Київський національний університет імені Тараса Шевченка
Філософський факультет
Кафедра теоретичної та практичної філософії

Соціальні основи штучного інтелекту (філософська оцінка)
Social bases of artificial intelligence (philosophical assessment)

Кваліфікаційна робота за освітньо-професійною програмою «Філософія»
за спеціальністю 033 Філософія
на здобуття освітнього рівня бакалавра філософії

Студент-виконавець:
Майборода Валерія Андріївна
_____
IV курс, 2 група, ОР – Бакалавр

Науковий керівник:
Бойченко Михайло Іванович
Доктор філософських наук

_____

Допущено до захисту:
на засіданні кафедри теоретичної і практичної філософії
протокол №____ від _____2022 р.

Зав. кафедри теоретичної і практичної філософії,
доктор філософських наук, професор
Шашкова Людмила Олексіївна_____

Київ - 2022

ЗМІСТ

Вступ

Актуальність теми дослідження. На сьогоднішній день людство активно починає впроваджувати технології штучного інтелекта. Не вперше і точно не в останній раз лунають сміливі та далекосяжні заяви про те, як нові технології змінять світ. Дослідження штучного інтелекту знаходиться на межі міждисциплінарності, його історія нерозривно пов'язана з когнітивними науками та філософією. Впродовж формування концепту штучного інтелекту, філософія була одним із вирішальним фактором його розвитку. У 21 сторіччі світ починає ділитись на дві категорії: штучного та природного, тож для людства стають нагальними такі питання, як моральні, соціальні та філософські базиси штучних витворів. Починаючи з середини минулого століття, надалі частіше виникають питання щодо того, чи може машина мати розум, тож це питання є не лише суто технічним, а й філософським.

Штучний інтелект і філософія мають більше спільного, ніж наука зазвичай має з філософією наук. Це тому, що штучний інтелект на людському рівні вимагає оснащення комп'ютерної програми деякими філософськими поглядами, особливо епістемологічними. Якщо програма має розуміти, що вона може, а що не може робити, її інженерам знадобиться підіймати питання свободи волі. Якщо програма має бути захищена від виконання неетичних дій, її розробникам доведеться сформувати позицію щодо того, що є етичним, а що ні.

Дослідження штучного інтелекту в контексті філософії не обмежується лише питанням: "Чи може машина мислити?" – це також питання щодо природи та поняття мислення, щодо впливу штучного інтелекту на соціум та визначення його меж у взаємодії з соціумом. На перший погляд ці питання

можуть здатись нескладними, але вони потребують невідкладного глибокого філософського осмислення.

Актуальність історико-понятійного дослідження філософських понять “мислення” та “штучний інтелект” підкріплена тим, що існує потреба прояснити та зрозуміти значення термінів, які ми вживаємо на постійній основі та якими описуємо реальність сьогодення.

Ця робота пропонує міждисциплінарний підхід до цифрового майбутнього у трьох частинах. По-перше, вона спирається на теорію суспільних наук, щоб встановити соціотехнічний підхід до майбутнього. Я перконана в тому, що те, як ми думаємо про майбутнє, в минулому і сьогоденні, має вирішальне значення для того, як виникне конкретне майбутнє. По-друге, у роботі розглядаються як ми можемо підійти до майбутнього штучного інтелекту по-різному, щоб кинути виклик детермінізму за допомогою спекулятивного проектування, інклюзивного розвитку потенціалу та громадського діалогу. По-третє, викладається філософське осмислення та історична реконструкція понять “мислення” та “штучний інтелект”. Ця робота є поєднанням опису явищ в реальному часі та її філософського осмислення – тому, на мою думку, вона є актуальною.

Об’єктом дослідження є сфера соціального застосування штучного інтелекту.

Предметом дослідження є безпосередньо дослідження соціальних, моральних та філософських передумов створення та використання штучного інтелекту.

Метою цієї роботи є з’ясування соціально-філософського підґрунтя створення, осмислення та застосування штучного інтелекту.

Відповідно до мети основними завданнями цього дослідження є:

- Надання характеристики соціальних базисів штучного інтелекту як напряму філософських досліджень.
- Відтворити реконструкцію розуміння понять “мислення” та “штучний інтелект” на прикладі робіт провідних сучасних дослідників штучного інтелекту.
- Створити цілісний погляд та усвідомити значення філософського переосмислення штучного інтелекту.

Основними методами, що були використані у проведенні дослідження є: системний підхід, абстрагування, узагальнення.

Структуру кваліфікаційної роботи складають вступ, три розділи, в перших двох розділах містяться три підрозділи, у третьому розділі міститься два підрозділи, висновки та список використаних джерел. Загальний обсяг роботи становить 31 сторінку.

Розділ 1. Штучний інтелект як соціальний феномен

1.1 Проблема визначення штучного інтелекту

Визначення поняття штучного інтелекту – це саме по собі глибоке філософське питання, і спроби систематично відповісти на нього утворюють базиси штучного інтелекту, як багату тему для аналізу та обговорення. Тим не менш, попередню відповідь можна дати: штучний інтелект – це область, присвячена створенню артефактів, здатних відображати в контрольованому, добре зрозуміле середовищі та поведінку, яку ми розглядаємо протягом тривалого періоду часу, вважаючи її розумною, або, загальніше, поведінкою, яку ми вважаємо основою того, до чого вона спрямована мати розум [15]. Звичайно, ця «відповідь» породжує додаткові запитання, зокрема, що саме являє собою розумну поведінку, що означає мати розум і як люди насправді керують розумом. Останнє питання є емпіричним, тож це скоріше стосується психології та когнітивної науки. Однак подібні питання є особливо доречними, адже будь-яке розуміння людської думки може допомогти людству створити машини, які працюють подібним чином. Дійсно, штучний інтелект і когнітивні науки розвивалася паралельними й нерозривно переплетеними шляхами: їхні історії йдуть поруч. Друге питання, те, що запитує, що є ознакою розумового, є філософським. Штучний інтелект надав йому значної актуальності, і ми побачимо, що філософські роздуми над цим питанням вплинули на хід створення та переосмислення самого поняття штучного інтелекту. Нарешті, звертаючись до першого запропонованого питання, традиційно проблемним було точно визначити, що вважати розумною поведінкою. Згодом були сформовані певні поведінкові тести, успішне проходження яких означало б наявність інтелекту. Найвідомішим з них є те, що стало відомим як тест

Тьюринга (ТТ) Аланом Тьюрингом в 1950 [24]. У ТТ жінка і комп'ютер перебувають у закритих приміщеннях, а людина (суддя), не знаючи, яка з двох кімнат містить якого учасника, задає питання електронною поштою (насправді, телетайпом, якщо використовувати оригінальний термін) з двох. Якщо на основі отриманих відповідей, суддя може зробити не менше ніж 50% помилок при винесенні вердикту щодо того, в якій кімнаті який гравець, ми говоримо, що комп'ютер, про який йде мова, пройшов тест Тьюринга. За словами Тьюринга, комп'ютер здатний для проходження ТТ слід оголосити мислячою машиною. Його твердження було суперечливим, одним з його критиків був Нед Блок, який стверджував, що не тільки поведінка організму є визначною у питанні, чи він є розумним. Ми також повинні розглянути, яким чином у організму з'являється інтелект. Тобто необхідно враховувати внутрішню функціональну організацію системи.

Інша критика ТТ полягає в тому, що він нереалістичний і, можливо, навіть перешкоджає прогресу штучного інтелекта. Тут необхідно розглянути аргументи Чаряка та Макдермотта, адже вони пропонують свою відповідь на питання "що таке штучний інтелект?". Штучний інтелект – це галузь, присвячена будівництву особи (або ж персоналії). Як написали Чарняк і Макдермотт у своєму класичному вступі до штучного інтелекту: "Кінцева мета штучного інтелекту, до досягнення якої ми далекі, — створити людину або, будучи більш скромними, тварину" [13, с. 7].

Чарняк і Макдермотт кажуть, що кінцевою метою не є створення чогось такого, що має людську подобу – їх розуміння штучного інтелекту — це так званий сильний штучний інтелект. Тут варто згадати за слова Хаугеланда: "Фундаментальна мета [досліджень штучного інтелекту] полягає не тільки в тому, щоб імітувати інтелект або зробити його подобу. Зовсім ні. Штучний інтелект хоче має бути справжнім: машини з розумом, у повному обсязі й

буквальному сенсі. Це не наукова фантастика, а справжня наука, заснована на глибокій теоретичній концепції; це сміливо: а саме той факт, що ми самі є комп'ютерами [15, с. 2].

Ця «теоретична концепція» людського розуму як комп'ютера послужила основою більшості на сьогоднішній день дослідження сильного штучного інтелекту. Вона стала відома як обчислювальна теорія розуму.

1.2 Сучасне розуміння штучного інтелекту

Класичні теоретичні дебати щодо штучного інтелекту зосереджені на питаннях, чи можливий штучний інтелект взагалі (часто кажучи як «чи можуть машини думати?») або чи можуть вони вирішити певні проблеми («Чи може машина робити х?»). Тим часом технічні системи штучного інтелекту почали масово прогресували і зараз є присутніми в багатьох аспектах нашого середовища [20]. Незважаючи на цей розвиток, є відчуття, що класична концепція штучного інтелекту є за своєю суттю дещо обмеженою, і її необхідно наново переосмислювати іншими методами, такими як: нейронними мережами, когнітивними науками, статистичними методами, універсальними алгоритмами, біологією та нейронауками тощо.

Штучний інтелект є унікальним серед інженерних предметів у тому, що він піднімає основні питання про природу обчислень, сприйняття, міркування, навчання, мови, дії, свідомістості, людства, життя тощо – і в той же час він вносить значний внесок у відповідь на ці запитання. Зараз ми знаходимось на тому етапі історичного розвитку, коли ми можемо більш чітко побачити існуючі альтернативи у вивченні штучного інтелекту [1].

З огляду на те, що ми зараз знаходимося, зв'язок між штучним інтелектом та когнітивною наукою потребує підлягає повторному обговоренню – у більшому масштабі це означає, що співвідношення між технічними продуктами та людьми повторно узгоджується. Від того, як ми

дивимося на перспективи штучного інтелекту, залежить як ми ставимося до себе і як ми дивимося на технічні продукти, які ми виробляємо: це також є однією з причин, чому теорія і філософія штучного інтелекту повинні розглядатися у перспективі різноманітності питань: від людського пізнання та аж до технічного функціонування.

1.3 Соціальний вимір штучного інтелекту

Сучасні вчені в галузі штучного інтелекту визнають існування соціального виміру та використовують кілька стратегій, щоб пояснити їх в рамках штучного інтелекту. Типовий підхід практикуючого інженера штучного інтелекту включає в себе розширений збір даних про виміри соціальних світів, наприклад, коли включається інформація про демографію пацієнтів (раса, етнічна приналежність та стать) та інші соціальні детермінанти здоров'я (житловий статус, стабільність харчування, соціальна підтримка). Проте ці зусилля практиків штучного інтелекту зазвичай не визнають всієї складності соціального життя, особливо коли вони суперечать критичним ідеям, запропонованим соціологами. Фахівці, які практикують штучний інтелект, схильні включати соціальне таким чином, щоб людські дані позиціонувалися як поодинокі, прямі та стабільні, абсолютно не враховуючи наявність таких явищ, як динаміка влади, системна дискримінація та соціальна нерівність. Підходи до науки про дані на регулярній основі не враховують втілену, взаємодіючу природу відмінності та її змінне значення в часі та місці. Більше того, варто зазначити, що невелика кількість експертів не можуть давати адекватну оцінку складності соціальних світів, що складаються з різноманітності професій та соціальних ієрархій. Незважаючи на цілеспрямовані зусилля, спрямовані на включення знань про соціальні світи в соціотехнічні системи, інженери штучного інтелекту продовжують демонструвати обмежене розуміння соціального, віддаючи пріоритет тому, що

може бути корисним для виконання завдань інженерії штучного інтелекту, але спрощуючи складність і існування соціальних світів.

Поява штучного інтелекту дає нам можливість запитати, як соціотехнічні зміни можуть допомогти створити кращий і більш справедливий світ. Спираючись на нашу глибоку взаємодію з соціальною структурою та нерівністю, соціологи можуть розвивати теоретичні та емпіричні ідеї, щоб надавати суспільству раціональне бачення про соціотехнічне майбутнє – і про те, як воно може вкоренитися – на рівнях індивідів, установ і суспільств. Те, як люди розгортають та інтерпретують системи штучного інтелекту, варіюється в залежності від інституційного та організаційного контексту, в якому вони впроваджуються. Алгоритмічні моделі можуть бути розроблені для оптимізації ефективності чи справедливості. Цифровими платформами можуть володіти і керувати корпорації, які прагнуть отримати прибуток, міста або ж люди, які їх використовують. Але саме моральні норми будуть формувати та обмежувати застосування штучного інтелекту в різних соціальних сферах, а державна політика визначатиме, як окремі особи та суспільства справляються з наслідками його розробки та впровадження.

Висновок до розділу №1

У першому розділі був розглянутий термін “штучний інтелект” та були описані проблеми у його використанні. Також була описана зміна парадигми у дослідженні штучного інтелекту, а саме з класичного питання філософії свідомості: “Чи може машина думати?” – до розуміння, що штучний інтелект має бути описаний на межі дисциплінарності, із застосуванням методів різних наук. Також були описані та окреслені можливості застосування штучного інтелекту у соціології.

Розділ 2. Межі та можливості соціального конструювання штучного інтелекту: філософська оцінка

2.1 Межі штучного інтелекту з погляду історично-філософської реконструкції

Швидкий розвиток штучного інтелекту породив безліч важливих етичних дискусій, які стають все більш помітними. Чи має штучний інтелект моральний статус і який моральний статус він має, є дуже суперечливими питаннями. З одного боку, багато відомих вчених стверджують, що штучний інтелект, як і механізми загалом, не мають морального статусу. З іншого боку, є також люди, які стверджують, що, якщо штучний інтелект має перцептивні характеристики, тож він повинен мати моральний статус.

Визначення рівня свідомості штучного інтелекту може допомогти пояснити його можливий моральний статус. Деякі вчені вважають, що штучний інтелект не стане самосвідомим, і що сильний штучний інтелект ніколи не існуватиме. Картезіанський підхід заперечував би здатність штучного інтелекту набувати свідомість. За Декартом існує два дуже надійних критерії для відрізнення людей (користуючись сучасною термінологією), від штучного інтелекту.

По-перше, він ніколи не зміг би використовувати слова чи інші сконструйовані знаки, як ми робимо, щоб виказати свої думки іншим. По-друге, хоча він міг би робити багато речей так само добре, як будь-хто з нас або навіть краще, він безпомилково зазнавав б невдачі в інших, виявляючи, що він діяв не з знання, а лише з розпорядження своїх органів.

Матеріальний світ Декарта – це світ, заснований на механічному погляді. Таким чином, нелюдську поведінку можна пояснити чисто механічними властивостями, які не вимагають наявності свідомості. Декартівський підхід

втілює принцип простоти, тобто ми повинні прагнути описати поведінку штучного інтелекту з найпростішим можливим поясненням.

Згадуючи про принцип простоти, також доречним буде згадати про принцип бритви Оккама, сформульований філософом Вільямом з Оккама, який стверджує, що сутності не повинні примножуватися без необхідності. Сучасною версією бритви Оккама в психології є «Канон Моргана», який стверджує, що поведінку штучного інтелекту можна пояснити без урахування внутрішньої свідомості. Цей принцип може застосовуватися і до людей. Проблема полягає в тому, що люди завжди демонструють складну та нову поведінку, яка є не простою реакцією на подразники, а результатом раціональної дедукції, отриманої з їхнього сприйняття світу. Більше того, люди мають мовну здатність висловлювати свої думки. Деякі програми штучного інтелекту, наприклад Siri, видають звуки. Проте ці звуки штучного інтелекту є просто механічно викликаною поведінкою, яка нагадує інших, а не вимовляє власну мову. Тільки люди можуть використовувати мову, щоб висловлювати свої думки [14].

Згідно з дуалізмом Декарта, матерія і розум є двома паралельними сутностями. Хоча всі люди тісно пов’язані зі своїми матеріальними тілами, вони не лише їхні тіла. Люди об’єднані в своїх душах або в нематеріальних сутностях, породжених людьми. Декарт вважав, що нематеріальні сутності можуть пояснити складність людської поведінки та мови. Штучному інтелекту для своєї поведінки не потрібні такі сутності – це більше схоже на рухому машину, а не на багатий розум.

2.2 Загрози штучного інтелекту

Підіймаючи питання про етичний статус штучного інтелекту, варто згадати про філософські аргументи загрози штучного інтелекту для суспільства. Пропоную розглянути, як на це питання відповідає Джеймс

Баррат у книзі “Останній винахід людства” [27]. Ідея Баррата полягає в тому, що інтелект можна вважати ключовою рисою, яка відрізняє людей від інших видів тварин. Ми, очевидно, не найсильніші звірі в джунглях, але завдяки нашому розуму ми вийшли на перше місце у природі. Тепер нашому домінанту загрожують створені нами створіння. Програмісти можуть розробляти штучний інтелект з таким рівнем інтелекту, що здатний перевищувати людський. Він може стати настільки потужним, що або вирішить всі наші проблеми, або вб’є нас усіх, залежно від того, як він розроблений.

Це підводить нас до області, де перетинається філософія з наукою: ми маємо шукати можливі вирішення цих питань з обох боків. Існує два основних типи опцій для захисту від штучного інтелекту. По-перше, ми можемо створити його безпечним. Здається, це величезний філософський і технічний виклик, але якщо це вдасться, це може вирішити багато світових проблем [6].

По-друге, ми не можемо створювати небезпечний штучний інтелект. У книзі докладно обговорюється економічний і військовий тиск, який штовхає штучний інтелект вперед. Необхідно вживати заходів, щоб люди не створювали небезпечний штучний інтелект.

Що не є варіантом, так це чекати, поки штучний інтелект вийде з-під контролю, а потім спробувати розпочати кампанію «війни світів» проти надінтелектуального штучного інтелекту. Штучний інтелект був би занадто розумним і надто потужним, щоб ми мали шанс проти них. Натомість нам потрібно зробити це завчасно. Це ці аргументи не враховують філософський аспект створення сильного штучного інтелекту [7].

У книзі 2014 року «Суперінтелект» оксфордський філософ Нік Бостром описує загрозу від штучного інтелекту та пропонує кілька сценаріїв судного дня. Але він вказує на аргумент з філософії етики. Якщо штучний інтелект має почуття, він не зможе хотіти взагалі нічого робити, не кажучи вже про те, щоб

діяти всупереч інтересам людства та боротися з людським опором. Бажання необхідне для будь-якої незалежної дії. І в ту хвилину, коли штучний інтелект хоче чогось, він житиме у всесвіті з нагородами та покараннями, включаючи покарання від нас за погану поведінку. Щоб вижити у світі, де панують люди, штучному інтелекту доведеться розвинути людське моральне відчуття, що деякі речі правильні, а інші – неправильні [3].

Щоб оцінити щось, сутність штучного інтелекту повинна вміти щось відчувати. Точніше, він повинен мати можливість чогось хотіти. Сучасному програмному забезпеченню не вистачає такої можливості, і програмісти не знають, як цього досягти.

Гарною ілюстрацією цьому є програма Google AlphaGo, яка нещодавно обіграла чемпіона світу з гри Go. Інженери Google використовували стратегію глибокого навчання, передаючи нейронну мережу мільйони ігор, у які грають люди, таким чином, щоб мережа моделювала поведінку експертів. Зрештою, AlphaGo могла просто «подивитись» на гру, що триває, і передбачити, яка сторона переможе, взагалі не роблячи жодного попереднього пошуку. Але AlphaGo поняття не має, що він грає в гру – у нього немає відчуття, що він щось робить. Він не відчуває задоволення від перемоги, не відчуває жалю від програшу. У нього немає можливості сказати: «Я б не хотів сьогодні грати в Go. Натомість пограймо в шахи». По суті, це всього лише трильйони логічних елементів.

І наразі це питання переводиться вже у область біології. У своїй книзі «Бути ніким» німецький філософ Томас Метцінгер [18] пише, що біологія має перевагу над людськими обчисленнями, оскільки вона може використовувати водну хімію для створення вишукано складних систем. Для цього використовується вода, яка одночасно є середовищем і розчинником. Вода дозволяє молекулам, що несуть інформацію, існувати у суспензії та сприяє їх

взаємодії. Це дозволяє створити декілька паралельних і надзвичайно чутливих петель зворотного зв'язку в молекулярному масштабі. Комп'ютери навіть близько не мають такої складності.

Тепер, припустимо, люди винаходять роботів такої природи, і вони починають відчувати відчуття. Як тільки система обробки інформації має відчуття, вона може мати моральну інтуїцію. У своїй книзі "The Expanding Circle" [23] філософ із Прінстона Пітер Сінгер зазначає, що собаки, дельфіни та шимпанзе демонструють, можливо, моральну поведінку, наприклад таку як взаємність та альтруїзм. Це також можна спостерігати у щурів, які вирішили врятувати потопаючого супутника замість того, щоб їсти шоколад. Якщо моральні інтуїції надають придатність, і якщо організми можуть передати ці інтуїції наступникам, то можна сказати, що вид знаходиться на шляху до формування самої моралі.

Тепер ось вирішальний момент. Як тільки вид з моральною інтуїцією набуває здатності міркувати, його моральність має тенденцію покращуватися з часом. Сінгер називає цей процес «ескалатором міркувань». Припустимо, ми живемо в примітивному племені людей. Якщо я скажу комусь, що можу мати більше горіхів, ніж інший, той запитаєте, чому це так. Щоб відповісти, я повинен дати вам причину – і це не може бути просто «тому що». Погані причини в кінцевому підсумку призведуть до повстання або колапсу суспільства. Навіть собаки розуміють поняття справедливості і перестануть співпрацювати з людьми, які до них ставляться несправедливо.

І як тільки починають наводити причини, їх можна поставити під сумнів. Чому наприклад начальник отримує більше горіхів? Це справді справедливо? «Міркування за своєю суттю є експансіоністським» – пише Сінгер. «Воно прагне універсального застосування» [23].

Філософ Джон Смарт стверджує: «Якщо мораль та імунітет є процесами розвитку, якщо вони неминуче виникають у всіх розумних колективах як тип гри з позитивною сумою, вони також повинні зростати в силі та масштабі в міру зростання обчислювальних можливостей кожної цивілізації». Справді, Джон Смарт вважає, що, враховуючи обробну здатність машин, штучний інтелект насправді був би «значно більш відповідальними, регульованими та самообмеженими, ніж люди» [24].

Замість жахів аморального, шаленого штучного інтелекту, це майбутнє, на яке варто чекати з нетерпінням. Точніше кажучи, це майбутнє, яке варто дозволити еволюції творити.

2.3 Соціальні межі використання штучного інтелекту

Розглянувши соціальні загрози штучного інтелекту, варто перейти до соціальних меж використання штучного інтелекту. Одним із прикладом соціальних меж штучного інтелекту є дискусія щодо того, чи слід дозволити військовим використовувати штучний інтелект, і якщо так, то в якій мірі. Штучний інтелект потенційно може як збільшити, так і зменшити смертність. Інший приклад: якщо штучний інтелект коли-небудь осягає почуття та мислення, чи варто йому надавати права, подібні до тих, які надаються людям чи тваринам? Це може включати право на життя, свободу та свободу вираження поглядів. Якщо це взагалі станеться, штучний інтелект, який має розум, має бути наділений правами. Крім того, ми ніколи не можемо бути впевненими, що він має розум, навіть якщо він має. [2]

Варто зазначити, що на фоні постійних філософських дискусій, алгоритми штучного інтелекту відіграють все більшу роль у сучасному суспільстві, хоча зазвичай їх не називають «штучним інтелектом». У майбутньому стає все більш важливим розробляти алгоритми штучного інтелекту, які не тільки є потужними та масштабованими, але й прозорими для

перевірки – це одна із багатьох соціально важливих властивостей. Деякі проблеми машинної етики дуже схожі на інші проблеми, пов’язані із проектуванням машин. Це передбачає нові проблеми програмування, але не нові етичні проблеми. Але коли алгоритми штучного інтелекту беруть на себе когнітивну роботу з соціальними вимірами – когнітивними завданнями, які раніше виконували люди, алгоритм штучного інтелекту має успадковувати соціальні вимоги [8].

Прозорість – не єдина бажана риса штучного інтелекту. Також важливо, щоб алгоритми штучного інтелекту, які беруть на себе соціальні функції, були передбачуваними для тих, ким вони керують. Щоб зрозуміти важливість такої передбачуваності, розглянемо аналогію. Правові принципи зобов'язують суддів дотримуватися попередніх прецедентів, коли це можливо. Для інженера ця перевага прецеденту може здатися незрозумілою – навіщо прив’язувати майбутнє в минуле, коли технології постійно вдосконалюються? Але одна з найважливіших функцій правової системи – бути передбачуваною, щоб, наприклад, можна було складати контракти, знаючи, як вони будуть виконуватися. Тож робота юридичної системи має на меті не лише оптимізацію суспільства, але й забезпечення передбачуваного середовища, в якому громадяни можуть оптимізувати своє власне життя.

Нік Бостром зазначає, що існує ще один соціальний критерій створення та роботи зі штучним інтелектом, а саме той, що хтось має нести відповідальність за кінцеві результати роботи штучного інтелекту, важливим питанням є хто саме: “ще один важливий соціальний критерій для роботи з організацією штучного інтелекту – вміння знайти людину, відповідальну за те, щоб щось зробити. Коли система штучного інтелекту виходить з ладу через поставлене завдання, хто бере на себе провину? Програмісти? Чи може кінцеві користувачі?” [5].

Інший набір етично-соціальних питань виникає, коли ми розглядаємо можливість того, що деякі майбутні системи штучного інтелекту можуть бути кандидатами на отримання морального статусу. Наші стосунки з істотами, які мають моральний статус, не є виключно інструментальною раціональністю: у нас також є моральні причини ставитися до них певним чином і утримуватися від поводження з ними якось іншим чином. Френсіс Камм запропонувала таке визначення морального статусу: Х має моральний статус = оскільки Х має моральне значення сам по собі, тож допустимо/неприпустимо робити з ним речі заради нього самого [17].

Наприклад, скеля не має морального статусу: ми можемо подрібнити її або піддати будь-якому поводженню, яке нам подобається, не турбуючись про саму скелю. З іншого боку, до людини треба ставитися не лише як до засобу, а й як до мети. Що саме означає розглядати людину як мету, це та теза, з якою різні етичні теорії розходяться, але це, безумовно, передбачає врахування її законних інтересів – надання ваги її добробуту, і це також може включати в себе прийняття суворих моральних обмежень у наших стосунках з нею, таких як заборона вбивати її, красти у неї чи робити різноманітні інші речі до неї або її власності без її згоди [21]. Більш стисло це можна висловити сказавши, що людська особистість має моральний статус.

Загальноприйнято, що нинішні системи штучного інтелекту не мають морального статусу. Ми можемо змінювати, копіювати, припиняти, видаляти або використовувати комп’ютерні програми за своїм бажанням; принаймні, що стосується самих програм [4]. Моральні обмеження, яким ми піддаємося в наших стосунках із сучасними системами штучного інтелекту, ґрунтуються на наших обов’язках перед іншими істотами, перед іншими людьми, а не на будь-яких обов’язках перед самими системами.

Висновки до розділу №2

У цьому розділі було розглянуті питання можливості існування штучного інтелекту, застосовуючи історичну реконструкцію. Також були описані загрози штучного інтелекту, а саме філософський погляд на майбутню інтеграцію штучного інтелекту в соціум. Були також описані моральні межі та етичні питання щодо застосування штучного інтелекту у побуті.

Розділ 3. Штучний інтелект у контексті філософії свідомості

3.1 Фізикаліський підхід у філософії свідомості

Розглянувши етично-соціальні питання можливості штучного інтелекту, необхідно перейти до найґрунтовнішого питання філософії свідомості та філософії штучного інтелекту: “Чи може бути наявна свідомість у штучного інетелекту?”.

Відповідаючи на це питання, варто згадати за існуючі до нього підходи. Умовно їх можна розділити на два напрямки. Перший з них – це фізикалізм.

Тож якщо розуміти штучний інтелект як повне відтворення інформації за допомогою людської мисленнєвої машини – то така система, мабуть, неможлива. Якщо говорити про комп’ютерне моделювання деяких інформаційних аспектів процесу мислення людини, то така система цілком можлива. Концепцію “сильного” штучного інтелекту в 20 столітті впроваджує американський філософ Джон Серл. У цьому сенсі штучний інтелект не тільки розглядається як модель розуму, але й насправді є таким розумом. Таким чином, передбачається, що принципової різниці між природним інтелектом (людина) і штучним інтелектом (машиною) немає [22].

Критерії виникнення штучного інтелекту дуже розпливчасті та умовні. У яких випадках ми можемо говорити, що ми фактично створили і вже використовуємо штучний інтелект, а в яких – що ми маємо лише складну комп’ютерну програму, але не створили сильний штучний інтелект? Є три альтернативні підходи.

Перший підхід: штучний інтелект виникає, коли поведінка обчислювальної машини не відрізняється від поведінки людини. У цьому підході критерієм є вищезгаданий тест Алана Тьюринга. У 1950 році Алан Тюрінг написав пророчу статтю про штучний інтелект [25], яка починається з фрази: "Розглянемо питання "Чи може машина думати?". Теоретично це

цілком можливо. Але це не означає, що машина має сильний інтелект і фактично починає мислити. Відомий розумовий експеримент Джона Серла (Китайська кімната) є хорошим аргументом того, що проходження тесту Тьюринга не є серйозним критерієм для того, щоб машина розвивала процес мислення, подібний до людського мислення [16].

Більше ніж півстоліття назад стало популярним порівнювати мозок та розум з комп'ютерами та програмами. Незважаючи на привабливість обчислювальної моделі розуму, може бути важко дати точне формулювання, що саме ми називаємо людською поведінкою. Справді, такі критики теста Тьюринга, як вже згаданий Джон Серл та Гіларі Патнем, стверджували, що будь-що, навіть камінь, можна розглядати як приклад будь-яких обчислень, а це означає, що твердження, що розум є комп'ютером є хибним.

Відповіддю на аргумент Серла і Патнема про те, що обчислювальна модель є порожньою, оскільки будь-яку довільну систему (наприклад, молекули в стіні) можна розглядати як приклад довільної комп'ютерної програми (наприклад, програма для обробки текстів) є такою, що аргумент ігнорує причинно-динамічний аспект будь-якого екземпляра програми. Фактично, тільки обережно конструйовані системи (а саме комп'ютери) матимуть контрфактичну поведінку, визначену обчислювальною моделлю. Ми можемо дати онтологічно надійний звіт, які системи є екземплярами певних програм, а які ні – цей звіт буде спиратися на ефективні ступені свободи цієї системи та динаміку, яка керує еволюцією цих ступенів свободи.

Фізика надає нам дуже вагомі докази істинності фізикалізму, і ця онтологія вимагає, щоб усі системи були механічними системами.

Загальноприйнято, що визначальною рисою фізикалізму є метафізична супервентність всіх фактів щодо фізичних фактів [11]. Таким чином, якщо світ ідентичний дійсному світ у всіх його фізичних фактах, він (якщо фізикалізм

істинний) повинен бути повністю ідентичним справжньому світу. Теза про метафізичну супервентність захоплює уявлення про те, що фізикалізм вимагає, щоб опис світу, який пропонує фізика, був повною, але це не означає, що аргументи інших наук (які не є сформульовані мовою фізики) є хибними – натомість такі розповіді можуть бути правдивими, якщо вони підтверджуються онтологічно більш фундаментальними фізичними фактами [12].

Вимога супервентності, на мою думку, цілком нормальна. Але вона має бути ближчою в поясненні, чому і як факти вищого рівня метафізично впливають на фізичні факти. Для цього нам потрібний більш істотний метафізичний каркас, який дозволить нам зрозуміти, як основні фізичні факти фіксують факти інших наук. Нам потрібен звіт про те, як властивості та сутності вищого рівня виникають із базових фізичних властивостей та сутностей. Існує форма емергентизму, яка несумісна з фізикалізмом, оскільки заперечує, що всі факти супроводжуються мікрофізичними фактами. Назвемо цю позицію сильниим емерджентизмом. Він стверджує, що поява спеціальних наукових властивостей передбачає вихід за межі області чисто фізичних властивостей і законів. Якщо емерджентні властивості є причинними, тоді виникнення певних явищ передбачає заперечення причинного замикання фізики. Слабкий емерджентизм, з іншого боку, сумісний з фізикалізмом і причинним закриттям фізики. Він просто стверджує, що складні системи мають певні особливості, які добре описані іншими науками [9].

Властивості вищого рівня не усуваються нашою фізикалістською онтологією, натомість вони є реальними складними ознаками фізичних систем.

3.2 Субстанційний дуалізм

Другий підхід до штучного інтелекту полягає у наступному: штучний інтелект – це коли машина має почуття, емоції, інтуїцію, креативність та інші специфічні якості людської свідомості. Іншими словами, це вже передбачає штучне відтворення не лише окремих інформаційних аспектів процесу мислення, а відтворення свідомості як такої. Але чи можна відтворити свідомість технічними засобами? Це вже питання субстанційного дуалізму.

Однією з основних проблем сфери штучного інтелекту є створення таких систем, які були б здатні до самосвідомості та усвідомлення навколишнього світу, самопізнання своїх внутрішніх станів та властивостей. По суті, проблема полягає в моделюванні такого штучного «Я», яке має рефлексію [10]. Під рефлексією я маю на увазі усвідомлене відображення змісту явищ суб'єктивної реальності. За допомогою рефлексії здійснюються самоідентифікація та спрямована активність «Я». Саме вона тягне за собою самореалізацію особистості, людина, вирішуючи свої життєві проблеми, схильна до відмови від усталеного образу свого «Я», прагне створення його рухливого і мінливого еквівалента, що розвивається з кожною новою життєвою ситуацією. Існуючи в інтервалі «тут і зараз», «Я» не усвідомлює і не споглядає себе повністю. З одного боку, це обумовлюється багатомірністю змістів. З іншого боку, важливим є те, що «Я» виявляється спрямованим у майбутнє, несе в собі потенції творчості як можливості нового «Я», що змінилося в результаті свого розвитку. "Я" завжди виходить за рамки "поточного сьогодення" [10].

Існуючі нині моделі штучного інтелекту, на жаль, поки що обмежені і недосконалі. Моделювання інтелектуальних систем, що мають рефлексію і самосвідомість, вимагає серйозних міждисциплінарних зусиль у сфері їх побудови. Хоча в науковому дискурсі можна зустріти роботи, де ментальні властивості приписуються комп'ютерам та іншим технічним пристроям, але

про адекватність таких концептуальних побудов говорити поки рано. Очевидно, що в проблематиці штучного інтелекту багато факторів є незрозумілими, є білі плями, які чекають на своє прояснення в майбутньому [12].

У сучасному розумінні поняття штучного інтелекту вводиться Джоном Маккарті. За Маккарті штучний інтелект – це наука та техніка для створення інтелектуальних машин, а особливо інтелектуальних комп'ютерних програм. Це пов'язано з завданням використання комп'ютерів для розуміння людського інтелекту, але штучний інтелект не повинен обмежуватися методами, що можуть бути спостереженими біологічно [19].

Під інтелектом Маккарті розуміє «обчислювальну» частину здатності досягати певної мети людиною. Різні види та ступеня інтелекту зустрічаються у людей, у багатьох тварин та деяких машин. На сьогоднішній день немає точного визначення поняття «інтелект», яке б не залежало від його зв'язку з людським інтелектом. Проблема полягає в тому, що ми поки що не можемо загалом охарактеризувати, які обчислювальні процедури ми можемо беззаперечно назвати інтелектом. Ми розуміємо деяку роботу механізмів інтелекту, але ж очевидно, що далеко не всі.

Питання «Чи розумна ця машина чи ні?» – неправильне питання. Інтелект включає в собі безліч працюючих механізмів, і дослідження штучного інтелекту наразі дійшло до тієї точки розвитку, що ми можемо змусити комп'ютери виконувати лише деякі з них. Якщо для виконання певних завдань потрібні добре зрозумілі на сьогодні механізми, то комп'ютерні програми можуть дати вражаючу продуктивність в цих завданнях. Такі програми слід вважати "розумними". Штучний інтелект не завжди симулює людський інтелект. З одного боку, ми можемо дізнатися про те, як зробити так, щоб машини вирішували проблеми, спостерігаючи за іншими людьми або

просто надати їй інструменти для спостереження нашими власними методами. З іншого боку, велика частина роботи пов'язана з вивченням проблем, які пов'язані з біологічним поглядом на інтелект, а не з вивченням ментальності людей або тварин. Дослідники штучного інтелекту вільні у використанні методів, які не є притаманними для мислення людей, або ж які вимагають набагато більше обчислювальної потужності, ніж люди.

Комп'ютерні програми мають достатню швидкість і пам'ять, але їх можливості відповідають тим інтелектуальним механізмам, які розробники достатньо добре розуміють, щоб складати їх в програми. Деякі здібності, які діти зазвичай не розвивають до підліткового віку, цілком можуть бути закладені в програму, а деякі здібності, які мають дворічні діти, все ще не реалізовані. Справа ускладнюється ще й тим, що когнітивним наукам досі не вдалося точно визначити повноту людських здібностей [8]. Цілком ймовірно, що організація інтелектуальних механізмів штучного інтелекту може успішно відрізнятися від організації інтелектуальних механізмів у людей.

Висновки до розділу №3

У цьому розділі були загально розглянуті основні концепції теорії свідомості: фізикалізм та субстанційний дуалізм, ці концепції протилежні одні одному, тож завдяки їх опису ми побачили основні погляди філософів на створення штучного інтелекту. Обидві концепції були проаналізовані, були викладені їх аргументи та проаналізовані їх сумісність з дійсністю.

Висновки

Отже, підводячи підсумки виконаної роботи, зауважу, що відповідно до мети дослідження, а саме: виявлення та окреслення соціальних базисів штучного інтелекту та створення цілісного уявлення застосування філософських понять “мислення” та “штучний інтелект” у сучасних дослідженнях штучного інтелекту, в розділах даної роботи виконувались відповідні завдання, які варто узагальнити.

В першому розділі був описаний важливий для даної роботи філософський аналіз – окреслення понять та виявлення соціального аспекта у дослідженні штучного інтелекту. В першому підрозділі було досліджено проблему понять – а саме поняття “штучний інтелект” у класичному розумінні та проблеми даного класичного визначення. Були описані: головний класичний аргумент визначення штучного інтелекту за Тьюрингом, а також основні контраргументи для прояснення його недосконалості. Тьюринг пропонує вважати розумною ту машину, яка складає враження у людини, що вона має справу з людиною, контраргументом до цієї тези є те, що розум людини не може бути зведений до поведінки або біології, адже людськість є більш комплексним явищем. В другому підрозділі було описано розуміння штучного інтелекту у сучасних провідних філософських тлумаченнях, де було виявлено слабкість класичного трактування штучного інтелекту та необхідність розглядати це явище з боку міждисциплінарності. В третьому підрозділі були розглянуті соціальне підґрунтя штучного інтелекту та на прикладах були продемонстровані соціальні аспекти використання штучного інтелекту та недосконалості його використання у соціальному вимірі.

У другому розділі розглядаються моральні межі та межі філософського конструювання штучного інтелекту. У першому підрозділі описується історична реконструкція поняття “інтелект” спираючись на філософію Рене

Декарта, з якої бере початок одна з провідних течій у філософії свідомості та філософії штучного інтелекта – субстанційний дуалізм. В розумінні Декарта матерія та розум є двома паралельними сутностями, тому створити машину, яка поєднує їх у собі ми не можемо. У другому підрозділі описано загрози штучного інтелекту з філософської точки зору на прикладі книг провідних інтелектуалів сучасності, були зазначено два типа загроз, з технічної сторони (створення надінтелектуальної машини, яка ставить під загрозу існування людства), а також з етичної сторони, аргументи якої полягають, що розумна машина (у разі виникнення) буде інтегрована у соціальне середовище, тож буде діяти згідно з соціальними інтересами. У третьому підрозділі описано соціальні межі штучного інтелекту, зазначена можливість появи морального статусу штучного інтелекту та необхідність наявності певних соціальних критеріїв.

У третьому розділі були розглянуті основні філософські концепції штучного інтелекту у контексті філософії свідомості. У першому підрозділі розглядається фізикаліський підхід до теорії свідомості, у цьому підрозділі описується підхід до людського мислення як до комп'ютера, також були наведені контраргументи та спростування цієї тези. У другому підрозділі описано підхід субстанційного дуалізму, а саме, що тіло та розум є різними категоріями, що людська свідомість не може бути зведена до обчислень, описується Я-концепт задля прояснення аргументів щодо існування душі, як окремої від тіла сутності.

Тож, у даній роботі були окреслені соціальні основи штучного інтелекту, були окреслені поняття штучного інтелекту у класичному розумінні, у сучасному та з погляду історичної реконструкції, були проаналізовані соціальні аспекти використання штучного інтелекту, етичні питання штучного

інтелекту, було доведено необхідність використання філософії та соціології при конструюванні штучного інтелекту.

Список використаних джерел

1. Berkeley, I. (2008) CajunBot: A Case Study in Embodied Cognition. In: Calvo, P., Gomila, A. (eds.) *Handbook of Cognitive Science: An Embodied Approach*. Elsevier B.V, Amsterdam.
2. Bechtel, W. (2008). *Mental Mechanisms*. Routledge, New York.
3. Bostrom, N. (2002). Existential Risks: Analyzing Human Extinction Scenarios. *Journal of Evolution and Technology 9*. URL: http://jetpress.org/volume9/risks.html
4. Bostrom, N. (2003). Astronomical Waste: The Opportunity Cost of Delayed Technological Development. *Utilitas 15*: 308-314.
5. Bostrom, N. (2004). The Future of Human Evolution. *Death and Anti-Death: Two Hundred Years After Kant, Fifty Years After Turing*, ed. Charles Tandy. Palo Alto, California: Ria University Press.
6. Bostrom, N. and Cirkovic, M. (eds.) (2007). *Global Catastrophic Risks*. Oxford: Oxford University Press.
7. Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, 352p.
8. Clark, A. (2001). *Mindware: An introduction to the philosophy of cognitive science*. Oxford University Press, Oxford.
9. Chalmers, D. J. (1996). Does a rock implement every finite-state automaton? *Synthese 108*, 310- 333.
10. Chalmers, D. J. (2010). *The character of consciousness*. Oxford University Press, Oxford.
11. Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. New York and Oxford: Oxford University Press.
12. Dreyfus, H. (1972). *What Computers Can't Do*. MIT Press.

13. Charniak, Eugene, McDermott, Drew. (1985). *Introduction to Artificial Intelligence.* Addison-Wesley Publishing Co, 1985, 701p.

14. Hayes, P., Berkeley, I., Bringsjord, S., Hartcastle, V., McKee, G., Stufflebeam, R. (1997). What is a computer? An electronic discussion. *The Monist* (80), 389—404.

15. Haugeland, J. (1989). *Artificial Intelligence: The Very Idea*. MIT Press, Cambridge.

16. Harre, R., Wang, H. (1999). Setting up a real 'Chinese room': an empirical replication of a famous thought experiment. *Journal of Experimental & Theoretical Artificial Intelligence 11*(2), 153—154.

17. Kamm, F. (2007). *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. Oxford: Oxford University Press.

18. Metzinger T. (2003). *Being No One: The Self-Model Theory of Subjectivity*. Cambridge, MA: The MIT Press, 699 pp.

19. McCarthy, J. (1979). Ascribing Mental Qualities to Machines. In: Ringle, M. (ed.) *Philosophical Perspectives in Artificial Intelligence*, pp. 161—195. Harvester Press, Sussex.

20. McTear, M. (ed.). (1988). *Understanding Cognitive Science*. Horwood Ltd., Chichester.

21. Robbins, P., Aydede, M. (eds.). (2009). *The Cambridge Handbook of Situated Cognition*. New York: Cambridge University Press.

22. Searle, J. (1980). Minds, Brains and Programs. *Behavioral and Brain Sciences 3*(3): 417-57.

23. Stinger P. (2011). *The Expanding Circle: Ethics, Evolution, and Moral Progress,* Princeton University Press, 227p.

24. Smart, J. C. C. (2013). *Philosophy and Scientific Realism,* Routledge, 174p.

25. Turing, A. (1950). Computing Machinery and Intelligence. *Mind 236*, 433—460.

26. Harre, R., Wang, H. (1999). Setting up a real ‘Chinese room’: an empirical replication of a famous thought experiment. *Journal of Experimental & Theoretical Artificial Intelligence 11*(2), 153—154.

27. Баррат Д. (2015). *Последнее изобретение человечества: Искусственный интеллект и конец эры Homo sapiens* / пер. с англ. М.: Альпина нон-фикшн, 304 с.