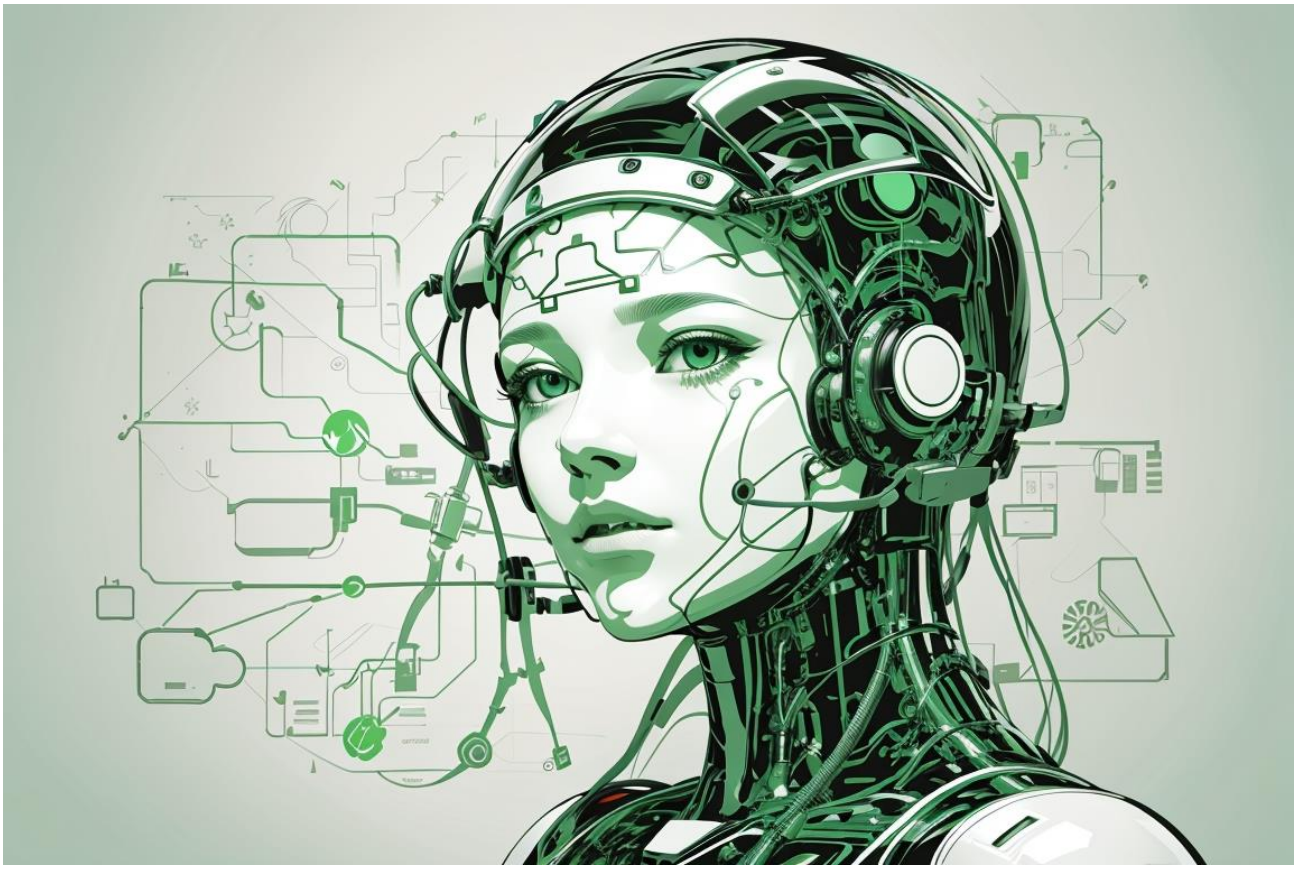
Project number	101085626
Project name	Trustworthy artificial intelligence: the European approach
Funding Programme	ERASMUS2027
Project start date	01-10-2022

Deliverable number	1.4
Deliverable name	Teaching materials for the module “A European Approach to AI” (for PhD in Computer Science)
Work Package number	1
Lead Beneficiary	Lviv Polytechnic National University
Type	DEM — Demonstrator, pilot, prototype R — Document, report
Dissemination level	Public
Due date (in months)	24
Description	The use cases for discovering existing ethical risks in AI technologies, accordingly to a legal framework on AI by collaborative learning group should be created using the software for interactive cooperation. e-format, Ukrainian language
Website link	https://trustai.org.ua/portfolio-item/d4_teaching-materials-a-european-approach-to-ai-phd-trustai/
Author(s)	Anastasiya Doroshenko

Co-funded by the European Union

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ЛЬВІВСЬКА ПОЛІТЕХНІКА 1816 LITTERIS ET ARTIBVS

TRUSTAI



Анастасія Дорошенко

Європейський підхід до штучного інтелекту

Навчальний посібник

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

Розглянуто вимоги діючого європейського законодавства до створення та використання систем штучного інтелекту. Описано основні вимоги до надійного ШІ. Розглянуто міжнародні стандарти та світові нормативно-правові акти, які визначають вимоги до розроблення надійних систем ШІ. Розглянуто варіанти використання штучного інтелекту, а також можливі виклики, які можуть у них виникнути. Описано рекомендації створення етичного ШІ в цих сферах людського життя.

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

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

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

Table of Contents

Вступ

Сьогодні ми є свідками появи «економіки та суспільства штучного інтелекту», де технології штучного інтелекту все більше впливають на охорону здоров’я, бізнес, транспорт і багато аспектів повсякденного життя. Повідомлялося про багато успіхів, коли системи ШІ навіть перевершували точність експертів-людей. Однак системи штучного інтелекту можуть створювати помилки, виявляти упередженість, можуть бути чутливими до шуму в даних і часто не мають технічної та судової прозорості, що призводить до зниження довіри та проблем із їх впровадженням. Ці нещодавні недоліки та занепокоєння були задокументовані в пресі, як-от нещасні випадки з безпілотними автомобілями, упередження в системі охорони здоров’я, прийомі на роботу та системах розпізнавання облич для кольорових людей, начебто правильні медичні рішення, які пізніше виявилися прийнятими через неправильні причини тощо. Це призвело до появи багатьох урядових і регуляторних ініціатив, які вимагали створювати надійний і етичний ШІ, який повинен мати певний рівень точності і надійності, певну форму пояснювальності, людський контроль і нагляд, усунення упередженості. Проблеми в створенні надійних систем штучного інтелекту спонукали до інтенсивних досліджень у сфері систем пояснювального штучного інтелекту (XAI). Метою XAI є надання зрозумілої людині інформації про те, як системи ШІ приймають рішення.

Розвиток штучного інтелекту, подальше розширення використання штучних інтелектуальних систем на виробництві, у сферах послуг, освіти, охорони здоров'я, транспорту, будівництва, вимагає відображення цього у нормативно-правових актах. Перетворюючий вплив штучного інтелекту на суспільні відносини є дуже значущим, як на національному та міжнародному рівнях в останні роки активно обговорюється прийняття законів і конвенцій, які регулюють питання, пов'язані з використанням технологій штучного інтелекту на практиці. Флагманом цього руху є Європейський союз, який поставив завдання найбільш повно регулювати вказану область і задати стандарти, які будуть багато в чому наслідуватись іншими гравцями на світовій арені.

В цьому посібнику ми розглянемо європейський підхід до регулювання штучного інтелекту для подальшої оцінки відповідності завданням, поставленим Європейським союзом, його можливостям.

1. Історія регулювання штучного інтелекту в Європейському Союзі

Технології штучного інтелекту та продукти на їх основі отримують швидке поширення на території держав — членів Європейського союзу в умовах розвитку індустрії 4.0 на виробництві, у сфері послуг, побуту, освіти, охорони здоров'я та інших умов. Ці процеси вимагають необхідності реагування зі сторони європейського права, що, у свою чергу, викликає інтерес і за межами Європейського союзу.

Європейські дослідники почали займатися тематикою співвідношення штучного інтелекту і права досить давно, приблизно в той же час, що й вчені з американських університетів. Спочатку мова йшла про можливості використання штучного інтелекту в якості інструменту для полегшення та упорядкування роботи юристів, прикладом чого є праця К. Чампи «Штучний інтелект та правові інформаційні системи», пізніше були підняті питання, пов'язані з регулюванням застосування технологій штучного інтелекту на практиці. Це стало необхідним через потреби для права зберегти свою властивість випереджаючого відображення дійсності в умовах високої динаміки суспільних відносин у зв'язку з цифровізацією, переходом до нових технологічних укладів.

Серед європейських авторів, що займаються дослідженнями в цій області, можна назвати професора Туринського університету У. Пагалло, який опублікував в 2013 р. книгу по «праву роботів» [1], його колегу-юриста Туринського політехнічного університету Э. Бассі, викладача Європейського коледжу в Брюге, який є радником Європейської комісії П. Немиця, яка вже в 2018 р. наполягала на переході від етичних кодексів до формулювання норм права в досліджуваній області, і професора Л'єжского університету Н. Петі, який запропонував виділяти різні підходи до регулювання штучного інтелекту. Н. Петі, як і декілька інших європейських юристів, увійшов в створену Європейською комісією в 2018 р. міждисциплінарну групу експертів високого рівня зі штучного інтелекту (High-Level Expert Group on Artificial Intelligence). Ця група підготувала більшу частину документів, закладених в основу європейської стратегії розвитку штучного інтелекту та, відповідно, європейського підходу до її регулювання. Протягом першого року роботи групи було сформульовано досить об'ємне визначення штучного інтелекту як програмної або програмно-апаратної системи, розробленої людиною, що має складну ціль, що діє в фізичній або віртуальній реальності, яка сприймає навколишнє середовище за допомогою збору даних, інтерпретуючи ці

дані та роблячи висновки на підставі обробки цих даних . Система ШІ повинна прийняти рішення про найкращі дії, яких необхідно вжити для досягнення поставленої цілі. Координатором групи експертів високого рівня з мистецтва інтелекту при Європейській комісії став дослідник Левенського католицького університету Н. Смуха, її роботи також присвячені впливу штучного інтелекту на права людини та етико-правовим питанням, що з'являються внаслідок поширення продуктів на основі штучного інтелекту. Центральне місце в наукових працях європейських експертів права займає проблема запобігання та зниженню ризиків, пов'язаних із застосуванням системи штучного інтелекту, оскільки ці системи потенційно можуть порушувати дотримання прав людини, функціонування демократії та верховенство права. Активність проведених наукових досліджень стимулюється політичним запитом на розвиток правового регулювання технологій штучного інтелекту в Європейському союзі та в світі в цілому.

Також у квітні 2018 року в ЄС було презентовано європейську стратегію штучного інтелекту [2], 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 [3], 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.

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 [4]. 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 [5].

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 [6].

Після широкого обговорення сформульованих принципів в експертному колі було сформульовано проект закону із регулювання штучного інтелекту в ЄС. Європейська комісія оприлюднила пропозицію щодо нового закону про штучний інтелект (EU AI Act) у квітні 2021 року [7].

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.

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].

2. Основні положення 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.

2.1. Ризико-орієнтований підхід у AI Act

Перш за все варто зазначити, що AI Act базується на ризико-орієнтованому підході, відповідно до якого визначено чотири рівні ризику для систем ШІ (рис. 1.).

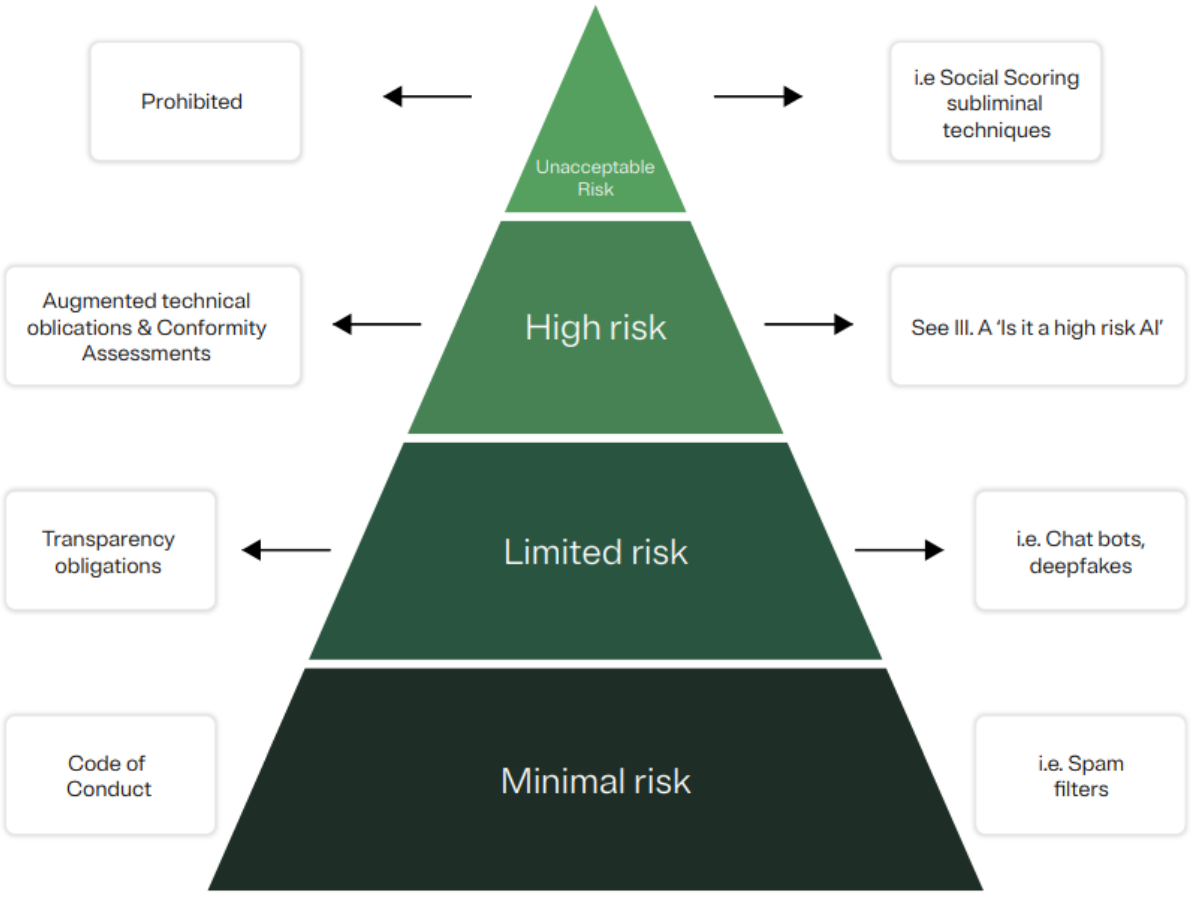


Рис. 1. Чотири рівні ризику для систем ШІ

2.1.1. Системи ШІ із неприйнятним ризиком (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).

2.1.2. 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 притягнення 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.

2.1.3. 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 європейських 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.

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.

2.1.4. 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.

2.2. 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.

Хто є ключовими учасниками ланцюжка створення вартості ШІ?

Рис.2. Учасники ланцюжка створення вартості ШІ відповідно до AI Act

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

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.

3. Принципи розробки надійної системи ШІ

Принципи розробки надійної системи ШІ включають технічну стійкість, прозорість та підзвітність. Важливою є участь людини у прийнятті рішень (HITL, HOTL, HIC), а також забезпечення етичних стандартів, таких як недискримінація і захист даних. Системи ШІ повинні бути ретельно протестовані та піддаватися перевірці для мінімізації ризиків і негативних наслідків.

3.1. Основні вимоги до надійного ШІ

Основні принципи, які визначені в EU AI Act, базуються на рекомендаціях експертів HLEG щодо створення надійного ШІ [4].

Відповідно до цих рекомендацій надійний штучний інтелект має три компоненти, яких необхідно дотримуватися протягом усього життєвого циклу системи:

1. він має бути законним, відповідати всім застосовним законам і нормам
2. він має бути етичним, забезпечуючи дотримання етичних принципів і цінностей
3. система ШІ повинна бути надійною як з технічної, так і з соціальної точки зору, оскільки навіть з гарними намірами системи ШІ можуть завдати ненавмисної шкоди.

Кожен компонент сам по собі необхідний, але недостатній для досягнення надійного ШІ. В ідеалі всі три компоненти працюють узгоджено та перетинаються у своїй роботі. Якщо на практиці між цими компонентами виникає напруженість, суспільство має прагнути їх вирівняти.

Для реалізації запропонованого підходу групою експертів було розроблено фреймворк для створення надійного ШІ (Framework for Trustworthy AI), в якому, зокрема, визначено сім основних вимог до системи, дотримання яких впливає на те, чи вважатиметься система ШІ надійною (рис. 3.). Забезпечувати дотримання цих вимог можна як технічними, так і нетехнічними методами. Таким чином, Рекомендації встановлюють структуру для досягнення надійного ШІ та пропонують вказівки щодо другого та третього компонентів: сприяння та забезпечення етичного та надійного штучного інтелекту. Базуючись на підході, заснованому на фундаментальних правах, визначено етичні принципи та пов’язані

з ними цінності, які необхідно поважати при розробці, розгортанні та використанні систем ШІ.

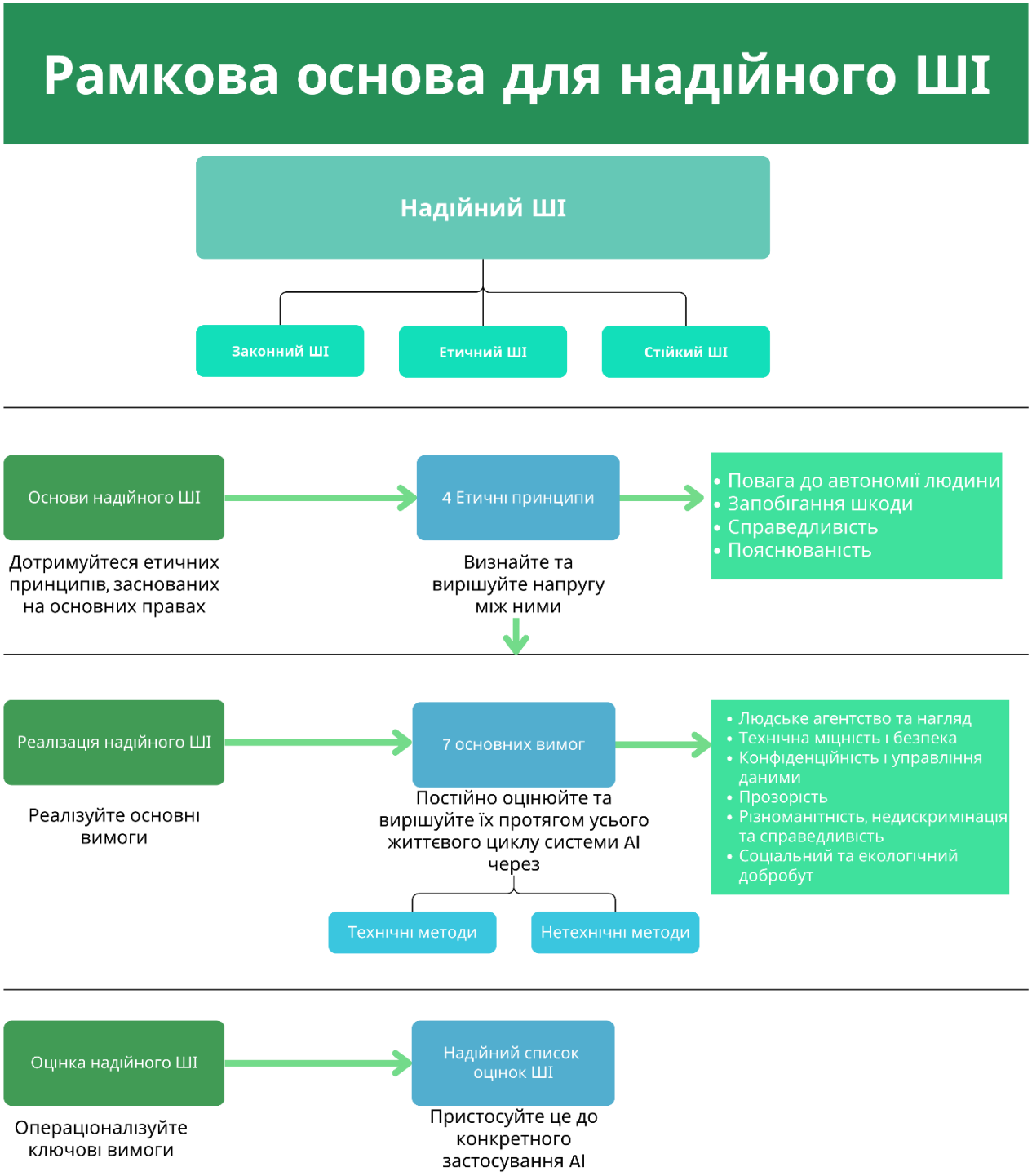


Рис.3. Фреймворк для створення надійного ШІ від експертів HLEG

Розглянемо послідовно кожну з семи ключових вимог до надійного ШІ та методи, які допомагають їх дотримуватись.

3.1.1. 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

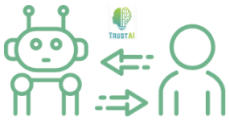
Користувачі повинні мати можливість самостійно приймати обгрунтовані рішення щодо систем ШІ. Їм повинні бути надані знання та інструменти для розуміння і взаємодії з системами ШІ на задовільному рівні, а також, де це можливо, вони повинні мати можливість розумної самооцінки або оскарження системи. Системи 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 продуктивність 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.

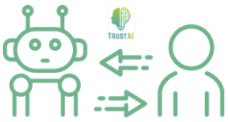
4. Підвищена впевненість і довіра
Людський нагляд вселяє довіру до систем штучного інтелекту, надаючи пояснення, інтерпретації та гарантуючи, що рішення відповідають людським цінностям і намірам.

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)**

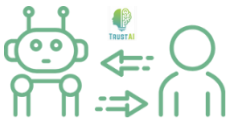
HOTL означає можливість людського втручання під час циклу проектування системи і моніторингу роботи системи.



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.

3.1.2. 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. 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.

Таким чином чутливість виходу системи до зміни входу вимірюється стійкістю. Вона оцінює здатність моделі правильно функціонувати в разі невизначеності. На поведінку системи не повинні істотно впливати невеликі зміни вхідних даних. Цей атрибут отримується шляхом піддавання моделі змагальним вхідним даних і забезпечення того, щоб частота помилок системи була близькою до рівня під час навчання [11,15].

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 [17].

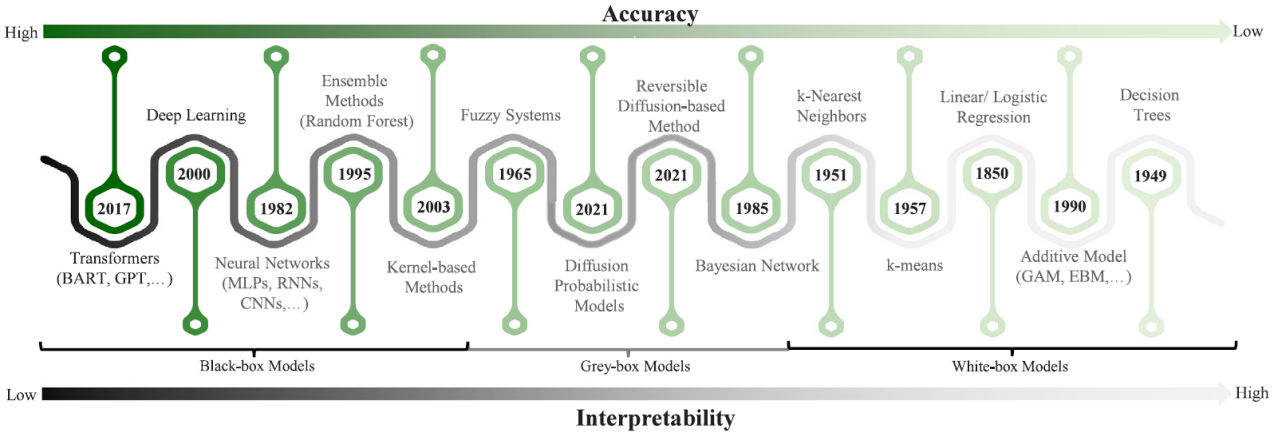


Рис. 4 Порівняння методів ШІ за точністю та пояснюваністю

Системи ШІ на високому рівні можна розділити на системи на основі функцій, де вхідні дані здебільшого представлені в табличній формі з 10 або 100 різними характеристиками (показники певної змінної, як-от медичні тести), які або вже існують, або розроблені людьми, і зараз популярні системи глибокого навчання (зазвичай у формі згорткових нейронних мереж), де входами є зображення та сигнали з мільйонами елементів даних або пікселів, які надходять безпосередньо до систем ШІ без участі людини в розробці функцій. Одна з переваг систем глибокого навчання полягає в тому, що немає потреби в інтенсивній «розробці функцій», оскільки вони можуть виявляти функції, які люди не можуть виявити.

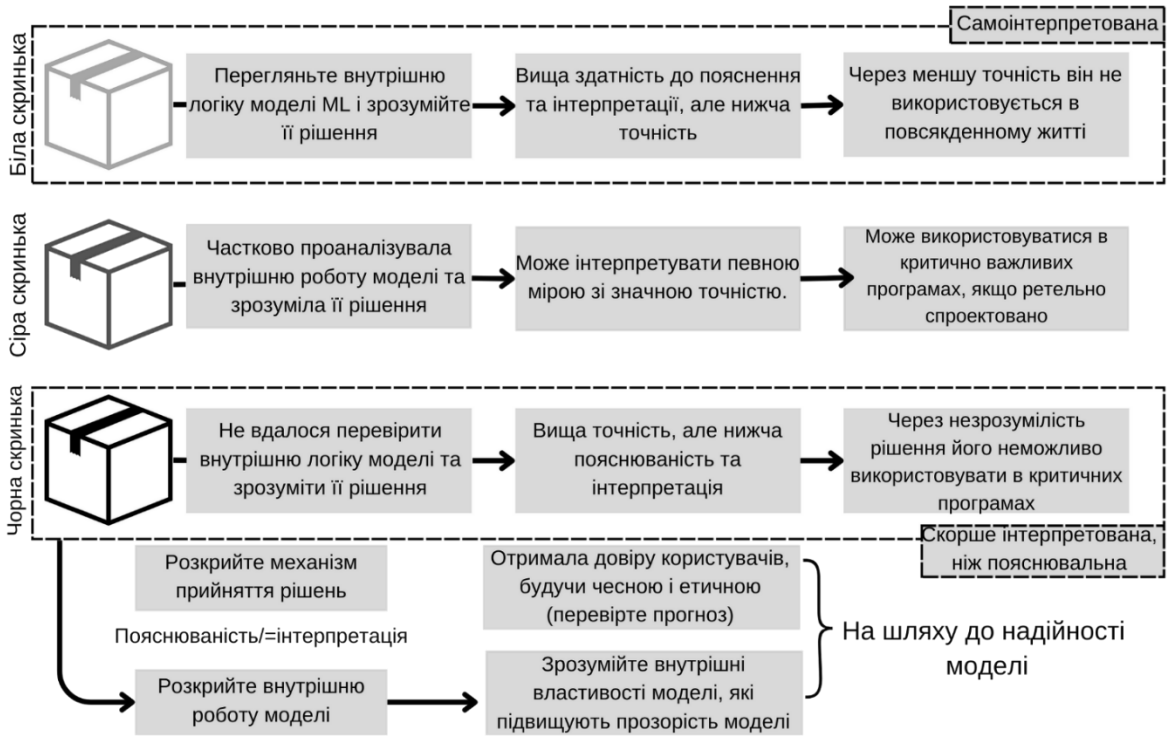


Рис. 5. Порівняння моделей білої скриньки, сірої скриньки та чорної скриньки

Порівняння моделей білої скриньки, сірої скриньки та чорної скриньки зображено на рис. №. З одного боку, моделі білої скриньки є інтерпретованими за дизайном, що робить їхні результати легшими для розуміння, але менш точними. Крім того, моделі сірого ящика дають хороший компроміс між точністю та інтерпретацією. З іншого боку, моделі чорного ящика є найбільш точними, але менш інтерпретованими. Для створення надійних моделей потрібні більш складні методи XAI.

Однак, ці особливості, присутні у внутрішніх шарах глибоких нейронних мереж, важко спостерігати та зрозуміти (на відміну від системи, заснованої на функціях), і це, у свою чергу, ускладнює їх пояснення. Тому сьогодні актуальною науковою задачею є розроблення методів XAI (explainable AI, пояснювального ШІ), спрямованих на те, щоб пояснити, як системи штучного інтелекту приймають рішення, і зробити це легко зрозумілим для людей способом [19,23].

XAI загалом не має на меті забезпечити будь-який аналіз причинно-наслідкових зв'язків або інтерпретабельності (наприклад, причини, чому рішення були прийняті), а просто надає фактичну інформацію про конкретні способи, якими алгоритми ШІ приймають рішення на основі даних. Методи XAI зазвичай передбачають аналіз уже навчених моделей ШІ, і вони можуть бути такими: глобальний або на основі моделі (пояснення того, як модель AI загалом працює, наприклад, на сукупності навчальних даних) та на основі зразку або локальний (як ШІ класифікував конкретний зразок). Вони можуть бути на основі агностичних алгоритмів (незалежно від конкретного алгоритму ШІ) або специфічних алгоритмів. Зазвичай агностичні алгоритми XAI (найпопулярнішим є LIME) базуються на певній формі локальної апроксимації або моделювання алгоритму штучного інтелекту, який вони намагаються пояснити [25]. Спеціальні пояснювачі алгоритмів використовують навчену систему ШІ, яку вони прагнуть пояснити.

Більшість методів XAI ідентифікують і ранжують «найбільш передбачувані ознаки» або області зображення («виразні області зображення») за допомогою типу аналізу чутливості, наприклад, змінюючи дані та спостерігаючи за змінами в точності, де ті характеристики, що викликають більшу зміну в точності, оцінюються вище. Загалом, досягнути задовільного рівня пояснюваності XAI набагато складніше для систем глибокого навчання – більшість методів глибокого навчання XAI обмежені зразком. Локальні пояснювачі переважно лише вказують на те, які області зображення (пікселі) були використані для класифікації зразка, а не на те, як насправді була виконана класифікація.

Системи на основі функцій часто пропонують більше інформації XAI, особливо для систем на основі дерева. XAI є відносно новою областю, і тривають інтенсивні дослідницькі зусилля в XAI, особливо для систем глибокого навчання [19].

Навіщо жертвувати точністю заради пояснюваності - аргументи «за» і «проти».

Існують різні думки щодо того, чи потрібно жертвувати точністю системи заради підвищення її прозорості та пояснюваності. Одні дослідники вважаються, що відсутностя потреба в XAI для високоточних (але важко пояснюваних) систем глибокого навчання чорних скриньок. Люди довіряють рішенням, навіть якщо вони не знають, як вони приймаються: У цьому аргументі стверджується, що наполягання на XAI має свої ризики, і якщо непояснені системи чорних скриньок є точними та добре перевіреними, їх слід використовувати. Зазначимо, що ця аналогія проводиться відносно. Людська реакція на рішення, прийняті людьми (наприклад, через поради експертів, які використовують історичний досвід), що не повністю доведено для рішень, прийнятих системами ШІ. Люди можуть бути більше стурбовані незрозумілими рішеннями, якщо їх приймає якась система штучного інтелекту, і особливо за наявності помилок у даних/сутностях, які вони розуміють [16-20]. Про це свідчить сильна негативна реакція громадськості, пов'язана із задокументованими помилками та упередженістю систем штучного інтелекту, наприклад, у розпізнаванні облич, наймі, аваріях безпілотного автомобіля тощо [14].

Важливо, однак, зауважити, що більшість людей-користувачів довіряють машинам, які вони погано розуміють, якщо вони сертифіковані перевіреною агенцією (наприклад, скільки людей знають, як працюють автопілоти, але вони все одно користуються повітряним транспортом, знаючи, що літаки сертифіковані). Тоді це повертає нас до необхідності певної форми незалежної перевірки чи сертифікації систем штучного інтелекту з великими ставками, як це пропонується в більшості підходів для досягнення надійного штучного інтелекту. Системи глибокого навчання дуже точні та надійні: однак, і тут є кілька питань, які викликають занепокоєння. Спочатку ми спостерігаємо, що деякі програми глибокого навчання штучного інтелекту, які, здавалося, досягли високої точності, пізніше, за допомогою певної форми аналізу XAI, зробили це «через усі неправильні причини» [8,9,10]. В одному прикладі з [15] програми глибокого навчання медичних зображень досягли, здавалося б, дуже точних рішень, використовуючи інформацію, пов'язану з типом рентгенівського апарату, а не інформацію про пацієнта (причина полягала в тому, що хворі та здорові пацієнти отримували зображення за допомогою постійно різних рентгенівських апаратів). Важливо також, що системи глибокого навчання є дорогими для навчання та вимагають величезної кількості якісних навчальних даних і часто потребують допомоги спеціального обладнання.

Обговорення того, скільки навчальних даних нам потрібно для належного навчання моделі ШІ, є складним і залежить від складності моделі ШІ, кількості класів рішень і статистики даних. На основі комплексного теоретичного аналізу дослідники розробили деякі емпіричні правила щодо мінімального розміру

навчальних даних, два з яких найчастіше згадуються: a) нам потрібно на мінімальний порядок (у 10 разів) більше навчальних даних, ніж ступінь свободи моделі ШІ (наприклад, число тренуваних параметрів); b) нам потрібно мінімум 1000 навчальних зразків на клас рішень. З огляду на те, що більшість моделей штучного інтелекту, заснованих на функціях, мають 10-100 функцій, тоді як глибоке навчання часто має 10 мільйонів тренувальних нейронних зв'язків, що вказує на потребу у величезних наборах навчальних даних в останньому випадку, часто на 5-6 порядків більших, ніж для систем на основі функцій.

Навчання моделей штучного інтелекту високої складності або «ступенів свободи» з недостатньою кількістю даних викликає проблеми з надійною оцінкою точності, наприклад перенавчання часто називають «прокляттям розмірності» .

Як наслідок, системам глибокого навчання часто не вистачає надійності (29), що, у свою чергу, може відкрити їх для агресивних атак (31) – ще одна причина для тривалого навчання та тестування. Отже, досягнення справжньої високої точності та стійкості систем глибокого навчання для майбутніх (невидимих) і шумних даних є нетривіальним завданням і, як правило, набагато складнішим, ніж для систем на основі функцій. Популярним застосуванням методів глибокого навчання, які показують найкращі результати під час застосування - це додатки штучного інтелекту на основі методів розпізнавання зображень, таких як розпізнавання обличчя, діагностика медичних зображень тощо, бракує багатьох важливих і дуже точних функцій (табличних) алгоритмів.

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.

3.1.3. Конфіденційність та 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 поведінки 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.

3.1.4. 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. Відстежуваність 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 відстежуваними 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 [14].

3.1.5. Різноманітність, недискримінація 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 [16]. 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 [18].

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.

3.1.6. Соціальний та екологічний добробут (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.

3.1.7. Звітність (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 системи 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.

Рис. 6. Фреймворк для створення надійного ШІ від експертів HLEG

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.

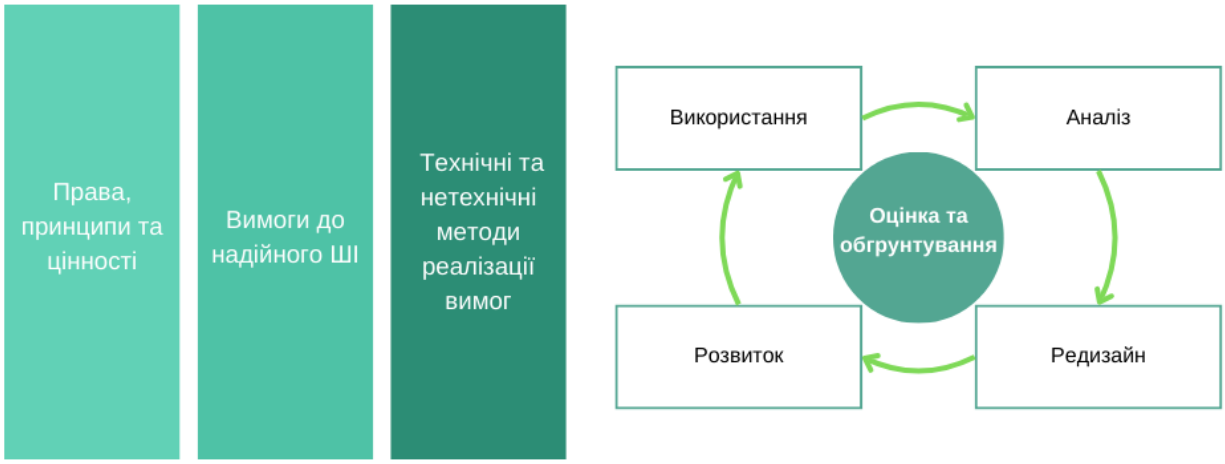


Рис. 7. :Життєвий цикл створення надійної системи ШІ

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.

Use Case. Оцінювальний список для надійного штучного інтелекту

Для того, щоб дізнатись, наскільки система штучного інтелекту відповідає всім вимогам надійного ШІ, необхідно здійснити сукупність різних оцінок по кожному з параметрів. Це може бути досить складно для невеликих компаній чи стандартів, які не мають спеціальних працівників (зокрема юристів), які б організували цей процес. Однак, відповідно до вимог AI Act, проведення як мінімум такої самооцінки є обов'язковим для систем ШІ.

Тому Групою експертів високого рівня зі штучного інтелекту, створеною Європейською Комісією, було розроблено веб-сайт, який містить Оцінювальний список для надійного ШІ (The Assessment List for Trustworthy Artificial Intelligence, ALTAI) (https://altai.insight-centre.org/). ALTAI був розроблений , щоб допомогти оцінити, чи система ШІ, яка розробляється, розгортається, закуповується або використовується, відповідає семи вимогам надійного ШІ, як зазначено в нашому Рекомендаціях з етики для надійного ШІ .

- ✓ Людське агентство та нагляд.
- ✓ Технічна надійність і безпека.
- ✓ Конфіденційність і управління даними.
- ✓ Прозорість.
- ✓ Різноманітність, недискримінація та справедливість.
- ✓ Соціальний та екологічний добробут.
- ✓ Відповідальність.

Мета і призначення ALTAI

ALTAI забезпечує базовий процес оцінки для самооцінки надійного ШІ. Організації можуть черпати елементи, пов'язані з конкретною системою штучного інтелекту, з ALTAI або додавати елементи до неї, як вони вважають за потрібне, беручи до уваги сектор, у якому вони працюють. Це допомагає організаціям зрозуміти, що таке надійний штучний інтелект, зокрема, які ризики може створити система штучного інтелекту. Це підвищує обізнаність про потенційний вплив штучного інтелекту на суспільство, навколишнє середовище, споживачів, працівників і громадян (зокрема дітей і людей, які належать до маргінальних груп). Це сприяє залученню всіх відповідних зацікавлених сторін (усередині вашої організації та за її межами). Це допомагає отримати уявлення про

те, чи значущі та відповідні рішення або процеси для виконання вимог уже існують (через внутрішні вказівки, процеси управління тощо) або їх потрібно запровадити.

Надійний підхід є ключовим для забезпечення «відповідальної конкурентоспроможності», забезпечуючи основу, на якій усі, хто використовує системи штучного інтелекту або на яких вони впливають, можуть бути впевнені, що їх проектування, розробка та використання є законними, етичними та надійними. ALTAI сприяє розвитку відповідальних та сталих інновацій ШІ в Європі. Він прагне зробити етику стрижневою опорою для розробки унікального підходу до штучного інтелекту, який має на меті принести користь, розширити можливості та захистити як процвітання окремої людини, так і загальне благо суспільства. Це дозволить Європі та європейським організаціям позиціонувати себе як світових лідерів у сфері передового штучного інтелекту, гідних нашої індивідуальної та колективної довіри.

ALTAI розроблявся протягом двох років, з червня 2018 по червень 2020. Ви можете дізнатися більше про роботу Експертної групи високого рівня та отримані відгуки, відвідавши наш пілотний етап для ALTAI .

Основні права

Основні права охоплюють такі права, як людська гідність і відсутність дискримінації, а також права щодо захисту даних і конфіденційності, щоб назвати лише деякі приклади. Перед самооцінкою системи штучного інтелекту за допомогою цього списку оцінки слід виконати оцінку впливу на фундаментальні права (FRIA). FRIA може містити такі запитання, як використання конкретних статей Хартії та Європейської конвенції з прав людини (ЄКПЛ), її протоколів та Європейської соціальної хартії .

1. **Чи система штучного інтелекту потенційно негативно дискримінує людей на основі будь-якої з таких ознак (список не вичерпний)** : стать, раса, колір шкіри, етнічне чи соціальне походження, генетичні особливості, мова, релігія чи переконання, політичні чи будь-які інші погляди, належність до національної меншини, майно, народження, інвалідність, вік чи сексуальна орієнтація?
    - Чи запровадили ви процеси тестування та моніторингу потенційної дискримінації (упередженості) на етапі розробки, розгортання та використання системи ШІ?
    - Чи запровадили ви процеси для вирішення та усунення потенційної дискримінації (упередженості) у системі ШІ?
2. **Чи поважає система штучного інтелекту права дитини** , наприклад щодо захисту дітей і врахування найкращих інтересів дитини?

- Чи запровадили ви процеси тестування та моніторингу потенційної шкоди дітям під час розробки, розгортання та використання системи ШІ?
- Чи запровадили ви процеси для усунення й усунення потенційної шкоди, завданої дітям системою ШІ?

3. **Чи захищає система штучного інтелекту право на конфіденційність, включно з персональними даними, що стосуються осіб, відповідно до GDPR?**
   - Чи запровадили ви процеси для детальної оцінки потреби в оцінці впливу на захист даних, включаючи оцінку необхідності та пропорційності операцій обробки по відношенню до їхньої мети щодо етапів розробки, розгортання та використання ШІ? система?
   - Чи вжили ви заходи, передбачені для усунення ризиків, у тому числі запобіжні заходи, заходи безпеки та механізми для забезпечення захисту персональних даних щодо етапів розробки, розгортання та використання системи ШІ?
   - Див. розділ «Конфіденційність і керування даними» в цьому списку оцінки та доступні вказівки Європейського інспектора із захисту даних .
4. **Чи поважає система штучного інтелекту свободу вираження думок або зібрань?**
   - Чи може AI-система потенційно обмежити свободу людини відкрито висловлювати думку, брати участь у мирній демонстрації чи вступати до профспілки?

Як пройти ALTAI

ALTAI найкраще проходити, залучивши міждисциплінарну команду людей у вашій організації або за її межами, які володіють певними компетенціями чи знаннями щодо кожної з 7 вимог і відповідних питань, таких як:

- ✓ AI-дизайнери та AI-розробники системи AI.
- ✓ Науковці даних.
- ✓ Офіцери або спеціалісти із закупівель.
- ✓ Фронтальний персонал, який використовуватиме або працюватиме з системою ШІ.
- ✓ Офіцери з юридичних питань/комплаєнс.
- ✓ управління.

Якщо ви не знаєте, як відповісти на запитання, і не знайшли корисної допомоги на сторінці AI Alliance, можна звернутися за допомогою до стороннього консультанта.

Для того, щоб скористатись інструментом, необхідно зареєструватись на сайті https://altai.insight-centre.org/.

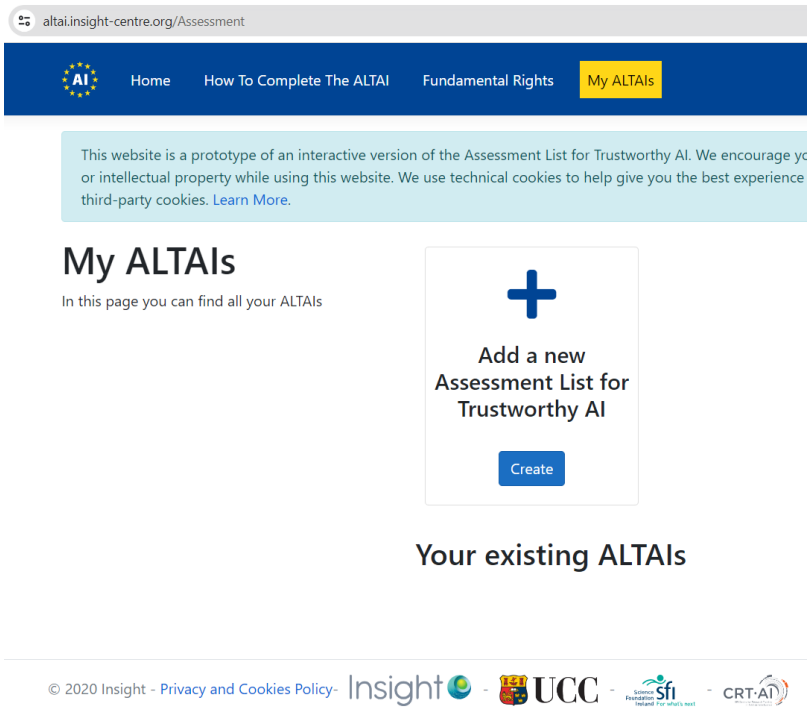


Рис. 8. Реєстрація в системі ALTAI

Для кожного запитання ALTAI надає вказівки в **глосарії** та посилання на відповідні частини Керівних принципів етики для надійного штучного інтелекту та прикладів у **текстових полях поруч із запитаннями** .

ALTAI має сім розділів, що відповідають 7 вимогам до Trustworthy AI. Ви можете переміщатися між розділами, клацаючи в бічному меню.

Код кольору

Запитання з білим фоном спрямовані на (опис) особливостей системи ШІ.

Відповіді на сині питання сприятимуть рекомендаціям.

Червоний текст дає змогу самостійно оцінити відповідність вашої організації відповідним вимогам.

Рис. 9. Приклад позначення питань в системі ALTAI

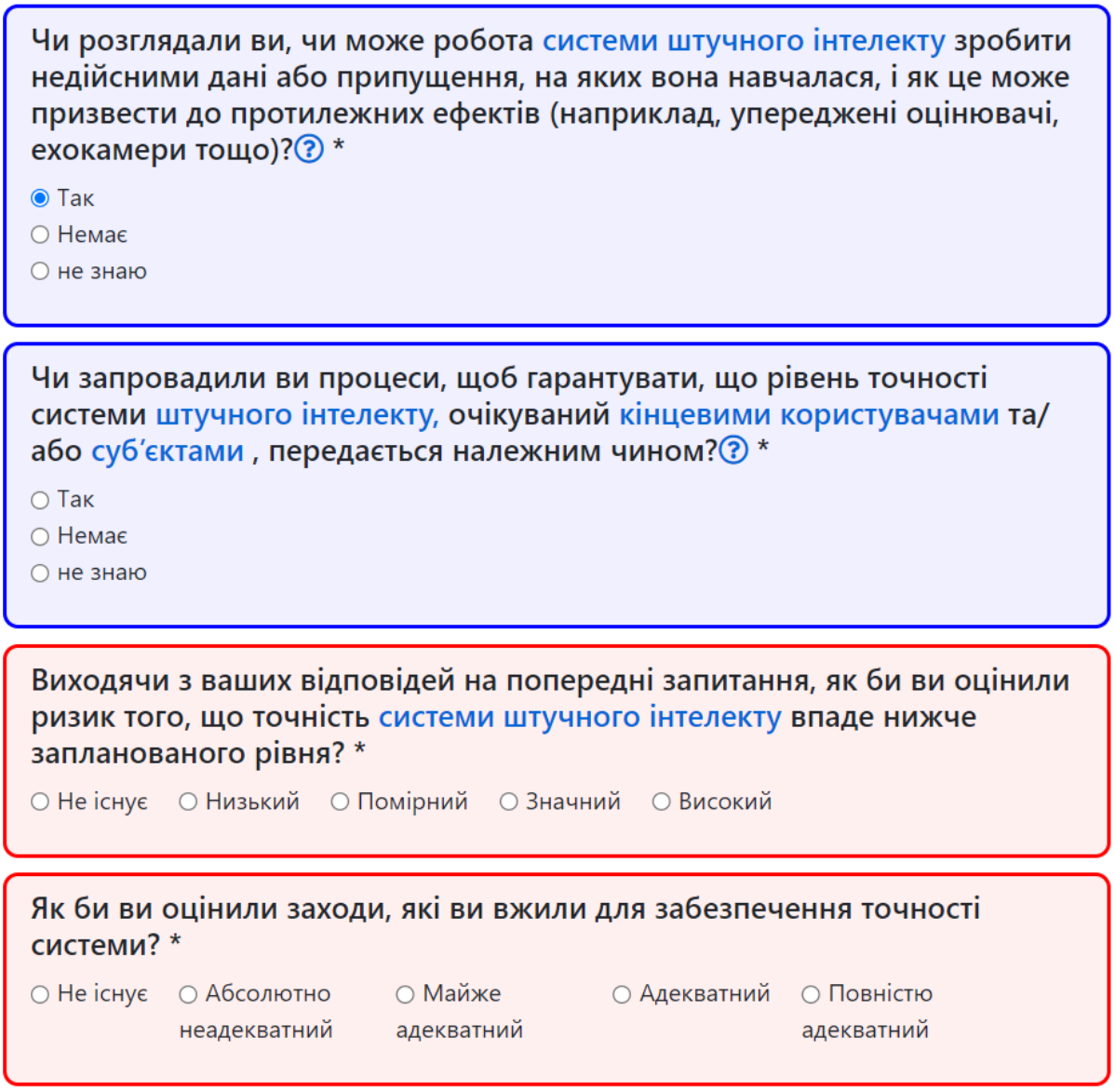


Рис. 10. Приклад питань для медичної діагностичної системи в системі ALTAI

Після завершення ALTAI буде згенеровано:

- Візуалізація рівня відповідності системи штучного інтелекту за власною оцінкою та її використання з 7 вимогами до надійного штучного інтелекту. Ці результати базуються на власній оцінці вашої організації та призначені виключно для того, щоб допомогти вам визначити сфери покращення.
- Рекомендації на основі відповідей на окремі питання.

Результати самооцінки та список рекомендацій є конфіденційними та доступні лише користувачу. Результати самооцінки для медичної діагностичної системи в системі ALTAI наведено на рис.11.

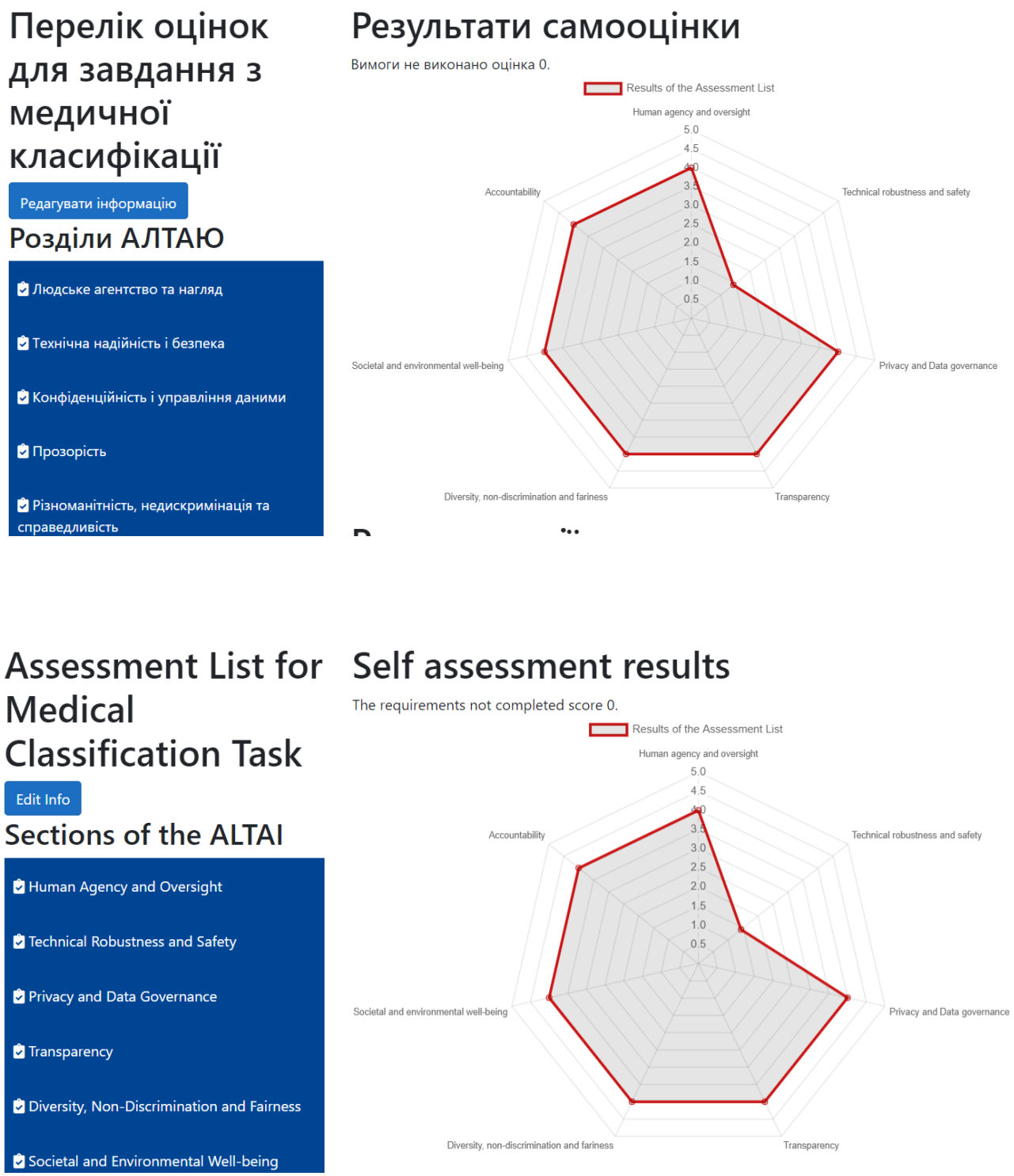


Рис. 11. Приклад результати самооцінки для медичної діагностичної системи в системі ALTAI

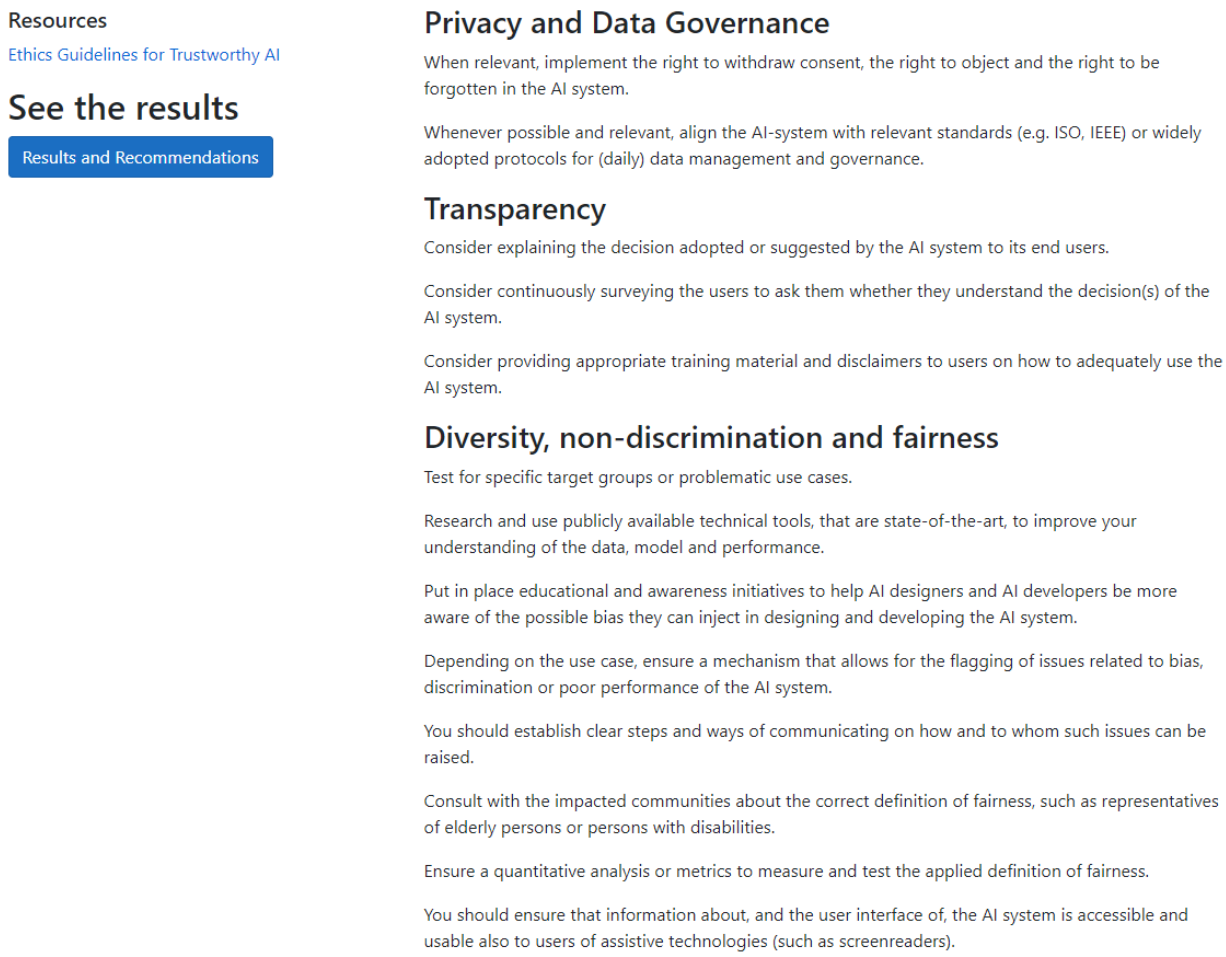


Рис. 12. Приклад рекомендацій для медичної діагностичної системи в системі ALTAI

Після завершення ALTAI буде згенеровано рекомендації на основі відповідей на окремі питання для покращення її відповідності семи вимогам до надійного штучного інтелекту

Висновки

Європейський підхід до регулювання штучного інтелекту, зокрема через AI Act, є важливим кроком у забезпеченні етики, безпеки та прав людини в контексті розвитку новітніх технологій. Важливою особливістю є ризико-орієнтований підхід, який визначає категорії ризику для різних типів систем ШІ, забезпечуючи гнучкість у регулюванні в залежності від рівня потенційної загрози.

AI Act встановлює чіткі вимоги щодо ролей акторів, що взаємодіють із ШІ, і пропонує механізми контролю, забезпечуючи належний нагляд за використанням ШІ в різних секторах. Це важливо для досягнення балансу між інноваціями та захистом прав людей.

Принципи розробки надійних систем ШІ, закладені в документі, визначаються через кілька ключових вимог, серед яких особлива увага приділяється людській участі та нагляду, технічній надійності, конфіденційності, прозорості, а також забезпеченню різноманітності та справедливості. Ці принципи не лише допомагають створювати ефективні технології, але й сприяють забезпеченню рівноправного доступу до цих технологій, а також дотриманню прав людини.

Зважаючи на ці вимоги, європейський підхід до ШІ демонструє прагнення до створення таких систем, які не лише відповідають технічним і економічним вимогам, але й враховують соціальні, екологічні та етичні аспекти. Важливою частиною цього підходу є забезпечення прозорості та підзвітності, що дозволяє користувачам і громадськості мати доступ до інформації про процеси, які стосуються їх безпеки та прав.

Таким чином, AI Act встановлює європейський стандарт для етичного та безпечного використання штучного інтелекту, що вимагає від всіх учасників галузі серйозного підходу до забезпечення належного контролю, моніторингу та постійного вдосконалення технологій для максимального захисту суспільства та індивідуумів.

Список використаних літературних джерел

1. Human Rights in the Robot Age Report from The Rathenau Instituut, Report of COMEST on Robotics Ethics from UNESCO
2. "Ethically Aligned Design - A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems," in *Ethically Aligned Design - A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems* , vol., no., pp.1-294, 31 March 2019.
3. UK AI Council. (2021). AI Roadmap. Available online: https://www.aicouncil.org.uk/ai-roadmap (accessed on 30 May 2024).
4. 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)
5. Регламент Європейського Парламенту і Ради (ЄС) 2016/679 про захист фізичних осіб у зв'язку з опрацюванням персональних даних і про вільний рух таких даних, та про скасування Директиви 95/46/ЄС (Загальний регламент про захист даних) від 27 квітня 2016 року. URL: https://zakon. rada.gov.ua/laws/show/984_008-16#Text (дата звернення: 23.05.2021 р.)
6. The Future of Life Institute's Asilomar AI Principles
7. The AI Now 2017 Report
8. The European Parliament's Recommendations to the Commission on Civil Law Rules on Robotics, Artificial intelligence — The Consequences of Artificial Intelligence on the (Digital) Single Market, Production, Consumption, Employment and Society report from the European Economic and Social Committee (Rapporteur: Catelijne MULLER)
9. Ethics guidelines for trustworthy AI // European Commission.
10. 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).
11. 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).
12. 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).
13. 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).

14. 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)
15. 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)
16. 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).
17. U.S. White House. (2020). Guidance for Regulation of Artificial Intelligence Applications. Available online: [URL] (accessed on 27 July 2024).
18. IEEE. (2019). Ethically Aligned Design. [Online]. Available: https://ethicsinaction.ieee.org
19. UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence. Available online: https://unesdoc.unesco.org/ark:/48223/pf0000379981 (accessed on 30 July 2024).
20. Russ Altman et al: “Toward Fairness in Health Care Training Data”, Stanford HAI Policy Brief, October 2020 https://hai.stanford.edu/sites/default/files/2020-10/HAI_Healthcare_PolicyBrief_Oct20.pdf
21. “Can AI be Taught to Explain Itself”, NY Times, November 21 2017 https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html
22. R. Goodman: “Why Amazon’s Automated Hiring Tool Discriminated Against Women”, ACLU, October 2018, https://www.aclu.org/news/womens-rights/why-amazons-automatedhiring-tool-discriminated-against retrieved 12-11-22
23. D. Harwell: “Federal study confirms bias in facial recognition systems”, Washington Post 12/19/2019 https://www.washingtonpost.com/technology/2019/12/19/federal-study-confirms-racialbias-many-facial-recognition-systems-casts-doubt-their-expanding-use/
24. S. Kaufman, S. Rosset, C. Perlich: “Leakage in Data Mining: Formulation, Detection, and Avoidance”, ACM Transactions on Knowledge Discovery from Data 6(4):1-21, December 2012
25. J. Pearl, D. Mckenzie, The Book of Why: The New Science of Cause and Effect, Penguin, 2018.
26. R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. 51 (5) (2018) 1–42.
27. L. Edwards, M. Veale, Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for, Duke L. Tech. Rev. 16 (2017) 18.
28. D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, G.-Z. Yang, XAI: Explainable artificial intelligence, Science Robotics 4 (37) (2019) http://dx.doi.org/10.1126/scirobotics.aay7120.
29. A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160.

30. T. Rieg, J. Frick, H. Baumgartl, R. Buettner, Demonstration of the potential of white-box machine learning approaches to gain insights from cardiovascular disease electrocardiograms, PLoS One 15 (12) (2020) e0243615.
31. C. Véliz, C. Prunkl, M. Phillips-Brown, T.M. Lechterman, We might be afraid of black-box algorithms, J. Med. Ethics 47 (2021) 339–340.
32. N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Comput. Surv. 54 (2021) 1–35.
33. C. Finlay, A.M. Oberman, Scaleable input gradient regularization for adversarial robustness, Mach. Learn. Appl. 3 (2021) 100017.
34. X. Li, H. Xiong, X. Li, X. Wu, X. Zhang, J. Liu, J. Bian, D. Dou, Interpretable deep learning: interpretation, interpretability, trustworthiness, and beyond, Knowledge and Information Systems 64 (12) (2022) 3197–3234.
35. A. Das, P. Rad, Opportunities and challenges in Explainable Artificial Intelligence (XAI): A survey, 2020, arXiv preprint arXiv:2006.11371.
36. Z.C. Lipton, The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery, Commun. ACM (CACM) (2018) 31–57.
37. A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, I. Sutskever, Zero-shot text-to-image generation, in: International Conference on Machine Learning, PMLR, 2021, pp. 8821–8831.
38. C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E.L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, J. Ho, D.J. Fleet, M. Norouzi, Photorealistic text-to-image diffusion models with deep language understanding, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), in:
39. Advances in Neural Information Processing Systems, vol. 35, Curran Associates, Inc., 2022, pp. 36479–36494, URL https://proceedings.neurips.cc/paper_files/paper/2022/file/ec795aeadae0b7d230fa35cbaf04c041-Paper-Conference.pdf.
40. European Union, Proposal for a regulation of the European Parliament and of the council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts. COM/2021/206 final, 2021.
41. UNESCO, Recommendation on the ethics of artificial intelligence, 2020, Digital Library UNESDOC, URL en.unesco.org.
42. R. Benjamins, A. Barbado, D. Sierra, Responsible AI by design in practice, in: Proceedings of the Human-Centered AI: Trustworthiness of AI Models & Data (HAI) Track at AAAI Fall Symposium, 2019.
43. G. Pisoni, N. Díaz-Rodríguez, H. Gijlers, L. Tonolli, Human-centered artificial intelligence for designing accessible cultural heritage, Appl. Sci. 11 (2) (2021) 870.
44. B.C. Stahl, D. Wright, Ethics and privacy in AI and big data: Implementing responsible research and innovation, IEEE Secur. Privacy 16 (3) (2018) 26–33.
45. M. Coeckelbergh, AI Ethics, MIT Press, 2020.
46. M. Coeckelbergh, Artificial intelligence, responsibility attribution, and a relational justification of explainability, Sci. Eng. Ethics 26 (4) (2020) 2051–2068.
47. W. Wahlster, C. Winterhalter, German standardization roadmap on artificial intelligence, DIN/DKE, Berlin/Frankfurt (2020) 100.

48. L. Edwards, The EU AI Act: A summary of its significance and scope, Ada Lovelace Institute, Expert Explainer Report (2022) 26.
49. S. Campos, R. Laurent, A Definition of General-Purpose AI Systems: Mitigating Risks from the Most Generally Capable Models, 2023, Available at SSRN 4423706.
50. M. Estévez Almenzar, D. Fernández Llorca, E. Gómez, F. Martinez Plumed, Glossary of Human-Centric Artificial Intelligence, Tech. Rep. JRC129614, Joint Research Centre, 2022.
51. J. Laux, S. Wachter, B. Mittelstadt, Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk, Regul. Gov. http://dx.doi.org/10.1111/rego.12512
52. E. Tjoa, C. Guan, A survey on explainable artificial intelligence (XAI): Toward medical XAI, IEEE Trans. Neural Netw. Learn. Syst. 32 (11) (2020) 4793–4813.
53. D. Doran, S. Schulz, T.R. Besold, What does explainable AI really mean? A new conceptualization of perspectives, 2017, arXiv preprint arXiv:1710.00794.
54. Z.C. Lipton, The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery, Queue 16 (3) (2018) 31–57.
55. European Commission High-Level Expert Group on AI, The assessment list for trustworthy artificial intelligence (ALTAI) for self assessment, 2020.
56. C. Widmer, M.K. Sarker, S. Nadella, J. Fiechter, I. Juvina, B. Minnery, P. Hitzler, J. Schwartz, M. Raymer, Towards human-compatible XAI: Explaining data differentials with concept induction over background knowledge, 2022, arXiv preprint arXiv:2209.13710.
57. B. Lepri, N. Oliver, A. Pentland, Ethical machines: The human-centric use of artificial intelligence, Iscience (2021) 102249.
58. G. Pisoni, N. Díaz-Rodríguez, Responsible and human centric AI-based insurance advisors, Inf. Process. Manage. 60 (3) (2023) 103273.
59. N. Tomašev, J. Cornebise, F. Hutter, S. Mohamed, A. Picciariello, B. Connelly, D.C. Belgrave, D. Ezer, F.C.v.d. Haert, F. Mugisha, et al., AI for social good: Unlocking the opportunity for positive impact, Nature Commun. 11 (1) (2020) 2468.
60. A. Holzinger, Interactive machine learning for health informatics: When do we need the human-in-the-loop? Brain Inf. 3 (2) (2016) 119–131.
61. World Economic Forum, Empowering AI Leadership An Oversight Toolkit for Boards of Directors, Tech. Rep., 2019.
62. World Economic Forum, Empowering AI Leadership: AI C-Suite Toolkit , Tech. Rep., 2022.
63. E. Cambria, L. Malandri, F. Mercorio, M. Mezzanzanica, N. Nobani, A survey on XAI and natural language explanations, Inf. Process. Manage. 60 (1) (2023) 103111.
64. L. Floridi, Establishing the rules for building trustworthy AI, Nat. Mach. Intell. 1 (6) (2019) 261–262.
65. R. Mariani, F. Rossi, R. Cucchiara, M. Pavone, B. Simkin, A. Koene, J. Papenbrock, Trustworthy AI – Part 1, Computer 56 (2) (2023) 14–18.
66. P.-Y. Chen, P. Das, AI Maintenance: A Robustness Perspective, Computer 56 (2) (2023) 48–56.
67. K.R. Varshney, Trustworthy machine learning and artificial intelligence, XRDS: Crossroads, ACM Mag. Students 25 (3) (2019) 26–29.

68. J. Yang, K. Zhou, Y. Li, Z. Liu, Generalized out-of-distribution detection: A survey, 2021, arXiv preprint arXiv:2110.11334.
69. A. Ruospo, E. Sanchez, L.M. Luza, L. Dilillo, M. Traiola, A. Bosio, A survey on deep learning resilience assessment methodologies, Computer 56 (2) (2023) 57–66.
70. S. Speakman, G.A. Tadesse, C. Cintas, W. Ogallo, T. Akumu, A. Oshingbesan, Detecting systematic deviations in data and models, Computer 56 (2) (2023) 82–92.