які відображають широкий людський досвід, таким чином запобігаючи надмірному представленню будь-якої окремої демографічної групи в навчальних наборах даних. Під час вибору моделі необхідно надати пріоритет алгоритмам, які є прозорими та здатними виявляти, коли вони генерують упереджені результати. Корисними можуть бути такі методи, як змагальність, коли моделі постійно перевіряються на сценарії, призначені для виявлення упередженості [30]. Постобробка передбачає критичну оцінку створеного штучним інтелектом контенту та, якщо необхідно, коригування результатів для виправлення помилок. Це може включати використання додаткових фільтрів або перенесення методів навчання для подальшого вдосконалення моделей. Регулярні аудити, безперервний моніторинг і включення циклів зворотного зв'язку мають важливе значення для того, щоб генеративні системи штучного інтелекту залишалися чесними та справедливими протягом тривалого часу. Ці зусилля мають підкріплюватись дотриманням етичних принципів штучного інтелекту, активним залученням різноманітних команд до розробки штучного інтелекту та сприянням міждисциплінарній співпраці для ефективного вирішення та пом'якшення упередженості штучного інтелекту [32-35].

Крім того, впровадження цих підходів вимагає ретельного розгляду етичних і суспільних наслідків. Наприклад, коригування прогнозів моделі для забезпечення справедливості може призвести до компромісів між різними формами упередженості та може мати непередбачені наслідки для розподілу результатів для різних груп [27]. Таблиця 3 ілюструє приклади таких стратегій, їхні проблеми та обмеження.

Таблиця 3

Різні підходи до пом'якшення зміщення ШІ та пов'язані з цим проблеми

Підхід	Опис	Приклади	Обмеження та виклики	Етичні міркування
Попередня обробка даних	Включає виявлення та усунення упереджень у даних перед навчанням моделі. Такі методи, як надмірна вибірка, недостатня вибірка або генерація синтетичних даних, використовуються для забезпечення репрезентативності даних для всього населення, включаючи історично маргіналізовані групи.	1. Перевибірка темношкірих людей у наборі даних розпізнавання облич 2. Збільшення даних для збільшення представництва в недостатньо представлених групах. 3. Змагальне упередження, щоб навчити модель бути стійкою до конкретних типів упереджень	1. Трудомісткий процес. 2. Може не завжди бути ефективним, особливо якщо дані, які використовуються для навчання моделей, уже є упередженими.	1. Можливість надмірного чи недостатнього представлення певних груп у даних, що може зберегти існуючі упередження або створити нові. 2. Проблеми конфіденційності, пов'язані зі збором і використанням даних, особливо для історично маргіналізованих груп.

Підхід	Опис	Приклади	Обмеження та виклики	Етичні міркування
Вибір моделі	Зосереджено на використанні методів відбору моделей, які віддають пріоритет справедливості. Дослідники запропонували методи, засновані на груповій або індивідуальній справедливості. Методи включають регулярізацію, яка карає моделі за створення дискримінаційних прогнозів, і ансамблеві методи, які поєднують кілька моделей для зменшення упередженості.	1. Вибір класифікаторів, що забезпечують демографічний паритет 2. Використання методів відбору моделей на основі групової справедливості або індивідуальної справедливості. 3. Регулювання для покарання за дискримінаційні прогнози. 4. Методи ансамблю для поєднання кількох моделей і зменшення зміщення.	Обмежений можливою відсутністю консенсусу щодо того, що є справедливістю.	1. Збалансування справедливості з іншими показниками ефективності, такими як точність або ефективність. 2. Можливість моделей зміцнювати існуючі стереотипи чи упередження, якщо критерії справедливості не розглядаються ретельно.
Постобробка рішень	Включає коригування виходу моделей AI для усунення упередженості та забезпечення справедливості. Дослідники запропонували методи, які коригують рішення, прийняті моделлю, для досягнення вирівняних шансів, гарантуючи, що помилкові позитивні та помилкові негативні результати рівномірно розподіляються між різними демографічними групами.	Методи постобробки, які досягають вирівняних шансів.	Може бути складним і потребувати великої кількості додаткових даних.	1. Компроміс між різними формами упередженості під час коригування прогнозів для справедливості. 2. Непередбачені наслідки для розподілу результатів для різних груп.

Однак, варто зазначити, що хоча наведені підходи орієнтовані на усунення упередженості в ШІ, вони також мають свої обмеження та проблеми.

Зокрема, однією з головних проблем є відсутність різноманітних і репрезентативних навчальних даних. Як згадувалося раніше, зміщення даних може призвести до зміщення результатів систем ШІ. Однак збір різноманітних і репрезентативних даних може бути складним завданням, особливо коли йдеться про чутливі або рідкісні події [53]. Крім того, можуть виникнути проблеми з конфіденційністю під час збору певних типів даних, наприклад медичних записів або фінансової інформації. Ці проблеми можуть обмежити ефективність розширення набору даних як підходу до пом'якшення [54].

Іншою проблемою є складність виявлення та вимірювання різних типів упередженості в системах ШІ. Алгоритмічне зміщення може бути важко виявити та кількісно оцінити, особливо коли алгоритми складні або непрозорі. Крім того, може бути важко виділити джерела зміщення, оскільки зміщення може виникати з кількох джерел, таких як дані, алгоритм і користувач. Це може обмежити ефективність алгоритмів з упередженням і механізмів зворотного зв'язку з користувачами в міру пом'якшення [55].

Крім того, підходи до пом'якшення можуть запровадити компроміс між справедливістю та точністю. Наприклад, одним із підходів до зменшення алгоритмічного зміщення є модифікація алгоритму, щоб переконатися, що він обробляє всі групи однаково [57]. Однак це може призвести до зниження точності для певних груп або в певних контекстах. Досягнення як справедливості, так і точності може бути складним завданням і потребує ретельного розгляду відповідних компромісів [58].

Нарешті, можуть існувати етичні міркування щодо того, як визначити пріоритетність різних типів упередженості та яким групам віддавати пріоритет у пом'якшенні упередженості. Наприклад, чи слід приділяти більше уваги упередженню, яке впливає на історично маргіналізовані групи, чи слід надавати рівну вагу всім типам упередженості? Ці етичні міркування можуть ускладнити розробку та впровадження підходів до пом'якшення упередженості [59].

Незважаючи на ці проблеми, усунення упередженості в штучному інтелекті має вирішальне значення для створення справедливих систем ШІ [58-60]. Постійні дослідження та розробка підходів до пом'якшення наслідків необхідні для подолання цих викликів і забезпечення використання систем ШІ на благо всіх людей і суспільства.

2.3. Справедливість у ШІ

Справедливість у штучному інтелекті є важливою темою, яка привернула багато уваги як в академічних, так і в промислових колах. За своєю суттю, справедливість у ШІ означає відсутність упередженості або дискримінації в системах ШІ, чого може бути складно досягти через різні типи упереджень, які можуть виникати в цих системах. У літературі пропонується кілька типів справедливості, включаючи групову справедливість, індивідуальну справедливість і контрфактичну справедливість. Хоча справедли-

вість і упередженість є тісно пов'язаними поняттями, вони відрізняються в важливих аспектах, зокрема тим, що справедливість за своєю суттю є заздалегідь визначеною метою, тоді як упередженість може бути ненавмисною. Досягнення справедливості в ШІ вимагає ретельного розгляду контексту та залучених зацікавлених сторін. Реальні приклади справедливості в ШІ демонструють потенційні переваги включення справедливості в системи ШІ.

2.3.1. Визначення справедливості в ШІ та його різних типах

Справедливість у штучному інтелекті – це складна та багатогранна концепція, яка є предметом численних дискусій як в академічній, так і в галузевій спільнотах. За своєю суттю справедливість означає відсутність упередженості або дискримінації в системах ШІ. Однак досягнення справедливості в штучному інтелекті може бути складним завданням, оскільки вимагає ретельного розгляду різних типів упереджень, які можуть виникнути в цих системах, і способів їх пом'якшення [52].

Існує кілька різних типів справедливості, які були запропоновані в літературі, включаючи групову справедливість, індивідуальну справедливість і контрфактичну справедливість. Детальне порівняння кожного з цих типів наведено в Таблиці 4.

Групова справедливість означає забезпечення рівного чи пропорційного ставлення до різних груп у системах ШІ. Далі це можна розділити на різні типи, такі як демографічний паритет, який гарантує рівномірний розподіл позитивних і негативних результатів між різними демографічними групами [54], поняття несправедливості, неоднакового поганого поводження, визначеного з точки зору частоти неправильної класифікації, або рівних можливостей, що гарантує, що справжні позитивні показники (TP – true positive, чутливість) і помилкові позитивні показники (FP- false positive, специфічність) однакові для різних демографічних груп [53].

Індивідуальна справедливість, з іншого боку, означає забезпечення того, щоб системи штучного інтелекту однаково ставилися до подібних людей, незалежно від їх членства в групі. Цього можна досягти за допомогою таких методів, як вимірювання на основі подібності або відстані, які спрямовані на те, щоб система штучного інтелекту однаково ставилася до осіб, схожих за своїми характеристиками чи атрибутами [67].

Контрфактична справедливість — це нова концепція, яка спрямована на те, щоб системи ШІ були справедливими навіть у гіпотетичних сценаріях. Зокрема, контрфактична справедливість має на меті гарантувати, що система штучного інтелекту прийме те саме рішення щодо особи, незалежно від її членства в групі, навіть якщо її атрибути будуть різними. Тобто, наприклад, рішення про те, чи надати людині кредит не повинно змінитись, якщо в атрибутах, що його описують, замінити чутливі параметри (такі як раса чи стать) на протилежні [68].

Таблиця 4

Характеристика різних типів визначень справедливості ШІ

Тип справедливості	Опис	Приклади
Групова справедливість	Забезпечує однакове чи пропорційне ставлення до різних груп у системах ШІ. Можна далі поділити на демографічний паритет, неоднакове погане поводження або рівні можливості.	**1. Демографічний паритет:** Позитивні та негативні результати рівномірно розподілені між демографічними групами. **2. Різноманітне неправильне поводження:** визначається з точки зору частоти неправильної класифікації. **3. Рівні можливості:** частота істинних позитивних результатів (чутливість) і частота хибних позитивних результатів (1-специфічність) однакові для різних демографічних груп.
Індивідуальна справедливість	Гарантує, що системи штучного інтелекту однаково обробляють подібних людей, незалежно від їх членства в групі. Може бути досягнуто за допомогою таких методів, як вимірювання на основі подібності або відстані.	Використання заходів, заснованих на схожості або відстані, щоб гарантувати, що система штучного інтелекту однаково ставиться до осіб зі схожими характеристиками або атрибутами.
Контрфактична справедливість	Прагне забезпечити справедливість систем ШІ навіть у гіпотетичних сценаріях. Зокрема, контрфактична справедливість спрямована на те, щоб система штучного інтелекту прийняла те саме рішення щодо особи, незалежно від її належності до групи, навіть якби її атрибути були різними.	Забезпечення того, щоб система штучного інтелекту приймала те саме рішення для людини, навіть якщо її атрибути відрізнялися.
Процедурна справедливість	Включає забезпечення справедливості та прозорості процесу прийняття рішень.	Впровадження прозорого процесу прийняття рішень у системах ШІ.
Причинно-наслідкова справедливість	Включає забезпечення того, щоб система не зберігала історичні упередження та нерівність.	Розробка систем штучного інтелекту, які уникають збереження історичних упереджень і нерівності.

Інші типи справедливості включають процедурну справедливість, яка передбачає забезпечення того, що процес, який використовується для прийняття рішень, є справедливим і прозорим, і причинно-наслідкову справедливість, яка передбачає забезпечення того, щоб система не зберігала історичні упередження та нерівність.

Важливо зазначити, що ці різні типи справедливості не є взаємовиключаючими та можуть збігатися на практиці. Крім того, різні типи справедливості можуть конфліктувати один з одним, і може знадобитися компроміс, щоб досягти справедливості в конкретних контекстах. Важливо зазначити, що досягнення справедливості в ШІ не є універсальним рішенням і вимагає ретельного розгляду контексту та зацікавлених сторін. Досягнення справедливості в системах штучного інтелекту часто вимагає тонкого розуміння цих різних типів справедливості та способів їх збалансування та пріоритетності залежно від предметної області [52].

2.3.2. Порівняння справедливості та упередженості в ШІ

Хоча справедливість і упередженість є тісно пов'язаними поняттями, вони мають важливі відмінності. Зміщення (bias) відноситься до систематичного та послідовного відхилення виходу алгоритму від справжнього значення або від того, чого можна було б очікувати за відсутності зміщення [69]. З іншого боку, справедливість у штучному інтелекті означає відсутність дискримінації чи фаворитизму щодо будь-якої особи чи групи на основі захищених характеристик, таких як раса, стать, вік чи релігія [55].

Однією з ключових відмінностей між справедливістю та упередженістю є те, що хоча упередженість може бути ненавмисною, справедливість за своєю суттю є свідомою та навмисною метою. Упередженість може виникнути через різні чинники, наприклад, упереджені дані чи алгоритмічний дизайн, але справедливість вимагає свідомих зусиль, щоб гарантувати, що алгоритм не дискримінує жодну групу чи людину. Інакше кажучи, упередженість можна розглядати як технічну проблему, тоді як справедливість є соціальною та етичною [52].

Інша відмінність полягає в тому, що упередження може бути як позитивним, так і негативним, тоді як справедливість стосується лише негативного упередження або дискримінації. Позитивне упередження виникає, коли алгоритм систематично надає перевагу певній групі чи індивідууму, тоді як негативне упередження виникає, коли алгоритм систематично дискримінує певну групу чи індивіда. Навпаки, справедливість стосується запобігання негативним упередженням або дискримінації щодо будь-якої групи чи особи [69].

Незважаючи на ці відмінності, справедливість і упередженість часто тісно пов'язані, і усунення упередженості є важливим кроком на шляху до досягнення справедливості в ШІ. Наприклад, усунення упередженості в навчальних даних або алгоритмах може допомогти зменшити ймовірність несправедливих результатів. Однак важливо визнати, що упередженість — не єдиний фактор, який може призвести до несправедливості, і досягнення справедливості може вимагати додаткових зусиль, окрім пом'якшення упередженості [70].

Загалом, розуміння відмінностей між справедливістю та упередженням є важливим для розробки ефективних стратегій пом'якшення упередженості та забезпечення справедливості в системах ШІ. Визнаючи ці відмінності та розробляючи алгоритми та системи, які віддають пріоритет справедливості, ми можемо гарантувати, що системи штучного інтелекту будуть використовуватися на благо всіх людей і груп, не зберігаючи та не загострюючи існуючу соціальну та економічну нерівність.

2.3.3. Реальні приклади проблем із справедливістю в ШІ

Існують різні реальні приклади систем ШІ, які мали проблему із справедливістю під час прийняття рішень. Одним із прикладів є система COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), яка використовується для прогнозування ймовірності рецидиву у злочинців. Дослідження показали, що система була упереджена щодо підсудних афроамериканців, оскільки вона мала більше шансів помилково передбачити, що вони вчинять повторний злочин, порівняно з білими підсудними [19]. Щоб вирішити цю проблему, Northpointe COMPAS було модифіковано, щоб створити «расово-нейтральну» версію алгоритму, яка досягла подібної точності при зменшенні расових упереджень [25].

Ще один приклад – використання штучного інтелекту в процесі найму. Дослідження показали, що системи найму ШІ можуть бути упередженими щодо жінок, оскільки вони з меншою ймовірністю будуть обрані на ролі, де домінують чоловіки [27]. Щоб вирішити цю проблему, деякі компанії впровадили інструменти «гендерного декодера», які аналізують оголошення про роботу та пропонують зміни для зменшення гендерних упереджень [66].

Дуже небезпечним є приклад зі сфери охороні здоров'я. Дослідження показали, що системи ШІ, які використовуються для прогнозування результатів охорони здоров'я, можуть бути упередженими щодо певних груп, таких як афроамериканці [55]. Щоб вирішити цю проблему, дослідники запропонували використовувати такі методи, як аналіз підгруп, щоб виявити та усунути упередження в даних, які використовуються для навчання моделей ШІ [47].

Ці реальні приклади демонструють потенційні переваги включення справедливості в системи ШІ. Усуваючи упередженість і забезпечуючи справедливість, системи штучного інтелекту можуть бути більш точними, етичними та справедливими та можуть сприяти соціальній справедливості та рівності.

2.3.4. Стратегії пом'якшення несправедливості в ШІ

Оскільки використання штучного інтелекту продовжує зростати, забезпечення справедливості в прийнятті ним рішень стає все більш важливим. Використання штучного інтелекту в критично важливих сферах, таких як охорона здоров'я, фінанси та законодавство, може значно вплинути на життя людей, тому вкрай важливо, щоб ці системи приймали справедливі та неупереджені рішення. Щоб вирішити цю проблему, було

розроблено різні підходи, включаючи групову справедливість та індивідуальну справедливість. Однак ці підходи не позбавлені обмежень і викликів, таких як компроміси між різними типами справедливості та труднощі визначення самої справедливості.

Розглянемо різні стратегії пом'якшення справедливості в штучному інтелекті, включаючи поточні підходи, виклики та сфери майбутніх досліджень. Розвиваючи краще розуміння цих стратегій пом'якшення, ми можемо працювати над створенням справедливих, неупереджених і справедливих систем ШІ для всіх.

Забезпечення справедливості в штучному інтелекті є складною сферою, яка постійно розвивається, у якій розробляються різні підходи для вирішення різних аспектів справедливості. Виникли два ключові підходи: групова справедливість та індивідуальна справедливість.

Групова справедливість пов'язана із забезпеченням справедливості систем ШІ по відношенню до різних груп людей, наприклад людей різної статі, раси чи етнічної приналежності. Групова справедливість спрямована на те, щоб система штучного інтелекту не дискримінувала будь-яку групу. Цього можна досягти за допомогою різних методів, таких як повторна вибірка, попередня обробка або постобробка даних, які використовуються для навчання моделі ШІ. Наприклад, якщо модель штучного інтелекту навчена на даних, які мають зміщення щодо певної групи, методи повторної вибірки можна використовувати для створення збалансованого набору даних, де кожна група представлена однаково. Інші методи, такі як попередня обробка або постобробка, можуть бути використані для коригування виходу моделі штучного інтелекту, щоб переконатися, що вона не ставить жодну групу [70].

З іншого боку, індивідуальна справедливість полягає в тому, щоб системи штучного інтелекту були справедливими відносно окремих людей, незалежно від їх членства в групі. Індивідуальна справедливість має на меті запобігти системі штучного інтелекту приймати рішення, які систематично є упередженими проти певних осіб. Індивідуальна справедливість може бути досягнута за допомогою таких методів, як контрфактична справедливість або причинно-наслідкова справедливість. Наприклад, контрфактична справедливість спрямована на те, щоб модель штучного інтелекту прийняла те саме рішення щодо особи, незалежно від раси чи статі.

Хоча групова та індивідуальна справедливість є важливими підходами до забезпечення справедливості в ШІ, вони не єдині. Інші підходи включають прозорість, підзвітність і можливість пояснення. Прозорість передбачає надання користувачам інформації про процес прийняття рішень у системі ШІ, тоді як підзвітність передбачає притягнення розробників системи до відповідальності за будь-яку шкоду, заподіяну системою. Пояснюваність передбачає прийняття рішень системи ШІ зрозумілими для користувачів [52].

Загалом, забезпечення справедливості в штучному інтелекті є складним і постійним викликом, який вимагає міждисциплінарного підходу із залученням експертів із таких галузей, як інформатика, право, етика та соціальні науки. Розробляючи та впроваджуючи низку підходів для забезпечення справедливості, ми можемо працювати над створенням систем ШІ, які є неупередженими, прозорими та підзвітними.

Хоча ці підходи показали гарні результати в забезпеченні справедливості в ШІ, вони не позбавлені обмежень і проблем. Одним із головних обмежень є можливість компро-

місів між різними типами справедливості. Наприклад, підходи до групової справедливості можуть призвести до нерівного ставлення до окремих осіб у групі, тоді як підходи до індивідуальної справедливості можуть не вирішувати системні упередження, які впливають на цілі групи. Крім того, може бути важко визначити, які типи справедливості є найбільш прийнятними для даного контексту, і як правильно їх збалансувати [42].

Іншою проблемою є складність визначення самої справедливості. Різні люди та групи можуть мати різні визначення справедливості, і ці визначення можуть змінюватися з часом [45]. Це може ускладнити розробку систем ШІ, які всі зацікавлені сторони вважають справедливими.

Крім того, багато сучасних підходів до забезпечення справедливості в штучному інтелекті ґрунтуються на статистичних методах і припущеннях, які можуть неточно відобразити складність людської поведінки та прийняття рішень. Наприклад, метрика групової справедливості може не враховувати взаємодію між різними аспектами ідентичності (наприклад, раси, статі та соціально-економічного статусу) і впливати на результати [71].

Нарешті, є занепокоєння щодо потенційних непередбачених наслідків і шкідливих результатів у результаті спроб забезпечити справедливість у ШІ. Наприклад, деякі дослідники виявили, що спроби пом'якшити упередженість в алгоритмах прогностичної роботи поліції можуть збільшити расові відмінності в арештах [67]. У таблиці 5 узагальнено ці стратегії.

Таблиця 5

Різні підходи до гарантування справедливості ШІ та пов'язані з цим проблеми

Підхід	Опис	Приклади	Обмеження та виклики
Групова справедливість	Забезпечує справедливість систем штучного інтелекту для різних груп людей, наприклад людей різної статі, раси чи етнічної приналежності. Націлений на те, щоб система штучного інтелекту не дискримінувала будь-яку групу. Може бути досягнуто за допомогою таких методів, як повторна вибірка, попередня обробка або постобробка даних.	1. Методи повторної вибірки для створення збалансованого набору даних. 2. Попередня обробка або постобробка для коригування виходу моделі ШІ.	1. Може призвести до нерівного ставлення до окремих осіб у групі. 2. Може не розглядати системні упередження, які впливають на індивідуальні характеристики. 3. Метрики групової справедливості можуть не враховувати перехресність.

Підхід	Опис	Приклади	Обмеження та виклики
Індивідуальна справедливість	Гарантує, що системи штучного інтелекту є справедливими по відношенню до окремих людей, незалежно від їх членства в групі. Має на меті запобігти системі штучного інтелекту приймати рішення, які систематично є упередженими проти певних осіб. Може бути досягнуто за допомогою таких методів, як контрфактична справедливість або причинно-наслідкова справедливість.	Контрфактична справедливість, що забезпечує однакове рішення незалежно від раси чи статі.	1. Може не розглядати системні упередження, які впливають на цілі групи. 2. Складно визначити, які типи справедливості підходять для даного контексту та як їх збалансувати.
Прозорість	Передбачає надання користувачам інформації про процес прийняття рішень у системі ШІ.	Зробити рішення та процеси системи AI зрозумілими для користувачів.	Різні визначення справедливості серед людей і груп і зміни визначень з часом.
Відповідальність	Передбачає притягнення розробників системи до відповідальності за будь-яку шкоду, заподіяну системою.	Розробники понесли відповідальність за несправедливі рішення систем ШІ.	Визначення відповідальності та усунення потенційної шкоди.
Пояснюваність	Передбачає створення зрозумілих для користувачів рішень системи ШІ.	Надання чітких пояснень рішень системи ШІ.	Розглядаючи складність людської поведінки та прийняття рішень.
Міждисциплінарність	Розглядає способи взаємодії різних аспектів ідентичності (наприклад, раси, статі та соціально-економічного статусу) та впливу на результати.	Розробка систем ШІ, які враховують взаємодію різних вимірів ідентичності.	Вирішення проблеми міжсекціональності та забезпечення справедливості в багатьох вимірах ідентичності.

Враховуючи описані виклики, очевидно, що розробка справедливого та неупередженого ШІ є важливим напрямом досліджень, що вимагає подальшого опрацювання. Майбутня робота потребуватиме вирішення цих проблем і продовження розробки нових підходів, які чутливі до нюансів справедливості та упередженості в різних предметних областях.

Існуючі джерела упереджень у системах штучного інтелекту та машинного навчання можуть глибоко впливати на суспільство. Особливо посилилися такі ризики у зв'язку із популярністю генеративного штучного інтелекту [72]. Зрозуміло, що ці потужні обчислювальні інструменти, якщо їх не ретельно розробити та перевірити, можуть увічнити та навіть посилити існуючі упередження, особливо ті, що пов'язані з расою, статтю та іншими чутливими характеристиками. Сьогодні вже відомо про численні приклади упереджених систем штучного інтелекту. Тому необхідно приділяти особливу увагу тонкощам генеративного штучного інтелекту, а також розробляти та впроваджувати комплексні стратегії для виявлення та пом'якшення відхилень у всьому спектрі процесу розробки штучного інтелекту.

Ефективними методами для створення надійних систем ШІ є такі стратегії, як надійне збільшення даних, застосування контрфактичної справедливості та забезпечення різноманітності та достатньої репрезентативності в наборах даних поряд із неупередженими методами збору даних, вибору відповідних моделей та алгоритмів для навчання системи ШІ, а також врахування етичних аспектів використання ШІ для збереження конфіденційності та необхідність прозорості, нагляду та постійної оцінки систем ШІ.

При машинному навчанні необхідно приділяти особливу увагу диверсифікації навчальних даних і розглядати нюанси упередженості в генеративних моделях, особливо тих, що використовуються для створення синтетичних даних і генерації контенту. Вкрай важливо розробити комплексні рамки та рекомендації для відповідального штучного інтелекту, які включають прозору документацію даних навчання, вибору моделей і генеративних процесів. Диверсифікація команд, які беруть участь у розробці та оцінці штучного інтелекту, є настільки ж важливою, наскільки вона приносить безліч точок зору, які можуть краще ідентифікувати та виправляти упередження [73].

Також важливим чинником для забезпечення надійності систем ШІ, що розробляються та використовуються, може бути створення надійних етичних і правових рамок, що регулюють системи штучного інтелекту. Саме правове регулювання має першорядне значення, гарантуючи, що конфіденційність, прозорість і підзвітність будуть основоположними елементами життєвого циклу розробки штучного інтелекту [68].

2.4. Узагальнення передумов виникнення надійного штучного інтелекту

Отже, серед передумов виникнення надійного штучного інтелекту (Trustworthy AI, НШІ) можна виділити такі ключові аспекти, що стосуються технологічних, соціальних, етичних та регуляторних сфер:

2.4.1. Технологічні досягнення

- **Алгоритми машинного навчання**

Штучний інтелект (ШІ) з'явився порівняно давно, в 50-ті роки 20 сторіччя і за цей час пережив декілька етапів як розквіту, так і занепаду – в перші десятиріччя формування ШІ науковці та суспільство доволі радикально змінювали своє ставлення до систем, що використовували ШІ. Для розуміння цього необхідно враховувати стан апаратного забезпечення, який був доступний для розробників систем ШІ на той момент часу: обчислювальні потужності були мізерні, порівняно із тим, що доступно сьогодні. Тим не менш саме в цей час розроблялись фундаментальні та класичні алгоритми та методи, які лежать в основі всіх сучасних систем розпізнавання образів, обробки зображень, методів машинного навчання тощо.

Відповідно вдосконалення алгоритмів машинного навчання, особливо поява методів глибокого навчання (deep learning), зробило ШІ більш точним та здатним до самостійного навчання на основі нових даних, адже залежно від типу задач та якості доступних даних системи, що використовують методи глибокого навчання, здатні досягати точності 95–99%.

- **Розвиток обчислювальних потужностей**

Зростання потужності комп'ютерів та хмарних технологій дозволило створювати більш складні моделі ШІ, здатні обробляти величезні обсяги даних за дуже короткий час, наближаючись до можливості до роботи в режимі реального часу. Це призводить до того, що прийняті ШІ рішення можуть миттєво впливати на життя та добробут користувачів, а, відповідно, розробникам необхідно бути впевненими в тому, що таке рішення є неризикованим, справедливим та позбавленим упередженості.

- **Великі дані**

Доступ до величезних обсягів даних став основою для навчання ефективних моделей ШІ, зокрема до появи такого напряму, як генеративний ШІ. Великі дані дали змогу знаходити нові та нетривіальні патерни та робити прогнози. Поява та висока популярність генеративного ШІ підняла питання авторських прав під час його навчання, адже, як відомо, Chat GPT компанії OpenAI навчався на всіх доступних даних, що є в інтернеті: книжках, статтях, відкритих бібліотеках та форумах тощо. Це призвело до низки судових позовів від авторів творів, які увійшли в цю навчальну вибірку (гільдія сценаристів Голлівуду та всесвітньо відома письменниця Джоан Роулінг [74]).

2.4.2. Етичні та соціальні виклики

Серед етичних та соціальних викликів щодо застосування систем ШІ можна визначити найважливіші, потенційні ризики для користувачів від яких є найбільшими:

- **Загроза приватності**

Використання ШІ може загрожувати конфіденційності особистих даних, що потребує створення механізмів захисту приватності. Так, робота з великими даними, про які ми згадували вище, пов'язана із новими викликами, що постають перед розробниками – законність збору та використання даних, зокрема – персональних даних користувачів. Особливості використанні чутливих даних – таких, як зображення людини, дані про її здоров'я, біометричні, генетичні. Це є важливою причиною для появи регуляторних актів, що мають визначати законність та етичність використання даних.

- **Дискримінація та упередження**

Моделі ШІ можуть повторювати та підсилювати існуючі соціальні упередження, що потребує спеціальних заходів для забезпечення справедливості.

- **Прозорість і пояснюваність**

Потреба у зрозумілих та пояснюваних рішеннях ШІ, щоб користувачі могли розуміти, як і чому було прийняте те чи інше рішення.

2.4.3. Регуляторні та політичні ініціативи

- **Міжнародні стандарти та рекомендації**

Розробка міжнародних стандартів та рекомендацій, таких як рекомендації ЮНЕСКО та EU AI Act, спрямовані на етичне використання ШІ.

- **Національні стратегії та закони**

Уряди багатьох країн розробляють національні стратегії та законодавство для регулювання використання ШІ, забезпечення його безпеки та етичності.

2.4.4. Інституційна підтримка та співпраця

- **Міжнародні організації**

Такі організації, як ООН, Європейський Союз та інші, активно працюють над створенням нормативно-правової бази та рекомендацій для етичного використання ШІ.

- **Академічні дослідження**

Університети та дослідницькі установи вивчають питання етики, прав людини та соціальних наслідків використання ШІ.

2.4.5. Громадська обізнаність та залученість

- **Просвіта суспільства**

Підвищення рівня обізнаності громадськості щодо потенційних ризиків та переваг ШІ. Для цього необхідно вжити інституційні заходи для підвищення ШІ-грамотності у всіх верствах суспільства: забезпечити базову освіту в галузі ШІ для всіх громадян, навчати їх критично та відповідально ставитися до свого вибору, прав та привілеїв у контексті ШІ та його впливу на повсякденне життя; інформувати громадян про те, як забезпечити конфіденційність персональних даних та контролювати свої дані та рішення; розвати міфи та ажіотаж навколо ШІ, інформуючи населення про межі його можливостей, а також про відмінності між ШІ та людським інтелектом; ретельно вбудовувати нові навички в галузі ШІ в існуючі базові компетенції, такі як медійна та інформаційна грамотність, та визначити найкращі способи об'єднання необхідних умінь, щоб уникнути навантаження навчальних програм.

- **Участь зацікавлених сторін**

Залучення різних груп, включаючи громадянське суспільство, бізнес та користувачів, до процесу обговорення та прийняття рішень щодо використання ШІ.

Ці передумови сприяють збільшенню досліджень в сфері надійного ШІ та створенню систем ШІ, які є безпечними, етичними та корисними для суспільства, забезпечуючи захист прав людини та сприяючи сталому розвитку.

САМОКОНТРОЛЬ

Питання для самоконтролю:

1. Як упередженість може виникнути на етапі збору даних?
2. Які основні проблеми виникають при використанні нерепрезентативних даних?
3. Як використання різних джерел даних може зменшити упередженість?
4. Як можна перевірити точність моделі в умовах можливого упередження?
5. Чим відрізняється упередженість у даних від упередженості в алгоритмах?
6. Які методи використовуються для зменшення стереотипізації в ШІ?
7. Як алгоритми можуть відтворювати соціальні нерівності через упередженість?
8. Що таке fairness-aware learning і як воно допомагає боротися з упередженням?
9. Які підходи до оцінки етичності алгоритмів застосовуються у сучасному ШІ?
10. Як компанії можуть забезпечувати етичне використання штучного інтелекту в продуктах і сервісах?

Контрольні запитання:

Перший рівень складності

Тестові питання:

1. **Що таке упередженість в алгоритмах штучного інтелекту?**
   a) Випадкові помилки в коді
   b) Статистичне відхилення даних
   c) Несправедливе поводження з певними групами людей
   d) Неправильна оптимізація алгоритму

2. Який основний ризик упереджених даних для моделі?
   a) Підвищення точності моделі
   b) Погіршення точності прогнозів
   c) Втрата даних
   d) Прискорення навчання

3. Що таке "відбіркове упередження"?
   a) Упередженість через неправильний вибір даних для навчання
   b) Упередженість у результатах через обмеження вибірки

c) Упередженість у даних через відсутність змінних

d) Упередженість через занадто велику кількість змінних

4. Який підхід може допомогти знизити упередженість у даних?

a) Підвищення кількості даних

b) Ручне видалення упереджених даних

c) Використання різноманітних джерел даних

d) Оптимізація швидкості алгоритму

5. Як перевірити наявність упередженості у моделі?

a) Аналізувати продуктивність моделі

b) Перевіряти точність на тестових даних

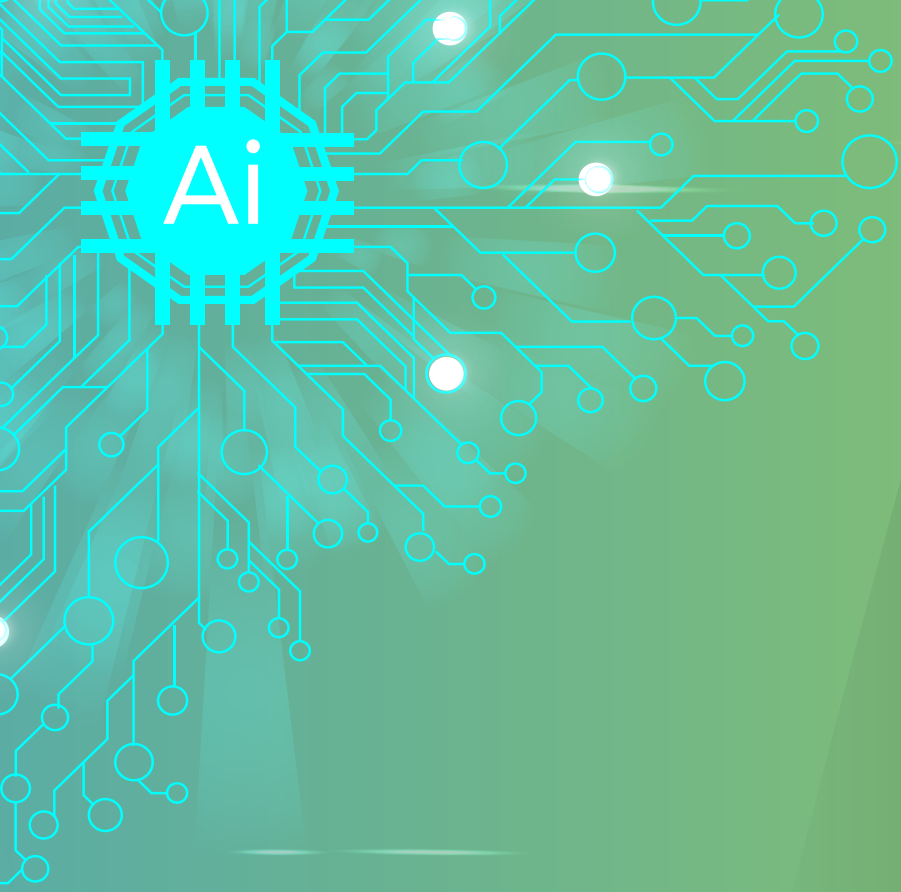
c) Оцінювати результати для різних груп користувачів

d) Вимірювати час виконання алгоритму

Другий рівень складності

Теоретичні питання

1. Що таке упередженість у штучному інтелекті та як вона може виникати?
2. Які існують види упередженості в навчальних даних?
3. Які етичні проблеми пов'язані з упередженістю в моделях штучного інтелекту?
4. Які методи існують для виявлення упередженості в алгоритмах?
5. Як упередженість може вплинути на прийняття рішень моделями ШІ?
6. Що таке "стереотипна упередженість" і як її уникнути в навчанні моделей?
7. Як процес збору даних може впливати на виникнення упередженості?
8. Які існують підходи до мінімізації упередженості в ШІ?
9. Наведіть приклади реальних випадків упередженості у штучному інтелекті.
10. Яким чином прозорість алгоритмів може допомогти знизити ризик упередженості?
11. Дайте визначення справедливості у штучному інтелекті.
12. Опишіть три різні типи справедливості, що використовуються в ШІ.
13. Як забезпечується групова справедливість у системах штучного інтелекту?
14. В чому полягає основна відмінність між демографічним паритетом та рівними можливостями?
15. Наведіть приклад контрфактичної справедливості в системах ШІ.
16. Як можна використовувати методи постобробки для забезпечення групової справедливості?
17. Яким чином процедурна справедливість впливає на прозорість рішень у ШІ?
18. Які основні виклики стоять перед досягненням справедливості в ШІ?
19. Чому усунення упередження є важливою складовою досягнення справедливості?
20. Наведіть приклад несправедливості в системах ШІ та можливі шляхи її усунення.



3

ОСНОВНІ РИЗИКИ В СИСТЕМАХ ШІ

3.1. Фактори ризику в штучному інтелекті

Під час концептуалізації, створення та впровадження системи штучного інтелекту важливо враховувати фактори ризику, пов'язані з даними, що використовуються, побудованою моделлю та обраними алгоритмами, тривалістю життя моделі [11-13]. Крім того, слід також враховувати суспільні наслідки та вплив (негативний чи позитивний; корисний чи шкідливий), що виникають протягом терміну використання системи ШІ.

З точки зору бізнесу існує чіткий зв'язок між визначеним ризиком у системі штучного інтелекту для певної предметної області та тим, наскільки користувачі довіряють рішенням, які вона приймає [27-30]. Більшість факторів ризику повинні бути добре задокументовані.

Упередження є одним із найскладніших факторів, оскільки слід враховувати упередженість, вбудовану в організаційну або виробничу культуру, особисту, несвідому та людську упередженість, а також упередженість у представленні даних [24]. Наприклад, дані, позначені людьми для навчання моделі, можуть бути суб'єктивними навіть серед експертів. Можливо, доведеться розробити різні моделі для різних статей, культур тощо, оскільки рідко можливо узагальнити моделі для всієї людської популяції на основі обмежених даних про навчання.

Справедливість полягає в тому, щоб ставитися до людей однаково через розробку моделей, які інкапсулюють моральні стандарти в процесі прийняття рішень (детальніше проблему справедливості в ШІ розглянуто в розділі2).

Зрозумілість потрібна, щоб усі зацікавлені сторони, включаючи людей, на яких впливають рішення автоматизованих систем, могли побачити та зрозуміти, як приймається рішення, а користувач знав, чому система прийняла саме таке рішення щодо нього [27].

Суспільний вплив, який може бути як потенційно корисним, так і шкідливим, компанія повинна враховувати не лише для пом'якшення репутаційної шкоди у разі судових скарг, але й для відповідності або перевищення мінімальних стандартів ділової етики. Підприємства повинні визначити, де відповідальність (завдання та зобов'язання) лежить у їхній структурі управління штучним інтелектом, і визначити підзвітність (нагляд і відповідальність) для ролей протягом життєвого циклу проектування, розробки чи розгортання. З огляду на зміни в законодавстві про штучний інтелект, науковці та промисловість, незалежно від розміру, потребують глибокого осмислення та консенсусу щодо цих факторів ризику, щоб оцінити ризик рішення штучного інтелекту як для окремих людей, так і для суспільства. Зараз проблема полягає в тому, щоб подолати розрив між принципами та практикою, тому є певна впевненість, що системи ШІ відповідають узгодженим принципам.

3.2. Приклади ризику в системах штучного інтелекту

Ризики для справедливості

Інструмент штучного інтелекту, який оцінює кредитоспроможність позичальників, навчається на неповних або упереджених даних, що змушує компанію пропонувати кредити особам на різних умовах залежно від раси чи статі.

Ризики для конфіденційності та нагляду

Підключені пристрої в домі можуть постійно збирати дані, включаючи розмови, потенційно створюючи майже повний портрет домашнього життя людини. Ризики конфіденційності посилюються, із збільшенням кількості осіб, які мають доступ до цих даних.

Ризики для прав людини

Генеративний ШІ може бути використаний для генерації deepfake із протиправною чи аморальною поведінкою, що потенційно може завдати шкоди репутації, стосункам і гідності суб'єкта.

Ризики для безпеки

Асистент штучного інтелекту на основі технології LLM (Large Language Model) рекомендує небезпечну активність, яку він знайшов в Інтернеті, не розуміючи та не повідомляючи контексту веб-сайту, де була описана ця діяльність. Користувач здійснює цю діяльність, завдаючи собі чи оточуючим фізичної шкоди.

Ризики для суспільного добробуту

Дезінформація, створена та поширена ШІ, може підірвати доступ до надійної інформації та довіру до демократичних інститутів і процесів.

Ризики для безпеки

Інструменти штучного інтелекту можна використовувати для автоматизації, прискорення та посилення впливу цілеспрямованих кібератак, підвищуючи серйозність загрози від зловмисників. Поява LLM дозволяє хакерам [примітка 48] з невеликими технічними знаннями чи навичками створювати фішингові кампанії з можливостями доставки зловмисного програмного забезпечення.

САМОКОНТРОЛЬ

Запитання для самоконтролю:

1. Які основні фактори ризику слід враховувати під час розробки системи ШІ?
2. Чому важливо оцінювати суспільні наслідки використання систем ШІ?
3. Як упередження впливають на дані, що використовуються для навчання моделей?
4. Яка роль справедливості в розробці систем штучного інтелекту?
5. Що таке зрозумілість у контексті прийняття рішень системами ШІ?
6. Які наслідки можуть виникнути через недотримання етичних стандартів у бізнесі?
7. Яким чином відповідальність у структурі управління штучним інтелектом впливає на його безпеку?
8. Чому важливо досягати консенсусу щодо факторів ризику в ШІ між науковцями та промисловістю?
9. Які приклади ризиків для справедливості можуть виникнути в системах ШІ?
10. Як ризики для безпеки можуть вплинути на користувачів технологій ШІ?

Контрольні запитання:

Перший рівень складності

Тестові питання:

1. **Який із наведених факторів ризику найбільше пов'язаний із упередженнями в даних?**
   A) Тривалість життя моделі
   B) Конфіденційність даних
   C) Культура організації
   D) Вибір алгоритму

2. **Який із наступних аспектів є ключовим для забезпечення справедливості в системах ШІ?**
   A) Висока продуктивність моделі
   B) Наявність великих обсягів даних
   C) Урахування моральних стандартів
   D) Використання складних алгоритмів

3. Що може бути наслідком відсутності зрозумілості в системах ШІ?
   A) Підвищення довіри користувачів
   B) Невірні рішення, прийняті системою
   C) Зменшення часу на навчання моделі
   D) Підвищення ефективності роботи

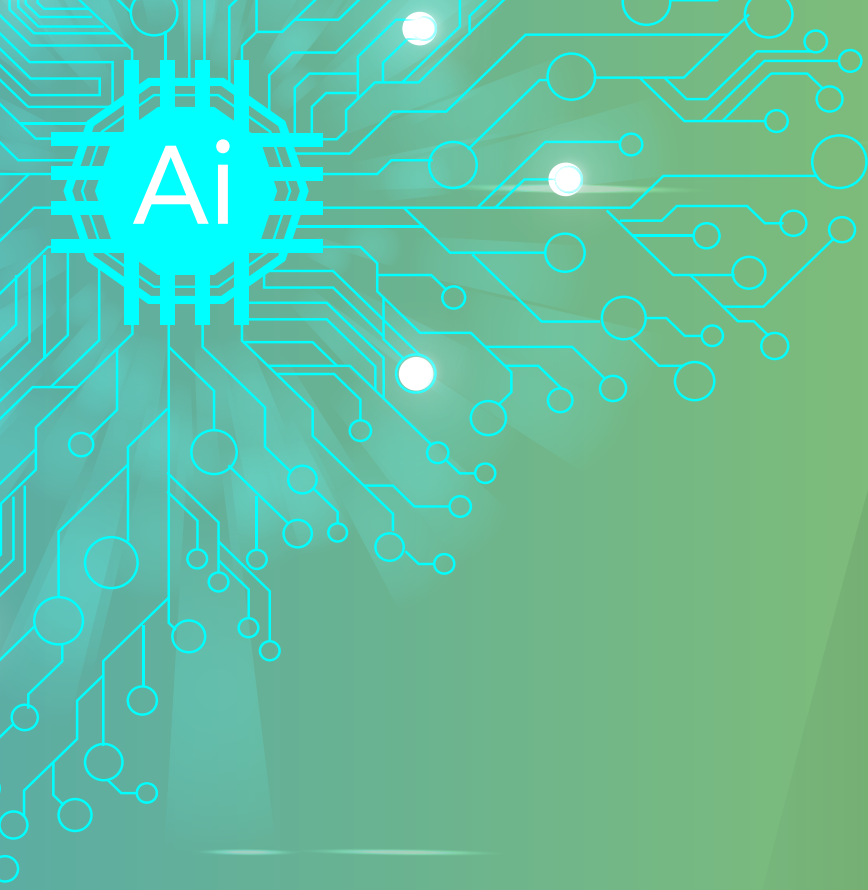
4. Який з ризиків найбільше пов'язаний із конфіденційністю даних?
   A) Генерація deepfake
   B) Неправильна оцінка кредитоспроможності
   C) Збір особистих даних через підключені пристрої
   D) Автоматизація кібератак

5. Яка з наведених стратегій може допомогти у пом'якшенні ризиків у системах ШІ?
   A) Використання закритих алгоритмів
   B) Розробка етичних стандартів для AI
   C) Уникнення публікації результатів
   D) Зосередження на швидкості розробки

Другий рівень складності

Теоретичні питання:

1. Які ризики пов'язані з упередженням в даних для навчання моделей ШІ?
2. Яким чином генерування deepfake може порушувати права людини?
3. Як підключені пристрої можуть загрожувати конфіденційності користувачів?
4. Які наслідки має дезінформація, створена штучним інтелектом, для суспільного добробуту?
5. Які заходи можна вжити для пом'якшення ризиків, пов'язаних з штучним інтелектом?
6. Як системи ШІ можуть впливати на довіру користувачів до прийнятих рішень?
7. Які проблеми можуть виникнути через недостатню зрозумілість алгоритмів прийняття рішень?
8. Яким чином використання ШІ може загрожувати безпеці користувачів?
9. Як можна визначити та оцінити ризики в системах штучного інтелекту?
10. Чому важливо враховувати культурні та соціальні аспекти під час розробки моделей ШІ?



4

ОСНОВНІ ПРИНЦИПИ НАДІЙНОГО ШІ

Надійний штучний інтелект (НШІ, Trustworthy Artificial Intelligence) – це підхід до розробки та використання систем штучного інтелекту, спрямований на створення технологій, які володіють високою стійкістю, безпекою, прозорістю та відповідальністю перед користувачами та суспільством загалом. Основна ідея полягає в створенні інтелектуальних систем, які можуть ефективно функціонувати в різних умовах, забезпечуючи при цьому безпеку, етичність та відповідальність в їх використанні. НШІ спрямований на вирішення питань безпеки даних, уникнення непередбачених помилок, пояснюваності прийнятих рішень та створення систем, які можуть взаємодіяти з людьми та іншими системами з високим рівнем довіри. Це важлива область розвитку штучного інтелекту, оскільки вона враховує не лише технічні аспекти, а й соціальні та етичні вимоги використання цих технологій.

Надійність штучного інтелекту стала особливо нагальним питанням із появою генеративного штучного інтелекту і систем розмовного штучного інтелекту, які повинні бути розроблені та розгорнутітак, щоб віддати пріоритет етичності, надійності, справедливості, прозорості та підзвітності. Оскільки межі між можливостями людини та штучного інтелекту стають все більш розмитими, надійний штучний інтелект прагне зберегти ці два світи розділеними, але гармонійними — роботи тут не для того, щоб взяти гору, вони тут, щоб допомогти. Використання ШІ в рамках надійної системи означає, що технології штучного інтелекту поважають людські цінності, захищають основні права та дотримуються етичних стандартів, водночас забезпечуючи послідовні та надійні результати. Чесність і прозорість процесів сприяє зміцненню довіри користувачів і добробуту суспільства, тому всі виграють, коли ми залишаємось чесними.

Надійний штучний інтелект має три основні компоненти, які необхідно враховувати протягом усього життєвого циклу системи ШІ:

1) бути законною (Законний ШІ), відповідати всім чинним законам і правилам;

2) бути етичною, тобто такою (Етичний ШІ), що забезпечує дотримання етичних принципів і цінностей;

3) бути стійкою (стійкий, надійний, стабільний ШІ) як з технічної, так і з соціальної точки зору, оскільки, навіть маючи добрі наміри, системи штучного інтелекту можуть завдати ненавмисної шкоди.

Кожен з цих трьох компонентів необхідний, але недостатній сам по собі для створення ШІ, що заслуговує на довіру. В ідеалі, всі три компоненти працюють гармонійно і перетинаються у своїй діяльності. Однак на практиці між цими елементами можуть виникати суперечності (наприклад, іноді обсяг і зміст чинного законодавства може не відповідати етичним нормам). Саме тому важливо працювати над тим, щоб усі три компоненти сприяли забезпеченню надійного ШІ [61].

4.1. Основні компоненти надійного ШІ

4.1.1. Законний ШІ

Сьогодні існує низка юридично обов'язкових правил на європейському, національному та міжнародному рівнях, які застосовуються або мають відношення до розробки, розгортання та використання систем штучного інтелекту. Такими документами є: Первинне право ЄС (Договори Європейського Союзу та Хартія основних прав), вторинне право ЄС (наприклад, Загальний регламент захисту даних, Директива про відповідальність за продукцію, Регламент про вільний потік неперсональних даних, антидискримінаційні директиви, законодавство про захист прав споживачів та Директиви про безпеку та гігієну праці), договори ООН з прав людини та конвенції Ради Європи (наприклад, Європейська конвенція з прав людини), Закон про штучний інтелект (EU AI Act), а також численні закони держав-членів ЄС [76-81].

Крім правил, що застосовуються горизонтально, існують різні галузеві правила, які застосовуються до конкретних застосувань ШІ (наприклад, Регламент про медичні вироби в секторі охорони здоров'я). Закон передбачає як позитивні, так і негативні зобов'язання, а це означає, що його слід тлумачити не лише з точки зору того, що не можна робити, але й з точки зору того, що слід робити і що можна робити. Відповідно закон не лише забороняє певні дії, але й дозволяє інші. У зв'язку з цим можна відзначити, що Хартія ЄС містить статті про «свободу ведення бізнесу» і «свободу мистецтв і наук», а також статті, що стосуються сфер, з якими ми більше знайомі в контексті забезпечення довіри до ШІ, таких як, наприклад, захист даних і недискримінація. Настанови не стосуються безпосередньо першого компонента надійного ШІ (законного ШІ), а мають на меті запропонувати рекомендації щодо розвитку та забезпечення другого і третього компонентів (етичного та стійкого ШІ). Хоча два останні компоненти певною мірою вже відображені в чинному законодавстві, їхня повна реалізація може виходити за рамки існуючих правових зобов'язань [75-77].

Ці настанови ґрунтуються на припущенні, що всі юридичні права та обов'язки, які застосовуються до процесів і діяльності, пов'язаної з розробкою, розгортанням і використанням систем ШІ, залишаються обов'язковими і повинні належним чином дотримуватися.

4.1.2. Етичний ШІ

Досягнення довіри до ШІ вимагає не лише дотримання законодавства, яке є лише однією з трьох його складових. Закони не завжди йдуть в ногу з технологічним розвитком, іноді можуть не відповідати етичним нормам або просто не підходять для вирішення певних питань. Для того, щоб системам ШІ можна було довіряти, вони також повинні бути етичними, тобто відповідати етичним нормам [76].

Етику в системах ШІ доцільно розглядати у двох основних аспектах: етичні принципи, які покладено в основу 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 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, 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 технологій ШІ 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 правильність 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 абсолютно різні аспекти етики ШІ: етичність прийняття рішень 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, необхідно мати гарантії, що система керується етичними міркуваннями. Тому так важливо закласти 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, зокрема: порушення приватності, можлива дискримінація, соціальне розшарування, проблеми працевлаштування тощо. Окремим питанням є тема професійної етики розробників систем ШІ, яка теж вимагає розгляду, і в перспективі створення етичних кодексів та рекомендацій для розробників.

Вже понад 40 країн затвердили або розробили національні стратегії з регулювання ШІ, адже у майбутньому людство не відмовиться від ШІ, а, швидше за все, ШІ буде все більше, ширше і активніше використовуватися в різних сферах. Важливо якнайшвидше задати етичні рамки, в яких він розвиватиметься, обмежити можливості його неетич-

ного застосування та спрямувати енергію розробників та ідеї законодавців у русло, що забезпечує максимальну безпеку та вигоду для суспільства.

4.1.3. Надійний(стабільний, стійкий) ШІ

Навіть якщо етична мета забезпечена, люди і суспільство повинні бути впевнені, що системи штучного інтелекту не завдадуть ненавмисної шкоди. Такі системи повинні працювати безпечно, надійно і безвідмовно, а для запобігання будь-яких ненавмисних негативних наслідків слід передбачити запобіжні заходи. Тому важливо забезпечити стійкість систем штучного інтелекту. Це необхідно як з технічної точки зору (забезпечення технічної надійності системи відповідно до конкретного контексту, наприклад, сфери застосування або фази життєвого циклу), так і з соціальної точки зору (з належним урахуванням контексту і середовища, в якому працює система).

4.2. Основні принципи надійного ШІ

4.2.1. Принцип поваги до автономії людини

Основні права, на яких ґрунтується ЄС, спрямовані на забезпечення поваги до свободи та автономії людини [77]. Люди, які взаємодіють з системами штучного інтелекту, повинні мати можливість зберігати повне і ефективне самовизначення, а також мати можливість брати участь у демократичному процесі. Системи штучного інтелекту не повинні невиправдано підпорядковувати, примушувати, обманювати, маніпулювати, обмежувати в правах людей або стежити за ними. Вони повинні бути розроблені так, щоб збільшувати, доповнювати і розширювати когнітивні, соціальні та культурні навички людини. Розподіл функцій між людиною і системами ШІ повинен відповідати принципам людиноцентричного дизайну і залишати людині можливість вибору. Це означає забезпечення людського нагляду за робочими процесами в системах ШІ. Системи штучного інтелекту також можуть докорінно змінити сферу праці. Вони повинні підтримувати людину в робочому середовищі і бути спрямовані на створення змістовної роботи.

4.2.2.Принцип запобігання шкоді

Системи штучного інтелекту не повинні ані спричиняти, ані посилювати шкоду або іншим чином негативно впливати на людей. Це передбачає захист людської гідності, а також психічної та фізичної недоторканності. Системи штучного інтелекту та середовища, в яких вони працюють, повинні бути безпечними та захищеними. Вони повинні бути технічно надійними та убезпечені від зловмисного використання. Вразливим групам населення слід приділяти більше уваги і залучати їх до розробки, розгортання і вико-

ристання систем штучного інтелекту. Особливу увагу слід також приділяти ситуаціям, коли системи ШІ можуть спричинити або посилити негативний вплив через асиметрію влади або інформації, наприклад, між роботодавцями і працівниками, бізнесом і споживачами, урядами і громадянами. Запобігання шкоді також передбачає турботу про природне середовище і всіх живих істот.

Слід визнати, що технології штучного інтелекту самі по собі не обов'язково забезпечують процвітання людей, навколишнього середовища та екосистем. Крім того, жоден із процесів, пов'язаних із життєвим циклом системи штучного інтелекту, не повинен перевищувати того, що необхідно для досягнення законних цілей або завдань, і повинен відповідати контексту. Якщо система ШІ гіпотетично здатна зашкодити людям, правам людини та основним свободам, громадам і суспільству загалом або навколишньому середовищу та екосистемам, необхідно впровадити процедуру оцінки ризику цієї системи ШІ та вжити всіх можливих заходів для запобігання виникненню такої шкоди.

Рішення про необхідність використання систем штучного інтелекту має бути обґрунтовано такими способами:

(a) обраний метод ШІ має бути адекватним і пропорційним для досягнення певної законної мети;
(b) обраний метод ШІ не повинен порушувати основоположних цінностей надійного ШІ;
(c) метод штучного інтелекту має відповідати контексту та базуватися на строгій науковій основі.

У випадках, коли прийняте рішення має незворотний вплив, яке важко скасувати або воно може включати рішення про життя чи смерть, остаточне рішення має приймати людина. Зокрема, системи штучного інтелекту не можна використовувати для соціального оцінювання чи масового спостереження

4.2.3. Принцип справедливості

Розробка, розгортання та використання систем штучного інтелекту повинні бути справедливими. У розділі 2 було розглянуто, що існують різні інтерпретацій справедливості, однак ми розглянемо визначення справедливості, яке має як матеріальний, так і процедурний вимір.

Матеріальний вимір справедливості передбачає зобов'язання забезпечити рівний і справедливий розподіл як вигод, так і витрат, а також гарантувати, що окремі особи і групи будуть вільні від несправедливої упередженості, дискримінації та маргіналізації. Якщо несправедливої упередженості можна уникнути, то системи ШІ можуть навіть підвищити соціальну справедливість. Слід також сприяти рівним можливостям у доступі до освіти, товарів, послуг і технологій. Крім того, використання систем ШІ ніколи не повинно призводити до обману людей або невиправданого обмеження їхньої свободи вибору. Крім того, справедливість передбачає, що фахівці, які застосовують ШІ, повинні поважати принцип пропорційності між засобами і цілями, а також ретельно продумувати, як збалансувати конкуруючі інтереси і цілі.

Процедурний вимір справедливості передбачає можливість оскаржувати рішення, прийняті системами ШІ і людьми, які ними керують, і домагатися ефективного відшкодування збитків. Для цього суб'єкт, відповідальний за рішення, повинен бути ідентифікований, а процеси прийняття рішень повинні бути зрозумілими.

4.2.4.Принцип зрозумілості

Зрозумілість має вирішальне значення для побудови та підтримки довіри користувачів до систем штучного інтелекту. Це означає, що процеси повинні бути прозорими, можливості та призначення систем ШІ повинні відкрито повідомлятися, а рішення, наскільки це можливо, повинні бути зрозумілими для тих, кого вони прямо чи опосередковано стосуються. Без такої інформації рішення не може бути належним чином оскаржене. Пояснити, чому модель згенерувала певний результат або рішення (і яка комбінація вхідних факторів сприяла цьому), не завжди можливо. Такі випадки називаються алгоритмами «чорної скриньки» і потребують особливої уваги. За таких обставин можуть знадобитися інші заходи забезпечення зрозумілості (наприклад, відстежуваність, можливість аудиту та прозоре інформування про можливості системи) за умови, що система загалом поважає основоположні права. Ступінь, до якого необхідна прозорість, значною мірою залежить від контексту та серйозності наслідків, якщо результат є помилковим або якоюсь мірою неточним.

4.3. Основні вимоги до надійного ШІ

Основні принципи ШІ, описані вище, повинні бути переведені в конкретні вимоги для створення ШІ, що заслуговує на довіру. Ці вимоги стосуються різних зацікавлених сторін, які беруть участь у життєвому циклі систем ШІ: розробників, компаній, що впроваджують системи ШІ, та кінцевих користувачів, а також широкої громадськості. Під розробниками ми маємо на увазі тих, хто досліджує, проектує та/або розробляє системи ШІ. Під впроваджувачами ми маємо на увазі державні та приватні організації, які використовують системи штучного інтелекту у своїх бізнес-процесах, а також для надання продуктів і послуг іншим. Кінцеві користувачі – це ті, хто безпосередньо чи опосередковано взаємодіє з системою штучного інтелекту. Нарешті, широке суспільство охоплює всіх інших, на кого системи штучного інтелекту безпосередньо чи опосередковано впливають.

Різні групи зацікавлених сторін відіграють різну роль у забезпеченні виконання вимог:

а) Розробники повинні впроваджувати та застосовувати вимоги до процесів проектування та розробки;

б) Впроваджувачі повинні переконатися, що системи, які вони використовують, а також продукти та послуги, які вони пропонують, відповідають вимогам;

в) Кінцеві користувачі та суспільство загалом повинні бути поінформовані про ці вимоги та мати можливість вимагати їх дотримання.

Наведений нижче перелік вимог до надійного ШІ не є вичерпним. Він включає системні, індивідуальні та суспільні аспекти:

4.3.1. Людська участь та нагляд (human agency and oversight)

Системи штучного інтелекту повинні підтримувати людську автономію та приймати рішення, керуючись принципом поваги до людської автономії. Це означає, що системи штучного інтелекту повинні сприяти розвитку демократичного, процвітаючого і справедливого суспільства, підтримуючи свободу дій користувачів, а також сприяти дотриманню фундаментальних прав і дозволяти здійснювати контроль з боку людини.

- **Дотримання фундаментальних прав людини**

Як і багато інших технологій, системи штучного інтелекту можуть як сприяти, так і перешкоджати дотриманню основоположних прав. Вони можуть приносити користь людям, наприклад, допомагаючи їм відстежувати свої персональні дані або підвищуючи доступність освіти, а отже, підтримуючи їхнє право на освіту. Однак, враховуючи охоплення і можливості систем штучного інтелекту, вони також можуть негативно впливати на основоположні права. У ситуаціях, коли такі ризики існують, слід проводити оцінку впливу на основоположні права. Це має бути зроблено до початку розробки системи і включати оцінку того, чи можуть ці ризики бути зменшені або виправдані, відповідно до норм демократичного суспільства для дотримання прав і свобод людей. Крім того, слід створити механізми для отримання зовнішнього зворотного зв'язку щодо систем ШІ, які потенційно порушують основоположні права.

- **Людська активність**

Користувачі повинні мати можливість самостійно приймати обгрунтовані рішення щодо систем ШІ. Їм повинні бути надані знання та інструменти для розуміння і взаємодії з системами ШІ на задовільному рівні, а також, де це можливо, вони повинні мати можливість розумної самооцінки або оскарження системи. Системи штучного інтелекту повинні допомагати людям робити кращий, більш усвідомлений вибір відповідно до їхніх цілей. Іноді системи ШІ можуть використовуватися для формування поведінки людини та впливу на неї за допомогою механізмів, які може бути важко виявити, оскільки вони можуть використовувати підсвідомі процеси, включаючи різні форми недобросовісної маніпуляції, обману та обумовленості, які можуть загрожувати індивідуальній автономії. Загальний принцип автономії користувача повинен бути центральним у функціонуванні системи. Ключовим для цього є право не бути об'єктом рішення, що ґрунтується виключно на автоматизованій обробці, коли це має юридичні наслідки для користувачів або аналогічним чином суттєво впливає на них.

- **Людський нагляд**

Людський нагляд допомагає гарантувати, що система ШІ не підриває автономію людини і не спричиняє інших негативних наслідків.

Нагляд може бути досягнутий за допомогою механізмів управління, таких як:

- **«людина в циклі» (HITL – human-in-the-loop)**

Human-in-the-Loop (HITL) – це підхід до розробки штучного інтелекту, який передбачає людський нагляд і контроль над процесом машинного навчання. HITL зазвичай використовується для перевірки та вдосконалення моделей машинного навчання, а також для обробки винятків і граничних ситуацій, з якими системі штучного інтелекту важко впоратися самостійно.

У HITL люди є невід'ємною частиною системи, контролюючи та контролюючи ШІ, забезпечуючи зворотний зв'язок і приймаючи рішення на основі результатів ШІ.

HITL означає можливість втручання людини в кожен цикл прийняття рішень у системі. Цей механізм сьогодні є ключовим та широко застосовуваним для LLM, в яких зворотний зв'язок від людини дозволяє покращувати якість роботи самої моделі. Окрім того, для певних класів задач постійна участь людини у контролі прийняття рішення системою є критично необхідною – у випадку військових, медичних, складних промислових задач.

Роль людей у процесі «людина в петлі» може бути різною, зокрема:

1. Анотація та маркування даних

Однією з найважливіших ролей людей у HITL є анотація та маркування даних. У контрольованому навчанні, коли алгоритми навчаються на даних із мітками, люди надають анотації або мітки для навчання моделей. Наприклад, під час розпізнавання зображень люди можуть позначати зображення, щоб ідентифікувати об'єкти, що допомагає алгоритмам навчитися точно розпізнавати ці об'єкти.

2. Очищення та попередня обробка даних

Втручання людини часто необхідне для очищення та попередньої обробки даних перед передачею їх у моделі машинного навчання. Люди можуть визначати та виправляти невідповідності, відсутні значення або помилки в наборі даних, гарантуючи, що модель навчається на високоякісних даних.

3. Вибір і налаштування алгоритму

Люди відіграють важливу роль у виборі відповідних алгоритмів машинного навчання, точному налаштуванні гіперпараметрів і конфігурації моделей на основі проблемної області та конкретних вимог. Їхній досвід допомагає оптимізувати продуктивність моделі.

4. Навчання та перевірка моделі

Тоді як алгоритми машинного навчання автоматизують навчання моделі, люди контролюють цей процес, вибираючи навчальні набори даних, перевіряючи про-

дуктивність моделі та приймаючи рішення щодо можливостей узагальнення моделі та потенційних упереджень.

5. Обробка граничних випадків і неоднозначностей

Люди чудово справляються з неоднозначними або складними ситуаціями, з якими можуть боротися алгоритми. Вони можуть обробляти граничні випадки, винятки або ситуації, які виходять за межі навчальних даних моделі, забезпечуючи надійність і адаптивність у сценаріях реального світу.

6. Постійний моніторинг і зворотний зв'язок

Навіть після розгортання та впровадження системи люди залишаються залученими в цикл шляхом моніторингу продуктивності моделі, виявлення упереджень і надання зворотного зв'язку. Цей постійний цикл зворотного зв'язку допомагає вдосконалювати моделі, підвищувати точність і гарантувати дотримання етичних міркувань.

Серед **переваг підходу «людина в циклі»** в машинному навчанні можна виокремити такі:

1. Підвищена точність моделі

Участь людини допомагає вдосконалювати моделі, підвищувати їхню точність і зменшувати кількість помилок, забезпечуючи розуміння контексту та досвід.

2. Етична розробка ШІ

Люди можуть оцінювати та пом'якшувати упередження, забезпечуючи справедливість, прозорість і етичне використання систем штучного інтелекту, що має вирішальне значення в чутливих програмах, таких як охорона здоров'я, фінанси та право.

3. Адаптивність до нових сценаріїв

Втручання людини дозволяє моделям адаптуватися до нових і непередбачених ситуацій, покращуючи їхні можливості узагальнення.

4. Підвищена впевненість і довіра

Людський нагляд вселяє довіру до систем штучного інтелекту, надаючи пояснення, інтерпретації та гарантуючи, що рішення відповідають людським цінностям і намірам.

Недоліки підходу «людина в циклі»

1. Вартість і час

Залучення людини може збільшити вартість і час, необхідні для машинного навчання, особливо в таких завданнях, як маркування даних, яке може бути трудомістким.

2. Суб'єктивність і упередженість

Самі люди можуть вносити упередження або суб'єктивність, впливаючи на якість позначених даних або прийняття рішень у циклі.

3. Масштабованість

Зі збільшенням обсягів даних масштабованість стає проблемою в управлінні людською участю, необхідною для анотації та контролю.

- «людина на зв'язку» (HOTL – human-on-the-loop)

HOTL означає можливість людського втручання під час циклу проектування системи і моніторингу роботи системи.

Human-on-the-Loop (HOTL) — це розширення HITL, яке залучає людей, які надають зворотний зв'язок системі штучного інтелекту для покращення її продуктивності з часом. HOTL зазвичай використовується, коли система штучного інтелекту досягла певного рівня продуктивності, але все ще потребує відгуку та втручання людини для продовження вдосконалення. У HOTL люди діють як тренери або вчителі для ШІ, надаючи позначені дані, виправляючи помилки та направляючи ШІ до кращих результатів.

HOTL часто використовується в автономних транспортних засобах, програмах виявлення шахрайства та медичної діагностики.

- «людина в управлінні» (HIC – human-in-command).

Human-in-command (HIC) – це можливість нагляду за загальною діяльністю системи штучного інтелекту (включно з її ширшим економічним, соціальним, правовим і етичним впливом) і можливість вирішувати, коли і як використовувати систему в будь-якій конкретній ситуації. Це може включати рішення не використовувати систему ШІ в конкретній ситуації, встановлення рівнів людського розсуду під час використання системи або забезпечення можливості скасувати рішення, прийняте системою.

Крім того, необхідно забезпечити, щоб державні виконавці мали можливість здійснювати нагляд відповідно до своїх повноважень. Механізми нагляду можуть знадобитися в різному ступені для підтримки інших заходів безпеки та контролю, залежно від сфери застосування системи ШІ та потенційного ризику. За інших рівних умов, чим менше людина може здійснювати нагляд за системою ШІ, тим більший обсяг тестування і суворіше управління потрібні.

4.3.2. Технічна надійність та безпека (technical robustness and safety)

Важливим компонентом досягнення надійного ШІ є технічна надійність, яка тісно пов'язана з принципом запобігання шкоді та включає стійкість до атак і безпеку, запасний план і загальну безпеку, точність, стійкість і відтворюваність. Технічна надійність вимагає, щоб системи ШІ розроблялися із застосуванням превентивного підходу до

ризиків – тобто, щоб вони надійно поводилися відповідно до призначення, мінімізуючи ненавмисну і неочікувану шкоду, а також запобігаючи неприйнятній шкоді [71]. Це також має стосуватися потенційних змін у робочому середовищі або присутності інших агентів (людських і штучних), які можуть взаємодіяти із системою у несприятливий для неї спосіб. Крім того, має бути забезпечена фізична і психічна недоторканність людини.

Необхідно уникати небажаної шкоди (ризиків для безпеки), а також вразливості до атак, їх слід розглядати, запобігати та усувати протягом усього життєвого циклу систем штучного інтелекту, щоб гарантувати безпеку людей, навколишнього середовища та екосистем. Безпечний і захищений штучний інтелект стане можливим завдяки розробці стійких структур доступу до даних із захистом конфіденційності, які сприятимуть кращому навчанню та перевірці моделей штучного інтелекту з використанням якісних даних.

- **Стійкість до атак і безпека**

Системи ШІ, як і всі програмні системи, повинні бути захищені від вразливостей, які можуть бути використані зловмисниками, наприклад, від хакерських атак. Атаки можуть бути спрямовані на дані (отруєння даних), модель (витік моделі) або базову інфраструктуру, як програмну, так і апаратну. Якщо систему ШІ атакують, наприклад, під час ворожих атак, дані, а також поведінка системи можуть бути змінені, що призведе до прийняття системою інших рішень або до її повного вимкнення. Системи та дані також можуть бути пошкоджені через зловмисні наміри або через вплив непередбачуваних ситуацій. Недостатні процеси безпеки також можуть призвести до помилкових рішень або навіть фізичної шкоди. Для того, щоб системи ШІ вважалися безпечними, необхідно враховувати можливі ненавмисні застосування системи ШІ (наприклад, програми подвійного призначення) і потенційні зловживання системою з боку зловмисників, а також вживати заходів для їхнього запобігання та пом'якшення наслідків [72].

- **Альтернативний план і загальна безпека**

Системи штучного інтелекту повинні мати засоби захисту, які забезпечують альтернативний план на випадок виникнення проблем. Це може означати, що системи ШІ переходять від статистичних процедур до процедур, заснованих на правилах, або що вони запитують людину-оператора перед тим, як продовжити свою дію. Необхідно гарантувати, що система буде робити те, що вона повинна робити, не завдаючи шкоди живим істотам або навколишньому середовищу [73]. Це включає в себе мінімізацію непередбачуваних наслідків і помилок. Крім того, слід розробити процеси для виявлення та оцінки потенційних ризиків, пов'язаних з використанням систем штучного інтелекту в різних сферах застосування. Рівень необхідних заходів безпеки залежить від величини ризику, який становить система ШІ, що, своєю чергою, залежить від можливостей системи. Якщо можна передбачити, що процес розробки або сама система становитимуть особливо високі ризики, вкрай важливо розробити і протестувати заходи безпеки заздалегідь.

- **Точність**

Точність стосується здатності системи ШІ робити правильні висновки, наприклад, правильно класифікувати інформацію за відповідними категоріями, або її здатності робити правильні прогнози, рекомендації чи рішення на основі даних або моделей. Чіткий і добре сформований процес розробки та оцінювання може підтримувати, пом'якшувати і виправляти непередбачувані ризики, пов'язані з неточними прогнозами. Якщо випадкових неточних прогнозів уникнути неможливо, важливо, щоб система могла вказати, наскільки ймовірні ці помилки. Високий рівень точності особливо важливий у ситуаціях, коли система ШІ безпосередньо впливає на людські життя [74].

- **Надійність і відтворюваність**

Також надзвичайно важливо, щоб результати систем штучного інтелекту були відтворюваними та надійними. Надійна система ШІ – це система, яка працює належним чином з різними вхідними даними і в різних ситуаціях. Це необхідно для ретельного вивчення системи ШІ та запобігання ненавмисній шкоді. Відтворюваність описує, чи демонструє експеримент зі штучним інтелектом однакову поведінку при повторенні в однакових умовах. Це дає змогу вченим і політикам точно описати, що роблять системи штучного інтелекту. Файли реплікації можуть полегшити процес тестування та відтворення поведінки.

4.3.3. Конфіденційність та управління(керування) даними (privacy, and data governance)

Конфіденційність та управління (керування) даними пов'язані із повагою до приватності, якості та цілісності даних, а також правилами доступу до даних.

- **Конфіденційність та керування даними (Privacy and Data Governance)**

З принципом запобігання шкоді тісно пов'язане право на недоторканність приватного життя – фундаментальне право, на яке особливо впливають системи ШІ. Запобігання шкоді, що може бути спричинена порушенням приватності, також вимагає адекватного управління даними, яке охоплює якість і цілісність використовуваних даних, їхню актуальність з огляду на сферу, в якій будуть застосовуватися системи ШІ, протоколи доступу до них і здатність обробляти дані таким чином, щоб захистити приватність.

- **Конфіденційність і захист даних**

Системи ШІ повинні гарантувати конфіденційність і захист даних протягом усього життєвого циклу системи. Сюди входить інформація, яку спочатку надає користувач, а також інформація, яка генерується про користувача в процесі його взаємодії з системою (наприклад, результати, які система ШІ генерує для конкретних користувачів, або те, як користувачі реагують на певні рекомендації). Цифрові записи людської поведін-

ки можуть дозволити системам штучного інтелекту робити висновки не лише про вподобання людини, але й про її сексуальну орієнтацію, вік, стать, релігійні чи політичні погляди. Щоб люди могли довіряти процесу збору даних, необхідно гарантувати, що зібрані про них дані не будуть використані для незаконної або несправедливої дискримінації [83].

Також варто пам'ятати, що перераховані типи даних (дані про здоров'я, біометричні, генетичні дані, інформація про сексуальну орієнтацію, вік, стать, релігійні чи політичні погляди) належать до чутливих даних відповідно до Загальноєвропейського регламенту із захисту персональних даних (GDPR), що накладає особливо суворі правила на їх збір та обробку.

- **Якість і цілісність даних**

Якість наборів даних, що використовуються, найбільше впливає на продуктивність систем штучного інтелекту. Коли дані збираються, вони можуть містити соціально сконструйовані упередження, неточності та помилки. Це необхідно враховувати перед початком навчання на будь-якому наборі даних. Крім того, необхідно забезпечити цілісність даних. Введення шкідливих даних у систему ШІ може змінити її поведінку, особливо в системах, що самонавчаються. Процеси і набори даних, що використовуються, повинні бути протестовані і задокументовані на кожному етапі, такому як планування, навчання, тестування і розгортання. Це також має стосуватися систем ШІ, які не були розроблені власними силами, а придбані деінде.

- **Доступ до даних**

У будь-якій організації, яка обробляє персональні дані (незалежно від того, чи є вона користувачем системи, чи ні), слід запровадити протоколи, що регулюють доступ до даних. Ці протоколи повинні визначати, хто може отримати доступ до даних і за яких обставин. Доступ до даних про особу має бути дозволений лише належним чином кваліфікованому персоналу, який має компетенцію та потребу в доступі до даних про особу.

4.3.4. Прозорість та відкритість (transparency)

Ця вимога тісно пов'язана з принципом зрозумілості і передбачає прозорість елементів, що мають відношення до системи ШІ: даних, системи та бізнес-моделей.

- **Відстежуваність**

Набори даних і процеси, на основі яких система ШІ приймає рішення, включно зі збором і маркуванням даних, а також алгоритми, що використовуються, повинні бути задокументовані відповідно до найкращих стандартів, щоб забезпечити відстежуваність і підвищити прозорість. Це також стосується рішень, прийнятих системою штучного інтелекту. Це дає змогу виявити причини помилкових рішень, що, своєю чергою,

може допомогти запобігти помилкам у майбутньому. Простежуваність полегшує аудит, а також пояснюваність [77].

- **Пояснюваність**

Пояснюваність стосується здатності пояснити як технічні процеси системи ШІ, так і пов'язані з ними людські рішення (наприклад, із сфери застосування системи). Технічна пояснюваність вимагає, щоб рішення, прийняті системою ШІ, могли бути зрозумілими і простежуваними людиною. Крім того, може знадобитися компроміс між покращенням пояснюваності системи (що може знизити її точність) або підвищенням її точності (за рахунок пояснюваності). Щоразу, коли система ШІ має значний вплив на життя людей, повинна бути можливість вимагати відповідного пояснення процесу прийняття рішень системою ШІ [83]. Таке пояснення має бути своєчасним і адаптованим до знань зацікавленої сторони (наприклад, неспеціаліста, регулятора або дослідника). Крім того, мають бути доступними пояснення того, якою мірою система ШІ впливає на процес прийняття рішень в організації та формує його, вибір дизайну системи та обґрунтування для її розгортання (таким чином забезпечуючи прозорість бізнес-моделі). Саме ця вимога надійного штучного інтелекту призвела до появи нового напряму досліджень – створення пояснюваних методів ШІ (Explainable AI).

- **Комунікація**

Системи штучного інтелекту не повинні видавати себе користувачам за людей; люди мають право бути поінформованими про те, що вони взаємодіють із системою штучного інтелекту. Це означає, що системи штучного інтелекту повинні бути ідентифіковані як такі. Крім того, для забезпечення дотримання основоположних прав слід передбачити можливість відмовитися від такої взаємодії на користь взаємодії з людиною, якщо це необхідно. Крім того, можливості та обмеження системи ШІ повинні бути доведені до відома фахівців-практиків або кінцевих користувачів у спосіб, що відповідає конкретному випадку використання. Це може включати інформування про рівень точності системи ШІ, а також про її обмеження [77].

4.3.5. Різноманітність (різномаїття), недискримінація (відсутність дискримінації) та справедливість (Diversity, non-discrimination and Fairness)

Для того, щоб досягти надійного ШІ, ми повинні забезпечити інклюзивність і різноманітність протягом усього життєвого циклу системи ШІ. Окрім врахування та залучення всіх зацікавлених сторін протягом усього процесу це також передбачає забезпечення рівного доступу через інклюзивні процеси проектування, а також рівне ставлення. Ця вимога тісно пов'язана з принципом справедливості, запобіганням несправедливої упередженості, доступністю та універсальним дизайном, а також участю зацікавлених сторін.

- **Уникнення несправедливої упередженості**

Набори даних, що використовуються системами ШІ (як для навчання, так і для роботи), можуть включати ненавмисні історичні упередження, неповні чи неточні моделі управління. Збереження таких упереджень може призвести до ненавмисного непрямого упередження і дискримінації певних груп або людей, що потенційно посилює упередження і маргіналізацію. Шкода також може бути спричинена навмисною експлуатацією упереджень або недобросовісною конкуренцією, наприклад, гомогенізацією цін за допомогою змови або непрозорого ринку [76]. Упередження, які можна ідентифікувати як дискримінаційні, слід усунути на етапі збору даних, якщо це можливо. Спосіб, у який розробляються системи ШІ (наприклад, програмування алгоритмів), також може містити елементи несправедливої упередженості. Цьому можна протидіяти, запровадивши процеси нагляду для аналізу та вирішення питань, пов'язаних з метою, обмеженнями, вимогами та рішеннями системи, у чіткий і прозорий спосіб. Крім того, наймання на роботу людей з різним досвідом, культурою та дисциплінами може забезпечити розмаїття думок, і це слід заохочувати [78].

- **Доступність та універсальний дизайн**

Зокрема, у сфері взаємодії бізнесу зі споживачами, системи повинні бути орієнтовані на користувача і розроблені таким чином, щоб усі люди могли користуватися продуктами або послугами штучного інтелекту, незалежно від їхнього віку, статі, здібностей або особливостей. Особливе значення має доступність цієї технології для людей з обмеженими можливостями, які присутні в усіх соціальних групах. Системи штучного інтелекту не повинні мати універсального підходу і повинні враховувати принципи універсального дизайну, орієнтуючись на якомога ширше коло користувачів і дотримуючись відповідних стандартів доступності. Це забезпечить рівний доступ і активну участь усіх людей в існуючих і нових видах людської діяльності, опосередкованих комп'ютером, а також у допоміжних технологіях.

- **Участь зацікавлених сторін**

Для того, щоб розробити системи штучного інтелекту, які заслуговують на довіру, доцільно консультуватися із зацікавленими сторонами, на яких система може безпосередньо чи опосередковано впливати протягом усього її життєвого циклу. Корисно регулярно отримувати зворотний зв'язок навіть після розгортання системи та створити довгострокові механізми участі зацікавлених сторін, наприклад, шляхом забезпечення інформування, консультацій та участі працівників протягом усього процесу впровадження систем ШІ в організаціях.

4.3.6. Соціальний та екологічний добробут (Добробут суспільства та довкілля (societal and environmental well-being))

Відповідно до принципів справедливості та запобігання шкоді, суспільство загалом, інші живі істоти та навколишнє середовище також повинні розглядатися як зацікавлені сторони протягом усього життєвого циклу системи ШІ. Слід заохочувати стійкість і екологічну відповідальність систем ШІ, а також сприяти дослідженням у галузі ШІ-рішень, спрямованих на вирішення глобальних проблем, таких як, наприклад, цілі сталого розвитку. В ідеалі системи штучного інтелекту повинні використовуватися на благо всіх людей, включно з майбутніми поколіннями.

- **Сталий та екологічно чистий ШІ**

Системи штучного інтелекту розробляються, щоб допомогти вирішити деякі з найгостріших суспільних проблем, але необхідно забезпечити, щоб це відбувалося в максимально екологічний спосіб. Процес розробки, розгортання та використання системи, а також весь ланцюжок її постачання повинні оцінюватися з цієї точки зору, наприклад, шляхом критичного аналізу використання ресурсів і споживання енергії під час навчання, щоб зробити вибір на користь менш шкідливих рішень. Слід заохочувати заходи, що забезпечують екологічність усього ланцюжка постачання систем ШІ.

Сьогодні ми бачимо виклик у виконанні цієї вимоги до систем ШІ, оскільки навчання та використання дуже складних моделей, що лежать в основі, наприклад, систем генеративного ШІ, вимагають значних обчислювальних ресурсів протягом довгого часу, а відповідно – значних енергетичних витрат.

- **Соціальний вплив**

Сьогодні ми спостерігаємо значний вплив соціальних систем ШІ на всі сфери нашого життя (в освіті, роботі, догляді чи розвагах). Він може змінити наше уявлення про соціальну активність або вплинути на наші соціальні стосунки та прив'язаність. Хоча системи штучного інтелекту можуть використовуватися для покращення соціальних навичок, вони також можуть спричиняти їхнє погіршення. Це також може вплинути на фізичний і психічний добробут людей. Тому вплив цих систем необхідно ретельно відстежувати і враховувати.

- **Суспільство і демократія**

Окрім оцінки впливу розробки, розгортання та використання систем ШІ на окремих осіб, цей вплив слід також оцінювати з точки зору суспільства, беручи до уваги його вплив на інститути, демократію та суспільство в загалом. Використання систем штучного інтелекту слід ретельно розглядати, особливо в ситуаціях, пов'язаних з демократичним процесом, включаючи не тільки прийняття політичних рішень, а й виборчий контекст.

4.3.7. Звітність (відповідальність, підзвітність, Accountability)

Ця вимога передбачає можливість аудиту (перевірки) мінімізацією та звітуванням про негативний вплив, компроміси та відшкодування.

- **Підзвітність (Відповідальність, Звітність)**

Вимога підзвітності доповнює вищезазначені вимоги і тісно пов'язана з принципом справедливості. Вона вимагає створення механізмів, що забезпечують відповідальність і підзвітність за системи ШІ та їхні результати як до, так і після їхніх розробки, розгортання та використання.

- **Можливість аудиту**

Можливість аудиту системи ШІ передбачає можливість оцінки алгоритмів, даних і процесів проектування. Це не обов'язково означає, що інформація про бізнес-моделі та інтелектуальну власність, пов'язану з системою ШІ, завжди повинна бути у відкритому доступі. Оцінка внутрішніми і зовнішніми аудиторами та наявність таких звітів про оцінку можуть сприяти підвищенню довіри до технології. У додатках, що впливають на основоположні права, в тому числі критично важливі для безпеки, системи ШІ повинні мати можливість проходити незалежний аудит.

- **Мінімізація та звітування про негативні впливи**

Необхідно забезпечити можливість звітувати про дії або рішення, які сприяють отриманню певного результату системи, і реагувати на наслідки такого результату. Виявлення, оцінка, документування та мінімізація потенційних негативних впливів систем ШІ є особливо важливими для тих, на кого вони впливають (безпосередньо чи опосередковано). Необхідно забезпечити належний захист інформаторам, неурядовим організаціям, профспілкам та іншим суб'єктам, які повідомляють про законні занепокоєння щодо системи ШІ. Для мінімізації негативного впливу може бути корисним використання оцінок впливу як до, так і під час розробки, розгортання та використання систем ШІ. Ці оцінки повинні бути пропорційними до ризику, який становлять системи штучного інтелекту.

- **Компроміси**

При реалізації вищезазначених вимог між ними можуть виникати протиріччя, що може призвести до неминучих компромісів. Такі компроміси повинні вирішуватися раціонально і методологічно в рамках сучасного рівня техніки. Це означає, що необхідно визначити відповідні інтереси і цінності, які зачіпає система ШІ, і що в разі виникнення конфлікту компроміси повинні бути чітко визнані й оцінені з точки зору їх ризику для етичних принципів, зокрема фундаментальних прав. У ситуаціях, коли етично прийнятні компроміси не можуть бути визначені, розробка, розгортання і використання сис-

теми ШІ не повинні продовжуватися в такій формі. Будь-яке рішення про компроміс має бути обґрунтованим і належним чином задокументованим. Особа, яка приймає рішення, повинна нести відповідальність за спосіб, у який приймається відповідний компроміс, і постійно переглядати адекватність прийнятого рішення, щоб забезпечити можливість внесення необхідних змін до системи, якщо це потрібно.

- **Відшкодування**

У разі несправедливого негативного впливу слід передбачити доступні механізми, які забезпечать адекватне відшкодування. Знання того, що відшкодування можливе, якщо щось пішло не так, є ключовим для забезпечення довіри. Особливу увагу слід приділяти вразливим особам або групам.

Підводячи підсумки, варто зазначити, що хоча всі вимоги є однаково важливі, при їх застосуванні в різних сферах і галузях необхідно враховувати контекст і потенційні суперечності між ними. Впровадження цих вимог має відбуватися протягом усього життєвого циклу системи ШІ і залежить від конкретного застосування. Хоча більшість вимог застосовуються до всіх систем ШІ, особлива увага приділяється тим, які прямо чи опосередковано впливають на людей. Тому для деяких застосувань (наприклад, у промисловості) вони можуть бути менш актуальними. Вищезазначені вимоги включають елементи, які в деяких випадках вже відображені в чинному законодавстві. Відповідальність за дотримання своїх правових зобов'язань несуть фахівці, які застосовують ШІ, як щодо горизонтально застосовних правил, так і щодо специфічного регулювання в конкретних галузях.

САМОКОНТРОЛЬ

Запитання для самоконтролю

1. Що таке «людина в циклі» (HITL) в контексті штучного інтелекту?
2. Які основні вимоги до надійного штучного інтелекту визначаються в керівництвах ЄС?
3. Чому важливо забезпечити людське управління (HIC) в автоматизованих системах?
4. Які аспекти технічної надійності є критично важливими для безпеки систем ШІ?
5. Які елементи включаються в концепцію справедливості в системах ШІ?
6. Як управління даними може вплинути на конфіденційність у системах ШІ?
7. Які ризики можуть виникнути внаслідок недостатньої різноманітності в даних для навчання моделей?
8. Чому важливо враховувати різноманіття в розробці систем ШІ?
9. Які приклади можуть ілюструвати дискримінацію в системах ШІ?
10. Яка роль технічних стандартів у забезпеченні надійності та безпеки ШІ?

Контрольні запитання:

Перший рівень складності

Тестові питання

1. Що означає термін «людина на зв'язку» (HOTL)?
   A) Людина, що контролює всю систему
   B) Людина, яка спостерігає за автоматизованими рішеннями
   C) Людина, що бере участь у прийнятті рішень
   D) Людина, що повністю відсторонена від системи

2. Який з наступних аспектів є критично важливим для технічної надійності ШІ?
   A) Висока вартість розробки
   B) Надійні алгоритми
   C) Відкритість даних
   D) Широкий спектр застосувань

3. Який із наступних принципів сприяє справедливості в системах ШІ?
   A) Автоматизація всіх процесів
   B) Використання обмежених даних

C) Упередженість у навчанні
D) Виключення певних груп населення з аналізу

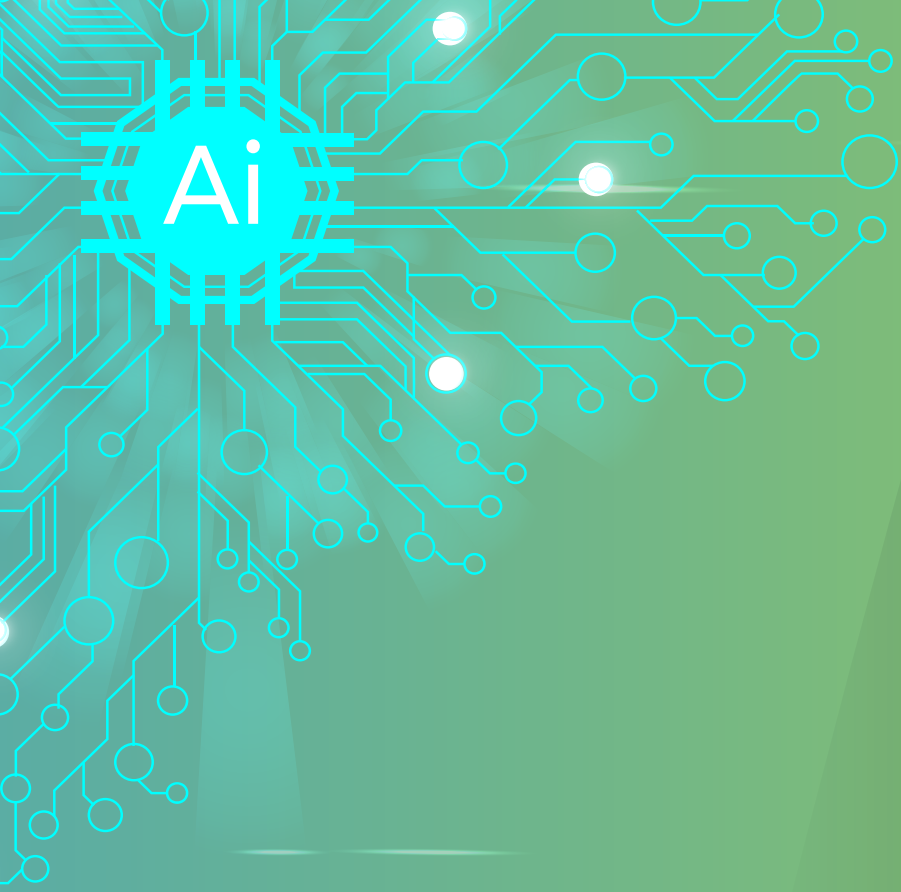
4. Яка роль «людина в управлінні» (HIC) в системах ШІ?
   A) Тільки контроль за безпекою
   B) Повна автономія системи
   C) Прийняття фінальних рішень
   D) Взаємодія з системою лише у випадку помилки

5. Що є наслідком відсутності різноманітності в даних для навчання моделей ШІ?
   A) Підвищення точності моделей
   B) Зменшення ризиків
   C) Дискримінація певних груп
   D) Зниження витрат на розробку

Другий рівень складності

Теоретичні питання

1. Яка основна мета впровадження HITL в системи ШІ?
2. Які основні ризики виникають через відсутність людського нагляду в системах ШІ?
3. Чим відрізняється HOTL від HITL?
4. Які принципи сприяють недискримінації у використанні систем ШІ?
5. Які етапи включає в себе управління даними в контексті штучного інтелекту?
6. Чому технічна надійність є критично важливою для успішної інтеграції ШІ в суспільство?
7. Як різноманіття може вплинути на точність і справедливість моделей ШІ?
8. Яка роль державного регулювання у забезпеченні етичного використання ШІ?
9. Як можуть бути реалізовані принципи справедливості в алгоритмах прийняття рішень?
10. Чому важливо розглядати суспільний вплив нових технологій у контексті їхнього впровадження?



5

ОЦІНКА СИСТЕМ ШІ, ЩО БАЗУЄТЬСЯ НА РИЗИКАХ

Підхід до регулювання, заснований на оцінці ризиків, виник у різних галузях економіки та технологій задовго до розвитку штучного інтелекту. Зокрема, ризико-орієнтований підхід є однією з ключових концепцій, впроваджених у стандарті ISO 9001:2015 [84]. Його основна мета полягає в інтеграції управління ризиками в усі процеси організації для забезпечення стабільної якості продукції або послуг. Відповідно до ISO 9001:2015, організації мають ідентифікувати та оцінювати ризики і можливості, що можуть вплинути на досягнення запланованих результатів, і вживати відповідних заходів для управління ними. Цей підхід дозволяє організаціям ефективніше реагувати на невизначеності та підтримувати безперервне вдосконалення.

Відповідно до визначення, наведеного у стандарті ISO 9001:2015: «Ризик – це «вплив невизначеності» на очікуваний результат». (ISO 9001:2015 - Системи управління якістю – Вимоги) [84].

5.1. Основні принципи ризико-орієнтованого підходу

Розглянемо основні принципи ризико-орієнтованого підходу:

1. Ідентифікація ризиків і можливостей

Одним із фундаментальних аспектів ризико-орієнтованого підходу є ідентифікація ризиків, які можуть негативно вплинути на здатність організації досягати своїх цілей, а також можливостей, що можуть сприяти поліпшенню процесів або результатів. Ризики можуть виникати в будь-якій частині системи управління якістю, починаючи від виробничих процесів і закінчуючи взаємодією з клієнтами.

2. Оцінка і пріоритезація ризиків

Після ідентифікації ризики оцінюються з точки зору ймовірності їх виникнення та потенційних наслідків. Оцінка дозволяє організації зрозуміти, які ризики мають найвищий потенціал впливу на якість продукту або послуги. Ризики з високою ймовірністю виникнення або значними негативними наслідками повинні бути пріоритетними для вжиття коригувальних дій.

3. Управління ризиками

Відповідно до ISO 9001:2015, організації мають розробити і впровадити заходи з управління ризиками для запобігання можливим загрозам або мінімізації їх впливу. Управління ризиками може включати превентивні заходи, адаптацію процесів, розробку планів дій на випадок непередбачених обставин та регулярний моніторинг ситуації.

Наприклад, у виробничих процесах це може означати впровадження систем контролю якості для запобігання браку або відстеження потенційних відхилень.

4. Інтеграція ризиків у процеси організації

Важливим аспектом ризико-орієнтованого підходу є його інтеграція в повсякденну діяльність організації. Це означає, що управління ризиками не повинно бути окремим процесом, а має стати невід'ємною частиною всіх процесів і прийняття рішень. Наприклад, під час планування нових проектів або запуску нових продуктів ризики повинні бути враховані на ранніх етапах, щоб уникнути можливих проблем у майбутньому.

5. Можливості як позитивний аспект управління ризиками

Крім управління негативними ризиками, стандарт ISO 9001:2015 підкреслює важливість виявлення можливостей — позитивних ризиків, які можуть покращити продуктивність або збільшити ефективність організації. Це може включати нові технології, вдосконалення процесів або нові ринки. Організації повинні вміти не лише мінімізувати загрози, але й активно використовувати можливості для розвитку.

Зв'язок між ризиками, плануванням і якістю

Ризико-орієнтований підхід відповідно до ISO 9001:2015 тісно пов'язаний із процесом планування в системі управління якістю. Організації повинні планувати свої дії на основі ризиків, які вони ідентифікують. Це включає:

- Визначення цілей та шляхів їх досягнення з урахуванням можливих ризиків.
- Визначення необхідних ресурсів для управління ризиками.
- Впровадження заходів для запобігання або зменшення впливу ризиків.
- Визначення методів оцінки результативності цих заходів.

Такий підхід дозволяє забезпечити стабільність і передбачуваність процесів, що сприяє поліпшенню якості продукції або послуг.

Документування та аналіз ризиків

ISO 9001:2015 не вимагає від організацій створювати формальні документи для управління ризиками, однак організаціям рекомендовано документувати основні ризики та заходи щодо їх усунення або мінімізації.

Цей процес може включати:

- Ведення реєстру ризиків.
- Оцінку ризиків у вигляді матриць або моделей ризиків.
- Розробку та моніторинг планів дій для управління ризиками.

Регулярний аналіз ризиків також є важливою частиною цього процесу. Організації повинні періодично переглядати свої ризики та можливості, зокрема в рамках внутрішніх аудитів, для того, щоб оцінити, чи дійсно заходи, які вживаються, є ефективними та достатніми.

- **Постійне вдосконалення та коригувальні дії**

Одним із ключових принципів ISO 9001:2015 є постійне вдосконалення, яке також стосується управління ризиками. Організації повинні постійно аналізувати свої ризики, результати управління ними та впроваджувати коригувальні дії у разі виявлення недоліків або нових загроз. Це забезпечує гнучкість та адаптивність системи управління якістю до мінливих умов [84].

Оскільки застосування ризико-орієнтованого підходу, впровадженого у стандарті ISO 9001:2015, стало важливим інструментом для забезпечення стабільної якості продукції та послуг в умовах невизначеності, який дав можливість організаціям не лише вчасно реагувати на загрози, але й використовувати можливості для розвитку та вдосконалення, сприяє підвищенню конкурентоспроможності та стійкості компаній в сучасному динамічному середовищі – цілком логічно виглядає вибір даного підходу для оцінки систем ШІ.

У сфері ШІ ризик-орієнтований підхід почав активно розроблятися в останні десятиліття, коли вплив технологій на суспільство став масштабнішим, а ризики від їх застосування можуть бути доволі критичними.

Конкретний ризик-орієнтований підхід до регулювання ШІ був вперше чітко запропонований Європейською комісією в рамках її Білої книги щодо штучного інтелекту, опублікованої у лютому 2020 року [77]. У цьому документі підкреслюється необхідність створення регуляторної рамки для ШІ, заснованої на аналізі та оцінці ризиків, щоб гарантувати як безпеку, так і відповідальне використання технологій. Європейська Комісія запропонувала класифікацію систем ШІ на низько-, середньо- і високоризикові для різних сфер застосування, і залежно від цього пропонувалося впроваджувати різні заходи контролю та нагляду.

Відповідно у контексті ризико-орієнтованого підходу до регулювання штучного інтелекту ризик – це потенційна загроза або негативний вплив, який може виникнути в результаті використання ШІ-системи. Ризики стосуються різних аспектів: від технічної безпеки до етичних та соціальних наслідків. За таким підходом ризики зазвичай визначаються як комбінація двох компонентів: ймовірності виникнення негативної події (оцінка того, наскільки вірогідним є виникнення негативного сценарію або небажаного результату) та серйозності її наслідків (наскільки серйозні наслідки цього сценарію для конкретних осіб, суспільства, економіки або держави).

5.2. Основні типи ризиків у ШІ

1. Технічні ризики

- Помилки або збої в роботі ШІ-системи, які можуть призвести до небажаних результатів (наприклад, аварії в автономних транспортних засобах).
- Некоректні або упереджені алгоритми, що можуть призвести до неточних або несправедливих рішень (наприклад, дискримінація за гендерними чи расовими ознаками).

2. Етичні ризики

- Використання ШІ для порушення прав людини (наприклад, системи стеження або автономна зброя).
- Проблеми з прозорістю рішень ШІ (відсутність пояснення щодо того, як і чому було прийнято певне рішення).

3. Правові ризики

- Відповідальність за дії систем ШІ (хто несе відповідальність у випадку, коли ШІ завдає шкоди, наприклад, у сфері автономних транспортних засобів).
- Відповідність ШІ-систем законодавству, зокрема щодо захисту даних або приватності (наприклад, дотримання GDPR в ЄС).

4. Соціальні ризики

- Вплив ШІ на ринок праці (автоматизація може призвести до скорочення робочих місць у певних секторах).
- Підрив соціальної довіри до технологій через непрозорі або несправедливі рішення ШІ.

5. Ризики кібербезпеки

- Можливість хакерських атак на системи ШІ, які можуть бути використані для маніпуляцій або саботажу.
- Використання ШІ для створення нових загроз, таких як deepfake або автоматизовані кіберзлочини.

Приклади ризиків у контексті ШІ:

- **Ризик упередженості:** Алгоритми ШІ, особливо системи машинного навчання, можуть навчатися на упереджених даних, що призводить до несправедливих або дискримінаційних рішень.
- **Ризик автономності:** Автономні системи, які приймають рішення без людського втручання, можуть діяти непередбачувано або небезпечно, наприклад, автономні транспортні засоби або дрони.
- **Ризик витоку даних:** ШІ-системи, які обробляють великі обсяги даних, можуть становити ризик для конфіденційності інформації, особливо якщо вони не відповідають вимогам кібербезпеки.

Оцінка та управління ризиками

За ризико-орієнтованим підходом до ШІ ризики класифікують на основі їх імовірності та наслідків, після чого для них розробляють відповідні стратегії управління. Це можуть бути:

- **Запобігання ризикам** (наприклад, шляхом адаптації алгоритмів або контролю якості).
- **Зменшення ризиків** (наприклад, через посилення кібербезпеки або аудит ШІ-систем).
- **Моніторинг ризиків** (регулярна перевірка роботи систем ШІ та їх впливу на користувачів).

Отже, ризик у контексті ШІ — це комплексний термін, що охоплює всі можливі загрози, пов'язані з технологічними, етичними, правовими та соціальними аспектами застосування штучного інтелекту, і його управління є критично важливим для безпечного розвитку та використання таких технологій.

Запитання для самоконтролю

1. Що таке ризико-орієнтований підхід і як він використовується в стандарті ISO 9001:2015?
2. Як визначаються ризики в системах управління якістю відповідно до ISO 9001:2015?
3. Яка мета управління ризиками у стандарті ISO 9001:2015?
4. Які основні етапи оцінки ризиків у стандарті ISO 9001:2015?
5. Що таке можливості в контексті управління ризиками за стандартом ISO 9001:2015?
6. Як впливає ідентифікація ризиків на якість продукції або послуг?
7. Як можна інтегрувати ризико-орієнтований підхід у щоденну діяльність організації?
8. Яким чином документи з управління ризиками допомагають організації у стандарті ISO 9001:2015?
9. Що таке коригувальні дії і як вони допомагають у постійному вдосконаленні?
10. Які основні типи ризиків виникають у штучному інтелекті?

Контрольні запитання:

Перший рівень складності

Тестові питання

1. Що таке ризик відповідно до ISO 9001:2015?
   a) Певний результат діяльності
   b) Вплив невизначеності на очікуваний результат
   c) Відсутність контролю над процесом
   d) Результат неочікуваних дій

2. Що є однією з основних цілей ризико-орієнтованого підходу у ISO 9001:2015?
   a) Мінімізувати витрати на виробництво
   b) Інтегрувати управління ризиками у всі процеси організації
   c) Забезпечити контроль за всіма співробітниками
   d) Підвищити швидкість виконання проектів

САМОКОНТРОЛЬ

3. Що таке можливості в контексті управління ризиками за ISO 9001:2015?
   a) Ризики, які можна уникнути
   b) Позитивні аспекти, що можуть покращити продуктивність
   c) Загрози, що можуть бути ліквідовані
   d) Фінансові ресурси для управління ризиками

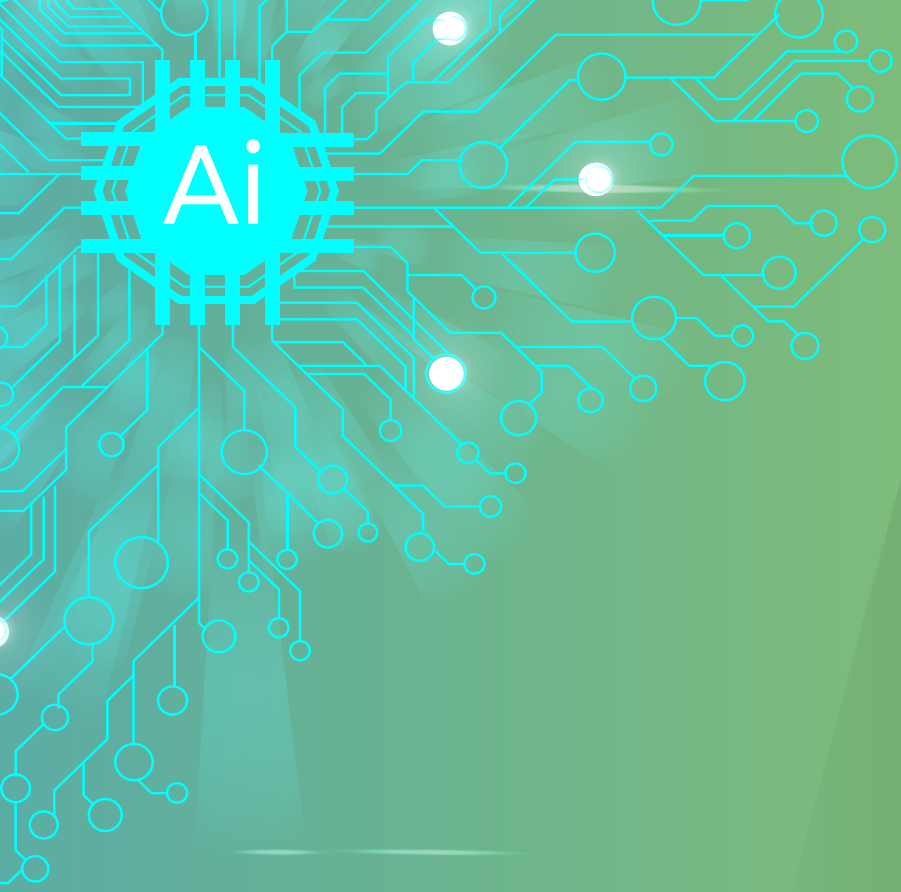
4. Яка дія є прикладом управління ризиками відповідно до ISO 9001:2015?
   a) Ігнорування можливих загроз
   b) Впровадження превентивних заходів
   c) Зниження кількості робочих годин
   d) Встановлення нових стандартів безпеки для виробничого процесу

5. Який з цих ризиків є прикладом технічного ризику у штучному інтелекті?
   a) Відсутність даних для навчання
   b) Упередженість алгоритмів
   c) Зменшення робочих місць через автоматизацію
   d) Недотримання законів про захист даних

Другий рівень складності

Теоретичні питання

1. Яке визначення ризику наведено у стандарті ISO 9001:2015?
2. Як проводиться пріоритезація ризиків відповідно до стандарту ISO 9001:2015?
3. Які приклади управління ризиками можна навести з виробничих процесів?
4. Як ризико-орієнтований підхід сприяє стабільній якості продукції?
5. Як постійне вдосконалення впливає на управління ризиками у стандарті ISO 9001:2015?
6. Які основні кроки необхідні для ефективного управління ризиками в організації?
7. Що таке можливості у стандарті ISO 9001:2015 і як їх можна використовувати?
8. Як здійснюється аналіз ризиків у рамках внутрішніх аудитів?
9. Як впровадження ризико-орієнтованого підходу може допомогти у уникненні збоїв у роботі ШІ-систем?
10. Як Європейська комісія пропонує класифікувати системи ШІ за рівнем ризику?



6

ЗАКОНОДАВЧЕ РЕГУЛЮВАННЯ ВИКОРИСТАННЯ СИСТЕМ ШІ

6.1. Історія розвитку законодавчого регулювання використання ШІ

Сьогодні ШІ розвивається надзвичайно швидко, його можливості та точність роботи настільки високі, що розробки із використанням ШІ зустрічаються вже у всіх сферах людського життя. Системи ШІ широко застосовуються для покращення сфери охорони здоров'я (наприклад, встановлення точнішого діагнозу, забезпечення кращої профілактики захворювань), підвищення ефективності сільського господарства, сприяння пом'якшенню наслідків зміни клімату та адаптації до нього, підвищення ефективності виробничих систем шляхом прогнозного обслуговування, підвищення безпеки громадян тощо.

Однак водночас штучний інтелект тягне за собою низку потенційних ризиків, таких як непрозоре прийняття рішень, гендерна чи інша дискримінація, втручання в наше приватне життя або використання в злочинних цілях.

Відповідно в суспільстві та урядах багатьох країн з'явилось розуміння того, що має бути розроблене законодавство, яке б регламентувало використання методів ШІ в різних системах, щоб захистити права та свободи громадян.

У 2019 році Асоціація обчислювальної техніки (Association for Computing Machinery, ACM) випустила нові етичні стандарти професійної поведінки, а IEEE опублікувала рекомендації щодо етичного проектування автономних та інтелектуальних систем, демонструючи зрушення серед професійних технологічних організацій у бік пріоритетності етики. Паралельно тисячі професіоналів у галузі технологій і соціальних науковців сформували міждисциплінарні комітети для розробки етичних принципів для проектування, розробки та використання технологій ШІ. Крім того, багато урядів і міжнародних організацій опублікували набори етичних принципів, зокрема Принципи OECP у 2019 році [79], Монреальська декларація у 2017 році [75], звіт Палати лордів Великобританії «ШІ у Великобританії: готові, бажають і здатні? » у 2018 р. [76] Група експертів високого рівня Європейської комісії (HLEG) у ШІ у 2018 році [77] та Пекінські принципи AI у 2019 році [78]. Загалом сьогодні існує понад 70 загальнодоступних наборів етичних принципів або рамок для штучного інтелекту, більшість з яких були опубліковані протягом останніх п'яти років.

Також у квітні 2018 року в ЄС було презентовано європейську стратегію штучного інтелекту, відповідно до якої, щоб пом'якшити виклики штучного інтелекту, ЄС повинен діяти як єдине ціле та визначити власний шлях, заснований на європейських цінностях, сприяти розвитку та розгортанню штучного інтелекту. Європейська Комісія прагне забезпечити науковий прорив, зберегти технологічне лідерство ЄС, однак, забезпечуючи при цьому повагу до прав та свобод громадян. Скоординований європейський підхід до людських та етичних наслідків штучного інтелекту, а також міркування щодо кращого використання великих даних для інновацій є основою регуляторного та інвестиційно-орієнтованого підходу з подвійною метою: сприяння поширенню ШІ та усунення ризиків, пов'язаних із певним використанням цієї нової технології. Ця стратегія була сформульована у Білій книзі штучного інтелекту [77], розробленій групою експертів із штучного інтелекту та оприлюдненій у лютому 2020 року. В цій книзі визначені варіанти

політики щодо досягнення етичності та справедливості систем ШІ, однак документ не стосується розробки та використання штучного інтелекту у військових цілях.

Однак, питання регулювання ШІ обговорюється та вирішується не лише в ЄС, але і в більшості інших розвинених країн світу. Так, у 2020 році Адміністративно-бюджетне управління США випустило Посібник з регулювання застосування штучного інтелекту [85], а у червні 2021 року Генеральна конференція Організації Об'єднаних Націй з питань освіти, науки та культури (ЮНЕСКО) представила проект тексту Рекомендації щодо етики штучного інтелекту, який зосереджується на орієнтованому на людину підході до штучного інтелекту, рекомендувавши, що «ШІ має бути корисним для людей, а не навпаки». [87]. Уряд Великої Британії надав оновлене резюме даних і етичних принципів штучного інтелекту, розроблених як державним сектором, так і урядом у 2020 році [88], яке включало спільну публікацію про керівні принципи розробки систем штучного інтелекту, розроблені разом із Всесвітнім економічним форумом [86], а також конкретні рекомендації та контрольний список для використання ШІ в охороні здоров'я. У 2021 році Рада зі штучного інтелекту Великобританії опублікувала дорожню карту щодо штучного інтелекту [88], подальші «посібники» щодо розробки систем і національну стратегію даних.

6.2. Регулювання використання систем ШІ у Європі

Правова база у сфері ШІ та технологій, що керуються даними, є відносно новою та швидко розвивається. Вперше питання щодо автоматизованого прийняття рішень було визначено у GDPR (General Data Protection Regulation, Загальноєвропейський регламент із захисту персональних даних) у 2018 році у статті 22, де було визначено серію заходів безпеки та інформаційних зобов'язань [89]. Зокрема, вона вимагає надання суб'єкту даних повноважень, як зазначено в Декларації 71, «не бути суб'єктом рішення, заснованого виключно на автоматизованій обробці, включно з профілюванням, яке створює юридичні наслідки щодо нього чи неї або подібним чином істотно впливає на нього чи неї», право попросити втручання людини, пояснення того, як було прийнято автоматизоване рішення «з урахуванням логіки». У пункті 71 зазначено, що контролер даних повинен використовувати відповідні математичні та статистичні процедури для профілювання, а дані мають бути точними, щоб мінімізувати ризик помилок [89].

6.2.1. Історія створення AI Act

Опублікована у 2018 році в ЄС стратегія розвитку штучного інтелекту, просувала людино-орієнтований підхід, зосереджений на повазі європейських цінностей і прав людини. Саме він увійшов в основу нормативно-правової бази щодо штучного інтелекту, яка містить рамки для оцінки ризику будь-якого продукту, послуги або системи ШІ [86].

Після широкого обговорення сформульованих принципів в експертному колі було сформульовано проект закону із регулювання штучного інтелекту в ЄС. Європей-

ська комісія оприлюднила пропозицію щодо нового закону про штучний інтелект (EU AI Act) у квітні 2021 року [83].

Комісія запропонувала:

- закріпити в законодавстві ЄС технологічно нейтральне визначення систем ШІ;
- прийняти набір правил, розроблених відповідно до підходу, що ґрунтується на оцінці ризику (заборонені системи штучного інтелекту, системи штучного інтелекту високого ризику, що підпадають під низку суворих вимог (наприклад, управління, тестування, навчання даних), системи штучного інтелекту, що становлять обмежений ризик, підлягають обмеженим вимогам до прозорості та системи штучного інтелекту, що представляють лише низький або мінімальний ризик.

У грудні 2022 року Рада ухвалила спільну позицію («загальний підхід») щодо акту про AI. У запропонованому тексті Ради, зокрема:

- звужено визначення системи ШІ;
- поширено на приватних акторів заборону на використання штучного інтелекту для соціального оцінювання;
- додано горизонтальний шар поверх класифікації високого ризику, щоб гарантувати, що системи штучного інтелекту, які не можуть спричинити серйозні порушення фундаментальних прав або інші значні ризики, не підпадатимуть під заборону використання;
- додано нові положення для врахування ситуацій, коли системи штучного інтелекту можна використовувати для багатьох різних цілей (штучний інтелект загального призначення);
- уточнено сферу дії закону про AI (наприклад, виключення з закону застосувань, що належать до задач національної безпеки, оборони та військових цілей зі сфери дії акту про AI) та положення, що стосуються правоохоронних органів;
- додано нові положення для підвищення прозорості та дозволу на скарги користувачів.
- суттєво змінено положення щодо заходів на підтримку інновацій (наприклад, введено поняття регуляторних пісочниць ШІ).

У парламенті ЄС обговорення проводилися під керівництвом Комітету з питань внутрішнього ринку та захисту прав споживачів та Комітету з питань громадянських свобод, юстиції та внутрішніх справ за процедурою спільного комітету. Проект регламенту викликав низку дискусій та заперечень як від окремих країн-членів ЄС, так і великих компаній, що працюють на ринку ЄС. Однак, після складної і довгої роботи було сформульовано переговорну позиція парламенту, прийняту в червні 2023 року, у який:

- внесено зміни до визначення систем ШІ, щоб узгодити його з визначенням, погодженим Організацією економічного співробітництва та розвитку (ОЕСР);

- суттєво змінено список заборонених в ЄС систем ШІ;
- додано вимогу про те, що системи повинні становити «значний ризик», щоб кваліфікуватись як високоризикові, і в певних випадках для проведення оцінки необхідно визначити вплив на фундаментальні права;
- у законі про штучний інтелект закріплено багаторівневий підхід до регулювання систем штучного інтелекту загального призначення, включаючи основні моделі штучного інтелекту, включаючи генеративні моделі штучного інтелекту (такі як Chat GPT), які створюють зображення, музику та інші твори мистецтва;
- створено Офіс AI, новий орган ЄС для підтримки узгодженого застосування акту AI, надання вказівок і координації спільних транскордонних розслідувань;
- погоджено, що дослідницька діяльність і розробка безкоштовних компонентів штучного інтелекту з відкритим кодом будуть значною мірою звільнені від дотримання правил закону про штучний інтелект.

Зустрічі трилогу відбувалися також в червні, липні, вересні, жовтні та грудні 2023 року та після тривалих переговорів Голова Ради та учасники переговорів Європейського парламенту досягли попередньої згоди щодо акту AI 9 грудня 2023 року. Парламент схвалив акт AI 13 березня 2024.

У фінальній версії EU AI Act:

- закріплює в законодавстві ЄС визначення систем штучного інтелекту відповідно до переглянутого визначення, а також містить визначення моделей загального призначення (GPAI, General-purpose AI - ШІ загального призначення).
- стосується насамперед постачальників і розробників, які вводять системи штучного інтелекту та моделі GPAI в експлуатацію або розміщують на ринку ЄС і які мають представництва або знаходяться в ЄС, а також розробників або постачальників систем штучного інтелекту, які створені в третій країні, коли продукція, вироблена їхніми системами, використовується в ЄС.
- підтримує підхід, що ґрунтується на оцінці ризику, запропонований Комісією, і класифікує системи штучного інтелекту за кількома категоріями ризику, із застосуванням різних ступенів регулювання.
- забороняє ширший спектр методів штучного інтелекту, ніж було запропоновано спочатку Комісією, через їх шкідливий вплив.
- визначає низку випадків використання, у яких системи штучного інтелекту слід вважати високоризикованими, оскільки вони потенційно можуть негативно впливати на здоров'я, безпеку людей або їхні основні права.
- визначає низку систем штучного інтелекту, які створюють обмежені ризики через їх недостатню прозорість (тобто глибокі фейки, синтетичний вміст), які підлягатимуть вимогам щодо інформації та прозорості.

- дозволяє використовувати системи, що представляють мінімальний ризик для людей (наприклад, спам-фільтри), які відповідають чинному законодавству (наприклад, GDPR).
- надає конкретні правила для моделей ШІ загального призначення і для моделей загального призначення з «можливостями високого впливу», які можуть становити системний ризик і мати значний вплив на внутрішній ринок. Винятки стосуються безкоштовних і відкритих моделей загального призначення.
- просить держави-члени створити регуляторні «пісочниці» та дозволити тестувати системи штучного інтелекту з високим ризиком у реальному світі, щоб полегшити розробку, навчання, тестування та перевірку інноваційних систем штучного інтелекту.

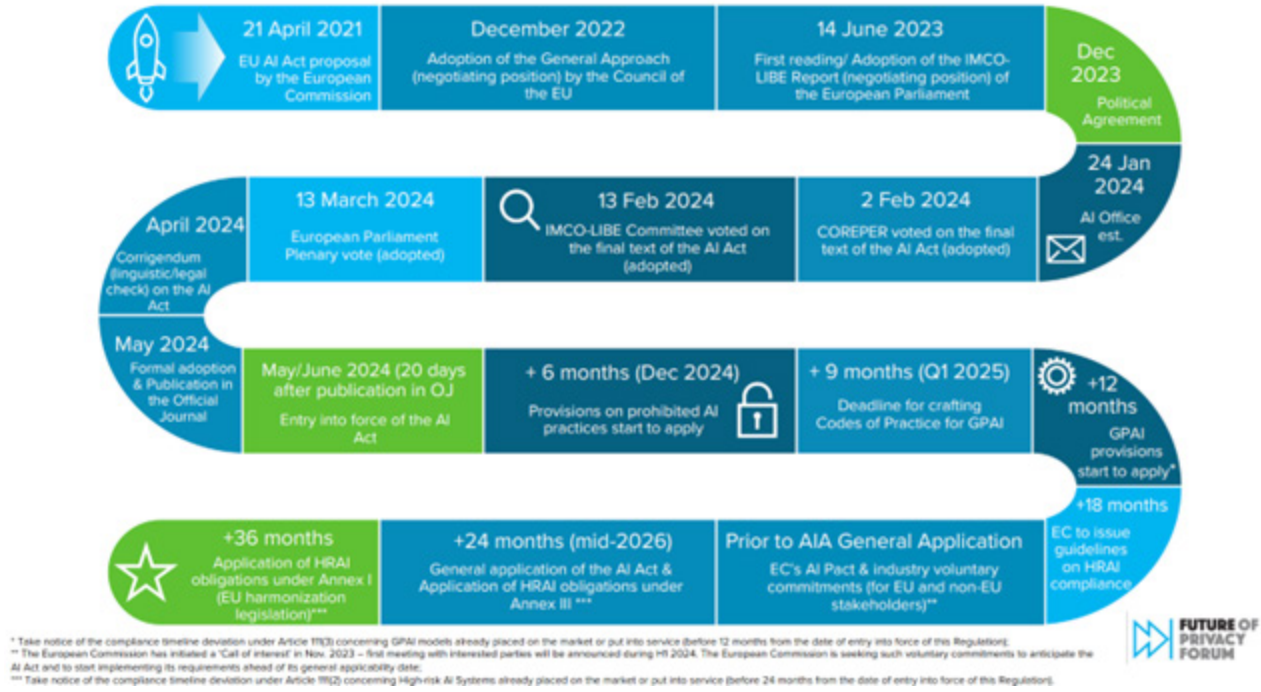


Рис. 6.1. Хронологія розробки та імплементації EU AI Act

Відповідальність за імплементацію закону про штучний інтелект буде нести низка учасників як на національному рівні, так і на рівні ЄС.

Застосування закону про штучний інтелект триватиме протягом двох років (починаючи з поступового припинення використання заборонених систем протягом шести місяців після набуття актом чинності) і вимагатиме від Європейської комісії видання різних імплементаційних, делегованих і вказівок.

Закон про ШІ (EU AI Act) був офіційно прийнятий парламентом під час його пленарної сесії 13 березнz 2024 року (було оголошено на сесії у квітні 2024 року). AI Act було опубліковано в Офіційному журналі ЄС 12 липня 2024 року. Він набув чинності в серпні 2024 року [83].

6.2.2. Імплементація Закону про ШІ: терміни та наступні кроки

На рис. 6.1 наведено основні дати, пов'язані з впровадженням Закону про AI. Відповідно розглянемо деякі **терміни** імплементації Закону про ШІ:

- **Через 6 місяців після набрання чинності:**
  - Входить в дію заборона на системи ШІ з неприйнятним ризиком AI (стаття 85).

- **Через 9 місяців після набрання чинності:**
  - Необхідно завершити роботу над правилами ШІ загального призначення (GPAI) (стаття 85).

- **Через 12 місяців після набрання чинності:**
  - Застосовуються правила GPAI (стаття 85).
  - Призначення компетентних органів держав-членів (стаття 59).
  - Щорічний огляд Комісії та можливі поправки щодо заборон (стаття 84).

- **Через 18 місяців після набрання чинності:**
  - Комісія видає імплементаційні акти, створюючи шаблон для плану постринкового моніторингу постачальників ШІ з високим рівнем ризику (стаття 6).

- **Через 24 місяці після набрання чинності:**
  - Будуть застосовуватись зобов'язання щодо систем штучного інтелекту високого ризику, конкретно перерахованих у Додатку III, який включає системи штучного інтелекту в біометрії, критичній інфраструктурі, освіті, працевлаштуванні, доступі до основних державних послуг, правоохоронних органах, імміграції та відправленні правосуддя (стаття 83).
  - Держави-члени мають запровадити правила щодо штрафів, включаючи адміністративні штрафи (стаття 53).
  - Органи влади держав-членів мають створити принаймні одну діючу регуляторну пісочницю ШІ (стаття 53).

- Перегляд Комісії та можливе внесення змін до списку систем ШІ високого ризику (стаття 84).

- **Через 36 місяців після набрання чинності:**
  - Застосовуються зобов'язання щодо систем штучного інтелекту високого ризику за Додатком II (стаття 85).
  - Зобов'язання щодо систем штучного інтелекту високого ризику, які не описані в Додатку III, але призначені для використання як компонент безпеки продукту, або штучний інтелект сам по собі є продуктом, і продукт повинен пройти оцінку відповідності третьою стороною відповідно до існуючих спеціальних законів ЄС, наприклад, про іграшки, радіообладнання, медичні пристрої для діагностики in vitro, безпеку цивільної авіації та сільськогосподарські транспортні засоби (стаття 85).

- **До кінця 2030 року:**
  - Зобов'язання набувають чинності для певних систем ШІ, які є компонентами великомасштабних ІТ-систем, встановлених законодавством ЄС у сферах свободи, безпеки та правосуддя, таких як Шенгенська інформаційна система. (стаття 83).

Вторинне право

- **Комісія може вводити делеговані акти, що визначають:**
  - Поняття та межі системи ШІ (стаття 82a).
  - Критерії, які звільняють системи ШІ від правил високого ризику (стаття 6).
  - Варіанти використання ШІ з високим ризиком (стаття 7).
  - Порогові значення, які класифікують моделі ШІ загального призначення як системні (стаття 52a).
  - Вимоги до технічної документації для систем ШІ високого ризику та GPAI (стаття 11).
  - Оцінки відповідності **(стаття 43)**
  - Декларація ЄС про відповідність **(стаття 48)**
  - Повноваження Комісії видавати делеговані акти тривають протягом початкового та продовжуваного періоду в п'ять років **(стаття 73)**

Управління **штучного інтелекту** має розробити кодекси практики, які охоплюють (але не обмежуються цим), зобов'язання постачальників моделей штучного інтелекту загального призначення. Кодекси практики повинні бути готові щонайпізніше через

дев'ять місяців після набрання чинності та повинні передбачати щонайменше тримісячний період до набуття чинності **(стаття 73)**.

- **Комісія може вносити підзаконні акти щодо:**
  - Затвердження кодексів практики для GPAI та генеративного водяного знака AI **(Стаття 52e)**
  - Створення наукової групи незалежних експертів **(Стаття 58b)**
  - Умови для оцінки AI Office на відповідність GPAI **(Стаття 68j)**
  - Правила експлуатації нормативних пісочних програм ШІ **(стаття 53)**
  - Інформація в реальних планах тестування **(Стаття 54a)**
  - Загальні специфікації (де стандарти не охоплюють правила) **(стаття 41)**

Інструкції комісії

- **Комісія може надати вказівки щодо:**
  - Через 12 місяців після набрання чинності: повідомлення про серйозні інциденти ШІ високого ризику **(стаття 62)**
  - Через 18 місяців після набрання чинності: практичне керівництво щодо визначення того, чи є система штучного інтелекту з високим ризиком, зі списком практичних прикладів випадків використання з високим і невисоким ризиком **(Стаття 6)**

- **Без конкретних часових рамок Комісія надасть вказівки щодо: (Стаття 82a)**

- Застосування визначення системи ШІ.
- Вимог до постачальників штучного інтелекту з високим ризиком.
- Заборон.
- Істотних модифікації.
- Прозорості розкриття інформації для кінцевих користувачів.
- Детальної інформація про зв'язок між Законом про AI та іншими законами ЄС.

- **Комісія повинна звітувати про свої делеговані повноваження не пізніше ніж через дев'ять місяців і раніше п'яти років після набрання чинності. (стаття 84)**

Закон про штучний інтелект [83] надає розробникам і постачальникам штучного інтелекту чіткі вимоги та зобов'язання щодо конкретних видів використання ШІ. Водночас регламент спрямований на зменшення адміністративного та фінансового тягаря для бізнесу, зокрема для малих та середніх підприємств (МСП).

Закон про штучний інтелект є частиною ширшого пакета політичних заходів для підтримки розвитку надійного штучного інтелекту, який також **включає інноваційний пакет штучного інтелекту та скоординований план щодо штучного інтелекту.** Разом ці заходи гарантують безпеку та фундаментальні права людей і компаній, коли мова йде про ШІ. Вони також сприяють сприйняттю, інвестиціям та інноваціям у ШІ в ЄС.

Закон про штучний інтелект є першою всеосяжною правовою базою щодо штучного інтелекту в усьому світі. Метою нових правил є сприяння надійному штучному інтелекту в Європі та за її межами, гарантуючи, що системи штучного інтелекту поважають основні права, безпеку та етичні принципи, а також ураховуючи ризики дуже потужних і впливових моделей штучного інтелекту.

6.2.3. Основні переваги від впровадження AI Act

Закон про ШІ гарантує, що європейські користувачі можуть довіряти тому, що може запропонувати AI. Хоча більшість систем штучного інтелекту майже не становлять ризику та можуть сприяти вирішенню багатьох суспільних проблем, певні системи штучного інтелекту створюють ризики, які ми повинні враховувати, щоб уникнути небажаних результатів.

Наприклад, часто неможливо з'ясувати, чому система штучного інтелекту прийняла рішення чи прогноз і вжила певну дію. Таким чином, може стати складно оцінити, чи була хтось несправедливо ущемлена, наприклад, у рішенні про прийом на роботу або в заявці на програму суспільного благоустрою.

Хоча чинне законодавство забезпечує певний захист, його недостатньо для вирішення конкретних проблем, які можуть створити системи ШІ.

Своєю чергою, AI Act дасть змогу:

- усунути ризики, створені спеціально програмами ШІ
- заборонити практику ШІ, яка створює неприйнятні ризики
- визначити список програм з високим ризиком
- встановити чіткі вимоги до систем штучного інтелекту для додатків із високим ризиком
- визначити конкретні зобов'язання розробників і постачальників додатків ШІ з високим рівнем ризику
- вимагати оцінки відповідності перед тим, як певну систему ШІ буде введено в експлуатацію або розміщено на ринку
- запровадити примусове виконання після того, як певну систему AI буде виведено на ринок
- створити структуру управління на європейському та національному рівнях

6.3. Основні положення EU AI Act

Розглянемо детальніше основні положення EU AI Act. Документ містить 13 основних розділів, 13 додатків (Annexes) та 180 уточнень (Recitals) [83]. Оскільки вивчення такого ґрунтовного документу є досить складним завданням, наведемо тут лише найважливіші його положення.

Перш за все варто зазначити, що AI Act базується на ризико-орієнтованому підході, відповідно до якого визначено чотири рівні ризику для систем ШІ (рис.6.2.)

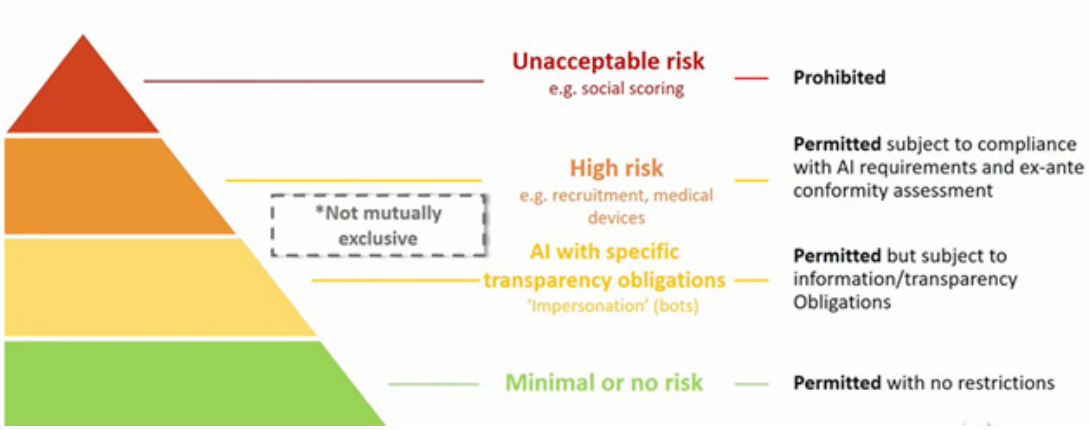


Рис. 6.2. Чотири рівні ризику для систем ШІ

6.3.1. Неприйнятний ризик (Unacceptable risk)

Усі системи штучного інтелекту, які вважаються явною загрозою безпеці, засобам існування та правам людей, заборонені, від соціальних оцінок урядів до іграшок із голосовою підтримкою, яка заохочує до небезпечної поведінки, належать, відповідно до AI Act до **заборонених системи штучного інтелекту (розділ II, ст. 5)**

До систем штучного інтелекту із неприйнятним ризиком відповідно до Закону про штучний інтелект належатимуть системи ШІ, які здійснюють:

- застосування **підсвідомих, маніпулятивних або оманливих методів** для спотворення поведінки та погіршення прийняття обґрунтованих рішень, завдаючи значної шкоди.
- **використання вразливостей,** пов'язаних з віком, інвалідністю чи соціально-економічними обставинами, для спотворення поведінки, завдаючи значної шкоди.
- **системи біометричної категоризації,** які визначають конфіденційні ознаки (раса, політичні погляди, членство в профспілці, релігійні чи філософські переконання, статеве життя чи сексуальна орієнтація), за винятком маркування

чи фільтрації законно отриманих наборів біометричних даних або коли правоохоронні органи класифікують біометричні дані.

- **соціальну оцінку,** тобто оцінку або класифікацію окремих осіб або груп на основі соціальної поведінки чи особистих рис, що спричиняє шкідливе або несприятливе ставлення до цих людей.
- **оцінку ризику вчинення особою кримінальних правопорушень** виключно на основі профілювання чи особистісних рис, за винятком випадків, коли використовується для посилення людських оцінок на основі об'єктивних фактів, які можна перевірити, безпосередньо пов'язаних із злочинною діяльністю.
- **складання баз даних розпізнавання облич** шляхом нецільового збирання зображень облич з Інтернету або записів із камер відеоспостереження.
- **передавання емоцій на робочих місцях або в навчальних закладах,** за винятком медичних причин або причин безпеки.
- **дистанційну біометричну ідентифікацію в реальному часі (RBI) у загальнодоступних місцях для правоохоронних органів,** за винятком випадків, коли:
  - розшук зниклих безвісти, жертв викрадення та людей, які постраждали від торгівлі людьми або сексуальної експлуатації;
  - запобігання суттєвій і безпосередній загрозі життю або передбачуваному теракту;
  - виявлення підозрюваних у вчиненні серйозних злочинів (наприклад, убивства, зґвалтування, збройне пограбування, незаконний обіг наркотиків і зброї, організована злочинність, екологічні злочини тощо).

Примітка щодо дистанційної біометричної ідентифікації (RBI) [83]:

Використання RBI в режимі реального часу з підтримкою штучного інтелекту дозволено лише **тоді, коли невикористання інструменту може завдати значної шкоди** та повинно враховувати права та свободи постраждалих осіб.

Перед розгортанням поліція має завершити **оцінку впливу на фундаментальні права та зареєструвати систему в базі даних ЄС,** однак у належним чином обґрунтованих випадках терміновості розгортання може розпочатися без реєстрації за умови, що вона буде зареєстрована пізніше без зайвої затримки.

Перед розгортанням вони також повинні отримати **дозвіл від судового органу або незалежного адміністративного органу,** хоча, у належним чином обґрунтованих випадках терміновості, розгортання може розпочатися без дозволу, за умови, що дозвіл буде подано протягом 24 годин. Якщо авторизацію відхилено, розгортання має бути негайно припинено, видаливши всі дані, результати та виходи.

6.3.2. Високий ризик (High risk)

Системи ШІ, визначені як високоризикові, включають технології ШІ, які використовуються в:

- критичній інфраструктури (наприклад, транспорт), які можуть поставити під загрозу життя та здоров'я громадян
- освітня або професійна підготовка, яка може визначати доступ до освіти та професійний курс чийогось життя (наприклад, підрахунок іспитів)
- компоненти безпеки продуктів (наприклад, застосування штучного інтелекту в хірургії за допомогою роботів)
- працевлаштування, управління працівниками та доступ до самозайнятості (наприклад, програмне забезпечення для сортування резюме для процедур найму)
- основні приватні та державні послуги (наприклад, кредитний рейтинг, що позбавляє громадян можливості отримати кредит)
- правоохоронні органи, які можуть втручатися в основні права людей (наприклад, оцінка достовірності доказів)
- управління міграцією, наданням притулку та прикордонним контролем (наприклад, автоматизована перевірка візових заяв)
- правосуддя та демократичні процеси (наприклад, рішення AI для пошуку судових рішень)

Системи штучного інтелекту з високим рівнем ризику підпадають під суворі зобов'язання, перш ніж їх можна буде вивести на ринок:

- адекватні системи оцінки та пом'якшення ризиків
- висока якість наборів даних, що подають систему, щоб мінімізувати ризики та дискримінаційні результати
- реєстрація діяльності для забезпечення відстеження результатів
- детальна документація, що містить всю необхідну інформацію про систему та її призначення, щоб органи влади могли оцінити її відповідність
- чітка та адекватна інформація для розробника
- належні заходи людського нагляду для мінімізації ризику
- високий рівень надійності, безпеки та точності

Усі системи дистанційної біометричної ідентифікації вважаються високоризиковими та підпадають під суворі вимоги. Використання дистанційної біометричної ідентифікації в загальнодоступних місцях для правоохоронних цілей у принципі заборонено. Окремі винятки суворо визначені та регламентовані, наприклад, коли це необхідно для пошуку зниклої дитини, для запобігання конкретній і безпосередній терористичній загрозі або для виявлення, встановлення місцезнаходження, ідентифікації чи притяг-

нення до відповідальності злочинця чи підозрюваного у вчиненні серйозного кримінального правопорушення. Таке використання підлягає дозволу судового чи іншого незалежного органу та відповідним обмеженням у часі, географічному охопленні та базах даних, у яких здійснюється пошук.

Розглянемо детальніше системи штучного інтелекту високого ризику (Розділ III) [83].

Відповідно до Закону про штучний інтелект деякі системи ШІ вважаються «високим ризиком». До постачальників цих систем висуватимуть додаткові вимоги.

Системи штучного інтелекту високого ризику відповідно до **статті 6** (2) – це системи штучного інтелекту, перелічені в будь-якій із наведених нижче сфер:

1. Біометрія, якщо її використання дозволено відповідним законодавством Союзу або національним законодавством:

(a) системи дистанційної біометричної ідентифікації. Це не стосується систем ШІ, призначених для використання для біометричної перевірки, єдиною метою якої є підтвердження того, що конкретна фізична особа є тією особою, за яку себе видає;

(aa) системи штучного інтелекту, призначені для використання для біометричної категоризації, відповідно до чутливих або захищених атрибутів або характеристик, заснованих на висновках цих атрибутів або характеристик;

(ab) Системи ШІ, призначені для використання для розпізнавання емоцій.

2. Критична інфраструктура:

(a) Системи штучного інтелекту, призначені для використання як компоненти безпеки в управлінні та експлуатації критичної цифрової інфраструктури, дорожнього руху та постачання води, газу, опалення та електроенергії.

3. Освіта та професійна підготовка:

(a) системи штучного інтелекту, призначені для визначення доступу чи вступу або для призначення фізичних осіб до навчальних закладів і закладів професійної підготовки на всіх рівнях;

(b) системи штучного інтелекту, призначені для оцінювання результатів навчання, зокрема, коли ці результати використовуються для керування процесом навчання фізичних осіб у закладах освіти та професійної підготовки на всіх рівнях;

(ba) системи штучного інтелекту, призначені для використання з метою оцінки належного рівня освіти, який особа отримає або до якої зможе отримати доступ, у контексті/в межах закладу освіти та професійної підготовки;

(bb) Системи штучного інтелекту, призначені для моніторингу та виявлення забороненої поведінки учнів під час іспитів у контексті/всередині закладів освіти та професійної підготовки.

4. Працевлаштування, управління працівниками та доступ до самозайнятості:

(a) системи штучного інтелекту, призначені для набору чи відбору фізичних осіб, зокрема для розміщення цільових оголошень про роботу, аналізу та фільтрації заявок на роботу та оцінки кандидатів;

(b) AI, призначений для прийняття рішень, що впливають на умови трудових відносин, просування по службі та припинення трудових договірних відносин, розподіл завдань на основі індивідуальної поведінки чи особистих рис або характеристик, а також для моніторингу та оцінки продуктивності та поведінки осіб у таких стосунках.

5. Доступ і використання основних приватних послуг і основних державних послуг і переваг:

(a) системи штучного інтелекту, призначені для використання державними органами або від імені державних органів для оцінки права фізичних осіб на основні пільги та послуги державної допомоги, включаючи медичні послуги, а також для надання, скорочення, скасування або повернення таких пільги та послуги;

(b) системи штучного інтелекту, призначені для оцінки кредитоспроможності фізичних осіб або встановлення їх кредитного рейтингу, за винятком систем штучного інтелекту, які використовуються з метою виявлення фінансового шахрайства;

(c) системи штучного інтелекту, призначені для оцінки та класифікації екстрених дзвінків фізичними особами або для використання для диспетчеризації чи встановлення пріоритету в диспетчеризації екстрених служб першого реагування, включаючи поліцію, пожежників і медичну допомогу, а також екстрену медичну допомогу системи сортування пацієнтів;

(ca) системи AI, призначені для використання для оцінки ризиків і ціноутворення стосовно фізичних осіб у разі страхування життя та здоров'я.

6. Правоохоронні органи, якщо їх використання дозволено відповідним законодавством Союзу або національним законодавством:

(a) системи штучного інтелекту, призначені для використання правоохоронними органами або від їх імені, або установами, агентствами, службами чи органами Союзу на підтримку правоохоронних органів або від їх імені для оцінки ризику фізичної особи стати жертвою кримінальних злочинів;

(b) системи штучного інтелекту, призначені для використання правоохоронними органами чи від їх імені, або установами, органами та агентствами Союзу на підтримку правоохоронних органів як поліграфи та подібні інструменти;

(d) системи штучного інтелекту, призначені для використання правоохоронними органами або від їхнього імені, або установами, агентствами, службами чи органами Союзу на підтримку правоохоронних органів для оцінки надійності доказів під час розслідування або судового переслідування кримінальних правопорушень ;

(e) системи штучного інтелекту, призначені для використання правоохоронними органами або від їх імені, або установами, агентствами, офісами чи органами Союзу на підтримку правоохоронних органів для оцінки ризику вчинення фізичною особою правопорушення чи повторного правопорушення не виключно на основі щодо профілювання фізичних осіб, як зазначено в **статті 3** (4) Директиви (ЄС) 2016/680, або для оцінки особистих рис і характеристик або минулої злочинної поведінки фізичних осіб або груп;

(f) системи штучного інтелекту, призначені для використання правоохоронними органами або від їх імені, установами, агенціями, службами чи органами Союзу на підтримку правоохоронних органів для профілювання фізичних осіб, як зазначено в **статті 3** (4) Директиви (ЄС) 2016/680 під час виявлення, розслідування або судового переслідування кримінальних правопорушень;

7. Управління міграцією, притулком і прикордонним контролем, якщо їх використання дозволено відповідним законодавством Союзу або національним законодавством:

(a) системи ШІ, призначені для використання компетентними державними органами як поліграфи та подібні інструменти;

(b) системи штучного інтелекту, призначені для використання компетентними державними органами або від їх імені, або агентствами, офісами чи органами Союзу для оцінки ризику, включаючи ризик безпеки, ризик нелегальної міграції або ризик для здоров'я, спричинений людині, яка має намір в'їхати або в'їхала на територію держави-члена;

(d) системи штучного інтелекту, призначені для використання компетентними державними органами або від їх імені, або агентствами, службами чи органами Союзу для надання допомоги компетентним державним органам у розгляді заяв про надання притулку, візи та дозволів на проживання та пов'язаних скарг щодо відповідності вимогам. фізичних осіб, які претендують на отримання статусу, включаючи відповідну оцінку достовірності доказів;

(da) системи штучного інтелекту, призначені для використання компетентними державними органами або від їхнього імені, включаючи агентства, офіси чи органи Союзу, у контексті міграції, притулку та управління прикордонним

контролем, з метою виявлення, розпізнавання чи ідентифікації фізичних осіб з виняток перевірки проїзних документів.

8. Відправлення правосуддя та демократичні процеси:

(a) системи штучного інтелекту, призначені для використання судовим органом або від його імені для надання допомоги судовому органу в дослідженні та тлумаченні фактів і закону, а також у застосуванні закону до конкретного набору фактів або використовуються подібним чином в альтернативних спорах резолюція;

(aa) системи штучного інтелекту, призначені для використання для впливу на результат виборів чи референдуму або поведінку фізичних осіб при голосуванні під час голосування на виборах чи референдумах. Це не включає системи штучного інтелекту, до результатів яких фізичні особи не піддаються безпосередньому впливу, наприклад інструменти, що використовуються для організації, оптимізації та структурування політичних кампаній з адміністративної та матеріально-технічної точок зору;

Правила класифікації систем штучного інтелекту високого ризику (ст. 6) [83].

Для того, щоб визначити, що система ШІ є високоризикованою потрібно переконатись чи вона:

- використовується як компонент безпеки або як продукт, на який поширюється дія законів ЄС у **Додатку II** ТА вимагає оцінювання відповідності третьою стороною відповідно до цих законів у Додатку II; **АБО**
- випадки використання **Додатку III** (додаток 1), за винятком випадків, коли:
  - система ШІ виконує вузьке процедурне завдання;
  - покращує результат раніше виконаної діяльності людини;
  - виявляє шаблони прийняття рішень або відхилення від попередніх шаблонів прийняття рішень і не призначений для заміни або впливу на раніше завершену оцінку людиною без належної перевірки людиною; або
  - виконує підготовче завдання до оцінювання, що стосується цілей випадків використання, перелічених у Додатку III.
- Системи штучного інтелекту завжди вважаються високоризиковими, якщо вони створюють профілі окремих осіб, тобто автоматизують обробку персональних даних для оцінки різних аспектів життя людини, таких як ефективність роботи, економічна ситуація, стан здоров'я, уподобання, інтереси, надійність, поведінка, місцезнаходження чи пересування.
- Постачальники, які вважають, що їхня система штучного інтелекту, яка входить в перелік з Додаткку III AI Act, не є високоризиковою, повинні задокументувати такуоцінку перед розміщенням її на ринку або введенням в експлуатацію.

Вимоги до постачальників систем штучного інтелекту високого ризику (ст. 8-25) [83].

Постачальники штучного інтелекту з високим рівнем ризику повинні:

- Створити **систему управління ризиками** протягом усього життєвого циклу системи ШІ з високим рівнем ризику;
- Здійснити **управління даними**, забезпечуючи, щоб набори даних навчання, перевірки та тестування були релевантними, достатньо репрезентативними та, наскільки це можливо, вільними від помилок і повними відповідно до запланованої мети.
- Скласти **технічну документацію** для демонстрації відповідності та надати органам влади інформацію для оцінки такої відповідності.
- Розробити їхню систему штучного інтелекту з високим рівнем ризику для **ведення записів**, щоб вона могла автоматично записувати події, важливі для визначення ризиків на національному рівні та істотних модифікацій протягом життєвого циклу системи.
- Надати **інструкції щодо використання** для наступних розгортачів, щоб забезпечити відповідність останніх.
- Розробити свою систему штучного інтелекту з високим рівнем ризику, щоб дозволити розгортачам здійснювати **нагляд з боку людини.**
- Спроектувати їхню систему штучного інтелекту з високим рівнем ризику, щоб досягти належного рівня **точності, надійності та кібербезпеки.**
- Встановити систему управління якістю для забезпечення відповідності.

Навчальні дані у системах ШІ високого ризику

Системи ШІ високого ризику під час навчання моделей ШІ мають дотримуватись критеріїв якості наборів даних для навчання, здійснювати їх валідацію та тестування. Такі набори даних мають забезпечувати їх репрезентативність для навчання системи ШІ та відсутність у них помилок (Стаття 10 частина 3 AI Act) [83].

Набори даних мають формуватися, враховуючи особливості контексту та середовища, в межах якого така система ШІ буде застосовуватись, а також поведінкових чи функціональних умов, де така система ШІ буде використовуватись (стаття 10 частина 4 AI Act). Також щодо наборів даних таких систем має проводитись перевірка для виявлення та усунення упереджень.

Увага до якості та репрезентативності даних у AI Act є критично важливою, оскільки вона може допомогти системам ШІ функціонувати більш справедливо та ефективно. Випадок з Amazon, який ми згадували вище, говорячи про упередженість системи ШІ, і тут буде показовим. Так, система була навчена на резюме, які компанія отримала протягом десяти років. Більшість цих резюме була від чоловіків, що демонструвало гендерний дисбаланс у технологічній індустрії. В результаті алгоритм навчився вважати

резюме чоловіків більш привабливими для технічних ролей. Цей випадок доводить, як важливо уникати упереджень у наборах даних, що використовуються для навчання систем ШІ. А також демонструє, чому подібні вимоги в AI Act є важливими для систем ШІ високого ризику.

Технічна документація у системах ШІ високого ризику

Розробники систем ШІ високого ризику зобов'язані регулярно оновлювати свою технічну документацію. Вона має бути розроблена до того, як система ШІ високого ризику буде розміщена на ринку або введена в експлуатацію. (стаття 11 частина 1).

Надання технічної документації відіграє важливу роль у тому, щоб компетентні органи могли мати всю необхідну інформацію про систему ШІ та здійснювати оцінку її відповідності вимогам. А також потенційно може зменшити ризик негативних наслідків використання систем ШІ, щодо яких не подали технічну документацію до введення в експлуатацію, якщо така система порушує вимоги Конвенції.

RECORD-KEEPING

Розробники систем ШІ високого ризику зобов'язані впроваджувати технічні засоби для автоматичного реєстрування подій (логів) протягом усього життєвого циклу системи. Це необхідно для забезпечення належного рівня відстежуваності її функціонування (стаття 12). Крім того, в певних системах ШІ високого ризику слід здійснювати запис періоду кожного використання системи, включно з датою і часом початку та закінчення кожного сеансу та реєстрацією вхідних даних.

Ці вимоги забезпечують прозорість та можливість аудиту систем ШІ високого ризику, що є критично важливим для гарантування їх безпечного та ефективного використання.

Прозорість та надання інформації щодо системи ШІ

Також AI Act визначає важливість прозорості та інформування користувачів систем ШІ високого ризику. Ці системи мають бути розроблені таким чином, щоб забезпечувати достатню прозорість у своїй роботі, дозволяючи користувачам правильно інтерпретувати результати їхньої діяльності. Системи мають супроводжуватися інструкціями із зрозумілою, стислою та доступною інформацією у відповідному цифровому форматі або в іншій формі (стаття 13 AI Act).

Інструкція повинна містити таку інформацію:

1. Особу та контактні дані постачальника та, за необхідності, його уповноваженого представника;
2. Характеристики, можливості та обмеження системи ШІ високого ризику, у тому числі її цільове призначення та рівень точності.

Також має надаватись інформація про людський контроль, необхідні ресурси для функціонування системи, її очікуваний термін функціонування, технічне обслуговування та оновлення.

Повна інформація про систему ШІ дозволяє користувачам і регуляторам забезпечувати її відповідальне використання. Це критично важливо у сферах, де рішення ШІ можуть мати значний вплив на життя та добробут людей: наприклад, у медичній діагностиці або кредитуванні. У системах ШІ, що використовуються, зокрема, для діагностики захворювань, критично важливо мати чіткі вказівки щодо точності, обмежень та умов, за яких результати системи ШІ можуть бути неточними. Це необхідно, аби зменшити ризики та уникнути діагностичних помилок, які можуть вплинути на лікування пацієнтів.

Людський нагляд

У системах ШІ високого ризику має забезпечуватись ефективний людський нагляд протягом періоду їх проєктування, розробки та використання (стаття 14 AI Act).

Також системи ШІ високого ризику мають відповідати вимогам точності, надійності та забезпечувати належний рівень кібербезпеки (стаття 15 AI Act).

6.3.3. Обмежений ризик

Обмежений ризик стосується ризиків, пов'язаних із недостатньою прозорістю використання ШІ. Закон про штучний інтелект вводить конкретні зобов'язання щодо прозорості, щоб забезпечити інформування людей, коли це необхідно, зміцнюючи довіру. Наприклад, під час використання систем штучного інтелекту, таких як чат-боти, люди повинні знати, що вони взаємодіють з машиною, щоб вони могли прийняти обґрунтоване рішення продовжити або відступити. Постачальники також мають забезпечити ідентифікацію створеного ШІ контенту. Крім того, текст, згенерований штучним інтелектом, опублікований з метою інформування громадськості про питання, що становлять суспільний інтерес, повинен бути позначений як штучно створений. Це також стосується аудіо- та відеоконтенту, що є глибокими фейками (deepfakes).

Також до цього рівня ризику належить ШІ загального призначення, якому в AI Act присвячено цілий розділ 5.

ШІ загального призначення (General-Purpose Artificial Intelligence models, GPAI)

Модель GPAI означає модель штучного інтелекту, у тому числі під час навчання з великою кількістю даних із використанням самонагляду в масштабі, яка демонструє значну загальність і здатна компетентно виконувати широкий спектр окремих завдань незалежно від того, як модель розміщена на ринку. і які можна інтегрувати в різноманітні подальші системи або програми. Це не поширюється на моделі штучного інтелекту, які використовуються перед випуском на ринок для досліджень, розробки та створення прототипів.

Система GPAI означає систему штучного інтелекту, яка базується на моделі штучного інтелекту загального призначення, яка може служити різноманітним цілям як для прямого використання, так і для інтеграції в інші системи штучного інтелекту.

Системи GPAI можна використовувати як системи штучного інтелекту високого ризику або інтегрувати в них. Постачальники систем GPAI повинні співпрацювати з такими постачальниками систем штучного інтелекту з високим ризиком, щоб забезпечити відповідність останніх.

Відповідно до AI Act (Розділ 5) розробники GPAI зобов'язані:

- Скласти **технічну документацію**, включаючи процес навчання та тестування та результати оцінювання.
- Підготувати **інформацію та документацію для надання подальшим постачальникам,** які мають намір інтегрувати модель GPAI у свою власну систему штучного інтелекту, щоб останні розуміли можливості та обмеження та мали змогу відповідати вимогам.
- Встановити політику **поваги до Директиви про авторське право.**
- Опублікувати **достатньо детальний опис вмісту, який використовується для навчання** моделі GPAI.

Моделі GPAI із вільною та відкритою ліцензією, чиї параметри, включаючи ваги, архітектуру моделі та використання моделі, є загальнодоступними, що дозволяє отримати доступ, використання, модифікацію та розповсюдження моделі, мають відповідати лише двом останнім зобов'язанням, зазначеним вище.

Моделі GPAI представляють системні ризики, коли сукупний обсяг обчислень, що використовуються для їх навчання, перевищує 10 25 операцій з плаваючою комою (FLOP). Провайдери повинні повідомити Комісію, якщо їх модель відповідає цьому критерію протягом 2 тижнів. Постачальник може надати аргументи, що, незважаючи на відповідність критеріям, його модель не створює системних ризиків. Комісія може вирішити самостійно або через кваліфіковане попередження від наукової групи незалежних експертів, що модель має високі можливості впливу, що робить її системною.

Окрім чотирьох вищезазначених зобов'язань, постачальники моделей GPAI **із системним ризиком** також повинні:

- Виконати **оцінювання моделі**, включаючи проведення та документування змагального тестування для виявлення та пом'якшення системного ризику.
- **Оцінити та пом'якшити можливі системні ризики,** включаючи їх джерела.
- Розробити політику відповідності законодавству ЄС щодо авторських прав та суміжних прав.
- **Відстежувати, документувати та повідомляти про серйозні інциденти** та можливі коригувальні заходи до Офісу AI та відповідних національних компетентних органів без зайвої затримки.
- Забезпечити належний рівень **захисту кібербезпеки.**

Усі постачальники моделі GPAI можуть продемонструвати дотримання своїх зобов'язань, якщо вони добровільно дотримуються кодексу практики до публікації єв-

ропейських узгоджених стандартів, дотримання яких призведе до презумпції відповідності. Тобто для того, щоб продемонструвати відповідність своїм зобов'язанням, розробникам систем з системними ризиками в AI Act рекомендується підписати кодекси поведінки. Ті, хто подібні кодекси підписувати не захоче, має продемонструвати належний **альтернативні адекватні засоби відповідності** для схвалення Комісії.

Кодекси практики

- Будуть враховувати міжнародні підходи.
- Охоплюватимуть, але не обов'язково обмежуватимуться вищевказаними зобов'язаннями, зокрема, відповідну інформацію, яку слід включити до технічної документації для органів влади та подальших постачальників, визначення типу та характеру системних ризиків та їх джерел, а також способи обліку управління ризиками для конкретних проблем у урахування ризиків через те, як вони можуть виникати та матеріалізуватись у всьому ланцюжку створення вартості.
- AI Office може запросити постачальників моделі GPAI, відповідні національні компетентні органи взяти участь у складанні кодексів, тоді як громадянське суспільство, промисловість, наукові кола, наступні постачальники та незалежні експерти можуть підтримати процес.

Відповідно до AI Act громадяни матимуть право подавати скарги на системи ШІ та отримувати пояснення щодо рішень, заснованих на системах високого ризику штучного інтелекту, які впливають на їхні права.

Перед розміщенням моделі ШІ загального призначення на ринку Європейського Союзу, постачальники, засновані в третіх країнах, повинні за письмовим дорученням призначити уповноваженого представника, який діятиме в ЄС.

6.3.4. Мінімальний ризик

Закон про штучний інтелект дозволяє використовувати штучний інтелект з мінімальним ризиком без жодних зобов'язань. Це включає такі програми, як відеоігри з підтримкою штучного інтелекту або фільтри спаму. Переважна більшість систем штучного інтелекту, які зараз використовуються в ЄС, підпадають під цю категорію.

6.3.5. Визначення ролей акторів відповідно до AI Act

Відповідно до Закону про штучний інтелект різні організації відіграють ключову роль, кожна з яких несе окрему відповідальність. Розуміння цих ролей є основоположним для навігації в нормативному ландшафті, визначеному цим законодавством, особливо тому, що вони несуть різний регуляторний тягар, враховуючи їхнє положення в ланцюжку створення вартості ШІ.

- **Постачальник (Provider)**: фізична або юридична особа, державний орган, агентство чи інший орган, який розробив систему штучного інтелекту для розміщення на ринку або ввів в експлуатацію під власним ім'ям чи торговою маркою.

У центрі ланцюжка створення вартості AI знаходиться особа, яка розробляє системи AI або моделі GPAI під власним ім'ям або торговою маркою, відома як « **Постачальник** ». Закон ЄС про штучний інтелект поширюється на розробників штучного інтелекту, коли їхні послуги, продукти чи результати потрапляють на ринок ЄС за таких сценаріїв:

- **розміщення штучного інтелекту на ринку,** що означає перше надання системи штучного інтелекту або моделі GPAI на ринку ЄС;
- **Введення в експлуатацію ШІ,** що означає постачання Системи ШІ для першого використання безпосередньо Розробнику або для власного використання на ринку ЄС;
- Виробництво **результатів штучного інтелекту,** що означає розробку системи штучного інтелекту, яка створює результати, що використовуються в ЄС, наприклад ті, які впливають на освіту, працевлаштування, безпеку продукції тощо жителів ЄС.

Постачальники, які відповідають одному з трьох критеріїв, повинні відповідати новим правилам незалежно від місця їх заснування та місцезнаходження, а також незалежно від того, платне чи безоплатне таке розповсюдження. Для постачальників будь-якої системи штучного інтелекту з високим ризиком Закон ЄС про штучний інтелект встановлює суворі вимоги щодо відповідності протягом усього життєвого циклу розробки та впровадження. Ці заходи включають комплексну документацію, оцінку відповідності та управління ризиками. Для певних систем штучного інтелекту з високим рівнем ризику Постачальник також повинен здійснити реєстрацію в базі даних ЄС перед їх використанням і розповсюдженням на ринку ЄС.

- **Розробник (Deployer):** фізична або юридична особа, державний орган, агентство чи інший орган, який використовує систему ШІ під своїм керівництвом.

«**Розробник** » відіграє вирішальну роль у забезпеченні систем штучного інтелекту для реальних додатків. Відповідно до Закону ЄС про штучний інтелект, «Розробник» означає « будь-яку фізичну або юридичну особу, державний орган, установу чи інший орган, який використовує систему штучного інтелекту під своїм керівництвом, за винятком випадків, коли система штучного інтелекту використовується під час особистої непрофесійної діяльності. «Нові правила застосовуються, якщо Розробник заснований або розташований у ЄС або якщо він керує системами штучного інтелекту, які виробляють результати, що використовуються в ЄС.

Остаточний текст нових правил вводить наступні вимоги до розробників, які працюють із системами ШІ високого ризику:

- **Використання та управління даними,** що вимагає моніторингу операцій AI Systems відповідно до їхніх інструкцій щодо використання та забезпечення відповідності введених даних запланованим цілям.

- **Навчання персоналу,** яке гарантує, що весь персонал, який працює з системами штучного інтелекту високого ризику, має достатню підготовку та кваліфікацію.
- **Відповідність нормативним вимогам**, що стосується вимог Закону ЄС про штучний інтелект щодо дотримання галузевого законодавства ЄС та проведення оцінки впливу на захист даних, серед іншого.
- **Повідомлення про інциденти,** що вимагає письмового повідомлення Постачальникам, Дистриб'юторам і відповідним органам у разі виявлення суттєвих збоїв у роботі систем штучного інтелекту високого ризику.

Нарешті, Розробник може бути повторно класифікований як «Постачальник» систем ШІ високого ризику, якщо він використовує Системи ШІ під власним ім'ям і брендом Розробника або іншим чином модифікує Системи ШІ в порушення інструкцій з використання або для непередбачених цілей. Подібним чином **імпортер** або **дистриб'ютор** (як визначено нижче) також підпадає під підвищені вимоги відповідності внаслідок неналежного брендування або несанкціонованих модифікацій.

- **Уповноважений представник (Authorized representative):** будь-яка фізична або юридична особа, розташована або зареєстрована в ЄС, яка отримала та прийняла доручення від постачальника виконувати свої зобов'язання від його імені.

« **Уповноважений представник** » відповідно до Закону ЄС про штучний інтелект розташований або заснований у ЄС і функціонує як посередник між постачальниками штучного інтелекту за межами ЄС, з одного боку, та європейськими органами влади та споживачами, з іншого боку. Постачальник за межами ЄС повинен призначити свого Уповноваженого представника в письмовій угоді (відомій як мандат) для виконання певних зобов'язань і процедур від його імені. Ці зобов'язання можуть включати перевірку відповідності Постачальника вимогам, подання документації до національного компетентного органу та збереження записів протягом десяти років.

- **Імпортер (Importer):** будь-яка фізична або юридична особа в ЄС, яка розміщує на ринку або вводить в експлуатацію систему штучного інтелекту, яка носить ім'я або товарний знак фізичної або юридичної особи, заснованої за межами ЄС.

У ланцюжку створення вартості ШІ імпортер є захисником перед тим, як певні системи ШІ з країн, що не входять до ЄС, вийдуть на ринок ЄС. Згідно з новими правилами, « **Імпортер** » — це юридична або фізична особа, розташована в ЄС, яка розміщує на ринку ЄС систему AI під назвою або торговою маркою Постачальника за межами ЄС. Імпортер також повинен виконати сувору належну перевірку та зобов'язання щодо ведення записів, перш ніж розмістити систему штучного інтелекту високого ризику на ринку ЄС.

Ці зусилля з перевірки можуть включати перевірку того, що Постачальник виконав оцінку відповідності, технічну документацію, призначення Уповноваженого представника та інші вимоги. Імпортер також повинен маркувати системи штучного інтелекту високого ризику своєю контактною інформацією та зареєстрованою торговою

маркою, подібно до процесу митного оформлення, який вимагає інформації про країну походження продукту, вміст і відповідні застереження.

- **Дистриб'ютор (Distributor):** будь-яка фізична або юридична особа в ланцюжку постачання, яка не є постачальником або імпортером, яка робить систему ШІ доступною на ринку ЄС.

« **Дистриб'ютор** » відповідно до Закону ЄС про штучний інтелект означає будь-яку фізичну або юридичну особу (окрім постачальника чи імпортера), яка *надає системи штучного інтелекту або моделі GPAI для розповсюдження чи використання на ринку ЄС* за оплату чи безкоштовно. Примітно, що дистриб'ютор не зобов'язаний бути заснованим або розташованим у ЄС, і не обов'язково бути першою стороною, яка випускає системи AI або моделі GPAI в ЄС. Будучи критично важливою ланкою в ланцюжку створення вартості штучного інтелекту, Дистриб'ютор повинен переконатися, що пов'язані постачальники та імпортери відповідають певним вимогам, і зобов'язаний співпрацювати з національними органами влади для перевірки відповідності. Якщо Дистриб'ютор підозрює невідповідність, він зобов'язаний вилучити застосовну Систему штучного інтелекту або модель GPAI з ринку ЄС, доки Постачальник або Імпортер не виконає необхідні коригувальні дії.

- **Виробник продукту (Product manufacturer):** виробник системи штучного інтелекту, який випускається на ринок, або виробник, який вводить в експлуатацію систему штучного інтелекту разом зі своїм продуктом і під власним ім'ям або торговою маркою.

Інновації штучного інтелекту перестали бути футуристичною концепцією, а швидко інтегрували системи штучного інтелекту в дизайн продуктів і розробку контенту. Доопрацьований Закон ЄС про штучний інтелект включив до сфери застосування **« виробників продукції », якщо вони надають, розповсюджують або використовують системи штучного інтелекту на ринку ЄС** разом зі своїми продуктами та під своїм власним ім'ям або торговою маркою. Інакше кажучи, включення систем штучного інтелекту в розробку продукту може також піддати виробника продукту чинності Закону ЄС про штучний інтелект, незалежно від його установи чи місця розташування.

Наприклад, якщо виробник автомобілів у США включає систему штучного інтелекту для моніторингу використання батареї, навігації чи підтримки функцій автономного керування та розповсюджує транспортний засіб під власним ім'ям або торговельною маркою в ЄС, такий виробник автомобілів є «Виробником продукції», що підпадає під дію Закон ЄС про штучний інтелект. Крім того, якщо така система штучного інтелекту класифікується як система штучного інтелекту високого ризику як компонент безпеки транспортного засобу, OEM бере на себе роль постачальника штучного інтелекту та відповідні зобов'язання щодо відповідності.

- **Оператор (Operator):** загальний термін, що стосується всіх наведених вище термінів (постачальник, розгортач, уповноважений представник, імпортер, дистриб'ютор або виробник продукту).

Хто є ключовими учасниками ланцюжка створення вартості ШІ?

Рис.6.3. Ключові учасники ланцюжка створення вартості ШІ відповідно до AI Act

6.3.6. Зобов'язання суб'єктів ШІ

1. Виробники продукції

Виробники продуктів відіграють ключову роль у забезпеченні відповідальної розробки та розгортання систем ШІ.

Відповідність технічним вимогам

Виробники продукції зобов'язані дотримуватися технічних вимог, викладених у нормативній базі. Це передбачає суворі процеси тестування та перевірки, щоб переконатися, що системи штучного інтелекту відповідають визначеним стандартам щодо безпеки, надійності та етичних міркувань.

Зобов'язання щодо документації та прозорості

Прозорість є ключовою в ландшафті ШІ. Виробники продуктів повинні ретельно

документувати дизайн, функції та потенційні ризики, пов'язані з їх системами ШІ. Ця документація не тільки сприяє дотриманню нормативних вимог, але й зміцнює довіру, надаючи зацікавленим сторонам чітке розуміння можливостей і обмежень системи ШІ.

2.Імпортери та дистриб'ютори

Імпортери та дистриб'ютори є ключовими посередниками в ланцюжку постачання штучного інтелекту, несучи чітку відповідальність.

Обов'язки щодо забезпечення відповідності

Імпортерам і дистриб'юторам доручено забезпечити відповідність систем ШІ, якими вони керують, нормативним стандартам, встановленим Законом про ШІ. Це передбачає ретельні оцінки та співпрацю з виробниками продуктів для вирішення будь-яких виявлених проблем, посилюючи зобов'язання постачати продукти штучного інтелекту, які відповідають встановленим критеріям, відповідно до статті 26 (Зобов'язання імпортерів).

Обов'язки щодо ведення документації

Стаття 12 (Облік) Закону про AI передбачає, що облік є важливим для підзвітності. Імпортери та дистриб'ютори зобов'язані вести ретельний облік систем ШІ, з якими вони мають справу. Це включає інформацію про походження, оцінки відповідності та будь-які вжиті коригувальні дії. Ці записи сприяють прозорій і підзвітній екосистемі ШІ.

3. Користувачі

Користувачі, будь то організації чи окремі особи, також несуть відповідальність за етичне використання технологій ШІ.

Дотримання вказівок і обмежень

Користувачі зобов'язані дотримуватися вказівок і обмежень, викладених у Законі про штучний інтелект. Це передбачає використання систем штучного інтелекту відповідно до етичних міркувань, суспільних цінностей і правових вимог. Користувачі відіграють вирішальну роль у забезпеченні того, щоб штучний інтелект розгортався відповідально та на благо суспільства. Наприклад, стаття 13 (Прозорість та надання інформації) у її нинішньому вигляді вимагає від користувачів:

Розуміння користувача :

- Користувач повинен мати змогу розуміти та належним чином використовувати систему ШІ.
- Це включає знання того, як працює система штучного інтелекту, і розуміння даних, які вона обробляє.

Пояснення постраждалим особам :

- Користувачі повинні мати можливість пояснити постраждалій особі рішення, прийняті системою штучного інтелекту, як зазначено в статті 68(c) Регламенту.

Більшість зобов'язань лягає на постачальників (розробників) систем ШІ високого ризику.

- Ті, які мають намір вивести на ринок або ввести в експлуатацію високоризикові системи ШІ в ЄС, незалежно від того, чи знаходяться вони в ЄС чи третій країні.
- А також постачальники з третіх країн, де в ЄС використовується вихід системи ШІ з високим ризиком.

Користувачі — це фізичні або юридичні особи, які розгортають систему штучного інтелекту на професійній основі, кінцеві користувачі не зачіпаються.

- Користувачі (розробники) систем штучного інтелекту з високим ризиком мають певні зобов'язання, хоча й менші, ніж постачальники (розробники).
- Це стосується користувачів, які перебувають у ЄС, і користувачів із третіх країн, де вихідні дані системи ШІ використовуються в ЄС.

Інструкції щодо оцінювання суб'єктів ШІ

Закон про штучний інтелект надає чіткі визначення та критерії для оцінки учасників ШІ. Виробники, імпортери, дистриб'ютори та користувачі керуються конкретними параметрами, що сприяє стандартизованому підходу. Ця чіткість оптимізує зусилля з дотримання вимог і сприяє спільному розумінню очікувань. Підприємства виграють від нормативної чіткості щодо ланцюга створення вартості ШІ.

Як це все працює на практиці для постачальників систем штучного інтелекту з високим ризиком?

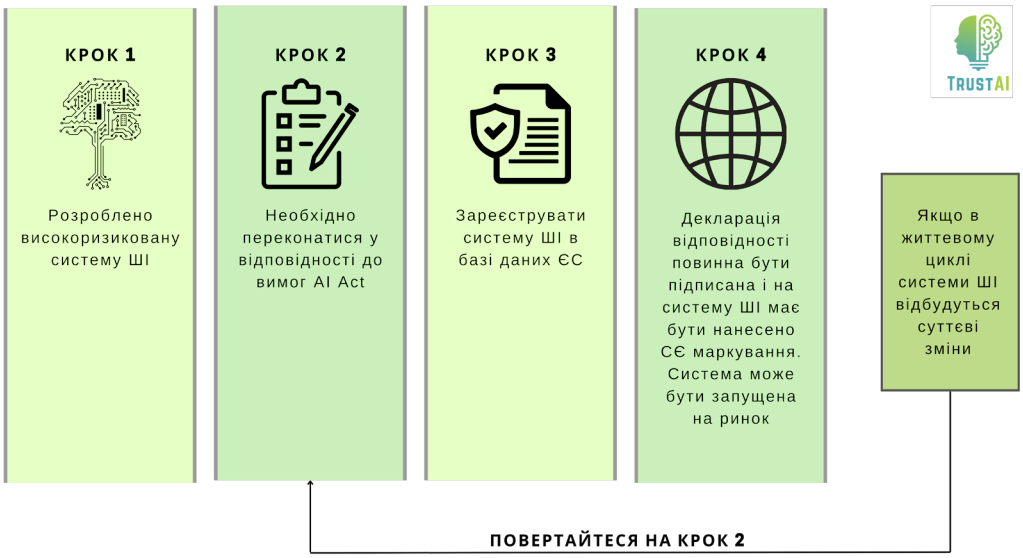


Рис. 6.4. Життєвий цикл систем ШІ з високим ризиком

Під час оцінювання базових моделей часто виникають сірі зони, які потребують детального розгляду. Законодавство визнає ці складності та заохочує динамічний підхід до вирішення нових проблем у технології ШІ. Наприклад, наразі ЄС погодився, що моделі штучного інтелекту загального призначення, які були навчені з використанням загальної обчислювальної потужності понад 10^25 FLOPs, вважаються такими, що несуть системні ризики, враховуючи, що моделі, навчені за допомогою більших обчислень, як правило, є потужнішими.. Ця адаптивність гарантує, що нормативна база залишається ефективною під час швидкого технічного прогресу.

Після того, як система штучного інтелекту виходить на ринок, органи влади відповідають за нагляд за ринком, розробники забезпечують людський нагляд і моніторинг, а постачальники мають систему постмаркетингового моніторингу. Постачальники та розробники також повідомлятимуть про серйозні інциденти та несправності.

Рішення для надійного використання великих моделей ШІ

Все більше і більше універсальні моделі ШІ стають компонентами систем ШІ. Ці моделі можуть виконувати та адаптувати незліченну кількість різних завдань.

Хоча моделі штучного інтелекту загального призначення можуть створювати кращі та потужніші рішення ШІ, важко контролювати всі можливості.

Закон про штучний інтелект вводить зобов'язання щодо прозорості для всіх моделей штучного інтелекту загального призначення, щоб краще зрозуміти ці моделі, а також додаткові зобов'язання щодо управління ризиками для дуже потужних і ефективних моделей. Ці додаткові зобов'язання включають самооцінку та пом'якшення системних ризиків, звітування про серйозні інциденти, проведення тестів і оцінок моделей, а також вимоги до кібербезпеки.

Законодавство, орієнтоване на майбутнє

Оскільки штучний інтелект є технологією, що швидко розвивається, регулювання має підхід, орієнтований на майбутнє, що дозволяє правилам адаптуватися до технологічних змін. Програми штучного інтелекту повинні залишатися надійними навіть після того, як вони були розміщені на ринку. Це вимагає від постачальників постійного управління якістю та ризиками.

Забезпечення та реалізація

Європейський офіс зі штучного інтелекту, створений у лютому 2024 року в рамках Комісії, наглядає за дотриманням та впровадженням Закону про штучний інтелект разом із державами-членами. Він спрямований на створення середовища, де технології ШІ поважають людську гідність, права та довіру. Він також сприяє співпраці, інноваціям і дослідженням у сфері ШІ серед різних зацікавлених сторін. Крім того, вона бере участь у міжнародному діалозі та співпраці з питань ШІ, визнаючи необхідність глобального узгодження управління ШІ. Завдяки цим зусиллям Європейський офіс штучного інтелекту прагне позиціонувати Європу як лідера в етичному та сталому розвитку технологій ШІ.

6.3.7. Покарання за Законом ЄС про штучний інтелект: висока ціна невиконання

Закон ЄС про штучний інтелект передбачає децентралізоване виконання положень шляхом надання державам-членам права встановлювати власні правила щодо покарань, включаючи адміністративні штрафи за порушення Закону. Таким чином, кожна держава-член повинна призначити принаймні один національний орган для нагляду за виконанням правил і ринковим наглядом.

Через структуру штрафів і покарань, розділених на рівні, що залежать від тяжкості порушення (Табл. 6), європейські інституції чітко пояснюють, як вони оцінюють системи ШІ та відповідні вимоги та зобов'язання. Рамки штрафів перевищують навіть штрафи, передбачені Загальним законом про захист даних (GDPR), які складають до 20 мільйонів євро.

Таблиця 6

Структура штрафів за останньою версією Європейського парламенту

Підхід	Максимальний штраф	Компанії (% від загального світового річного обороту за попередній фінансовий рік)
Стаття 71(3) недотримання заборони штучного інтелекту	40 мільйонів євро	7%
Стаття 71(3a) недотримання вимог щодо управління даними та прозорості, що стосуються систем ШІ з високим рівнем ризику	20 мільйонів євро	4%
Невідповідність статті 71(4) іншим вимогам, що стосуються систем штучного інтелекту високого ризику	10 мільйонів євро	2%
Стаття 71(5) надання неправдивої або оманливої інформації національним компетентним органам	5 мільйонів євро	1%

У принципі, будь-яка організація, яка зобов'язана виконувати вимоги та зобов'язання Закону ЄС про штучний інтелект, може стати предметом порушення, якщо такі вимоги та зобов'язання порушуються. Сюди входять постачальники, які як фізичні чи

юридичні особи, органи влади, установи чи інші органи розробляють системи штучного інтелекту або розробляють їх, виводять на ринок або вводять в експлуатацію. Крім того, одержувачами штрафів можуть бути виробники продуктів, імпортери, постачальники або розробники систем ШІ. Треті сторони також можуть бути покарані, хоча Закон ЄС про штучний інтелект не визначає, хто має бути включений. Згідно зі статтею 28 Закону ЄС про штучний інтелект треті сторони можуть вважатися постачальниками за певних умов і, таким чином, також взяти на себе зобов'язання постачальника.

Закон ЄС про штучний інтелект вимагає максимальної старанності від усіх сторін, які беруть участь у життєвому циклі системи високого ризику ШІ. Вимоги до цих систем тривалі, а ставки високі, що підкреслює, що високий ризик не означає високу винагороду. Акт ЄС про штучний інтелект наразі проходить Триалог між трьома основними європейськими інституціями, що є неофіційним етапом звичайної європейської законодавчої процедури, і це, ймовірно, призведе до значних змін у остаточному спільному тексті. Чи вплинуть ці зміни також на розмір штрафів – невідомо. Однак слід зазначити, що вже існують досить серйозні загрози штрафів, які були ще більше збільшені за деякі дії в останній позиції Європейського парламенту, а у випадку статті 71 (3) Закону про AI виходять далеко за рамки GDPR.

Регламент встановлює чотирирівневу структуру штрафів за порушення:

- Найбільші штрафи накладаються за порушення заборони окремих систем штучного інтелекту, до 40 000,00 євро або 7% обороту.
- Найменші штрафи передбачені за надання невірної, неповної або оманливої інформації, до 5 000 000 євро або 1% річного світового обороту.
- Штрафи за невідповідність можуть бути накладені на постачальників, розгортачів, імпортерів, дистриб'юторів і уповноважені органи

6.3.8. Наступні кроки імплементації AI Act

Закон про штучний інтелект набув чинності 1 серпня та буде повністю застосований через 2 роки, за деякими винятками: заборони набудуть чинності через шість місяців, правила управління та зобов'язання для моделей загального призначення AI стануть застосовними через 12 місяців, а правила для систем штучного інтелекту, вбудованих у регульовані продукти, будуть застосовуватися через 36 місяців. Щоб полегшити перехід до нової нормативної бази, Комісія запустила Пакт про штучний інтелект, добровільну ініціативу, яка спрямована на підтримку майбутнього впровадження та запрошує розробників штучного інтелекту з Європи та за її межами завчасно виконати основні зобов'язання Закону про штучний інтелект.

Як вплине прийняття AI Act на українські компанії?

У все більш взаємопов'язаному світі, де все більша кількість продуктів включатиме одну або декілька систем штучного інтелекту, українські компанії теж можуть потрапи-

ти під дію AI Act та, у випадку невідповідності вимогам, отримати штраф. Тож як визначити, чи повинна компанія відповідати вимогам AI Act? Для цього їм потрібно поставити собі такі запитання:

(i) **Ролі.** Які наші ролі в ланцюжку створення ШІ? Чи є наша компанія постачальником, розробником, імпортером чи дистриб'ютором?

(ii) **Продукти.**

- Чи є ми виробником продукції, яка використовує системи штучного інтелекту в своїх продуктах?
- Чи буде наша продукція розповсюджуватися, використовуватися чи іншим чином забезпечувати використання в ЄС?

(iii) **Ризики.** Чи надаємо ми системи ШІ, які класифікуються як «Заборонені системи ШІ», «Системи ШІ високого ризику» або «ШІ загального призначення, побудовані на моделях GPAI»?

Відповідно, в нових реаліях, визначених Законом ЄС про штучний інтелект, кожен учасник відіграє важливу роль у проактивному управлінні ризиками протягом життєвого циклу системи ШІ. Відповідність суворим вимогам нових правил стане спільною відповідальністю в ланцюжку створення вартості ШІ. Враховуючи розмір і стратегічні цінності ринку ЄС, американські компанії не можуть дозволити собі зволікати з дотриманням Закону ЄС про штучний інтелект або ризикувати втратою своєї частки ринку.

6.4. Регулювання використання систем ШІ в Україні

Сьогодні в Україні IT-сфера є одним з ключових гравців на ринку -станом на середину 2024 року IT залишається другим за значущістю експортно орієнтованим сектором економіки. За результатами першого півріччя IT-експорт становив 11,5% від усього експорту і майже 38% від експорту послуг. Серед ключових країн-партнерів України є США, Велика Британія, Мальта, Кіпр, Німеччина та інші країни ЄС. Вже багато років українські IT-фахівці успішно розробляють різноманітні продукти для компаній з цих країн. Останнім часом серед цих рішень все більше систем, що мають вбудований ШІ. Більше того, багато українських стартапів відомі у всьому світі саме завдяки створенню успішних розробок у сфері генеративного ШІ, розпізнавання образів, комп'ютерного зору (Respeecher, Reface, Deus Robotics тощо). Деякі з них мають понад 100 мільйонів користувачів.

Однак, варто розуміти, що штучний інтелект, як одна з ключових технологій XXI століття, не лише трансформує суспільство, економіку та всі сфери життя, а й піднімає низку етичних, правових і соціальних питань, що вимагають уваги на рівні регулювання. Такі країни як США, Велика Британія, і, звісно, Європейський Союз, вже давно активно працюють над створенням правових рамок для контролю та розвитку ШІ.

Так Сполучені Штати Америки обрали багаторівневий і децентралізований підхід до регулювання ШІ. Федеральний уряд через низку агентств і відомств розробляє рекомендації та стандарти для розвитку ШІ в різних галузях. Одним із ключових документів, що визначає національну політику у сфері ШІ, є Національна стратегія США з розвитку штучного інтелекту (National AI Initiative), затверджена в 2020 році. Стратегія орієнтована на стимулювання інновацій, підвищення безпеки та етики використання ШІ, а також на просування американських інтересів на глобальному рівні.

Крім цього, в США активно працює така організація, як Національний інститут стандартів і технологій (NIST), який розробляє рекомендації щодо етичного та безпечного використання ШІ. Одним з інструментів для регулювання є NIST AI Risk Management Framework, що пропонує набір методологій для управління ризиками, пов'язаними з використанням ШІ. Також американське законодавство включає низку актів, які регулюють специфічні аспекти використання ШІ. Наприклад, Закон про алгоритмічну справедливість і підзвітність (Algorithmic Accountability Act) зобов'язує компанії аналізувати свої алгоритми для запобігання дискримінації та забезпечення справедливості.

Велика Британія, яка також активно впроваджує регуляторні рамки для ШІ, акцентує увагу на балансі між інноваціями та регуляцією. У квітні 2023 року уряд Великої Британії опублікував «Білу книгу з регулювання ШІ» (AI Regulation White Paper), яка окреслила ключові принципи національного підходу. Основна стратегія полягає в тому, щоб уникнути жорсткого регулювання, яке могло б стримати інновації, і замість цього надати гнучкі рамки для саморегулювання. У Великій Британії активно застосовується «bottom-up» підхід. Це дозволяє галузі ШІ самостійно розробляти стандарти та практики з урахуванням інтересів бізнесу та суспільства. Водночас британські регулятори наголошують на необхідності етичного використання ШІ та захисту прав громадян. Особливу увагу у Великій Британії приділяють питанням етики та прозорості. Наприклад, у 2022 році було запроваджено «Кодекс етики для ШІ» (AI Ethics Code), який містить рекомендації для розробників щодо етичного використання технологій ШІ.

В Україні також формується стратегія регулювання штучного інтелекту, створено ініціативні робочі групи фахівців, які разом з урядовцями виробляють концепцію розвитку ШІ в Україні. Зокрема, у червні 2024 року Міністерство цифрової трансформації презентувало розроблену Білу книгу з регулювання ШІ в Україні: бачення Мінцифри — аналітичний матеріал, що має на меті запропонувати підхід до регулювання технологій штучного інтелекту в Україні.

Основним відкритим питанням є те, як розробити таку систему регулювання ШІ в Україні, яка буде достатньо ефективною для забезпечення балансу між інноваціями та захистом прав громадян?

Запропонований Bottom-up підхід було розроблено із врахуванням трьох основних цілей держави у сфері ШІ:

Ціль 1. Підтримка конкурентоспроможності бізнесу, адже, враховуючи важливість розвитку IT-сектору для України, буде нераціонально занадто гальмувати її розвиток, виставляючи дуже суворі обмеження та вимоги.

Ціль 2. Захист прав людини. Водночас необхідно забезпечувати, щоб використання розроблених систем ШІ не порушувало існуючих законодавчих актів щодо дотримання прав людини та захисту персональних даних користувачів.

Ціль 3. Дотримання вимог Європейського Союзу у зв'язку із політикою та намірами Євроінтеграції України. Відповідно, на момент вступу в Європейський Союз за умовами приєднання Україна повинна відповідати існуючим законодавчим нормам та вимогам діючого законодавства. Оскільки в ЄС вже затверджений EU AI Act, то логічно внутрішнє регулювання ШІ в Україні максимально наблизити до прийнятого в ЄС.

Bottom-up підхід

Рис. 6.5. Структурна схема Bottom-up підходу до імплементації AI Act в Україні

Після багатьох років роздумів і консультацій, український уряд затвердив стратегію регулювання ШІ, яка базується на індуктивному підході, а також на концепції «bottom-up», тобто саморегулювання індустрії. Цей підхід подібний до того, що активно використовується в деяких західних країнах, таких як Велика Британія, і підходить для динамічної галузі, що розвивається швидкими темпами. Основна ідея цього підходу полягає у тому, щоб не встановлювати жорстких законодавчих рамок на ранніх етапах розвитку технології, надаючи бізнесу певну свободу у саморегулюванні.

Варто зазначити, що український підхід підтримує ідею гармонізації з європейським законодавством, зокрема, з майбутнім Законом ЄС про ШІ (AI Act), однак це потребує часу та поступової інтеграції. Саморегулювання покликане дати бізнесу час і можливість адаптуватися до змін, одночасно створюючи умови для розвитку відповідальних практик у сфері штучного інтелекту.

Основні цілі української стратегії включають підтримку конкурентоспроможності галузі, захист прав громадян, а також забезпечення гармонізації з європейським законодавством. Перший етап реалізації цієї стратегії передбачає використання «позазаконодавчих» інструментів, таких як методологія оцінювання систем ШІ, регуляторна пісочниця і маркування продуктів.

Інструменти українського регулювання

Одним із ключових інструментів української стратегії є методологія оцінювання систем ШІ. Це дозволить ідентифікувати ризики, пов'язані з використанням певних систем, і визначити їхній вплив на права людини. Такий підхід схожий до підходу, що використовується у США в рамках NIST AI Risk Management Framework.

Другим важливим інструментом є «регуляторна пісочниця», яка створить середовище для тестування нових систем ШІ перед їх широким впровадженням. Пісочниця дозволить визначити відповідність продуктів регуляторним вимогам і оцінити їхній вплив на суспільство. У США подібні ініціативи також існують, зокрема, через програму фінансування досліджень у сфері ШІ, яка підтримує стартапи та компанії, що займаються інноваціями у галузі.

Додатково, важливим елементом є маркування ШІ-продуктів, що дозволить підвищити прозорість систем ШІ і забезпечити належний рівень підзвітності розробників. Це також схоже на практики, що використовуються у Великій Британії, де маркування й сертифікація продуктів ШІ допомагає користувачам оцінити етичність та безпеку використання технології.

Якщо порівняти український підхід до регулювання ШІ з досвідом США та Великої Британії, можна побачити кілька спільних рис, а також деякі суттєві відмінності. Українська стратегія, подібно до американської та британської, орієнтована на саморегулювання та поступову адаптацію законодавчих норм. Водночас, українське законодавство більше орієнтоване на євроінтеграцію та гармонізацію з європейськими стандартами, зокрема, з AI Act.

У США система регулювання є більш розосередженою, але водночас чітко окреслює етичні стандарти та принципи використання ШІ. Велика Британія, як і Україна, зосереджується на гнучкому підході до регулювання, однак британська система більш розвинена в аспектах етики та прозорості використання ШІ.

Україна, зі свого боку, лише на початку шляху до створення комплексного регулювання ШІ, що відображено в Білій книзі. Це означає, що на початковому етапі впроваджуватимуться позазаконодавчі інструменти, як-от маркування та регуляторні пісочниці, але остаточне законодавче закріплення принципів використання ШІ відбудеться після гармонізації з законодавством ЄС.

З огляду на амбітність планів, перед Україною стоять численні виклики. Зокрема, важливо забезпечити належний захист персональних даних, особливо з огляду на те,

що багато ШІ-систем працюють із великими обсягами чутливої інформації. У цьому контексті необхідно вдосконалити законодавство, що регулює захист даних, і привести його у відповідність до Загальноєвропейського регламенту про захист даних (GDPR).

Іншим викликом є необхідність розробки чітких критеріїв для оцінки ризиків, пов'язаних з використанням ШІ. Важливо визначити, які системи вимагають більш суворого контролю, а які можуть діяти у режимі саморегулювання. Це дозволить створити більш диференційований підхід до регулювання і забезпечити ефективний контроль за ризиковими технологіями.

Український уряд також має налагодити ефективну співпрацю з бізнесом та науковою спільнотою для розробки стандартів використання ШІ. Це допоможе створити довіру між суспільством, розробниками та регуляторами і сприятиме більш відповідальному впровадженню нових технологій.

САМОКОНТРОЛЬ

Питання для самоконтролю:

1. Які основні розділи містить EU AI Act?
2. Скільки рівнів ризику для систем штучного інтелекту передбачає EU AI Act?
3. Що таке системи штучного інтелекту з неприйнятним ризиком згідно з EU AI Act?
4. Наведіть приклади систем штучного інтелекту з неприйнятним рівнем ризику.
5. Які методи класифікуються як маніпулятивні або оманливі в системах ШІ?
6. Які вимоги встановлюються для систем штучного інтелекту з високим ризиком?
7. У яких випадках дозволено використання дистанційної біометричної ідентифікації в реальному часі?
8. Які зобов'язання мають постачальники систем штучного інтелекту з високим ризиком?
9. Чому важливо враховувати вразливості користувачів під час розробки ШІ?
10. Які сфери діяльності особливо підпадають під регулювання систем штучного інтелекту з високим ризиком?

Контрольні запитання:

Перший рівень складності

Тестові питання:

1. Скільки рівнів ризику для систем штучного інтелекту визначає AI Act?
   a) Три
   b) Чотири
   c) П'ять
   d) Два

2. Що означає термін «неприйнятний ризик» у контексті AI Act?
   a) Системи ШІ, що не пройшли тестування
   b) Системи ШІ, які заборонені через ризик для прав та безпеки людей
   c) Системи ШІ, які потребують доопрацювання
   d) Системи ШІ, що використовуються лише урядами

3. У яких випадках дозволено використання дистанційної біометричної ідентифікації в реальному часі?

a) Для моніторингу співробітників

b) Для запобігання терактам та зникненню людей

c) Для контролю руху транспорту

d) Для оцінки кредитоспроможності

4. Які системи вважаються високоризиковими за AI Act?

a) Системи, що використовуються в правоохоронній діяльності

b) Системи, що використовуються для створення відеоігор

c) Системи для обробки персональних даних у соціальних мережах

d) Системи, що використовуються для прогнозування погоди

5. Які вимоги висуває AI Act до систем із високим ризиком?

a) Повна заборона

b) Обов'язкове тестування та документація

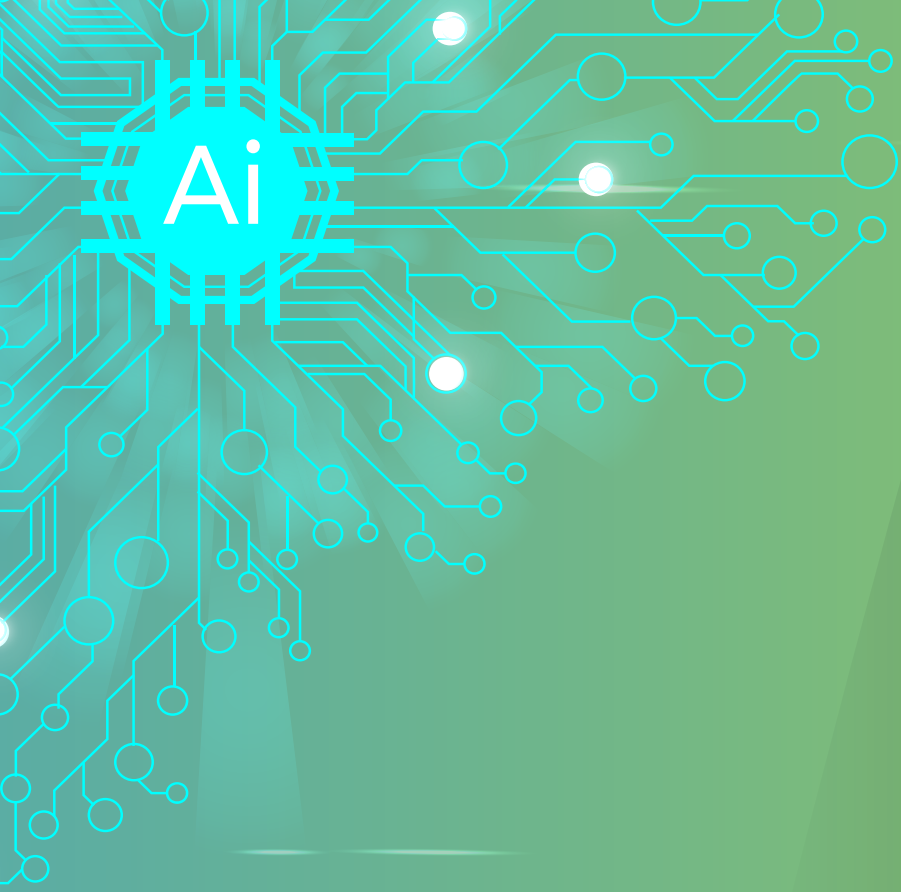
c) Відсутність будь-яких вимог

d) Використання лише для військових потреб

Другий рівень складності

Контрольні питання:

1. Які основні сфери застосування AI Act у сфері штучного інтелекту?
2. Як визначаються системи з високим та неприйнятним рівнем ризику в AI Act?
3. Які приклади систем із неприйнятним ризиком визначені в AI Act?
4. Які обмеження та вимоги встановлюються для систем біометричної ідентифікації?
5. Які особливості оцінки ризиків у критичній інфраструктурі згідно з AI Act?
6. Що таке «соціальна оцінка» і чому вона заборонена AI Act?
7. Як визначається законність використання дистанційної біометричної ідентифікації у публічних місцях?
8. Які зобов'язання щодо документації та нагляду висуває AI Act для систем високого ризику?
9. Яким чином AI Act регулює використання ШІ у сферах освіти та професійної підготовки?
10. Які заходи з мінімізації ризиків повинні впроваджувати постачальники високоризикових систем ШІ?



7

ПЕРСПЕКТИВИ РОЗВИТКУ СИСТЕМ НАДІЙНОГО ШІ

7.1. Технічні методи реалізації надійного ШІ

Для реалізації вимог до надійного ШІ, які було описано у попередніх розділах, можна використовувати як технічні, так і нетехнічні методи. Вони охоплюють всі етапи життєвого циклу системи ШІ. Оцінка методів, що використовуються для реалізації вимог, а також звітування та обґрунтування змін у процесах впровадження, повинні відбуватися на постійній основі. Системи ШІ постійно розвиваються і діють у динамічному середовищі. Тому реалізація надійного ШІ є безперервним процесом.

Наведені нижче методи та підходи можуть бути як взаємодоповнювальними, так і альтернативними один одному, оскільки різні вимоги можуть спричиняти потребу в різних методах реалізації. Розглянемо деякі з найбільш поширених та ефективних методів, які можуть допомогти у впровадженні надійного штучного інтелекту.

Розглянемо технічні методи забезпечення надійності ШІ, які можна застосувати на етапах проектування, розробки та використання системи ШІ. Перелічені нижче методи різняться за рівнем зрілості.

Архітектура для надійного ШІ

Відповідно до вимог до надійного ШІ можна сформувати процедури та/або обмеження, які повинні бути закріплені в архітектурі системи ШІ. Цього можна досягти за допомогою набору правил «білого списку» з поведінки або станів, яких система повинна завжди дотримуватися, обмежень «чорного списку» на поведінку або стани, які система ніколи не повинна порушувати, а також комбінації цих або більш складних доказових гарантій щодо поведінки системи. Моніторинг дотримання системою цих обмежень під час роботи може бути забезпечений окремим процесом.

Системи штучного інтелекту зі здатністю до навчання, які можуть динамічно адаптовувати свою поведінку, можна розглядати як недетерміновані системи, що можуть демонструвати неочікувану поведінку. Вони часто розглядаються через теоретичну призму циклу «відчуття-план-дія». Адаптація цієї архітектури для забезпечення надійного ШІ вимагає інтеграції вимог на всіх трьох етапах циклу:

1. **На етапі «відчуття»** система повинна бути розроблена таким чином, щоб вона розпізнавала всі елементи середовища, необхідні для дотримання вимог;

2. **На етапі «план»** система повинна розглядати тільки ті плани, які відповідають вимогам;

3. **На етапі «дія»** дії системи повинні бути обмежені поведінкою, яка реалізує вимоги.

Описана вище архітектура є загальною і дає лише недосконалий опис більшості систем штучного інтелекту. Тим не менш, вона дає опорні точки для обмежень і політик, які повинні бути відображені в конкретних модулях, щоб в результаті отримати загальну систему, яка заслуговує на довіру і сприймається як така.

Зображена вище архітектура є загальною і дає лише недосконалий опис більшості систем ШІ. Тим не менш, вона дає опорні точки для обмежень і політик, які повинні бути відображені в конкретних модулях, щоб у результаті отримати загальну систему, яка заслуговує на довіру і сприймається як така.

Етика та верховенство права за задумом (X-by-design)

Методи забезпечення цінностей за задумом забезпечують точний і чіткий зв'язок між абстрактними принципами, яких система повинна дотримуватися, і конкретними рішеннями щодо їх реалізації. Ідея про те, що дотримання норм може бути впроваджене в дизайн системи ШІ, є ключовою для цього методу. Компанії несуть відповідальність за потенційний вплив своїх систем ШІ на користувачів з самого початку її розробки. Відповідно, мають дотримуватись всі існуючі норми до системи ШІ (законодавчі, етичні, стандарти тощо), щоб запобігти негативним наслідкам. Різні концепції «by-design» вже широко використовуються, наприклад, для дотримання вимог GDPR: privacy-by-design і security-by-design. Відповідно, щоб завоювати довіру, ШІ повинен бути безпечним у своїх процесах, даних і результатах, а також бути стійким до ворожих даних і атак. Він повинен реалізовувати механізм безпечного вимкнення і дозволяти відновлювати роботу після вимушеного вимкнення (наприклад, внаслідок атаки).

Методи пояснення

Щоб системі можна було довіряти, ми повинні розуміти, чому вона поводиться певним чином і чому вона дає ту чи іншу відповідь. Ціла галузь досліджень – пояснювальний ШІ (Explainable AI, XAI) – намагається вирішити цю проблему, щоб краще зрозуміти механізми, що лежать в основі системи, і знайти рішення. Сьогодні це все ще залишається викликом для систем ШІ, заснованих на нейронних мережах. Процеси навчання нейронних мереж можуть призвести до того, що параметри мережі набувають числових значень, які важко співвіднести з результатами. Більше того, іноді невеликі зміни в значеннях даних можуть призвести до кардинальних змін в інтерпретації, що може призвести до того, що система, наприклад, сплутає шкільний автобус зі страусом. Ця вразливість також може бути використана під час атак на систему. Використання методів XAI під час розробки системи ШІ є життєво важливим не лише для пояснення поведінки системи користувачам, але й для розгортання надійної технології.

Тестування та валідація

Через недетерміновану та контекстно-залежну природу систем штучного інтелекту традиційного тестування недостатньо. Помилки в концепціях і уявленнях, що використовуються системою, можуть проявитися лише тоді, коли програма застосовується до достатньо реалістичних даних. Отже, для перевірки і валідації опрацювання даних та оцінювання прийнятого рішення необхідно ретельно контролювати базову модель як під час навчання, так і під час розгортання на предмет її стабільності, надійності і роботи в добре зрозумілих і передбачуваних межах. Необхідно переконатися, що резуль-

тати процесу планування відповідають вхідним даним, а рішення приймаються таким чином, щоб можна було перевірити процес, що лежить в його основі.

Тестування та валідація системи повинні відбуватися якомога раніше, щоб забезпечити належну поведінку системи протягом усього її життєвого циклу і особливо після розгортання. Вони повинні охоплювати всі компоненти системи ШІ, включаючи дані, попередньо навчені моделі, середовище і поведінку системи загалом. Дані важливо перевірити на відсутність упередженості та зміщення, збалансованість та коректність. Важливо також правильно визначити метрики, аналіз яких буде визначати адекватність та точність роботи системи.

Процеси тестування повинні розроблятися і виконуватися якомога більш різноманітною групою людей. Необхідно розробити кілька метрик, щоб охопити категорії, які тестуються, з різних точок зору. Можна розглянути можливість проведення змагального тестування довіреними і різноманітними «червоними командами», які навмисно намагаються зламати систему, щоб знайти вразливості, а також запроваджувати винагороди за виправлення помилок, які заохочують сторонніх осіб виявляти і негайно повідомляти про помилки і слабкі місця системи. Нарешті, необхідно переконатися, що результати або дії узгоджуються з результатами попередніх процесів, порівнюючи їх з раніше визначеними політиками, щоб переконатися, що вони не порушуються.

Показники якості обслуговування

Для систем штучного інтелекту можна визначити відповідні показники якості обслуговування, щоб забезпечити базове розуміння того, чи були вони протестовані та розроблені з урахуванням міркувань безпеки та захисту. Ці показники можуть включати заходи для оцінки тестування і навчання алгоритмів, а також традиційні програмні метрики функціональності, продуктивності, зручності використання, надійності, безпеки і ремонтопридатності.

7.2. Нетехнічні методи реалізації надійного ШІ

Розглянемо також різноманітні нетехнічні методи, які можуть відігравати важливу роль у забезпеченні та підтримці надійного ШІ. Їх також слід оцінювати на постійній основі.

Регулювання

Як уже згадувалося вище, законодавство, що підтримує довіру до ШІ та регулює процес створення та використання надійного ШІ, вже існує сьогодні у багатьох країнах. Важливо зазначити і те, що існуюче регулювання розглядає питання надійного ШІ з усіх точок зору – як з технічного, так і з погляду етики та моралі використання ШІ.

Кодекси поведінки

Організації та зацікавлені сторони можуть створити свою хартію корпоративної

відповідальності вимогам надійного ШІ, визначити ключові показники ефективності (KPI), кодекси поведінки або документи внутрішньої політики, щоб додати прагнення до надійного ШІ. Організація, яка працює над системами ШІ або з ними, може задокументувати свої наміри, а також підкріпити їх стандартами певних бажаних цінностей, таких як фундаментальні права, прозорість і уникнення шкоди.

Стандартизація

Стандарти, наприклад, на дизайн, виробництво та бізнес-практики, можуть функціонувати як система управління якістю для користувачів ШІ, споживачів, організацій, науково-дослідних установ та урядів, пропонуючи можливість визнавати та заохочувати етичну поведінку під час прийняття рішень. Окрім звичайних стандартів, існують регуляторні підходи: системи акредитації, професійні етичні кодекси або стандарти для дизайну, що відповідає основним правам людини. Нинішніми прикладами є, наприклад, стандарти ISO або серія стандартів IEEE P7000, але в майбутньому може з'явитися етикетка «Надійний ШІ», яка підтверджуватиме, посилаючись на конкретні технічні стандарти, що система, наприклад, відповідає вимогам безпеки, технічної надійності та прозорості.

Сертифікація

Оскільки не можна очікувати, що кожна людина здатна повністю зрозуміти роботу та наслідки систем ШІ, можна розглянути можливість створення організацій, які можуть засвідчити широкій громадськості, що система ШІ є прозорою, підзвітною та справедливою. Така сертифікація застосовуватиме стандарти, розроблені для різних сфер бізнесу та методів ШІ, які належним чином узгоджені з промисловими та суспільними стандартами для різних предметних областей. Однак, сертифікація ніколи не може замінити відповідальність. Тому вона повинна доповнюватися системами підзвітності, включно з відмовами від відповідальності, а також механізмами перевірки та відшкодування.

Підзвітність через системи управління

Організації повинні створити внутрішні та зовнішні системи управління, що забезпечують підзвітність за етичні аспекти рішень, пов'язаних із розробкою, впровадженням та використанням систем ШІ. Це може включати, наприклад, призначення особи, відповідальної за етичні питання, пов'язані з системами ШІ, або створення внутрішньої чи зовнішньої комісії з питань етики. Серед можливих функцій такої особи, комісії або ради - забезпечення нагляду та надання консультацій. Як зазначалося вище, специфікації та органи сертифікації також можуть відігравати певну роль у цьому процесі. Необхідно забезпечити канали зв'язку з галузевими та/або громадськими наглядовими групами для обміну найкращими практиками, обговорення дилем або повідомлення про нові етичні проблеми. Такі механізми можуть доповнювати, але не можуть замінити правовий нагляд (наприклад, у формі призначення відповідального за захист даних або еквівалентних заходів, передбачених законодавством про захист даних).

Освіта та обізнаність для формування етичного мислення

Надійний ШІ заохочує усвідомлену участь усіх зацікавлених сторін. Комунікація, освіта і навчання відіграють важливу роль як для забезпечення широкого розповсюдження знань про потенційний вплив систем ШІ, так і для усвідомлення людьми того, що вони можуть брати участь у формуванні суспільного розвитку. Це стосується всіх зацікавлених сторін, наприклад, тих, хто бере участь у створенні продуктів (дизайнерів і розробників), користувачів (компаній і приватних осіб) та інших груп, на які впливають системи ШІ (тих, хто може не купувати і не використовувати системи ШІ, але за яких вони ухвалюють рішення, і суспільства загалом). Базову грамотність у сфері штучного інтелекту слід розвивати в усьому суспільстві. Передумовою для навчання громадськості є забезпечення належних навичок і підготовки фахівців з етики в цій сфері.

Участь зацікавлених сторін і соціальний діалог

Переваги від використання систем штучного інтелекту є вражаючими, тому важливо забезпечити їхню доступність для всіх. Це вимагає відкритого обговорення та залучення соціальних партнерів і зацікавлених сторін, зокрема широкої громадськості. Багато організацій вже покладаються на групи зацікавлених сторін для обговорення використання систем штучного інтелекту та аналізу даних. До складу цих груп входять різні члени, такі як юристи, технічні експерти, фахівці з етики, представники споживачів і працівники. Активне залучення до участі та діалогу щодо використання та впливу систем штучного інтелекту сприяє оцінці результатів і підходів і може бути особливо корисним у складних випадках.

Різноманітність та інклюзивні команди розробників

Різноманітність та інклюзивність відіграють важливу роль при розробці систем ШІ, які будуть застосовуватися в реальному світі. Оскільки системи ШІ виконують все більше завдань самостійно, дуже важливо, щоб команди, які проектують, розробляють, тестують і підтримують, розгортають і закуповують ці системи, відображали різноманітність користувачів і всього суспільства. Це сприяє об'єктивності та врахуванню різних поглядів, потреб і цілей. В ідеалі, команди мають бути різноманітними не лише за статтю, культурою, віком, але й за професійним досвідом та набором навичок.

САМОКОНТРОЛЬ

Питання для самоконтролю:

1. Які технічні методи реалізації надійного ШІ існують?
2. Чому важливо дотримуватися правил білого та чорного списку в архітектурі ШІ?
3. Як етапи «відчуття», «план» та «дія» пов'язані з надійним ШІ?
4. Що таке метод «етика за задумом» і як він застосовується до ШІ?
5. Як пояснювальний ШІ (XAI) впливає на надійність систем ШІ?
6. Чому тестування та валідація є важливими для надійності системи ШІ?
7. Які показники якості обслуговування використовуються для систем ШІ?
8. Як регулювання сприяє надійності ШІ?
9. Яка роль сертифікації у впровадженні надійного ШІ?
10. Чому різноманітність команд розробників важлива для створення надійних систем ШІ?

Контрольні запитання:

Перший рівень складності

Тестові питання:

1. Який з методів відноситься до технічних для забезпечення надійного ШІ?
   a) Регулювання
   b) Тестування та валідація
   c) Участь зацікавлених сторін
   d) Освіта та обізнаність

2. На якому етапі життєвого циклу ШІ система повинна розпізнавати всі елементи середовища?
   a) Відчуття
   b) Планування
   c) Дія
   d) Тестування

3. Який підхід передбачає впровадження етичних норм у дизайн ШІ?
   a) Пояснювальний ШІ
   b) Архітектура за задумом

c) Кодекси поведінки

d) Показники якості обслуговування

4. Що забезпечує система сертифікації для ШІ?

a) Відповідність регуляторним нормам

b) Пояснення результатів користувачам

c) Валідацію моделі

d) Моніторинг якості обслуговування

5. Чому важливо мати різноманітні команди розробників у процесі створення ШІ?

a) Для зменшення вартості розробки

b) Для підвищення різноманітності точок зору

c) Для поліпшення продуктивності алгоритмів

d) Для зниження складності моделювання

Другий рівень складності

Теоретичні питання:

1. Опишіть основні етапи життєвого циклу надійної системи ШІ.
2. Що таке обмеження «білого» і «чорного» списку і як вони впливають на поведінку ШІ?
3. Як реалізується надійність ШІ на етапах відчуття, планування та дії?
4. Які основні принципи концепції «за задумом» і як їх можна застосувати до етичного дизайну ШІ?
5. Що таке Explainable AI (XAI) і чому він є важливим для надійного ШІ?
6. Як різноманітні групи людей можуть вплинути на процес тестування ШІ?
7. Які нетехнічні методи реалізації надійного ШІ можна використовувати для його впровадження?
8. Опишіть роль сертифікації та стандартизації у забезпеченні надійного ШІ.
9. Які ключові виклики виникають при тестуванні систем ШІ?
10. Як залучення соціальних партнерів і зацікавлених сторін може допомогти у формуванні надійного ШІ?

ВИСНОВКИ

Сьогодні штучний інтелект розвивається надзвичайно швидко, володіє фантастичними здібностями, вміє вирішувати найрізноманітніші задачі: він без проблем не лише може перетворити текст на голос, а голос на текст, впізнати вас на фотографії чи відео, за лічені секунди лише за одним фото ідентифікувати будь яку рослину, перекласти текст безпосередньо на зображенні з однієї мови на будь-яку іншу, але й створити нову пісню чи картину, написати вірш чи реферат, згенерувати відео, на якому до вас звертатиметься відомий вчений чи актор, який вже давно помер. Більше того – на виробництвах безліч задач сьогодні виконують роботи, а безпілотні автомобілі вже дуже скоро можуть стати звичним явищем на вулицях міст.

Такий прогрес та поширення систем ШІ безперечно орієнтований на те, щоб полегшити та покращити життя людей, вирішити економічні та екологічні проблеми, зробити суспільство інклюзивнішим та вирішити багатьох інших важливих проблем. Однак, як розробникам систем штучного інтелекту, так і користувачам таких систем ніколи не варто забувати про те, що чим більше влади ми надаємо ШІ, тим більш вагомі виклики може спричиняти його використання. І мова зараз аж ніяк не про повстання машин з фантастичних фільмів (поки що це питання не на часі), а про дуже банальні скорингові системи, які класифікують людей і вирішують: кому надати кредит – а кому ні, кого взяти на роботу, а кого – ні, хто благонадійний громадянин, а хто – ні…. Яка буде вага помилкового рішення, прийнятого такою системою? Чи не зруйнує вона чиюсь долю, чи не позбавить можливості отримати гарну освіту? Чи не поставить хибний діагноз?

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 надійності можуть виникати в системах ШІ, а також детально описано різноманітні підходи, завдяки яким ці ризики можна зменшити чи загалом позбавитись їх.

ГЛОСАРІЙ

AI-in-the-loop — це переосмислення концепції « людини в циклі », яка визнає важливість людей у процесі прийняття рішень за допомогою AI. Він визнає цінність людського досвіду та здатності приймати рішення, одночасно використовуючи сильні сторони штучного інтелекту для підтримки та допомоги в процесі прийняття рішень. AI-in-the-loop наголошує на тому, що штучний інтелект має покращувати роботу людей, але люди завжди повинні залишатися в центрі прийняття рішень.

Алгоритмічне зміщення (Algorithmic Bias) - виникає, коли алгоритми, які використовуються в моделях машинного навчання, мають властиві зміщення, які відображаються на їхніх результатах. Це може статися, коли алгоритми базуються на упереджених припущеннях або коли вони використовують упереджені критерії для прийняття рішень.

Вузький штучний інтелект (Artificial Narrow Intelligence) зосереджений на конкретному, окремому або цілеспрямованому завданні та не має функціональних можливостей саморозширення для вирішення незнайомих проблем.

Генеративний штучний інтелект (Generative AI) – це штучний інтелект, який може створювати новий контент, використовуючи наявний текст, аудіо чи зображення.

Глибокі нейронні мережі (Deep Neural Networks, DNN) – моделі з багатьма прихованими шарами, кожен з яких відповідає за виявлення складніших патернів.

Зміщення даних (Data Bias) виникає, коли дані, які використовуються для навчання моделей машинного навчання, є нерепрезентативними або неповними, що призводить до упереджених результатів. Це може статися, коли дані збираються з упереджених джерел або коли дані неповні, не мають важливої інформації або містять помилки

Контрольоване навчання (Supervised learning – навчання із вчителем) - під час контрольованого навчання алгоритм навчається на попередньо визначеному наборі даних, де кожен приклад має вхідні та відповідні вихідні дані, для того, щоб робити прогнози щодо нових, невідомих даних.

Humans-in-the-Loop (HITL) — це підхід у розробці штучного інтелекту, який передбачає людський нагляд і контроль над процесом машинного навчання. HITL зазви-

чай використовується для перевірки та вдосконалення моделей машинного навчання, а також для обробки винятків і граничних ситуацій, які системі штучного інтелекту важко впоратися самостійно.

Humans-on-the-Loop (HOTL) — це розширення HITL, яке залучає людей, які надають зворотній зв'язок системі штучного інтелекту для покращення її продуктивності з часом. HOTL зазвичай використовується, коли система штучного інтелекту досягла певного рівня продуктивності, але все ще потребує відгуку та втручання людини для продовження вдосконалення. У HOTL люди діють як тренери або вчителі для ШІ, надаючи позначені дані, виправляючи помилки та направляючи ШІ до кращих результатів

Людиноорієнтований дизайн (Human-Centered Design, HCD) — це підхід до проектування продуктів, систем і послуг, який визначає пріоритетність потреб, бажань і обмежень користувачів. Це передбачає співпереживання кінцевим користувачам і розуміння їх контексту, поведінки та вподобань для створення ефективних, ефективних і задовільних рішень.

Неконтрольоване навчання (Unsupervised learning – навчання без вчителя) - навчальні дані за такого підходу не мають будь-яких попередньо визначених міток або вихідних даних – алгоритм навчається виявляти приховані залежності, структури чи групи в даних.

Навчання з підкріпленням (Reinforcement learning) - Створюється агент, який може взаємодіяти з навколишнім середовищем і навчатися методом проб і помилок. У процесі навчання та функціонування агент отримує зворотний зв'язок у формі винагород або штрафів під час виконання дій, які дозволяють йому навчатися та покращувати ефективність з часом.

Прозорість (Transparency) – у штучному інтелекті це не лише можливість пояснення, а й забезпечення того, щоб люди розуміли, що робить ваша система штучного інтелекту та як вона це робить.

Пояснюваність (Explainability) — це здатність системи штучного інтелекту надавати користувачам чіткі та зрозумілі пояснення процесів прийняття рішень. Це важливо для створення довіри до систем штучного інтелекту, виявлення та виправлення помилок і упереджень, а також забезпечення відповідності систем штучного інтелекту людським цінностям і етичним принципам.

Реактивні машини — це системи штучного інтелекту без пам'яті та призначені для виконання дуже специфічних завдань. Оскільки вони не можуть згадати попередні результати чи рішення, вони працюють лише з наявними на даний момент даними.

Рекурентні нейронні мережі (RNN) і **LSTM** (Long Short-Term Memory) – архітектури, що використовуються для послідовних даних, таких як часові ряди або природна мова.

Сильний ШІ (Strong AI), або **загальний ШІ** (Artificial General Intelligence, AGI) — це гіпотетична концепція системи, яка може мати інтелект, аналогічний людському, тобто вона здатна виконувати будь-які завдання, що вимагають інтелекту, в різних доменах, включаючи ті, до яких вона не була спеціально навчена.

Слабкий штучний інтелект (Weak Artificial Intelligence) зосереджений на конкретному, окремому або цілеспрямованому завданні та не має функціональних можливостей саморозширення для вирішення незнайомих проблем.

Соціальне упередження (Social Bias) стосується створених людиною упереджень, таких як стереотипи, які можуть відображатися в системах ШІ.

Статистичне зміщення (Statistical Bias) означає систематичну помилку в прогнозах системи штучного інтелекту, яка виникає через зміщення даних або алгоритмів.

Упередженість користувачів (User Bias) виникає, коли люди, які використовують системи штучного інтелекту, свідомо чи несвідомо вносять у систему власні переконання чи упередження. Це може статися, коли користувачі надають упереджені навчальні дані або коли вони взаємодіють із системою у спосіб, який відображає їхні власні упередження.

Штучний інтелект (Artificial Intelligence, AI) – це технічна та наукова галузь, присвячена розробленій системі, яка генерує такі результати, як контент, прогнози, рекомендації чи рішення для заданого набору цілей, визначених людиною.

ШІ загального призначення (General-purpose AI (GPAI)) - відноситься до систем штучного інтелекту, які мають широкий спектр можливих застосувань, як передбачуваних, так і непередбачених розробниками. Вони можуть бути застосовані для багатьох різних завдань у різних областях, часто без суттєвих змін і тонкого налаштування.

ШІ з обмеженою пам'яттю - ця форма штучного інтелекту може згадувати минулі події та результати та контролювати конкретні об'єкти чи ситуації в режимі реального часу. Штучний інтелект з обмеженою пам'яттю може використовувати дані минулого та поточного моментів, щоб прийняти рішення про напрямок дій, який, швидше за все, допоможе досягти бажаного результату.

Штучний суперінтелект (Artificial Superintelligence, ASI)— це тип ШІ, який перевершує людський інтелект і може виконувати будь-які завдання краще, ніж людина. Системи ASI не тільки розуміють людські почуття та досвід, але також можуть викликати власні емоції, переконання та бажання, подібні до людських.

Список використаних літературних джерел

1. ISO/IEC. Information Technology – Artificial Intelligence – Artificial Intelligence Concepts and Terminology (ISO/IEC 22989:2022); International Organization for Standardization (ISO), Geneva, Switzerland, 2022. https://www.iso.org/standard/74296.html
2. Russell, S.; Norvig, P. Artificial Intelligence: A Modern Approach, 4th ed.; Pearson: London, UK, 2020.
3. Kurzweil, R. The Singularity is Near: When Humans Transcend Biology; Penguin Books: New York, NY, USA, 2005.
4. Sarangi, S.; Sharma, P. Artificial Intelligence: Evolution, Ethics, and Public Policy; Routledge: London, UK, 2023.
5. IBM. Responsible AI in Healthcare: Addressing Bias in Algorithms. Available online: https://www.ibm.com/think/topics/responsible-ai-healthcare (accessed on 16 May 2024).
6. Bringsjord, S. Making AI Safe for Humans: A Conversation With Siri. Available online: https://www.researchgate.net/publication/301880679_Making_AI_Safe_for_Humans_A_Conversation_With_Siri (accessed on 16 May 2024).
7. Wu, Dan & Huang, Yeman. (2021). Why Do You Trust Siri?: The Factors Affecting Trustworthiness of Intelligent Personal Assistant. Proceedings of the Association for Information Science and Technology. 58. 366-379. 10.1002/pra2.464.
8. Joshi, A. Artificial Intelligence and Human Evolution: Contextualizing AI in Human History; Apress: New York, NY, USA, 2023.
9. Grigorescu, S., Trasnea, B., Cocias, T., & Macesanu, G. A Survey of Deep Learning Techniques for Autonomous Driving. Journal of Field Robotics, 2020, 37(3), 362-386. doi:10.1002/rob.21918.
10. Shladover, S. E. Connected and Automated Vehicle Systems: Introduction and Overview. Journal of Intelligent Transportation Systems, 2018, 22(3), 190-200. doi:10.1080/15472450.2017.1336053.
11. Litman, T. Autonomous Vehicle Implementation Predictions: Implications for Transport Planning. Victoria Transport Policy Institute, 2020. Available online: https://www.vtpi.org/avip.pdf (accessed on 15 April 2024).
12. Dixit, V. V., Chand, S., & Nair, D. Autonomous Vehicles: Disengagements, Accidents, and Risk Implications. Accident Analysis & Prevention, 2020, 141, 105501. doi:10.1016/j.aap.2020.105501.
13. Kaplan, A., & Haenlein, M. Siri, Siri, in My Hand: Who's the Fairest in the Land? On the Interpretations, Illustrations, and Implications of Artificial Intelligence. Business Horizons, 2019, 62(1), 15-25. doi:10.1016/j.bushor.2018.08.004.
14. Boden, M. A. AI: Its Nature and Future; Oxford University Press: Oxford, UK, 2016.
15. Risks of Artificial Intelligence. United Kingdom, CRC Press, 2016.
16. Mitchell, T. Machine Learning; McGraw-Hill: New York, NY, USA, 1997.
17. Goodfellow, I.; Bengio, Y.; Courville, A. Deep Learning; MIT Press: Cambridge, MA, USA, 2016.
18. Sutton, R. S.; Barto, A. G. Reinforcement Learning: An Introduction, 2nd ed.; MIT Press: Cambridge, MA, USA, 2018.

19. Zhao, Yang & Liu, Guoliang & Tian, Guohui & Luo, Yong & Wang, Ziren & Zhang, Wei & Li, Junwei. (2017). A Survey of Visual SLAM Based on Deep Learning. Robotics. 39. 889-896. 10.13973/j.cnki.robot.2017.0889.
20. Flach, P. Machine Learning: The Art and Science of Algorithms That Make Sense of Data; Cambridge University Press: Cambridge, UK, 2012.
21. Schmidhuber, J. Deep Learning in Neural Networks: An Overview; Elsevier: Amsterdam, The Netherlands, 2015.
22. Alan Turing, «Computing Machinery and Intelligence», Mind, vol. LIX, no. 236, October 1950, pp. 433—460
23. Turing, Alan (October 1950), Computing Machinery and Intelligence, Mind (англ.), LIX (236): 433—460, doi:10.1093/mind/LIX.236.433, ISSN 0026-4423
24. Searle, John (2009), Chinese room argument, Scholarpedia, 4 (8): 3100, Bibcode:2009SchpJ...4.3100S, doi:10.4249/scholarpedia.3100
25. Angwin, J.; Larson, J.; Mattu, S.; Kirchner, L. Machine bias. In Ethics of Data and Analytics; Auerbach Publications: Boca Raton, FL, USA, 2016; pp. 254–264. [Google Scholar]
26. Mittelstadt, B.D.; Allo, P.; Taddeo, M.; Wachter, S.; Floridi, L. The ethics of algorithms: Mapping the debate. Big Data Soc. 2016, 3, 2053951716679679. [Google Scholar] 27. Dastin, J. Amazon scraps secret AI recruiting tool that showed bias against women. In Ethics of Data and Analytics; Auerbach Publications: Boca Raton, FL, USA, 2018; pp. 296–299. [Google Scholar]
28. Ferrara, E. The butterfly effect in artificial intelligence systems: Implications for AI bias and fairness. arXiv 2023, arXiv:2307.05842. [Google Scholar] [CrossRef]
29. Schwartz, R.; Vassilev, A.; Greene, K.; Perine, L.; Burt, A.; Hall, P. Towards a Standard for Identifying and Managing Bias in Artificial Intelligence; NIST Special Publication: Gaithersburg, MD, USA, 2022; Volume 1270, pp. 1–77
30. Ferrara, E. Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models. First Monday 2023, 28. [Google Scholar]
31. Kleinberg, J.; Mullainathan, S.; Raghavan, M. Inherent trade-offs in the fair determination of risk scores. In Proceedings of the Innovations in Theoretical Computer Science (ITCS), Berkeley, CA, USA, 9–11 January 2017.
32. Buolamwini, J.; Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency, New York, NY, USA, 23–24 February 2018; pp. 77–91.
33. Eubanks, V. Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor; St. Martin's Press: New York, NY, USA, 2018.
34. Berk, R.; Heidari, H.; Jabbari, S.; Kearns, M.; Roth, A. Fairness in Criminal Justice Risk Assessments: The State of the Art. Sociol. Methods Res. 2018, 47, 175–210. [Google Scholar] [CrossRef]
35. Friedler, S.A.; Scheidegger, C.; Venkatasubramanian, S.; Choudhary, S.; Hamilton, E.P.; Roth, D. A comparative study of fairness-enhancing interventions in machine learning. In Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 29–31 January 2019; pp. 329–338. [Google Scholar]
36. Murphy, K. P. Machine Learning: A Probabilistic Perspective; MIT Press: Cambridge, MA, USA, 2012.

37. Chollet, F. Deep Learning with Python, 2nd ed.; Manning Publications: Shelter Island, NY, USA, 2021.
38. Bishop, C. M. Pattern Recognition and Machine Learning; Springer: Berlin, Germany, 2006.
39. Russell, S. J.; Dewey, D.; Tegmark, M. Research Priorities for Robust and Beneficial Artificial Intelligence; MIT Press: Cambridge, MA, USA, 2015.
40. Witten, I. H.; Frank, E.; Hall, M. A. Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed.; Elsevier: Amsterdam, The Netherlands, 2011.
41. Rumelhart D.E., Hinton G.E., Williams R.J., Learning Internal Representations by Error Propagation. In: Parallel Distributed Processing, vol. 1, pp. 318—362. Cambridge, MA, MIT Press. 1986.
42. Kleinberg, J.; Lakkaraju, H.; Leskovec, J.; Ludwig, J.; Mullainathan, S. Human decisions and machine predictions. Q. J. Econ. 2018, 133, 237–293. [Google Scholar] [PubMed]
43. O'Neil, C. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy; Broadway Books: New York, NY, USA, 2016. [Google Scholar]
44. Donovan, J.; Caplan, R.; Matthews, J.; Hanson, L. Algorithmic Accountability: A Primer; Data & Society: New York, NY, USA, 2018. [Google Scholar]
45. Ezzeldin, Y.H.; Yan, S.; He, C.; Ferrara, E.; Avestimehr, S. Fairfed: Enabling group fairness in federated learning. In Proceedings of the AAAI 2023—37th AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023. [Google Scholar]
46. Asan, O.; Bayrak, A.E.; Choudhury, A. Artificial intelligence and human trust in healthcare: Focus on clinicians. J. Med. Internet Res. 2020, 22, e15154. [Google Scholar] [CrossRef] [PubMed]
47. Yan, S.; Kao, H.T.; Ferrara, E. Fair class balancing: Enhancing model fairness without observing sensitive attributes. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, Online, 26 June–31 July 2020; pp. 1715–1724. [Google Scholar]
48. Caliskan, A.; Bryson, J.J.; Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. Science 2017, 356, 183–186. [Google Scholar] [CrossRef] [PubMed]
49. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. arXiv 2016, arXiv:1602.04938.
50. Lauer, D.(2021). You cannot have AI ethics without ethics. AI and Ethics,(1), 21–25. https://doi.org/10.1007/s43681-020-00013-4
51. Mittelstadt, B.D.; Allo, P.; Taddeo, M.; Wachter, S.; Floridi, L. The ethics of algorithms: Mapping the debate. Big Data Soc. 2016, 3, 2053951716679679. [Google Scholar] [CrossRef]
52. Ananny, M.; Crawford, K. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media Soc. 2018, 20, 973–989. [Google Scholar] [CrossRef]
53. Zafar, M.B.; Valera, I.; Gomez Rodriguez, M.; Gummadi, K.P. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In Proceedings of the 26th International Conference on World Wide Web, Perth, Australia, 3–7 May 2017; pp. 1171–1180. [Google Scholar]

54. Kamiran, F.; Calders, T. Data preprocessing techniques for classification without discrimination. Knowl. Inf. Syst. 2012, 33, 1–33. [Google Scholar] [CrossRef]
55. Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; Zemel, R. Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, Cambridge, MA, USA, 8–10 January 2012; pp. 214–226. [Google Scholar]
56. Barocas, S.; Selbst, A.D. Big data's disparate impact. Calif. Law Rev. 2016, 104, 671–732. [Google Scholar] [CrossRef]
57. Selbst, A.D.; Boyd, D.; Friedler, S.A.; Venkatasubramanian, S.; Vertesi, J. Fairness and abstraction in sociotechnical systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 29–31 January 2019; pp. 59–68. [Google Scholar]
58. Angwin, J.; Larson, J.; Mattu, S.; Kirchner, L. Machine bias. In Ethics of Data and Analytics; Auerbach Publications: Boca Raton, FL, USA, 2016; pp. 254–264. [Google Scholar]
59. Obermeyer, Z.; Powers, B.; Vogeli, C.; Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019, 366, 447–453. [Google Scholar] [CrossRef]
60. Ferrara, E. GenAI against humanity: Nefarious applications of generative artificial intelligence and large language models. arXiv 2023, arXiv:2310.00737. [Google Scholar] [CrossRef]
61. European Commission. Ethics Guidelines for Trustworthy AI. Commission Communication. 2019. Available online: https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1 (accessed on 15 May 2024).
62. Sweeney, L. Discrimination in online ad delivery. Commun. ACM 2013, 56, 44–54. [Google Scholar] [CrossRef]
63. Ananny, M.; Crawford, K. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media Soc. 2018, 20, 973–989. [Google Scholar] [CrossRef]
64. Kleinberg, J.; Ludwig, J.; Mullainathan, S.; Rambachan, A. Algorithmic fairness. AEA Pap. Proc. 2018, 108, 22–27. [Google Scholar] [CrossRef]
65. Schwartz, R.; Vassilev, A.; Greene, K.; Perine, L.; Burt, A.; Hall, P. Towards a Standard for Identifying and Managing Bias in Artificial Intelligence; NIST Special Publication: Gaithersburg, MD, USA, 2022; Volume 1270, pp. 1–77.
66. Crawford, K.; Calo, R. There is a blind spot in AI research. Nature 2016, 538, 311–313. [Google Scholar] [CrossRef]
67. Ferguson, A.G. Predictive policing and reasonable suspicion. Emory LJ 2012, 62, 259. [Google Scholar] [CrossRef]
68. Wachter, S.; Mittelstadt, B.; Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harv. J. Law Technol. 2018, 31, 841–887. [Google Scholar] [CrossRef]
69. Žliobaitė, I. Measuring discrimination in algorithmic decision making. Data Min. Knowl. Discov. 2017, 31, 1060–1089. [Google Scholar] [CrossRef]
70. Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; Huq, A. Algorithmic decision making and the cost of fairness. In Proceedings of the 23rd ACM SIGKDD International

Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, 13–17 August 2017; pp. 797–806. [Google Scholar]
71. Crenshaw, K. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. In Feminist Legal Theories; Routledge: London, UK, 1989; pp. 23–51. [Google Scholar]
72. Nicoletti, L.; Bass, D. Humans Are Biased: Generative AI Is Even Worse. Bloomberg Technology + Equality, 23 June 2023. [Google Scholar]
73. Chauhan, P.S.; Kshetri, N. The Role of Data and Artificial Intelligence in Driving Diversity, Equity, and Inclusion. Computer 2022, 55, 88–93. [Google Scholar] [CrossRef]
74. The Times (2023). AI is a Threat to the Creative World, Says JK Rowling's Agent. Available online: https://www.thetimes.com/culture/books/article/ai-jk-rowling-agent-threat-creative-world-industry-vrw5mndxb (accessed on 14 May 2024)
75. L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, E. J. M. Vayena (2018), «AI4People —An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations", Minds and Machines 28(4): 689-707.
76. High-Level Expert Group on Artificial Intelligence (AI HLEG). (2018). Ethical Guidelines for Trustworthy AI. [Online]. Available: https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1
77. European Commission, 2020d. White Paper on Artificial Intelligence–A European approach to excellence and trust. URL: https://ec.europa.eu/info/publications/ white-paper-artificial-intelligence-european-approach-excellence-and-trust_en. (Accessed 20 November 2023).
78. 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).
79. 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).
80. 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).
81. 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).
82. 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)
83. 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)

84. 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).
85. U.S. White House. (2020). Guidance for Regulation of Artificial Intelligence Applications. Available online: [URL] (accessed on 27 July 2024).
86. IEEE. (2019). Ethically Aligned Design. [Online]. Available: https://ethicsinaction.ieee.org
87. UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence. Available online: https://unesdoc.unesco.org/ark:/48223/pf0000379981 (accessed on 30 July 2024).
88. UK AI Council. (2021). AI Roadmap. Available online: https://www.aicouncil.org.uk/ai-roadmap (accessed on 30 May 2024).
89. European Parliament and Council. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data. Official Journal of the European Union, L 119, 1–88. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679 (accessed on 30 May 2024).

Додаток. Варіанти використання ШІ

Незаборонена біометрія: системи дистанційної біометричної ідентифікації, за винятком біометричної перевірки, яка підтверджує, що особа є тим, ким вона себе видає. Біометричні системи категоризації, які визначають конфіденційні або захищені атрибути чи характеристики. Системи розпізнавання емоцій.

Критична інфраструктура: компоненти безпеки в управлінні та експлуатації критичної цифрової інфраструктури, дорожнього руху та постачання води, газу, опалення та електроенергії.

Освіта та професійно-технічна підготовка: системи штучного інтелекту, що визначають доступ, вступ або призначення до освітніх та професійно-технічних навчальних закладів усіх рівнів. Оцінювання результатів навчання, у тому числі тих, які використовуються для керування процесом навчання студента. Оцінка відповідного рівня освіти особи. Контроль та виявлення забороненої поведінки учнів під час контрольних робіт.

Працевлаштування, управління працівниками та доступ до самозайнятості: системи штучного інтелекту, які використовуються для найму чи відбору, особливо цільових оголошень про роботу, аналізу та фільтрації заявок та оцінки кандидатів. Підвищення по службі та розірвання контрактів, розподіл завдань на основі особистих рис або характеристик і поведінки, а також контроль і оцінка продуктивності.

Доступ і використання основних державних і приватних послуг: системи штучного інтелекту, які використовуються державними органами для оцінки права на отримання пільг і послуг, включаючи їх розподіл, скорочення, скасування або відновлення. Оцінка кредитоспроможності, крім випадків виявлення фінансового шахрайства. Оцінка та класифікація екстрених викликів, у тому числі визначення пріоритетів для поліції, пожежників, медичної допомоги та термінових служб сортування пацієнтів. Оцінка ризиків і ціноутворення в страхуванні здоров'я та життя.

Правоохоронні органи: системи ШІ, які використовуються для оцінки ризику особи стати жертвою злочину. Поліграфи. Оцінка достовірності доказів під час кримінального розслідування або судового переслідування. Оцінка індивідуального ризику вчинення правопорушення або повторного вчинення злочину не виключно на основі профілювання чи оцінки рис особистості чи злочинної поведінки в минулому. Профілювання під час виявлення злочинів, розслідування чи судового переслідування.

Управління міграцією, притулком та прикордонним контролем: Поліграф. Оцінка нелегальної міграції або ризиків для здоров'я. Розгляд заявок на отримання притулку, візи та дозволу на проживання, а також пов'язаних скарг щодо відповідності вимогам. Виявлення, розпізнавання або ідентифікація осіб, крім перевірки проїзних документів.

Відправлення правосуддя та демократичні процеси: системи штучного інтелекту, які використовуються для дослідження та тлумачення фактів і застосування закону до конкретних фактів або для альтернативного вирішення спорів. Вплив на результати виборів і референдумів або поведінку при голосуванні, за винятком результатів, які безпосередньо не взаємодіють з людьми, як-от інструменти, що використовуються для організації, оптимізації та структурування політичних кампаній.


Ai