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

ШТУЧНИЙ ІНТЕЛЕКТ: ФІЛОСОФСЬКИЙ АНАЛІЗ ІСТОРІЇ ТА СУЧАСНИХ ВИКЛИКІВ

Кваліфікаційна робота за напрямом підготовки 033 «Філософія»
на здобуття кваліфікаційного рівня бакалавра філософії

Студент-виконавець:
Адаменко Богдан Володимирович
_____
IV курс
Науковий керівник:
Лактіонова Анна Валеріївна
доктор філософських наук, доцент кафедри
теоретичної та практичної філософії
КНУ ім. Тараса Шевченка
_____

Допущено до захисту:
на засіданні кафедри теоретичної і практичної філософії
протокол №____ від _____ 2020 р.
Зав. кафедри теоретичної і практичної філософії,
доктор філософських наук, професор
Шашкова Людмила Олексіївна _____

Київ-2020

ЗМІСТ

Актуальність роботи. Філософи завжди залучені у власне сучасне, що сповнене новими викликами, яких раніше не знала історія. Актуальний стрімкий розвиток наук та технологій — це саме той виклик з яким мають справу сучасні філософи. Більшість технологій, які зараз використовуються людьми у повсякденному житті, засновані на роботі систем, які є прикладами ШІ (штучного інтелекту). Ми живемо у фантастичних романах минулого та у мріях візіонерів попередніх століть. ШІ був такої мрією, проте став нашим сьогоденням. Однак, чим насправді є “штучний інтелект”, залишається відкритим питанням. Сфера його досліджень та застосування виходить за рамки однієї галузі чи науки. Це — мультидисциплінарна тема, яка однаково важлива, як для математиків, логіків, нейронауковців, так і для філософів. Актуальність та відповідальність філософів, що досліджують сферу ШІ, постає у визначенні проблемних аспектів даної теми, а також аналізі впливів, які “штучний інтелект”, як технологія та ідея, здійснює на людину, у тому числі розгляд значення, яке дана технологія матиме для людства у майбутньому.

Метою роботи є аналіз і систематизація знань, а також філософська рефлексія стосовно ідеї та визначення “штучного інтелекту”. Для досягнення поставленої мети були окреслені наступні **завдання:**

- Проаналізувати передумови виникнення та окреслення ідеї “штучного інтелекту” в історико-контекстуальному вимірі.
- Систематизувати критичні реакції філософів на концепцію “штучного інтелекту”.
- Сформувати власне визначення “штучного інтелекту” та розглянути приклади “парадоксальних” проблем, які несе в собі сфера досліджень ШІ.
- Узагальнити знання щодо етичної проблематики, яка виникає в контексті розгляду ШІ як автономної системи. Переглянути “етику

даних”, “машинну етику”, а також концепцію відповідальності в межах створення технологій на основі ШІ.

- Сформувати власні висновки на основі розгляду ШІ як технології, що без філософського осмислення та проблематизації несе можливі ризики у своїх впливах.

Об’єктом дослідження є філософський доробок дослідників теми “штучного інтелекту”.

Предметом дослідження є розгляд ідеї “штучного інтелекту” в проблематизації викликів, які провокує дана технологія.

Методи дослідження. У ході даного дослідження були використані такі загальнонаукові та філософські методи:

- Аналітичний метод, як один із головних методів даної роботи, що допоміг в історико-філософській реконструкції ідеї “штучного інтелекту”, а також основних підходів та критик даної сфери досліджень.
- Герменевтичний метод, що забезпечив більш глибоке проникнення в сутність досліджуваної теми, її тлумачення та інтерпретації на основі першоджерел з філософського осмислення ШІ.
- Компаративний метод, застосування якого було зумовлене зверненням до доробку більш ніж як одного дослідника, що обумовило необхідність порівняння та аналізу різних позицій та підходів.

Використання вказаної методології дало змогу комплексно проаналізувати та систематизувати отримані в ході роботи результати задля чіткого і глибинного розкриття теми дипломної роботи.

Ступінь опрацювання. Наголошуючи на актуальності даного дослідження, слід зазначити, що в українському філософському середовищі ідея “штучного інтелекту” знаходить своє відображення в роботах багатьох дослідників, зокрема в працях Олени Комар та Дмитра Сепетия, які

розглядають ШІ в контексті філософії свідомості. Починаючи власне дослідження ідейних витоків “штучного інтелекту”, наш послідовний історико-ідейний аналіз був заснований на лініях міркувань Вільяма Рамзея, Селмера Брінгсйорда та Вінсента Мюллера, що були поєднані в цілісну ідею та доповнені оригінальними статтями та дослідженнями як і піонерів даної галузі (Алан Тюрінг, Джон МакКарті, Марвін Мінський, Аллан Ньюел, Герберт Саймон) так і їхніми послідовниками (Гіларі Патнем, Пітер Норвіг, Стюарт Расел) та основними філософськими критиками (Г’юберт Дрейфус, Нед Блок, Джон Серль). Сучасні виклики у сфері досліджень та застосування “штучного інтелекту” були виявлені та проаналізовані на основі робіт Віргінії Дігнум, Дональда Готербарна, Мерел Нурман, Ендрю Фінберга та інших. В контексті розгляду питань етики та відповідальності було залучено доробок українського філософа Анатолія Єрмоленка.

Сформульовані завдання і загальна логіка дослідження зумовили **структуру даної роботи,** яка складається зі вступу, двох розділів, висновків та списку використаної літератури та джерел (40 найменувань).

Загальний обсяг дипломної роботи становить 47 сторінки (без списку використаної літератури та джерел)

РОЗДІЛ 1. Історія та концептуальні передумови ідеї “Штучного інтелекту”

В першій частині роботі ми спробуємо відновити лінію міркування стосовно ШІ, показати історичні початки, розвиток, мотивацію основних дослідників та натхненників даної теми, цим самим окресливши ідейне ядро “штучного інтелекту”. Також ми розглянемо три важливі критики зі сторони філософії. Це буде першим кроком, який в результаті допоможе сформувати власне визначення та підхід для подальшого викладу ідей другого розділу даної роботи.

1.1 Алан Тюрінг, як ідейний провісник “штучного інтелекту”

Алан Тюрінг (Alan Turing), математик, логік і криптограф, окрім того, що заклав основи інформатики, розробив теоретичну модель роботи сучасних комп’ютерів та здійснив науковий вклад в теорію алгоритмів, розпочав активні дебати стосовно питання “Чи може машина мислити?”. Хоча сам А. Тюрінг не дав відповідь на власне запитання, він стимулював запитування, наукові та філософські пошуки в руслі ідей про “мислячі машини”. До його ідей зверталися більшість авторів, які будуть згадані далі.

Замість пошуку відповідей на питання “Чи може машина мислити?”, А. Тюрінг, власне, пропонує пограти у своєрідну мовну гру “в мислення” — так звану “гру в імітацію”. Популярність тесту Тюрінга (ТТ) пояснюється як науковими, так і філософськими положеннями. Науковий інтерес до ТТ обумовлений тим, що науковому співтовариству випала нагода організувати раціональний дискурс щодо питання створення машин, здатних до мислення. А. Тюрінг вважав питання “Чи може машина мислити?” безглуздим. Неможливо, на його думку, міркувати не тільки з приводу того, що таке “мислення”, а й з приводу того, що таке “машина”. Сама “гра в імітацію” відразу розділяє дві сутності “людину” та “машину”. А. Тюрінг

пропонує відкинути метафізичні спекуляції і псевдонаукові запитання, а на їх заміну пропонує своєрідний тест на мислення, точніше на його імітацію.

“Гра в імітацію” подається А. Тюрінгом у вигляду тесту, який був запропонований у статті 1950 року “Обчислювальні машини та розум” (Computing Machinery and Intelligence) [38]. За тестом пізніше закріпилося прізвище автора. Вийшли сотні статей і об'ємних монографій, присвячені ТТ. Він став однією з найбільш обговорюваних тем у філософії штучного інтелекту, філософії свідомості, когнітивних та комп'ютерних наук. Розглянемо його базові положення. Грають чоловік (А), жінка (В) і суддя (С). Стать судді є несуттєвою. Суддя ізольований від А і В. Він знаходиться в окремій кімнаті. Завдання судді визначити хто з гравців жінка. Завдання чоловіка і жінки — переконати суддю в тому, що саме він / вона жінка. Засобом спілкування є телеграф. Суддя ставить питання в письмовій формі, використовуючи природну мову. Відповіді він отримує в тій же формі. Питання можуть бути на будь-яку тему: від математики до поезії, від погоди до шахів. За задумом А. Тюрінга, нове запитання з приводу мислення машини слід поставити так: “Якщо машина займе місце гравця В, то чи буде суддя помилятися так само часто, як і при грі з чоловіком і жінкою?”. Тобто замість жінки буде грати машина. В майбутніх інтерпретаціях ТТ ігнорується стать гравців — грають машина (А), людина (В) і суддя (С). Така форма гри в імітацію, власне і отримала назву “тест Тюрінга”. Мета С визначити, хто з гравців машина, а хто людина. Саме на цьому етапі відбувається заміна питання: “Чи може машина мислити?”, питанням: “Чи може машина грати в імітацію?”

ТТ через свою інтуїтивну зрозумілість та відсутність чітко вираженої філософської позиції у самого А. Тюрінга, спричинив активне обговорення. Почавши з нечіткого питання, хоча й перетворивши його в більш інтуїтивно зрозумілий тест, А. Тюрінг відкрив дуже широке поле для інтерпретацій.

Біґевіористи знаходили підтвердження своїх ідей в тому, що мислення — це лиш зовнішня поведінка, яку й демонструє машина в ТТ. У біґевіоризмі предметом психологічного аналізу виступає не якась ментальна структура суб'єктивної реальності, а об'єктивно фіксуються параметри поведінки (реакції), які визначаються зовнішніми впливами (стимулами). Філософський успіх ТТ зумовлений тим, що він поклав початок могутньому напрямку в сучасній філософії, відомому під назвою «функціоналізм», до якого ми ще звернемося в наступних підрозділах.

1.2 Народження “штучного інтелекту”

Визначення “штучний інтелект” (artificial intelligence) з’явилося в 1955 році, під час підготовки до літньої конференції в Дартмутському коледжі, яка відбулася 1956 році. Офіційно його творцем вважається Джон МакКарті (John McCarthy), який міркував про інтелект як про математичну абстракцію. Його ідеєю було створення “штучного” аналога інтелекту людини. Важливо згадати, що Д. Маккарті наголошував на тому, що штучний інтелект не має подібності з поведінкою людини, окрім хіба що натяків на виконання задач, які властиві людям. Лиш Ален Ньюел (Allen Newell) та Герберт Саймон (Herbert Simon) з Інституту Карнегі з усіх учасників конференції займалися поведінкою людини. Вони на той час вже були відомими дослідниками, котрі поєднали соціологію з когнітивістикою. В той час як решта учасників були радше мрійниками-теоретиками, А. Ньюелл та Г. Саймон вже могли представити практичну програму яка, за їхніми словами, була здатна “міркувати” [7, 125-127].

Дартмутська конференція є ключовою подією для сучасних дослідників сфери штучного інтелекту. Важливість конференції полягає у двох її ключових здобутках. Перше — це саме введення терміну “штучний інтелект”, який, як бачимо, міцно вкорінився. Друге, А. Ньюелл та Г. Саймон

представили комп’ютерну програму “Теоретик Логіки” (Logic Theorist). Дана програма вважалася важливим кроком в поданні міркувань людського рівня через конкретні обрахунки. Загалом конференція заклала фундамент окремої сфери знань. Аналіз ідейних планів конференції [13] дає зрозуміти, чому дослідження присвячені ШІ не можна було маркувати як такі, що належать до теорії рішень, теорій управління чи дослідження операцій або ж математики. По-перше, ШІ розпочав розвиток ідей моделювання таких людських здатностей як використання природної мови, творчість та навчання. По-друге, це була абсолютно нова методологія, яка розвинула напрям комп'ютерних наук і в якій розпочалися спроби створення машин, що діють автономно в складних та мінливих середовищах. Учасники конференції вважали, що когнітивні здатності людини можна описати за допомогою логіки та символьно змоделювати їх з допомогою комп’ютера. Даний підхід зараз окреслюється у визначенні Символьний/Логічний ШІ (Symbolic/Logic AI). Нижче спробуємо показати передумови даної ідеї [14, 40].

Основу для власних міркувань спеціалісти ШІ віднаходять в ідеях Арістотеля [14, 36]. Філософ вважав раціональність важливою характеристикою людської свідомості. Дедуктивні міркування, виражені в формі силогізмів, були рисою такої раціональності, а також основним інструментом для всієї науки. Система силогізмів Арістотеля призначена для проведення правильних міркувань, так що кожен може на їх основі випрацювати логічні висновки механістично, при наявності початкових засновків. Тобто, силогізми Арістотеля — це те, що ми зараз називаємо правилами логічного виводу. Найбільш глибоким вкладом Арістотеля в осердя ШІ є його ідея формалізму — уявлення про те, що певні патерни міркувань дійсні завдяки своїй синтетичній формі, незалежно від змісту. Дане поняття надалі укорінилося в обчислювальній теорії свідомості.

Як бачимо, першим рівнем підходу до штучного інтелекту є дедуктивна логіка та ідеї обчислюваності. Зважаючи на те значення, яке відіграла дедукція у філософії, сама ідея розумних машин була часто рівносильною до ідеї таких машин, які можуть робити логічні висновки. Від самого початку сфера досліджень ШІ пов'язана з теоріями автоматичного доведення теорем (Automated theorem proving), в яких ключову роль відіграє механізація дедукції (mechanizing deduction). Розглянемо проблеми механізації дедукції, які проявили себе протягом досліджень у сфері ШІ [14, 36].

Перший рівень проблеми пов’язаний з перевіркою доведень (proof cheking). Технічно дана проблема є найбільш простою і стосується логіки першого порядку. З появою перших комп’ютерів перевірка математичних доведень була найбільш цікавим їх застосуванням.

Другий рівень стосується пошуку доведень (proof discovery) і стосується теорії алгоритмів. А. Тюрінг відомий не лише своїм тестом: в першу чергу, він — математик, результати досліджень в теорії рекурсивних функцій якого (разом з Алонзо Черчем (Alonzo Church) в 1936 році) показали, що не існує такої машини Тюрінга, яка б могла визначити, чи є валідною формула логіки першого порядку, і це постає алгоритмічно нерозв'язною проблемою. Немає такого загального механістичного методу, який би давав змогу приймати правильне рішення за скінченний проміжок часу. Для більш простих формалізмів ще на ранніх етапах дана проблема була розв'язана. Вище згадувана програма “Теоретик Логіки” А. Ньюелла і Г. Саймона, що була представлена на першій Дартмутській конференції, змогла довести 38 із 52 теорем пропозиційної логіки, які були подані у праці Бертрана Рассела (Bertrand Russell) та Альфреда Вайтгеда (Alfred Whitehead) “Principia Mathematica”. Як зазначали автори програми, в основі її побудови була закладена імітація процесів людського мислення. Більшою мірою дане

твердження можна сприймати як метафоричне, проте самі дослідники вважали, що це було першим знаком того, що штучний інтелект, в основі якого буде лежати символьно-логічна імітація людського мислення, точно з'явиться. Інші теоретики даної сфери вважали, що закладення в основу подібних програм “людського мислення” є стримуючим фактором. Дане твердження вже відсилає нас до двох підходів ШІ: перший, який розглядає ШІ як науку, що досліджує людське мислення; другий, для якого сфера ШІ — суто інженерія, побудова інтелектуальних систем, діяльність яких не повинна нагадувати внутрішню когнітивну діяльність людини.

На третьому рівні даної проблеми розглядається генерація припущень (conjecture generation). Припущення не з'являються нізвідки. Маючи інформацію як знання, математики, наприклад, висувають гіпотези і намагаються їх довести, інколи навіть успішно. Процес відкриттів, так само як формування нових концепцій, є однією з творчих активностей людського інтелекту. Складність імітації такого процесу обчислювальними засобами є головною причиною, чому ШІ не здійснює тут поступ. Дослідники ШІ намагалися і намагаються моделювати “відкриття як інтелектуальний процес” обчислювальними методами, та навіть з появою нейронних мереж та машинного навчання це залишається проблемою [14, 37-38]. Головне в цій проблемі, що, більшою мірою, вищі когнітивні процеси мають цілісний характер. Як ми побачимо далі, маніпуляції із символами не можуть пояснити такі важливі якості людини як судження, уява чи інтуїція. Більш детально розглянемо це в частині, присвяченій критиці ШІ.

1.3 Обчислювальна теорія свідомості

Хоча й “відкриття” штучного інтелекту відбулося у 1956 році, а філософи століттями розмірковували про мислячі машини та інтелект, головні концептуальні витоки ШІ знаходяться на перехресті двох найбільш

важливих інтелектуальних розробок XX століття. Це теорія алгоритмів (theory of computability), яка розроблялася такими відомими вченими як Алан Тюрінг, Алонзо Черч, Стівен Кліні (Stephen Kleene) та Курт Гедель (Kurt Gödel). А також це “когнітивна революція”, яка розпочалася в середині 1950-х і яка здійснила крок від біґевіоризму до когнітивної психології. Зазвичай когнітивну революцію пов'язують з роботами 1950-х років Джорджа Міллера (George Miller) та Ноама Чомскі (Noam Chomsky), особливо з критичного огляду теорії мови Берреса Скіннера (Burrhus Skinner). Б. Скіннер в 1957 році опублікував книгу “Вербальна поведінка” (Verbal Behavior), дослідження мови в якій перебували в рамках біґевіористичного підходу. Ноам Чомскі у своїй рецензії на дану роботу показав, що біґевіористична теорія не дозволяє зрозуміти витоки творчої діяльності, вираженої з допомогою мови. Впливовим був аргумент “бідність стимулу” (poverty of stimulus) Н. Чомскі, який показував, що швидкість та ефективність вивчення мов дітьми пов'язана не із зовнішніми стимулами, а з наявністю вроджених ментальних правил та репрезентацій, що кодують мовну компетенцію.

Також когнітивну революцію передбачали ще в 40-х роках XX століття Воррен МакКалох (Warren McCulloch), Уолтер Піттс (Walter Pitts) та інші піонери кібернетики, які вказували на подібність між людським мисленням та обробкою інформації. Вони запропонували модель штучного нейрона, беручи за основу три ідейні джерела: знання основ фізіології та будови нейронів у мозку; формальний аналіз логіки висловлювань Б. Рассела і А. Вайтгеда; а також теорія обчислень А. Тюрінга [8, 54].

Зі сторони психології, Едвард Толмен (Edward Tolman), проводячи експериментами з навігації щурів в лабіринтах, доводив існування когнітивних карт (cognitive maps). Сильним аргументом існування ментальних репрезентацій (mental representations) вважаються

експериментальні результати, отримані Джорджем Сперлінгом (George Sperling) в 1960, котрі стосувалися пам'яті, і того, що люди зберігають більше інформації, чим вони можуть повідомити. Біґевіоризм, своєю чергою, недовіряв всьому тому, що не піддається спостереженню безпосередньо в зовнішніх проявах. Біґевіоризм залишався домінуючою інтелектуальною теорією включно до кінця 1950-х років, допоки своєї інтелектуальної значущості не набув "когнітивний" підхід [14, 41].

Після перших кроків в теоретизації "ментальної репрезентації", дослідники А. Ньюелл та Г. Саймон ввели "метафору комп'ютера" ("computer metaphore"). Мислячи за аналогією, що комп'ютери теж, подібно до людей, зберігають та структурують дані в пам'яті та розв'язують проблеми, здійснюючи маніпуляції з цими даними, виконуючи певні інструкції, ними було зроблено перші кроки до обчислювальної теорії свідомості. Свідомість, в даній теорії, розглядається як "синтаксичний двигун" (syntactic engine). Психічні процеси тут — це послідовності знаків ментальних репрезентацій, що виражають пропозиційний зміст певних думок.

Продовженням обчислювальної теорії свідомості являється функціоналізм машини Тюрінга, вперше описаний Гіларі Патнемом (Hilary Putnam) в 1960 році [34]. Дана ідея, просунувши "когнітивну революцію" у філософських колах (більшою мірою), підважила біґевіоризм та окреслила перспективи розвитку ШІ. В основі функціоналізму лежить ідея, що ментальні стани не знаходяться в мозку, а радше в ролі тих станів, яку вони відіграють у ментальному житті; особливо в причиново-наслідкових зв'язках, які мають справу зі стимулами (вхід), поведінкою (вихід) та іншими ментальними станами. Свідомість для функціоналістів — це велика машина Тюрінга, робота якої визначається наборами інструкцій між переходами від різних функціональних станів.

В даному контексті цікаво згадати і про гіпотезу системи фізичних символів (phisical symbolic system hipothesis), яку висловили А. Ньюелл та Г. Саймон в 1976 році [32]. Згідно з гіпотезою, фізичні символьні системи мають необхідні і достатні засоби для загальної інтелектуальної дії. Система фізичних символів — це машина, що із часом розробляє розгорнуту колекцію (набір) символьних структур. Структура символа — це набір символьних токенів[1] (tokens), які пов'язані з будь-якими фізичними образами. Наприклад, токен "знаходитися поруч з іншим" [14, 43].

На противагу типам (type), які є абстрактними та унікальними, токени — це конкретні дані, що складаються, наприклад, із пікселів світла на екрані комп'ютера, електронних ланцюгів крапок та тире, сигналів диму, сигналів рук, звукових хвиль і т.п. [39].

Згідно обчислювальної теорії свідомості, складні думки репрезентуються складними символьними структурами, таким самим чином як природна мова та формальна логіка. Складні речення рекурсивно будуються із простих компонентів. Складні думки ж набувають свого значення із комбінацій простих компонентів. Наприклад, ментальна репрезентація такої комплексної думки як: "Всі люди смертні", в собі містить компоненти ментальних репрезентацій для таких термінів як: "всі", "людина", "смерть". Ці компоненти певним чином з'єднуються разом (наука, на думку авторів теорії, має пояснити, як символьні операції відбуваються у мозку) і формують комплексну думку, про те що всі люди смертні.

Проблемою тут постає те, як же все-таки ці "компоненти" набувають своїх значень. Якби дана проблема була розв'язана, то всі існуючі аргументи проти обчислювальної теорії свідомості втратили б свою силу. Машини тоді б "думали", виконуючи обчислення над формальними символьними

[1]Калькування терміну "токен" з англійської мови пов'язане з проблемою неперекладності. Неможливість знаходження відповідника в українській мові затвердило вживання даної кальки серед філософів та науковців.

структурами, і навіть більше — претендували на позасимвольне розуміння (extra-symbolic understending).

В даному контексті розглянемо також ряд теорій щодо “натуралізації змісту”, “натуралізації семантики” та “натуралізації інтенційності”, ціль яких була представити фізично пояснення того, яким чином токени ментальних символів набувають значень. Нижче стисло розглянемо три спроби натуралізації пояснення значень.

Теорія інформації (information theory) — суть даної теорії — це коваріація. Ідея в тому, що якщо кількість *X* систематично коваріюється із кількістю *Y*, то тоді *X* несе (вміщує) інформацію про *Y*. Прояснимо на прикладі. Ми бачимо дим. Дим “означає” вогонь в цьому димі, і несе інформацію про вогонь. “Означати”, за Полом Грайсом (Paul Grice) є природним значенням (natural meaning), і відправляє нас до теорії інформаційної семантики. Оскільки дим і вогонь систематично коваріюються, то тут можна сказати що один відноситься до іншого. Інтенціональність в даній теорії можна розглядати не як ментальну здатність, а як природне явище.

Еволюційна теорія (evolutionary theories) — інтенціональні стани постають в даній теорії адаптацією. Зміст “значення” інтенціонального стану лиш просто функція, тобто ціль, якій вона послуговується.

Концептуально-рольова семантика (comceptuale-role semantic). В даній теорії значення лінгвістичного елементу (виразу, речення та ін.) — це спосіб, в який цей елемент використовується носієм мови. Значення “mentalese symbol” *S* є фіксацією тієї ролі, яку *S* відіграє в пізнавальному житті людини. Дану теорію можна вважати функціоналізмом стосовно ментальних станів. Концептуально-рольова семантика розглядає значення символа цілковито внутрішньо, залежним лиш від відношень з іншими символами.

Відсутність зв'язку між mentalese (чи символами комп'ютера) та світом, лягло в основу критики Джеррі Фодора (Jerry Fodor). Критика Д. Фодора (1978) провіщала аргумент Китайської кімнати Джона Серля (John Searle). Перший писав про те, що комп'ютерні моделі не забезпечують семантичну теорію, якщо під семантичною теорією розуміти відношення мови та світу. Д. Фодор пише, що машина може скомпілювати "Люсі принесла десерт?", але не матиме жодної ідеї (уявлення), що в реченні запитується чи принесла Люсі десерт [22].

Обчислювальна теорія свідомості, функціоналізм машини Тюрінга та гіпотеза систем фізичних символів хоча й мають суперечності стосовно ідеї вродженості (innateness), разом характеризують те, що зараз називають класичним ШІ. Джон Хаугеланд (John Haugeland) у своїй книзі 1985 року [25] підсумував дані підходи як "старий добрий штучний інтелект" (good old-fashioned AI).

1.4 Критика з точки зору філософії

Означивши головні теоретичні аспекти ШІ, ми вже частково прослідкували за рухом думки дослідників даної сфери. На цьому етапі варто провести перше чітке термінологічне розрізнення між двома версіями ШІ: сильною та слабкою. Вперше ці два терміни вжив філософ Дж. Серль у статті 1980 року "Розуми, мозки і програми" (Minds, Brains, and Programs) для того, щоб прояснити: "Яке психологічне та філософське значення нам треба надавати спробам імітації людських когнітивних здібностей за допомогою комп'ютера?" [9]. Слабка версія ШІ (Weak Artificial Intelligence), згідно з філософом, це інструмент, який, наприклад, допомагає нам точніше перевіряти та формулювати певні гіпотези, при цьому така комп'ютерна програма не володіє когнітивними здатностями. Сильна версія ШІ (Strong Artificial Intelligence) здатна перебувати в когнітивних станах та мати

свідомість. З того, що було згадано в попередніх підрозділах, найбільш відповідними до ідеї сильного ШІ є ідеї А. Тюрінга, Г. Патнема та обчислювальна теорія свідомості загалом.

Розглянемо три основні критичні зауваги в сторону сильної версії ШІ, які допомогли змінити ситуацію в спільноті дослідників і вказали на нові напрями досліджень ШІ. Найбільш відомою атакою на дану концепцію вважається мисленнєвий експеримент Дж. Серля відомий під назвою “аргумент китайської кімнати” (АКК). Емпіричним базисом для атаки Дж. Серля слугували практичні напрацювання в області ШІ — роботи Роджера Шенка (Roger Schank) і його колег з Єльського університету в області розуміння текстів (оповідань). Мета програми Р. Шенка — це моделювання людської здатності розуміти тексти. Факт розуміння пов'язується зі здатністю людини відповідати на питання про зміст тексту в умовах неявно представленої в тексті інформації про відповідь, тобто в умовах відсутності можливості встановлення відповідності між питанням і відповіддю без залучення позатекстової інформації. Програма Р. Шенка, вважає Дж. Серль, це реалізація тесту Тюрінга. Машина проходить ТТ якщо на її вхід подають текст і вона відповідає на запитання таким же чином, яким, як очікується, буде відповідати і людина. Прихильники сильного ШІ стверджують, що в даній послідовності запитань і відповідей машина не просто моделює людську здатність до розуміння. Машина в прямому сенсі розуміє текст. Більш того, машина і її програма пояснюють людську здатність розуміти текст і здатність людини осмислено відповідати на питання про даний текст. Проти такої позиції Дж. Серль і пропонує “аргумент китайської кімнати”. АКК заснований на мисленнєвому експерименті, в якому Дж. Серль сам грає головну роль. Уявімо наступне. Дж. Серль в середині кімнати. За межами кімнати — носії китайської мови, які не знають, що всередині — Джон Серль, який і в реальному житті, так і у

кімнаті, не знає китайської мови, проте вільно володіє англійською. Китайці через отвір надсилають до кімнати картки. На картках китайською мовою написані запитання. З кімнати, завдяки таємній роботі Дж. Серля в ній, відправляються картки носіям мови. Те, що відправляється назовні, Дж. Серль формує працюючи з книгою правил. Дана книга являється довідковою таблицею (інструкцією), в якій навпроти того, що Дж. Серль отримав, показано, яку відповідь відправити. Також подані відповіді оцінює на осмисленість сторонній спостерігач (подібно до "судді" в тесті Тюрінга). Крім текстів китайською, Дж. Серлю дають і тексти англійською, ставлять питання англійською і вимагають відповідей англійською мовою. Через деякий час суддя переконується, що відповіді на питання поставлені китайською абсолютно не відрізняються від відповідей, які суддя отримує від справжніх носіїв китайської мови. З точки зору спостерігача, відповіді англійською та відповіді китайською однією мірою якісні. Якщо відновити схему "перекладу" в її "китайській частині", то виходить, що для видачі осмислених відповідей ніякого розуміння з боку Дж. Серля і не потрібно. Раз так, то немає такого "розуміння" і у машини (включаючи програму Шенка). Тобто програма Р. Шенка аж ніяк не наближає нас до пояснення людської здатності розуміти, на відміну від думки прихильників сильного ШІ. Адже не можна назвати "розумінням" сукупність формалізованих операцій над сукупністю символів. Підсумувати аргумент Джона Серля можна наступним чином:

1. Синтаксису не достатньо для семантики.
2. Комп'ютерні програми повністю визначаються їх формальною чи синтаксичною структурою.
3. Свідомості притаманні ментальні змісти; ці змісти є семантичними.

Нед Блок (Ned Block) завдавав удару по машинному функціоналізму у 1978 році. Він пропонує розглянути наступний мисленнєвий експеримент

[15]. Уявімо, що кожен мешканець Китаю буде симулювати роботу людського мозку протягом однієї години. В кожного мешканця є радіо передавач, що працює у дві сторони. Представимо, що кожен мешканець Китаю — це окремий нейрон або будь-яка інша частина мозку, яку ми вважаємо атомарною. Люди через радіо приєднані до штучного тіла, від якого вони можуть отримувати сенсорні стимули, і якому вони можуть передавати вихідні сигнали для генерування такої психічної поведінки, наприклад, як підняття руки. Згідно з машинним функціоналізмом, з цього можна зробити наступний висновок. Якщо китайці правильно б симулювали таблицю переходів (transition table) і зважаючи на те, що вони відповідним чином пов'язані між собою з входами та виходами, то вони фактично складають свідомий розум. Але це постає для нас контрінтуїтивним. Хоча подана система на певному описовому рівні ізоморфна до мозку, проте такі речі як біль, почуття, переконання, бажання і т.п. дійсним чином не перебувають в даній системі.

Історично першою була критика Г'юберта Дрейфуса (Hubert Dreyfus), яку він детально описав у своїй книзі 1972 року "Чого не можуть комп'ютери: межі штучного інтелекту" (What Computers Can't Do: The Limits of Artificial Intelligence) [3]. Г. Дрейфус, американський філософ, який став відомим, в першу чергу, завдяки своїм напрацюванням з феноменології. Саме завдяки його творам американські філософи, які більшою мірою орієнтовані на аналітичну філософську традицію, стали знайомитися з феноменологією Едмунда Гуссерля. Г. Дрейфус, беручи за основу результати епістемологічних досліджень у філософії XX ст., котрі були отримані Е. Гуссерлем у феноменології "життєсвіту", Мартіном Хайдеггером в "Бутті та часі", а також Людвігом Вітґенштайном у "Філософських дослідженнях", стверджує пріорітет практичного, повсякденного рівня буття перед формалізованими теоретичними

конструкціями. Діяльність людини в повсякденному світі не є лиш просто складною для формалізації, але й взагалі унеможливлює проведення формально-логічних зведень. Г. Дрейфус робить висновок, що дослідження у сфері ШІ, в межах якої здійснюється спроба логічної формалізації людського мислення з метою імітації на комп'ютерах, приречена на провал. Критика Г. Дрейфуса — це суміш емпіричних та філософських аргументів. Емпіричні зауваги Г. Дрейфуса постають більшою мірою у вигляді звинувачень у сторону дослідників ШІ в тому, що вони не змогли виконати власні цілі і ще досі не наблизилися до створення ШІ. Дана сторона критики є достатньо слабкою, бо сфера досліджень ШІ — молода галузь знань, тому повернемося до філософських аргументів. Г. Дрейфус вказує на те, що наша здатність розуміти себе та інших людей лежить у феноменологічному вимірі, і не піддається пропозиційній кодифікації в стилі "старого доброго ШІ". Феноменологічний рівень не може бути схопленим системою заснованою на правилах. В той час як Марвін Мінський (Marvin Minsky) писав, що поведінку людини визначають правила, Г. Дрейфус вказував, що неможливо створити правил для всіх ситуацій, які людина переживає у своєму досвіді. Діяльність людини залежить від контексту, в якому ми постійно оцінюємо релевантність певних об'єктів і фактів [3, 240]. Г. Дрейфус пише про важливість здатності людини розділяти істотне від несуттєвого. Людина без особливих зусиль звертається до власного досвіду та знань в залежності від актуальної ситуації, в якій вона перебуває, як того вимагає її постійний взаємозв'язок зі світом. Г. Дрейфус звертає увагу, що в той час як комп'ютерна програма замкнута у "мікросвіті", то людина — ні. Досвід людини залишається відкритим і не зводиться до "мікросвітів". Навіть зараз проблема релевантності (problem of relevance) залишається однією з ключових технічних викликів у розробці ШІ як у сильній, так і у слабкій його версії.

1.5 Відхід від ідей “старого доброго штучного інтелекту”

В час коли були висунуті попередні філософські критичні зауваги, виникли і технічні складнощі у розробці класичного ШІ. Одна з таких проблем — проблема фреймів. Найбільш загальне визначення даної проблеми полягає в наступному. Це проблема визначення умов, при яких переконання має оновлюватися після того, як була здійснена певна дія. В технічному аспекті, це міркування про обрахунок ситуації (situation calculus). Більшою мірою дана проблема постала для формальних систем, які не можуть визначити, на якій саме релевантній (актуальній) інформації зосереджуватися. Концептуальні та інженерні проблеми, які з’явилися в межах розробок “старого доброго ШІ” в поєднанні з розчаруванням розробок у сфері експертних систем, підготували ґрунт для альтернативних підходів.

В 1980-х роках починає активно розвиватися “конекціонізм”. Основи даного підходу були закладені у працях В. МакКалоха й У. Піттса 40-х років XX століття, котрі запропонували модель штучного нейрона. Отже, основним концептуальним та інженерним інструментом конекціонізму є “нейронні мережі” [36]. Концепція нейронних мереж засновується на припущенні, що пізнавальні здатності людини пов'язані з результатами взаємодії великої кількості простих елементів чи юнітів (тобто нейронів), які здійснюють певні процеси обробки. Ідея покладається на те, що мозок складається з величезної кількості таких юнітів, і що разом вони здатні здійснювати складні когнітивні обробки такі як сприйняття, мова, моторні навички та інші [18, 187-189]. В конекціонізмі “репрезентації” є патернами активації наборів юнітів, котрі здійснюють обробку. Обробка відбувається шляхом розповсюдження активацій серед блоків обробки (вузлів) через їхній взаємозв'язок. Зв’язки між вузлами можуть бути “збуджуючими” чи “гальмуючими”, що залежить від призначення ваги — позитивного чи

негативного цифрового значення кожному зв’язку. В нейронних мережах немає центрального блоку (юніту) чи якихось явно закодованих інструкцій, котрі б визначали поведінку системи. Лише окремі вузли, які мають малу кількість цілковито локальних даних, які вони отримують від сусідів. При цьому пошкодження одного з вузлів не впливає на всю систему. Вона просто продовжує працювати згідно з принципом “паралелізму”. Паралелізм тут означає, що вузли можуть здійснювати обрахунки одночасно, і через зв’язки у мережі інформація розповсюджується також одночасно, на противагу системам заснованих на маніпуляціях із символами, в яких найменше відхилення від програмного курсу може мати катастрофічні наслідки. В неперервності, як зазначають прихильники даної теорії, полягає подібність нейронних мереж з пізнавальними здатностями людини.

Розглянемо ще один альтернативний підхід, який розробляв Родні Брукс (Rodney Brooks). Він дійшов до висновку, що системи спроектовані інженерами виключно на детальному символьному уявленні світу надто когнітивно неправдоподібні. Р. Брукс змінює акцент від символьних задач (символьного порядку) до буцімто простіших задач сприйняття та моторики [17]. Такі дослідники як Р. Брукс стверджували, що лише повністю втілені агенти (embodied agent) можуть виконувати такі класи задач, які потім можна визначити, як штучні агенти (artificial agents). Дослідження Символьного ШІ просувалися згори донизу (top-down), починаючи з символьно-логічних моделей раціонального мислення та вищих когнітивних здатностей людини. “Втілення” не грало важливої ролі. Р. Брукс та його колеги вважають, що дослідження ШІ мають починатися з вивчення та моделювання сенсорних та моторних систем людини. За їх словами, як тільки ми зробимо крок в розумінні таких простих та буденних систем — загадка інтелекту буде вирішена. Це амбітний план, в якому не останню роль відіграє питання

існування ментальних репрезентацій, хоча деякі дослідники вважають підхід “втіленого ШІ” поверненням до біґевіоризму.

1.6 Штучний інтелект. Проблема визначення

До цього моменту ми розглянули основні теоретичні та концептуальні ідеї ШІ. Проте історичний екскурс, поки що, не наблизив нас до визначення. Як видно з попередніх підрозділів, від самого початку думки дослідників розділилися на два підходи: перший, який розглядає ШІ як дисципліну, що досліджує людське мислення з метою його повного відтворення аж до самоусвідомлення комп’ютерною програмою себе (сильна версія ШІ); другий, який розглядає сферу ШІ виключно як інженерний підхід у створенні інтелектуальних систем, що лиш імітують когнітивні здатності людини (слабка версія ШІ). Проте проблема із визначенням пов’язана і ще з одним важливим аспектом, а саме з масовою культурою. Початково сфера створення ШІ була сферою мрійництва дослідників, які намагалися обґрунтувати власні фантазії математикою, логікою та кібернетикою загалом. Концепція обчислювальної теорії свідомості, в якій йдеться про те, що свідомість людини можна відтворити обчислювальними засобами, заклала уявлення про свідомі машини. В масовій культурі це було крайнє радикалізовано та вульгаризовано. Наприклад, такі відомі фільми як “Космічна Одіссея 2001 року” (2001: A Space Odyssey), “Термінатор”, (The Terminator), “Матриця” (The Matrix), “Я, Робот”, “Вона” (Her) та ін. сформували уявлення про те, що ШІ має свідомість, і що його поведінка та вчинки нічим не відрізняються від відповідних у людини. Це породило також ряди невиправданих страхів стосовно ШІ, аж до формування хибних уявлень та визначень. Так чим же насправді є “штучний інтелект”?

Консенсус у визначенні на цей час відсутній, проте варто розглянути позицію стосовно цього питання спеціалістів сфери досліджень штучного

інтелекту, а саме Пітера Норвіга (Peter Norvig) та Стюарта Рассела (Stuart Russell). Вони пропонують формувати визначення ШІ з точки зору кінцевих цілей досліджень: створення систем, що думають як людина; створення систем, що поводять себе як люди; створення систем, що думають раціонально; створення систем, що діють (поводять) себе раціонально [8, 35-36]. Отже, П. Норвіг та С. Рассел класифікують визначення ШІ слідуючи двом основам: відносяться вони до мислення чи до поведінки; співвідносяться вони до діяльності людини чи до раціональної діяльності. При цьому автори вказують, що раціональність для них це “правильне виконання всіх дій”, при умові, що попередньо відомо, що є правильним. Тобто, в таких умовах визначення ШІ залежить від підходу та відповідних цілей, яких намагаються досягти дослідники. Наприклад, П. Норвіг та С. Рассел розглядають ШІ, як сферу, присвячену розробці оптимальних програм для інтелектуальних агентів (intelligent agents), в умовах часової та просторової обмеженості машин, що реалізують дані програми. На основі даного визначення вони розвивають власну теорію інтелектуальних агентів, котрі застосовуються у визначених, обмежених умовах [35]. Тобто можемо прослідкувати, що початково сфера ШІ (“старий добрий ШІ”) ставила собі велику ціль створення програм, які б цілковито не відрізнялися від людини. Дана мрія піонерів галузі не справдилася, підтвердженням цього є ряд змістовних філософських критик показаних вище. Після, як показує власне позиція П. Норвіг та С. Рассел, дослідники зосередилися на більш вузьких завданнях: створення програм для вирішення конкретно поставлених завдань.

Перед тим, як ми подамо власне визначення, яким ми будемо послуговуватись в наступних розділах, варто звернути увагу і на певну неточність у використанні терміну “штучний інтелект”, яка пов’язана з аспектами використання слів в українській та англійській мовах.

Англійською даний термін подається як “artifficial inteligence”, при цьому слово “inteligence”, не є відповідником для “інтелект”, бо в англійській мові використовується і слово “intelect”. Якщо звернутися до словника філософії [12], то інтелект визначається як вища пізнавальна здатність мислення. “Inteligence”, слідуючи за визначенням самого творця терміну “штучний інтелект” Д. МакКарті [26] — це обчислювальна частина здатності досягати цілі у світі. Отже, “inteligence” не передбачає розуміння та осмисленості. Дана термінологічна плутанина стала однією з причин негативної конотації визначення ШІ[2].

Надалі ми будемо послуговуватися наступним визначенням. “Штучний інтелект” — це технологія, заснована на роботі комп’ютерних програм, що обмежено імітують деякі когнітивні здатності людини без їх розуміння. Якщо ми розглянемо ШІ під таким кутом, уникаючи гіперболізованих страхів, які більшою мірою засновані на хибних знаннях стосовно ШІ, то ми зможемо перейти до продуктивної філософської дискусії стосовно того, як дана технологія впливає та впливатиме на наше життя.

[2] Варто зазначити, що в межах даної роботи, ми окремо не аналізуємо термін "інтелект", однак його аналіз може слугувати самостійною темою історико-філософських досліджень [як приклад 2; 10;]. Ми залишаємо за собою право проведення порівняльного дослідження даного терміну, в контексті визначення "штучного інтелекту", в майбутньому.

РОЗДІЛ 2. Сучасні виклики у сфері досліджень та застосування “Штучного інтелекту”

2.1 П’ять парадоксів в розробках штучного інтелекту

Підсумувавши в попередньому розділі результати філософських розвідок основ ідеї “штучного інтелекту” та сформувавши на їх основі власне актуальне визначення в межах даної роботи, ми перейшли до визначення ШІ як технології. Це відкриває для нас поле для більш глибокої концептуальної проблематизації ШІ. Для цього ми залучаємо напрацювання з критичної теорії техніки, сформульовані та узагальнені американським філософом Ендрю Фінбергом (Andrew Feenberg) в його роботі “Десять парадоксів техніки” (Ten Paradoxes of Technology) [11]. Імплементуючи наші попередні напрацювання стосовно досліджень ШІ в теорію Е. Фінберга, ми спробуємо окреслити п’ять парадоксів в розробках штучного інтелекту, узагальнивши та доповнивши оригінальні ідеї автора.

1. Парадокс частини і цілого. Е. Фінберг описує даний парадокс в розумінні того, що складні цілісності на позір походять від своїх частин, але в реальності частини віднаходять своє походження у цілісності, до якої вони належать. Порівнявши дану думку з ідеями описаними в попередніх підрозділах ми бачимо, що дослідники ШІ зіткнулися з даним парадоксом в межах двох проблем.

Перша проблема проявила себе в рамках обчислювальної теорії свідомості, згідно з якою складні думки набувають свого значення із комбінацій простих компонентів. Слідуючи за положеннями даної теорії, ці компоненти формують цілісну думку, з’єднуючись певним чином разом. Проте розглядаючи лише частини, дослідження в межах Символьного ШІ, лишали позаду цілісність, яка оформлює та надає змісту таким компонентам. Розглядаючи лише синтаксис, дослідники ШІ нехтували семантикою. До прикладу Дж. Хаугеланд писав: “Якщо ви потурбуєтеся про синтаксис,

семантика сама про себе потурбується”. Проте ігнорування походження синтаксису з семантики призвело до проблем в межах досліджень ШІ, які яскраво виражені в критиці Джона Серля.

Друга проблема в розробках ШІ, яку ми вбачаємо в межах даного парадоксу, пов'язана з теорією конекціонізму. Дана теорія, яка розвинулася в межах когнітивних наук та тісно пов'язана зі сферою розробки ШІ, розглядає та моделює такі вищі когнітивні здатності людини, як наприклад “бачення”, “слух” чи “навчання” через ідею того, що вони є результатом взаємодії нейронів в мозку людини, і що їх можна точно відтворити, створивши відповідну штучну (комп'ютерну) нейронну мережу. Задля розкриття глибинних передумов даної проблеми слід звернутися до концептуальних основ нейронаук, які і стимулювали розвиток та заклали зі свого боку теоретичну базу для побудови нейронних мереж. Для цього варто згадати критичну працю “Філософські основи нейронаук” (Philosophical foundations of neuroscience) авторства філософа Пітера Гакера (Peter Hacker) та нейронауковця Макса Беннетта (Max Bennett) [31]. Філософські дослідження концептуальних основ нейронауки, яке проводять автори, направлене на розкриття та прояснення концептуальних істин, які являються умовою змістовності описання когнітивних нейронаукових відкриттів та теорій. В центрі їх дослідження знаходяться концепції “мереологічної помилки” (mereological fallacy) та концептуальної плутанини (conceptual confusion) в нейронаукових дослідженнях.

Випадки мереологічної помилки та випадки концептуальної плутанини мають спільну основу: беззмістовне приписування психологічних атрибутів мозку. Причиною цього, як пишуть М. Беннетт і П. Гакер, є засилля в когнітивних нейронауках Картезіанської концепції дуалізму. Перші два покоління нейронауковців двадцятого століття були загальновідомими Картезіанськими дуалістами, серед них, до прикладу, Чарлз Шеррінгтон

(Charles Sherrington), Джон Екклс (John Eccles) та Вайлдер Пенфілд (Wilder Penfield). Третє покоління залишило структуру Декарта, відкинувши субстанційний дуалізм, проте зберігши структурний. “Зараз нейронауковці приписують мозку такий самий набір ментальних предикатів, що й Декарт розуму, зберігши Декартове уявлення про зв'язок між думкою і дією, досвідом та його об'єктами” [31, 130-131]. Сприйняття ідей Декарта легітимізувало приписування психічних атрибутів мозку. Прикладами такого приписування можуть слугувати ідеї Френсіса Кріка (Francis Crick) про те, що в мозку відбувається багаторівнева інтерпретація візуальних сцен, Джеральда Едельмана (Gerald Edelman), що мозок категоризує, розрізняє і об'єднує все у глобальну карту, Коліна Блекмора (Colin Blakemore), в теорії якого знання знаходиться в нейронах, які оцінюють ймовірність зовнішніх подій, а мозок своєю чергою отримує ці знання шляхом індуктивного розмірковування. Також варто згадати теорію Джона Захарі Янга (John Zachary Young), згідно з якою мозок ставить питання, шукає відповіді та будує гіпотези. Гіпотези ж формуються через сигнали зі сітківки ока [31, 16-17].

Спираючись на ці та інші приклади, автори заявляють, що в подібному приписуванні психологічних атрибутів мозку немає жодного сенсу, бо, по-перше, в нас немає концептуальної основи для подібного роду приписування, й, по-друге, використовуючи терміни психології, нейронауковці змінюють їх вихідні значення, що й веде до концептуальної плутанини. “Ми здатні розпізнати, коли одна людина ставить запитання й коли інша на нього відповідає. Проте чи є в нас концепція того, як це для мозку запитувати чи відповідати на запитання?” Як наслідок, в розробках нейронауковців може виникнути мереологічна помилка. “Мереологія — це логіка стосунків між частиною і цілим. Помилку нейронауковців, що полягає в приписуванні атрибутів певним частинам живої істоти, які логічно можуть

бути застосовані тільки до живої істоти як цілого, ми називаємо “мереологічною помилкою” в нейронауках.” [31, 23]. Форма даної помилки була вказана ще Арістотелем в його вченні про душу: “Говорити, що душа гнівається, це те ж саме, що сказати, що душа займається ткацтвом чи будує будинок. Бо ж краще, напевно, не говорити, що душа співчуває або ж вчиться, або ж розмірковує, а говорити, що людина це робить душею” [1]. М. Беннетт і П. Гакер називають це “принципом Арістотеля”. Тож базою для розробки їхньої концепції слугує “принцип Арістотеля”, доповнений поглядами Л. Вітґенштайна стосовно того, що “тільки про людину і про те, що схоже (поводить себе як) на людину, можна сказати: у неї є відчуття; вона бачить; сліпа; чує; глуха; є свідомою чи несвідомою” [40, para. 281].

Узагальнюючи, ми бачимо, що слідом за нейронауковцями, конекціоністи хибують у спільний спосіб. Розгляд нейронних активностей в мозку не здатен прояснити діяльність цілісної людини як такої. Так само, як і спеціалісти з розробки нейронних мереж, які вдають, що звертаються до концепцій роботи людського мозку, насправді лиш редукують когнітивні здатності людини до роботи нейронних мереж. Оманливі імплікації у сферах розробки ШІ лиш посилюють парадокс частини і цілого. З першого парадоксу більшою мірою висновуються всі наступні.

2. Парадокс очевидного. Даний парадокс походить від попереднього. Його загальне формулювання за Е. Фінбергом: найочевидніше є найбільш прихованим. Наприклад, після тесту Тюрінга здавалося очевидним, що коли ми створимо комп'ютерну програму здатну проходити ТТ, то ми створимо інтелектуальну машину. Але це “очевидне” породило ряд дискусій, доповнень та інтерпретацій оригінального ТТ, тож очевидне постійно вислизає з нашого поля зору і потребує розгляду в більш широкому контексті проблематики.

3. Парадокс походження. Часом здається, що технології не пов'язані зі своїм минулим. В буденному житті вони постають для нас самодостатніми і зазвичай ми не маємо жодного уявлення про те, як вони з'явилися та розвивалися. Як пише Е. Фінберг: "Адекватне пояснення будь-якого пристрою полягає у простежуванні каузальних зв'язків між його частинами". І таке пояснення є згідно з автором хибним і, однозначно, недостатнім. Парадокс походження формулюється наступним чином: за всім раціональним стоїть забута історія. Однак, теперішнє засліплює бачення минулого, якщо не докладати критичних зусиль для прояснень. Присутність минулого у теперішньому видається нам непримітною. Освітлені вказівники "Вихід" над дверима сприймаються нами як даність. Їх появу пояснює забута трагічна історія, яка сталася в 1903 році в Чиказькому театрі "Ірокез", коли 600 людей загинули, намагаючись знайти вихід під час пожежі. Зазвичай ми не замислюємося над походженням таких буденних речей. В нашому розумінні вони перебувають там, де їм і місце. Та це лиш часткове пояснення, яке для філософів є недостатнім. "Штучний інтелект" теж для більшості постає самодостатнім. Проте незнання історико-концептуального розвитку ШІ стає причиною хибних уявлень та теорій, що стане очевидним в наступних підрозділах.

4. Парадокс рами. Даний парадокс походить від попереднього. Його формулювання є наступним: не ефективність пояснює успіх, а успіх пояснює ефективність. Хоча на перший погляд це здається контрінтуїтивним, бо технології ми розглядаємо через критерій успішності їх застосування. Ефективність технологій є мірилом їх цінності. Е. Фінберг обґрунтовує даний парадокс звертаючись до історії техніки. Більшість технологій, які ми воліємо називати ефективними, є наслідком початкового вибору, який надав успішній технології привілейованого статусу стосовно її альтернатив. Цей вибір здійснюють соціальні актори, що мають достатню владу для

здійснення даного вибору, при цьому до нього застосовуються різні критерії: економічні, соціальні, політичні, іноді технічні. Парадокс рами звертає увагу на те, що ефективність не пояснює присутність певних технологій в нашому житті. Це може нам прояснити лише вивчення історії техніки.

5. Парадокс дії. Е. Фінберг виводить його метафорично, трактуючи третій закон Ньютона, згідно з якого будь-яка дія породжує рівнозначну протидію. І як зазначає автор: “Кожна наша дія повертається до нас у формі зворотнього зв’язку з Іншим.” Парадокс постулює те, що діючи, ми стаємо об'єктами дії, і те, що люди здатні впливати лише на ту систему, частиною якої вони є самі. Тобто ми перебуваємо під дією техніки, яку самі й створили. Техніка приховує від нас три взаємозумовленості технічної дії: побічні ефекти техніки, зміни у значенні нашого світу та зміни у нашій ідентичності. Повертаючись до ШІ, в контексті такого впливу актуальним постає питання Джуліана Фрідленда (Julian Fridlend), яке він формулює наступним чином: “Як нам розробляючи (та використовуючи) розумні машини уберегти себе від уподібнення цим машинам?” Тож звернемося до міркувань Д. Фрідленда стосовно того, як технології на основі ШІ можуть впливати на людину [23].

Застосунки, які встановлені на наших смартфонах чи комп'ютерах, все частіше розробляються на основі технологій машинного навчання. По ходу того, як ми делегуємо багато задач ШІ, ми стаємо більш пасивними, менш рефлексивними та менш відповідальними за результат. Найпростіші приклади. Нам потрібно кудись піти, ми відкриваємо карти, і програма будує нам маршрут. Один маршрут. Можливо, цей маршрут і є оптимальним, та це вже не ми вибираємо шлях, яким іти. Так, це усуває фактор стресу, нам не потрібно комунікувати з іншими людьми та запитувати дорогу. Начебто нічого страшного, та це призводить до декваліфікації та погіршення

орієнтації в просторі. Проте, навчаючись орієнтуватися в новому некомфортному середовищі, налагоджувати в ньому взаємодії, ми отримуємо контроль, можливо й побічний, та все ж над власною свідомістю. Другий приклад — це рекомендаційні системи підбору фільмів. ШІ підбирає нам релевантні фільми, засновані на наших уподобаннях. Ми не витрачаємо часу на таку рутинну справу, але це погіршує наше "інтелектуальне тертя" та відповідальність за вибір. Додамо, що всі застосунки розробляються для того, щоб утримувати нашу увагу та направляються на те, щоб задіювати наш автономний розум, а не рефлексивний, як це розрізнює Деніел Канеман (Daniel Kahneman) [6]. Людина техніки, як зазначає Д. Фрідленд, стає менш критичною, легше піддається впливам, та найгірше — "оштучнюючи" себе, нівелює поняття відповідальності [23]. Тож хибно вважати, що ми можемо впливати на світ без наслідків для себе. Варто згадати і про приклад, який наводить Е. Фінберг для підсилення тези, що технології здатні змінювати значення нашого світу. Залізниці, автомобілі та літаки радикально змінили наше досвідне переживання відстані. Просторові координати нашого життя, які ми маємо на увазі під словами "близько" та "далеко", абсолютно відмінні від тих, які були до появи даних винаходів.

2.2 Етика штучного інтелекту

П'ять парадоксів техніки, які ми сформулювали в попередньому підрозділи на основі ідей Е. Фінберг, відкривають поле для більш детального розгляду його окремих пунктів. Штучний інтелект як технологія матиме значний вплив на розвиток людства в майбутньому. Хибно говорити про комп'ютерну революцію, яка більшою мірою застосовує ШІ як інструмент, так, неначе вона вже відбулася. Людство знаходиться лише на початку осмислення та застосування ШІ. Останнє своєю чергою піднімає ряд

фундаментальних питань про те, як нам використовувати дану технологію, які ризики можуть виникнути і як нам їх попередити. Такі технології як атомна енергетика чи автомобілі спричинили етичні та політичні дискусії, як правило лише після того, як завдали шкоди. Проте, приступаючи до розгляду етичних питань та проблем у сфері ШІ, ми знаходимося в ситуації певної невизначеності. В нас не було чітких прецедентів для того, щоб вибудувати теоретичний підхід до етики ШІ, тому розгляд даної проблематики характеризується можливістю уникнення проблем у майбутньому, хоча вплив даної технології відчутний вже зараз. Для прояснення етичної проблематики у сфері досліджень та застосування технологій ШІ, ми залучаємо останні напрацювання з етики ШІ описані Вінсентом Мюллером (Vincent C. Müller) в роботі "Етика штучного інтелекту та робототехніки" ("Ethics of Artificial Intelligence and Robotics") [29].

Сучасна розробка систем на основі технології ШІ у своїй основі послуговується конекціоністськими моделями нейронних мереж. Дослідники полишили спроби побудови комплексно завершеного ШІ та перейшли до створення ШІ для виконання чітко визначених завдань. З цією метою створюються нейронні мережі під певні цілі. Розробка такого ШІ пов'язана з технологією машинного навчання, яке потребує збору та обробки даних. Тож в даному контексті варто розглянути ряд важливих проблемних моментів.

Дані та ШІ. Як вже зазначалося в попередньому підрозділі, застосунки для смартфонів та комп'ютерів розробляються на основі методів машинного навчання, які засновані на роботі нейронних мереж, які здатні видобувати шаблони (паттерни) із заданих наборів даних. З використанням таких методів "навчання" фіксується закономірність в даних і вони позначаються таким чином, що це здається корисним для рішень, які приймає система. Використовуючи додатки, шукаючи інформацію в

Інтернеті та навіть просто пересуваючись вулицями, ми мимовільно продукуємо дані для роботи подібних систем. Одна з етичних проблем ШІ полягає в тому, що обробка даної інформації відбувається з метою маніпулювання нашою поведінкою. Найбільше зараз це відчутно в онлайн-рекламі. Збираючи інформацію про нашу “поведінку” в Інтернеті та на її основі “навчаючи” нейронні мережі, маркетологи створюють “dark patterns”. Це свого роду лабіринти, ціллю яких є привести нас до покупки певного товару або ж заплутати нас так, щоб ми більше часу використовували застосунки. Такого роду маніпуляції є бізнес-моделлю в ігорному бізнесі, але зараз вони активно використовуються в онлайн-продажах. Постає проблематика збору та використання подібних даних, що окреслюється темою “етики даних” (data ethics).

Конкретною проблемою постає те, що методи машинного навчання ШІ використовують для свого “навчання” велику кількість даних. Це означає, що часто існує компроміс між конфіденційністю та правами на дані і технічною якістю продукту. Останнє означає, що з однієї сторони якість реклами, яка персоналізовано відображається для нас, або ж точність побудови маршруту картами нерозривно пов’язано з тими даними, які збираються програмами для покращення подібного роду роботи. Це впливає на послідовну оцінку практики порушення конфіденційності. Дана тема тісно пов'язана з політикою і правом, що, на жаль, не є темою даного дослідження. Проте коротко зазначимо, що в даній проблематиці є свої злети та падіння: громадянські свободи та захист прав особистості знаходяться під впливом лобістів, спецслужб та інших державних та недержавних організацій. Державні та бізнес-суб’єкти збільшили свої здатності втручатися в приватне життя і маніпулювати людьми з допомогою технології штучного інтелекту. Й, на жаль, продовжуватимуть це робити для просування своїх власних конкретних інтересів, якщо це не буде

стримуватися політикою спрямованою на інтереси суспільства в цілому. В Європейському союзі лиш нещодавно запровадили “Загальний регламент про захист даних”, що посилив захист конфіденційності [29].

Одними з центральних питань “етики даних” чи “етики великих даних” (big data ethics) постає проблема “непрозорості” та “упереджень”. Як вже зазначалося вище, більшість систем ШІ використовують методи машинного навчання, в основі яких лежать нейронні мережі, які “видобувають” паттерни із заданих наборів даних. З використанням таких методів “навчання” фіксуються закономірності в даних і вони маркуються таким чином, що це здається корисним для рішень, які приймає система, однак програміст насправді не знає, які саме закономірності в даних використовувала система. “Навчаючись”, система розвивається на основі нових даних, які вона видобуває, або ж через внутрішній зворотній зв'язок (система встановлює власний ступінь релевантності різних закономірностей, які пасують або не пасують для її подальшої роботи). Тобто така програма постійно змінює модель власної роботи. Це означає, що результати “непрозорі” для програмістів і для людей, які використовують застосунки на основі технології ШІ. Крім того, якість роботи програми залежить від якості даних, з якими вона “працює”. Якщо програма на вході мала “упереджені” дані, наприклад дані поліції про колір шкіри підозрюваного, то результати роботи програми будуть значною мірою упереджені. Деякі дослідники стверджують, що етичні проблеми з ШІ є результатом технічних “ярликів”, які використовує ШІ. Упередженість, як правило, проявляється, коли людина формує неправильні судження, тому що на неї впливають характеристики, які насправді не мають відношення до питання (чи проблеми), яке розглядається. Одна із таких форм упередження — це когнітивна особливість людини, яка часто не проявляється явно. Окрім соціального феномену наукового упередження, когнітивна система людини, як правило,

піддається різним видам “когнітивних упереджень”, наприклад “підтверджувальне упередження” (confirmation bias). Люди схильні інтерпретувати інформацію так, щоб вона підтверджувала їх власні переконання в те, у що людина вже вірить. Виникає питання, чи сама система ШІ може мати такі когнітивні упередження. На нашу думку, не може. Це напряму пов’язане з творцем такої системи і з його власними переконаннями.

Варто зважати, що будь-який набір даних є “неупередженим” лише для одного роду проблем, тому просте створення набору даних пов’язане з небезпекою, що воно може бути використаним для іншого типу проблем, в якому може проявитися упередження саме для такого типу. Машинне навчання на основі таких даних не тільки не зможе розпізнати упередження, а й може систематизувати і автоматизувати його в “історичне упередження”. Проблема з такими системами знаходиться не лише в “упереджених даних”, а й в тому, що люди починають надмірно довіряти подібним системам. Технічні зусилля направлені на уникнення систематичних упереджень в системах штучного інтелекту перебувають на ранній стадії розгляду.

ШІ як автономна система. Існує декілька визначень “автономії” в контексті обговорення автономних систем. Сильне визначення використовують у філософських дебатах, в яких автономію визначають основою індивідуальності та відповідальності [19]. В даному контексті відповідальність означає автономію, але не навпаки. Тому ми можемо говорити про системи, які мають певні ступені автономії, не піднімаючи питання відповідальності. Слабке, або ж більш технічне визначення поняття “автономія” означає, що система в деякому ступені автономна у відношенні людського контролю [28]. Тут віднаходиться паралель з проблемами “упередженості” та “непрозорості” в ШІ, оскільки автономія стосується питання хто контролює і хто відповідальний. Питання автономії піднімає

питання не лише технічних аспектів побудови автономних систем, а й питання права та безпеки.

Розглянемо приклад автономної системи на базі ШІ, яка наразі знаходиться серед найбільш обговорюваних, а саме автономні (безпілотні) транспортні засоби. Як очікується, автономні транспортні засоби можуть великою мірою зменшити шкоду, яку наразі завдають водії. Мова тут йдеться не лише про забруднення навколишнього середовища та логістику, найважливішим в даному контексті являється те, що приблизно один мільйон людей гине щорічно на дорогах. Постає питання, як повинен себе "поводити" автономний автомобіль, і як слід розглядати питання ризиків та відповідальності при побудові таких систем. В даному контексті часто розглядають відому "Проблему вагонетки", при дослідженні теми відповідальності ШІ. Однак, "Проблема вагонетки" не передбачає опису реальних етичних проблем і їх вирішення з допомогою "правильного" вибору. Дана проблема використовується як теоретичний інструмент для дослідження етичних інтуїцій та теорій, проте, як зазначає В. Мюллер, насправді перед водіями на дорогах не постає такої дилеми [29]. Більш розповсюдженими етичними проблемами водіння є перевищення швидкості, ризиковий обгін, недотримання дистанції та інші, що постають в межах класичних проблем, які пов'язані з переслідуванням власних інтересів. Все це підпадає під дії правових норм. Таким чином, програмування автомобілів згідно з "правилами", а не для "досягнення максимальної корисності" зводиться до стандартної проблеми програмування етичних машин. Проблемним тут постає не лише вибір програміста, якою етикою послуговуватися (кантіанською чи утилітаристською), а й тема розробки "машинної етики" (Machine ethics) та проблема визначення "штучного морального агента" (Artificial moral agents).

Машинна етика — це етика для машин, для “етичних машин”. Для машин як суб’єктів, а не як таких, що використовуються людиною в якості об’єктів. На сучасному етапі дослідження даної теми не зрозуміло, чи вона має охоплювати всю етику ШІ, чи бути її частиною. Експлікуючи подібні міркування, сумнівним постає висновок, що коли машини діють етично значимим способом, то для цього потрібно розробити машинну етику. Деякі дослідники, зокрема Стів Торранс (Steve Torrance), використовують більш широке визначення машинної етики, яка в їхньому розумінні пов’язана із забезпеченням того, щоб “поведінка” машин у відношенні до користувачів, а також, можливо, щодо інших машин була етично прийнятною [37, 115]. Дане визначення також може включати питання безпечності продукту. Інші автори хоча й виглядають доволі амбіційними, однак послуговуються більш вузьким визначенням даного поняття. Зокрема Віргінія Дігнум (Virginia Dignum) пише про те, що ШІ у своїх міркуваннях має бути здатним враховувати соціальні цінності, моральні та етичні міркування, а також має зважувати відповідні ціннісні пріоритети, які мають різні зацікавлені сторони в різноманітних мультикультурних контекстах [20]. Також В. Дігнум наголошує на прозорості систем ШІ, а також на важливості того, щоб система самостійно пояснювала власні міркування. В наявних дискусіях, присвячених машинній етиці, формується припущення, що машини здатні, в певному сенсі, бути етичними агентами, такими, що є відповідальними за власні дії, тобто визначатися як “автономні моральні агенти” (autonomous moral agents). Якщо співвіднести машинну етику та теорію морального агента, то в певному субстанційному сенсі таких агентів можна назвати “штучними моральними агентами”, що мають права та обов’язки. Проте ми з вами маємо розуміти та пам'ятати, що саме програмісти чи користувачі систем ШІ приймають моральні рішення.

Прихильники теорії “автономних моральних агентів” стверджують, що машинна етика виходить за рамки того, як інженери усвідомлюють цінності, які вони закладають при створенні продуктів на основі ШІ. Джеймс Мур (James Moor) для того, щоб додатково визначити, що означає для комп’ютера приймати етичні рішення, проводить розрізнення трьох типів етичних агентів: неявні етичні агенти (implicit ethical agents), явні етичні агенти (explicit ethical agents) та повні етичні агенти (full ethical agents) [27]. Перший тип агентів — це комп’ютери, в основу роботи яких закладена етика розробника. Ці агенти запрограмовані (сконструйовані) так, щоб дотримуватися норм і цінностей контексту, в якому вони розроблені чи будуть використовуватися. Явні етичні агенти — це комп’ютерні програми, як визначає Д. Мур, що можуть на основі етичної моделі визначити, що було б правильним вчинити в тій чи іншій ситуації, враховуючи певні дані на вході. Етична модель може бути заснована на традиційних етичних теоріях, таких як кантіанська чи утилітаристська, в залежності від уподобань творців програми. Явні етичні агенти здатні “приймати етичні рішення” від імені своїх творців. Визначення такого роду агентів більшою мірою подібне до визначення “автономних моральних агентів”. Повні етичні агенти можуть формувати етичні судження та обґрунтовувати їх так само, як і люди. Д. Мур зазначає, що хоча в цей момент розвитку технологій немає такого ШІ, який би можна було назвати повністю етичним агентом, однак це постає питанням, чи існує взагалі можливість його створення в майбутньому.

2.3 Переосмислення концепту моральної відповідальності

Експлікуючи попередні міркування, наявним постає завдання перегляду визначення відповідальності. Традиційні філософські дискусії щодо моральної відповідальності були зосереджені на людському компоненті моральної дії. Моральна відповідальність стосується дій

людини, її намірів та наслідків. Узагальнюючи, людина або група людей несуть моральну відповідальність, коли їх вільна дія має морально значливий результат. Приписування моральної відповідальності встановлює зв'язок між людиною (групою людей) і тим, що може впливати на її дію. Людину (групу людей) дії якої спричинили певні наслідки, називають агентом. Людину (групу людей), які знаходяться під впливом дій агентів, називають пацієнтами [33]. Встановлення зв'язків між агентом та пацієнтом у моральному відношенні може здійснюватися ретроспективно і перспективно. Тобто, приписування відповідальності залучає пояснення того, хто був винен, і кого потрібно покарати. У розгляді перспективи, ми визначаємо обов'язки та дії людини, яких вона має дотримуватися чи виконувати у майбутньому. Проте нечітким залишається визначення того, при яких обставинах варто приписувати моральну відповідальність. Дана ситуація спричинила дискусії не лише у відношенні моральної відповідальності щодо діяльності людей. Продукти, в основу роботи яких закладена технологія ШІ, викликали ряд нових питань. Вище були розглянуті питання "автономії" та "непрозорості" в роботі ШІ, коли проблематичним постає визначення того, хто відповідальний — творець чи програма. Алгоритми, які закладаються в нейронні мережі, через внутрішню "непрозору" для самого програміста роботу системи постійно змінюються самою системою, тож програмісти стають нездатними обґрунтувати результати, яких "самостійно" досягла програма. Це спонукає дослідників говорити про певний ступінь автономії таких систем, аж до приписування їм відповідальності. В даному контексті варто розглянути ідеї Дональда Готтербарна (Donald Gotterbarn), який пропонує переосмислити те, як приписується моральна відповідальність [24], а також їх рецепцію в роботі Мерел Нурман (Merel Noorman) [33].

Беручи до розгляду спеціалістів з обчислювальної техніки, Д. Готтербарн прослідковує можливості обходу питання їх відповідальності знайшовши тих, хто може її понести за них. Д. Готтербарн вбачає дану можливість у двох поширених хибних уявленнях щодо відповідальності. Перша хиба стосується того, що обчислення розглядають як етично нейтральну практику. Слідуючи думці Д. Готтербарна, хибною є думка, що технологічні артефакти і методи їх створення є етично нейтральними. Варто враховувати широкий контекст, в якому дані технології використовуються чи можуть бути використаними. Якщо інженер буде знаходитися лише в рамках створення нової технології, то таке вузьке фокусування може мати фатальні наслідки. Д. Готтербарн наводить як приклад програміста, завданням якого було написати програму для Рентген апарату, що мала піднімати та опускати даний пристрій. Зосередившись лише на написанні програми, і не врахувавши обставини і можливі наслідки її роботи, діяльність програміста спричинила смерть людини. Д. Готтербарн наголошує, що комп'ютерні спеціалісти несуть моральну відповідальність за врахування непередбачуваних обставин, навіть якщо це не є основною частиною їх роботи. Розробка та використання технологічних артефактів — це моральна діяльність, і вибір одного інженерного рішення, а не іншого, має реальні наслідки [24].

Друга хиба пов'язана з тим, що питання відповідальності виникає лиш тоді, коли намагаються визначити винного, якщо щось пішло не так. Комп'ютерні спеціалісти умовно прийняли модель відповідальності, яка стосується протиправних дій. Така модель дає змогу ухилитися від відповідальності. Відстань між розробником та наслідками використання технології, яку він створив, може слугувати аргументом на користь того, що немає безпосереднього причинно-наслідкового зв'язку. Також розробники можуть апелювати, що вони були частиною команди, і їх вклад є незначним,

або ж їх можливості були обмеженими та диктувалися компанією. Згідно з Д. Готтербарном, така модель спонукає спеціалістів дистанціюватися від відповідальності та звинувачення.

Дані хиби стосуються особливого уявлення про відповідальність, в якому акцент робиться на уникненні звинувачень та звільненні від відповідальності. Експлікуючи, Д. Готтербарн позначає це як “негативну відповідальність” і відділяє її від “позитивної відповідальності”. Остання своєю чергою характеризується необхідністю враховувати наслідки діяльності стосовно інших (людей). В контексті позитивної відповідальності, професіоналізм інженерів полягає у тому, скільки зусиль вони докладають задля мінімізації небажаних наслідків. Основна увага тут приділяється не уникненню відповідальності за безвідповідальну поведінку, а тому, що потрібно зробити на благо майбутнього. Д. Готтербарн спонукає комп’ютерних спеціалістів прийняти позитивну концепцію відповідальності, бо саме в ній підкреслюються обов’язок практиків сфери технологій враховувати наслідки їх діяльності та зведення до мінімуму можливої шкоди. Творці технологій несуть моральну відповідальність за те, щоб створений ними продукт не завдавав шкоди.

Попри те, що програміст, створивши нейронну мережу, може бути не здатним пояснити, як програма прийшла до певних висновків, хибно розглядати такий технологічний артефакт як “автономного морального агента”. Причиною “непрозорості” роботи систем ШІ є людина. Відповідальність цілковито і повністю лежить на творцях систем ШІ. В даному контексті важливо згадати моральний імператив для техніко-технологічної діяльності, сформульований філософом Анатолієм Єрмоленком: “Чини так, щоб максима твоєї технічної дії, проекту, винаходу, технології та виробництва, які впроваджуватимуться у життя, а також наслідки та побічні наслідки, що здогадно випливатимуть з їх застосування,

могли б бути без примусу прийняті всіма, кого це стосується, в процесі арґументації, в якому не залишався б поза увагою жодний раціональний аргумент!" [4, 326] Також А. Єрмоленко, погоджуючись з іншими класичними авторами комунікативної практичної філософії, етики відповідальності [про це див. 5], наголошує на принципі універсалізації ("для всіх", в тому числі і для майбутніх поколінь), щодо техніки та технологій. Однак аргументація залежить від контексту, в якому її наводять. Наявними постають питання обізнаності користувачів та творців не лише стосовно внутрішньо технічної роботи програм ШІ, а й ідейних передумов створення таких артефактів.

Штучний інтелект відображає саме ті цінності та цілі, які в нього закладають творці. Він буде діяти і вже діє згідно з тими цілями, для яких був створений. Прихильники концепції "автономних моральних агентів" значною мірою хибують, коли в "непрозорості" системи вбачають її самостійність. Проте саме "непрозорість" роботи деяких нейронних мереж має спонукати нас до розмислів, для яких цілей їх застосовувати, якщо потім не буде змоги пояснити причинно-наслідкові зв'язки. Етика та відповідальність тут відіграє важливу роль не для систем ШІ, а для тих, хто їх створює, а також керується результатами роботи подібних програм для прийняття рішень. Людиноцентричність має знаходитися в центрі таких побудов, тому варто уникати будь-якої невизначеності при поширенні та впровадженні подібних технологій. Делегуючи деяку важливу мисленнєву роботу ШІ, варто зважати на те, що небезпека полягає не в тому, що ШІ стає "розумнішим" і його предикативні здатності в деяких питаннях перевершують можливості людини, а в тому, що люди очікують від ШІ більшого, аніж те, на що він здатний, і можуть покладатися на програму в ситуаціях, в яких цього краще уникнути. Завжди варто пам'ятати про "штучність" таких програм, що вказує на їх інструментальність. Штучний

інтелект — це технологічний інструмент. Як для будь-якого інструменту, мають бути розроблені правила використання, а творцями має бути сприйнята “позитивна відповідальність” стосовно створення подібних технологій.

ВИСНОВКИ

У ході роботи нами було виконано поставлені на початку завдання. Підсумовуючи, на початку ми окреслили ідею “штучного інтелекту” в історико-контекстуальному вимірі. Як погоджуються більшість зі згаданих дослідників, Алан Тюрінг був першим, хто спровокував запитування щодо “мислячих машин”. Те, що означив А. Тюрінг, надовго закріпилося в розробках дослідників, а саме той момент, що “мислення” або ж “інтелект” людського рівня можна зімітувати за допомогою логіко-математичних операцій. Дана ідея була підхоплена учасниками Дартмутської конференції, на якій вперше прозвучав сам термін “штучний інтелект”. Ідея “обчислюваності” ментальних здатностей людини була редукована до свідомості. Дослідники А. Ньюел та Г. Саймон почали розглядати свідомість людини за аналогією з комп’ютером. Буцімто, він так само зберігає дані в пам’яті і здійснює маніпуляції із символами подібного до того, як це відбувається в людській свідомості та мозку. Комплекс ідей, які розглядають свідомість як “синтаксичний двигун”, а ментальні здатності людини як оперування з логічними та математичними символами, називають “старим добрим штучним інтелектом”. Дані підходи, які були засновані на хибних уявленнях про людину та її свідомість, застосовуючи методи біґевіоризму та функціоналізму, зазнали нищівної філософської критики.

В контексті критики ШІ, було розглянуто аргументи Дж. Серля, Н. Блока та Г. Дрейфуса, які по-різному, проте в схожий спосіб підійшли до ШІ. Загалом, автори наголошують на важливості семантики і розгляду людини та її життєвих практик як цілісності, яку неможливо імітувати та відтворювати за допомогою обчислень. Візії піонерів галузі щодо комплексного відтворення вищих когнітивних здатностей людини в їх цілісності за допомогою обчислень були відкинуті як не виправдані. Однак когнітивна революція, яка спричинила активні дослідження роботи мозку,

стала причиною конекціоністської концепції ШІ. Калькуючи роботу нейронів, було зроблено поступ в розробці комп'ютерних нейронних мереж. Дослідники ШІ перейшли від спроб побудови “завершеного” (сильна версія) ШІ, який претендував на розуміння та усвідомлення людського рівня, до побудови ШІ як програми для виконання чітко визначених задач. В термінах Дж. Серля, в дослідженнях відбувся перехід від “сильної версії ШІ” до активної розробки “слабкої версії ШІ”. В даному контексті, пропонуємо розглядати “штучний інтелект” як технологію, що заснована на роботі комп’ютерних програм, які обмежено імітують деякі когнітивні здатності людини без їх розуміння для досягнення наперед визначених розробниками цілей. Проблема в тому, що не всі наперед визначення є досконалими. ШІ як технологія впливає на людину і, як зазначає Е. Фінберг: “Кожна наша дія повертається до нас у формі зворотнього зв’язку...”.

В такому ключі постає питання, хто відповідальний за таку дію, якщо використовується ШІ. Незважаючи на очевидність відповіді, використання нейронних мереж розмило відповідь на дане запитання. Нейронні мережі, які наразі є невіддільною частино ШІ, (аж до синонімічного вживання даних термінів) у внутрішній своїй роботі постають для інженерів, які їх створюють, “непрозорими”. Такі системи працюють з неймовірною кількістю даних і в алгоритмі їх побудови закладене внутрішнє визначення релевантності певних даних. Система постійно змінює свою роботу, перебираючи та аналізуючи дані на вході, тож відповідь стає неочевидною для творця такої програми. Не можна просто встановити причинно-наслідкові зв'язки поглянувши на те, як програма працювала. Для людини це стає неможливим через величезну кількість внутрішніх операцій. З “непрозорості”, як було побачено, хибно експлікувалося визначення автономності таких систем аж до приписування їм відповідальності.

Розглядаючи ШІ як “автономного морального агента”, дослідники почали розробляти етику для подібних програм щоб встановити, чим для них є прийняття рішень. Подібні міркування значною мірою нагадують ідеї, в яких маніпуляції символами програмою сприймалися за усвідомлене діяння. Ігнорування цілісного характеру людської дії та семантичного характеру мисленнєвої діяльності показує, що історія в дослідженнях ШІ повторюється. В “непрозорості” роботи нейронних мереж починають вбачати можливість свідомості. Така оманлива можливість розмиває межі відповідальності творців подібних програм, які “непрозорість” редукують до автономії та відповідальності. Однак творці цим самим ігнорують власну відповідальність за “дії” подібних програм, забуваючи про “штучність” того, що вони створюють. Завжди варто пам’ятати, що причиною “непрозорості” системи є сама людина. І хоча тяжко прояснити шлях причинно-наслідкових зв’язків, яким “рухалася” програма, ми точно знаємо, що причиною створення програми була робота людини. У філософському осмисленні важливою постає концепція “позитивної відповідальності”, запропонована для творців ШІ, в якій йдеться про моральний обов’язок практиків сфери технологій враховувати та аналізувати можливі наслідки їх діяльності. Створюючи та застосовуючи технології ШІ, варто бути прозорливими, робити зупинки для осмислення питання “навіщо” і того, які це буде мати наслідки у майбутньому.

В контексті даної роботи, важливою постає роль філософів. Їх завданням у сфері ШІ постає детальний аналіз тих “ідей” минулого, якими керувалися дослідники “старого доброго штучного інтелекту”, та які мають ідейний вплив на сучасників. “Штучність” — невіддільна ознака сфери досліджень ШІ, яка вказує на інструментальність даних теоретико-практичних побудов. ШІ — це технологічний інструмент, який не лише полегшує життя, а й допомагає порозмислити над людиною, над тією

роллю, яку вона відіграє в створенні ШІ. Від самого початку розробок “штучного інтелекту”, це була сфера мрій, і саме філософи, хоча й з допомогою критики, та все ж примирили людей із мріями минулого. Вони показали, що ШІ — це не “замість”, а “для” людини. Повторимось, штучний інтелект є інструментом, а для кожного інструмента мають бути правила створення та правила використання. Якими вони будуть — це і є завданням для роботи філософів сьогодення, а дана робота є лиш відправною точкою для подальших розробок.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ ТА ДЖЕРЕЛ

1. Аристотель. О душе [Електронний ресурс] / Аристотель // Мысль. – 1976. – Режим доступу до ресурсу: http://psylib.org.ua/books/arist01/index.htm.
2. Де Лібера А. INTELLECTUS Інтелект, здатність розуміти, значення, сенс / Ален Де Лібера // Європейський словник філософій: Лексикон неперекладностей / Ален Де Лібера. – Київ, 2011. – (ДУХ І ЛІТЕРА). – С. 96–107.
3. Дрейфус Х. Чего не могут вычислительные машины. Критика искусственного разума / Хьюберт Дрейфус. – Москва, 2010. – 340 с. – (Либроком).
4. Єрмоленко А. Соціальна етика та екологія. Гідність людини — шанування природи. / Анатолій Єрмоленко. – Київ, 2010. – 416 с. – (Лібра).
5. Йонас Г. Принцип відповідальності. У пошуках етики для технологічної цивілізації / Ганс Йонас. – Київ, 2001. – 400 с. – (Лібра).
6. Канеман Д. Мислення швидке й повільне / Деніел Канеман. – Київ, 2017. – 480 с. – (Наш Формат).
7. Маркофф Д. Homo Roboticus? Люди и машины в поисках взаимопонимания / Джон Маркофф. – Москва, 2017. – 408 с. – (Альпина Паблишер).
8. Рассел С. Искусственный интеллект. Современный подход / С. Рассел, П. Норвиг. – Москва, 2007. – 1408 с. – (Вильямс).
9. Серль Д. Розуми, мозки і програми / Джон Серль // Антологія сучасної аналітичної філософії, або жук залишає коробку / Джон Серль. – Львів, 2014. – (Літопис). – С. 229–255/
10. Фає Е. ІНТЕЛЕКТ, ІНТЕЛІГУВАТИ / Емануель Фає // Європейський словник філософій: Лексикон неперекладностей (Том другий) / Емануель Фає. – Київ, 2011. – (ДУХ І ЛІТЕРА). – С. 81–83.

11. Фінберг Е. Десять парадоксів техніки / Ендрю Фінберг // Антологія сучасної філософії науки, або усмішка ASIMO / Ендрю Фінберг. – Львів, 2017. – (ЛНУ ім. І. Франка). – С. 258–281.
12. Шинкарук В. Філософський енциклопедичний словник / Володимир Шинкарук. – Київ, 2002. – 742 с. – (Абрис).
13. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955 / J. McCarthy, M. Minsky, N. Rochester, C. Shannon. // AI Magazine. – 2006. – №27. – С. 12–14.
14. Arkoudas K. Philosophical foundations / K. Arkoudas, S. Bringsjord // The Cambridge handbook of artificial intelligence / K. Arkoudas, S. Bringsjord. – Cambridge, 2014. – (Cambridge University Press). – С. 34–63.
15. Block N. Troubles with functionalism / Ned Block // Perception and Cognition: Issues in the Foundations of Psychology / Ned Block. – Minneapolis, 1978. – (University of Minnesota Press). – С. 261–325.
16. Bringsjord S. Artificial Intelligence [Електронний ресурс] / Selmer Bringsjord // Stanford Encyclopedia of Philosophy. – 2018. – Режим доступу до ресурсу: https://plato.stanford.edu/entries/artificial-intelligence/#LogiBaseAISomeSurgPoin.
17. Brooks R. Intelligence Without Reason [Електронний ресурс] / Rodney Brooks // MIT Artificial Intelligence Laboratory. – 1991. – Режим доступу до ресурсу: https://dspace.mit.edu/bitstream/handle/1721.1/6569/AIM-1293.pdf.
18. Carter M. Minds and computers: An introduction to the philosophy of artificial intelligence: An introduction to the philosophy of artificial intelligence / Matt Carter. – Edinburgh, 2007. – 222 с. – (Edinburgh University Press).
19. Christman J. Autonomy in Moral and Political Philosophy [Електронний ресурс] / John Christman // Stanford Encyclopedia of Philosophy. – 2015. – Режим доступу до ресурсу: https://plato.stanford.edu/archives/spr2018/entries/autonomy-moral/.

20. Dignum V. Ethics in Artificial Intelligence: Introduction to the Special Issue / Virginia Dignum. // Ethics and Information Technology. – 2018. – №20. – С. 1–3.
21. Floyd J. Wittgenstein's Diagonal Argument: A Variation on Cantor and Turing / Juliet Floyd // Epistemology versus Ontology. Logic, Epistemology, and the Unity of Science / Juliet Floyd. – Dordrecht, 2012. – (Springer). – С. 25–44.
22. Fodor J. Tom Swift and his procedural grandmother [Електронний ресурс] / Jerry Fodor // Cognition. – 1978. – Режим доступу до ресурсу: http://www.nyu.edu/gsas/dept/philo/courses/mindsandmachines/Papers/tomswift.pdf.
23. Friedland J. AI can help us live more deliberately [Електронний ресурс] / Julian Friedland // MIT Sloan Management Review. – 2019. – Режим доступу до ресурсу: https://www.academia.edu/39520151/AI_Can_Help_Us_Live_More_Deliberately .
24. Gotterbarn D. Informatics and professional responsibility / Donald Gotterbarn. // Science and Engineering Ethics. – 2001. – №7. – С. 221–230.
25. Haugeland J. Artificial Intelligence: The Very Idea / John Haugeland. – Cambridge, 1985. – 302 с. – (MIT Press).
26. McCarthy J. What is Artificial Intelligence? [Електронний ресурс] / John McCarthy // Stanford University. – 2007. – Режим доступу до ресурсу: http://www-formal.stanford.edu/jmc/whatisai.pdf.
27. Moor J. The Nature, Importance, and Difficulty of Machine Ethics / James Moor. // IEEE Intelligent Systems. – 2006. – №21. – С. 18–21.
28. Müller V. Autonomous Cognitive Systems in Real-World Environments: Less Control, More Flexibility and Better Interaction / Vincent Müller. // Cognitive Computation. – 2012. – №4. – С. 212–215.

29. Müller V. Ethics of Artificial Intelligence and Robotics [Електронний ресурс] / Vincent Müller // The Stanford Encyclopedia of Philosophy. – 2020. – Режим доступу до ресурсу: https://plato.stanford.edu/entries/ethics-ai/#OpacAISyst
30. Müller V. Philosophy and Theory of Artificial Intelligence 2017 / Vincent Müller. – Leeds, 2018. – 320 с. – (Springer Nature Switzerland AG).
31. Neuroscience and Philosophy. Brain, Mind, and Language / M. Bennett, D. Daniel, H. Peter, J. Searle. – New York, 2009. – 232 с. – (Columbia University Press).
32. Newell A. Computer science as empirical inquiry: Symbols and search / A. Newell, H. Simon. // Communications of the Association for Computing Machiner. – 1976. – №19. – С. 113–126.
33. Noorman M. Computing and Moral Responsibility [Електронний ресурс] / Merel Noorman // Stanford Encyclopedia of Philosophy. – 2018. – Режим доступу до ресурсу: https://plato.stanford.edu/entries/computing-responsibility/#RetConMorRes.
34. Putnam H. Minds and machines [Електронний ресурс] / Hilary Putnam // New York University Press. – 1960. – Режим доступу до ресурсу: https://philpapers.org/archive/PUTMAM.pdf.
35. Russell S. Rationality and intelligence / Stuart Russell. // Artificial Intelligence. – 1997. – №94. – С. 57–77.
36. Sun R. Connectionism and neural networks / Ron Sun // The Cambridge handbook of artificial intelligence / Ron Sun. – Cambridge, 2014. – (Cambridge University Press). – С. 108–127.
37. Torrance S. Machine Ethics and the Idea of a More-Than-Human Moral World / Steve Torrance // Machine Ethics / Steve Torrance. – Cambridge, 2011. – (Cambridge University Press). – С. 115–137.

38. Turing A. Computing machinery and intelligence [Електронний ресурс] / Alan Turing // Mind. – 1950. – Режим доступу до ресурсу: https://academic.oup.com/mind/article/LIX/236/433/986238.
39. Wetzel L. Types and Tokens [Електронний ресурс] / Linda Wetzel // Stanford Encyclopedia of Philosophy. – 2006. – Режим доступу до ресурсу: https://plato.stanford.edu/entries/types-tokens/.
40. Wittgenstein L. Philosophical Investigations [Електронний ресурс] / Ludwig Wittgenstein // Basil Blackwell. – 1958. – Режим доступу до ресурсу: https://static1.squarespace.com/static/54889e73e4b0a2c1f9891289/t/564b61a4e4b04eca59c4d232/1447780772744/Ludwig.Wittgenstein.-.Philosophical.Investigations.pdf.