МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ТЕРНОПІЛЬСЬКИЙ НАЦІОНАЛЬНИЙ ПЕДАГОГІЧНИЙ УНІВЕРСИТЕТ ІМЕНІ ВОЛОДИМИРА ГНАТЮКА

Історичний факультет
Кафедра філософії та суспільних наук

Кваліфікаційна робота

ФІЛОСОФІЯ ШТУЧНОГО ІНТЕЛЕКТУ: СВІДОМІСТЬ, ІНТЕЛЕКТ ТА СУЧАСНІ СИСТЕМИ

Спеціальність 033 Філософія
ОПП «Аналітика суспільних процесів»

Здобувача другого (магістерського)
рівня вищої освіти
Олексійка Юрія Руслановича

НАУКОВИЙ КЕРІВНИК:
кандидат філософських наук, доцент
Морська Наталія Львівна

РЕЦЕНЗЕНТ:
доктор філософських наук, професор
Чолач-Гончарук Тетяна Вікторівна

Тернопіль 2025

ЗМІСТ

ВСТУП

Актуальність теми дослідження. Стрімкий розвиток технологій штучного інтелекту (ШІ), зокрема поява та масова інтеграція великих мовних моделей (Large Language Models, LLM), таких як GPT, Claude та Gemini, стали визначальною подією не лише для технічної сфери, а й для сучасної філософії. Ці системи демонструють безпрецедентну здатність до генерації текстів, ведення діалогу та вирішення когнітивних завдань, що раніше вважалися прерогативою людини. Це породжує ілюзію розуміння та суб'єктності, ставлячи перед філософією фундаментальні запитання про природу свідомості, межі пізнання та статус розуму у цифрову епоху.

Актуальність теми зумовлена необхідністю філософського переосмислення класичних категорій «знання», «істина», «творчість» та «відповідальність» у контексті функціонування алгоритмічних систем. Традиційні дихотомії, такі як «суб'єкт – об'єкт» або «мислення – обчислення», виявляються недостатніми для опису феноменів, що виникають на перетині людської інтенціональності та машинної статистики. Особливої гостроти набуває проблема так званих «галюцинацій» ШІ, які є не просто технічними помилками, а свідченням нової епістемологічної ситуації, де правдоподібність заміщує істину, а синтаксична когерентність симулює семантичне розуміння.

У сучасному науковому дискурсі домінує тенденція до антропоморфізації ШІ або, навпаки, до його суто інструментального тлумачення. Проте обидва підходи спрощують онтологічну складність феномену. Необхідний глибокий філософський аналіз, який би поєднав історико-філософську традицію дослідження свідомості з розумінням архітектури сучасних нейромереж. Дослідження метафізичного статусу ШІ – як інструмента, агента чи «квазі-суб'єкта» – є критично важливим для побудови етичних та правових рамок взаємодії людини з технологіями майбутнього.

Зв'язок роботи з науковими програмами, планами, темами. Магістерська робота виконана в межах науково-дослідної роботи кафедри філософії та суспільних наук Тернопільського національного педагогічного університету імені Володимира Гнатюка «Духовно-моральні, культурно-етичні та соціально-економічні засади розвитку України.» і відповідає пріоритетним напрямам дослідження за спеціальністю 033 «Філософія».

Мета і завдання дослідження. Метою роботи є з'ясування онтологічного та епістемологічного статусу сучасних систем штучного інтелекту (LLM) через порівняльний аналіз машинної імітації когнітивних процесів та феноменальної свідомості людини.

Для досягнення поставленої мети необхідно вирішити такі завдання:

1. Проаналізувати еволюцію філософських підходів до природи розуму та свідомості в класичній (Р. Декарт, І. Кант), феноменологічній (Е. Гуссерль, М. Мерло-Понті) та аналітичній (Дж. Серл, Д. Чалмерс) традиціях.
2. Дослідити архітектуру та принципи функціонування великих мовних моделей (трансформери, RLHF) як об'єкта філософського аналізу.
3. Здійснити критичний аналіз функціоналізму в контексті проблеми «Розуміння» та аргументу «Статистичного папуги».
4. Розкрити сутність проблеми «вкоріненості» (grounding) та відсутності «тілесного» досвіду у мовних моделях.
5. Визначити епістемологічну природу «галюцинацій» ШІ та відмінність між машинним «знанням» (як статистичними вагами) і людським переконанням.
6. Проаналізувати феномен машинної творчості та визначити можливості створення онтологічно нового у статистичних системах.
7. Встановити метафізичний статус ШІ (інструмент, агент, квазі-суб'єкт) та окреслити етико-правові проблеми відповідальності за дії автономних алгоритмів.

Об’єктом дослідження є феномен штучного інтелекту, реалізований у формі великих мовних моделей (LLM) та генеративних нейромереж.

Предметом дослідження є онтологічні, епістемологічні та етичні виміри функціонування свідомості, знання та агентності у штучних інтелектуальних системах.

Методи дослідження. У роботі використано комплекс філософських та загальнонаукових методів. Історико-філософський метод дозволив простежити еволюцію понять «розум» і «свідомість» від класики до сучасності. Феноменологічний метод застосовано для аналізу структур інтенціональності, тілесності та суб’єктивного досвіду, відсутність яких є ключовою відмінністю ШІ від людини. Герменевтичний метод використано для інтерпретації текстів, генерованих ШІ, та аналізу їх смислового наповнення. Метод компаративного аналізу дав змогу зіставити людське пізнання та машинне навчання. Структурно-функціональний метод застосовано для аналізу архітектури нейромереж. Також використано метод мисленнєвого експерименту (аналіз кейсів «Китайська кімната», «Jailbreaks») для верифікації філософських гіпотез.

Наукова новизна одержаних результатів полягає у комплексному філософському осмисленні феномену великих мовних моделей як «квазі-суб’єктів», що формують нову онтологічну реальність.

- Вперше систематизовано аргументи проти наявності розуміння у LLM через поєднання феноменологічної критики безтілесності та технічного аналізу ймовірнісної генерації.
- Удосконалено епістемологічне визначення «галюцинацій» ШІ не як помилки, а як структурної властивості ймовірнісного моделювання, що продукує «симулякри знання».
- Набуло подальшого розвитку розуміння агентності в технічних системах через розмежування функціональної автономності та моральної відповідальності.

Практичне значення одержаних результатів. Матеріали магістерської роботи можуть бути використані при розробці навчальних курсів з філософії техніки, етики штучного інтелекту, філософії свідомості, а також у практичній діяльності розробників ШІ для врахування етичних обмежень систем. Висновки дослідження сприяють формуванню критичного мислення користувачів щодо взаємодії з генеративними моделями та розуміння меж їхньої компетентності.

Апробація результатів дослідження. Результати дослідження були апробовані у збірнику «Магістерський науковий вісник» випуск №44, 2025, «ФІЛОСОФІЯ МОЖЛИВОСТЕЙ ШТУЧНОГО ІНТЕЛЕКТУ»**;** випуск №45, 2025, «ФІЛОСОФІЯ ШІ: ГАЛЮЦИНАЦІЇ, УПЕРЕДЖЕННЯ, РИЗИКИ ТА СТРАТЕГІЇ ЇХ МІНІМІЗАЦІЇ» та у виступах на Всеукраїнській науковій конференцій магістрантів «Нехай не гасне світ науки».

Структура роботи. Магістерська робота складається зі вступу, трьох розділів, висновків до кожного розділу, загальних висновків та списку використаних джерел. Загальний обсяг роботи становить 77 сторінок.

РОЗДІЛ 1.

ТЕОРЕТИКО-МЕТОДОЛОГІЧНІ ЗАСАДИ ТА ТЕХНІЧНИЙ БАЗИС ДОСЛІДЖЕННЯ СВІДОМОСТІ

1.1. Еволюція філософських підходів до природи розуму та свідомості

1.1.1. Класичні дихотомії: Проблема «душі і тіла» (Декарт, *res cogitans*); Трансценденталізм (Кант) та роль апріорних структур розуму.

Проблема співвідношення душі й тіла, свідомості й матерії є однією з центральних у філософії та залишається ключовою навіть у контексті сучасних наукових досліджень штучного інтелекту. Зародившись у давньогрецькій натурфілософії, ця дихотомія набуває системного вираження у Новий час, зокрема у працях Рене Декарта, який закладає основи дуалістичної отології свідомості.

Декарт у «Meditationes de prima philosophia» формолює принципову відмінність між двома субстанціями: мислячою (res cogitans) та протяжною (res extensa). Мисляча субстанція – це носій свідомості, саморефлексії та інтенціональних актів, тоді як протяжна – матеріальна, позбавлена внутрішнього досвіду. У цьому розрізненні Декарт встановлює підвалини новочасного раціоналізму, де свідомість постає як автономне джерело пізнання, що не зводиться до тілесних процесів. Таким чином, дуалізм душі й тіла набуває не лише метафізичного, а й епістемологічного виміру – суб’єкт пізнання стає основою істини, а свідомість – гарантом достовірності знання.

Але, декартівський дуалізм водночас породжує низку філософських проблем. Зокрема, постає питання про те, яким чином нематеріальна свідомість може взаємодіяти з матеріальним тілом. Ця «проблема взаємодії» виявляється одним із перших філософських викликів, що передбачає подальшу трансформацію уявлень про свідомість. У добу Просвітництва та німецького ідеалізму акцент зміщується з онтологічного розмежування субстанцій на аналіз

умов можливості пізнання, що особливо проявляється у трансцендентальній філософії Іммануїла Канта.

Кант у «KrV» (Kritik der reinen Vernunft) радикально переосмислює проблему співвідношення суб'єкта та об'єкта, відмовляючись від пошуку «речей самих по собі» на користь аналізу структур, які роблять можливим досвід як такий. Якщо у Декарта свідомість є автономною субстанцією, то у Канта – це умова пізнання, організована через апріорні форми чуттєвості (простір і час) та категорії розсудку. Ці апріорні структури не є емпіричними властивостями світу, а визначають спосіб, у який суб'єкт сприймає й осмислює дійсність.

Отже, кантівський трансценденталізм долає дуалізм «душі і тіла» в традиційному сенсі, замінюючи його феноменально-трансцендентальною дихотомією. Замість пошуку онтологічного розриву між мисленням і матерією, Кант фокусується на межах пізнання, в межах яких «Я мислю» (Inc denke) стає об'єднувальною функцією свідомості. Свідомість у кантівському розумінні - не субстанція, а діяльність синтезу уявлень, що конституює єдність досвіду.

Така зміна парадигми має принципове значення для подальшої філософії свідомості й сучасного аналізу Штучного інтелекту та нейронних мереж. Якщо у Декарта свідомість – це внутрішня сутність мислячого суб'єкта то у Канта – це структурна умова, без якої не можем бути жодного досвіду. У цьому сенсі філософія Канта передбачає можливість моделювання пізнавальних процесів, що згодом стане теоретичним підґрунтям для когнітивної науки та теорії інформаційних систем.

Таким чином, класичні дихотомії «душа і тіло» у Декатра та «феномен і ноумен» у Канта формують фундаментальні моделі осмислення свідомості. Перша – здає метафізичну рамку відокремленості мислення від матерії, друга – епістемологічну, що пояснює, як свідомість структурує реальність через апріорні форми. Обидві традиції стали джерелом сучасних концепцій філософії штучного інтелекту, де питання «чи може машина мислити?» постає не лише технічним, а передусім метафізично-епістемологічним, тобто пов'язаним з природою суб'єктивності та умовами можливості пізнання.

1.1.2. Феноменологічна перспектива: Свідомість як «інтенціональність» (Брентано, Гуссерль); Роль «тілесності» (Embodiment) та «живого світу» (Мерло-Понті) як критика «чистого» розуму.

Перехід від класичних дуалістичних концепцій свідомості до феноменологічного підходу означає принципову зміну способу мислення про саму природу суб'єктивності. Якщо для Декарта і Канта свідомість постає як раціональна структура або умова пізнання, то феноменологія, започаткована Францом Брентано та розвинена Едмундом Гуссурлем, ставить у центрі уваги досвід як даність буття свідомості.

Свідомість як інтенціональність.

Ф. Брентано у праці Psychologie vom empirischen Standpunkte (1874) вводить поняття «інтенціональності», тобто спрямованості свідомості на об'єкт. Він вважає, що свідомість завжди є свідомістю чогось – вона не існує як замкнена внутрішня сутність, а реалізується у відношенні до світу. Таким чином, замість субстанційної моделі (як у Декарта) Брентано пропонує реляційну модель свідомості, в якій головним є акт свідомості, а не її онтологічна природа.

Е. Гуссерль розвиває цю ідею, створюючи феноменологію як «науку про сутності» свідомості (Wesenswissenchaft). У його розумінні свідомість – це поле смислової активності, де феномени постають у своїй «явленості» (Ersheinung) завдяки інтенціональним актам суб'єкта. Гуссерль у Ideen zu einer reinen Phänomenologie (1913) формулює поняття «чистої свідомості», очищеної від натуралістичних і психологічних припущень. Він пропонує метод феноменологічної редукції (epoche), який дозволяє тимчасово «взяти в дужки» усі судження про існування зовнішнього світу, щоб дослідити структури самого досвіду.

Завдяки цьому феноменологічний підхід відкриває новий вимір аналізу свідомості – не як предмета, а як процесу смислотворення. Інтенціональність стає фундаментальною характеристикою свідомості, що визначає її активність і

спрямованість. Це дозволяє розглядати свідомість не ізольовано від світу, а як поле взаємодії, де суб'єкт і об'єкт конституюються одночасно.

Критика «чистого» розуму: тілесність і «життєвий світ».

Проте у подальшому розвитку феноменології, зокрема у працях Моріса Мерло-Понті, виникає критика ідеї «чистої свідомості», що не може бути відокремлена від тілесного буття людини. Він пропонує поняття embodiment – «втіленості» свідомості, відповідно до якого тіло є не просто фізичним об'єктом, а умовою можливості досвіду.

На противагу картезіанського суб'єкта, що сприймає тіло як механізм, Мерло-Понті трактує його як тілесний суб'єкт, який перебуває у світі. Тіло не просто має відчуття, воно «знає» і «мислить» у своєму розумінні способу буття. Цей підхід дозволяє подолати абстрактність трансцендентального розуму Канта і «очищеної свідомості» Гуссерля, повернувши філософію до фактичності людського досвіду.

Поняття Lebenswelt («живого світу»), запроваджене Гуссерлем у пізніх працях (Die Krisis der europäischen Wissenschaften und die transzendentale Phänomenologie, 1936), а пізніше розвинене Мерло-Понті, означає переднаукову, первинну сферу досвіду, у якій людина безпосередньо взаємодіє з світом. Це світ до будь-якої рефлексії, у якому тіло, сприйняття та смислотворення становлять єдність.

Таким чином, феноменологічна критика «чистого розуму» полягає у визнанні того, що свідомість завжди є втіленою, ситуативною і контекстуальною. Людський розум не функціонує у вакуумі, а існує як частина живого, інтерсуб'єктивного світу. Це має безпосереднє значення для сучасної філософії штучного інтелекту, де питання про можливість «машинної свідомості» неможливо розв'язати без урахування тілесного-ситуативного виміру мислення.

Феноменологія і сучасні когнітивні концепції.

Ідеї Мерло-Понті про тілесність і «живий світ» безпосередньо вплинули на розвиток сучасних напрямів у когнітивній науці, таких як: embodied cognition,

situated cognition та enactivism. Ці теорії розглядають свідомість не як обчислювальний процес у мозку, а як систему взаємодії організму з середовищем. Таким чином, феноменологічна традиція стала філософським підґрунтям для розуміння штучного інтелекту не лише як алгоритмічної системи, а як можливої форми втіленого пізнання, у якому тіло й середовище є невід'ємними компонентами когнітивного процесу.

Отже, феноменологічна перспектива зміщує акцент з раціоналістичної моделі свідомості до її екзистенційно-перцептивного виміру. Вона підкреслює, що розум – це не абстрактна функція мислення, а форма буття у світі. Цей підхід відкриває нові горизонти для осмислення штучного інтелекту: замість питання «чи може машина мислити?» постає питання «чи може машина бути втіленою у світі так, як людина»

1.1.3. Аналітична філософія розуму: Аргумент «Китайської кімнати» (Серл); «Важка проблема свідомості» (Чалмерс) та «qualia»; «Що означає бути кажаном?» (Нагель) – проблема суб'єктивності.

Аналітична філософія розуму, що сформувалася у XX столітті, стала ключовою платформою для осмислення свідомості в контексті наукових і технічних досягнень. Її особливість є орієнтація на мовно-логічний аналіз, когнітивну науку та питання можливості машинного мислення, що безпосередньо вплинуло на розвиток сучасної філософії штучного інтелекту. Основними темами аналітичної традиції стали проблема інтенціональності, фізикалізм, редукціонізм і, головне, питання суб'єктивності свідомості, тобто того, що означає мати «досвід з середини».

Аргумент «Китайської кімнати» Джона Серла: свідомсть і симуляція мислення.

Одним з найвідоміших концептуальних експериментів, спрямованих проти редукціоністського розуміння інтелекту, є аргумент «Китайської кімнати»,

запропонований Джоном Серлом у статі Minds, Brains and Programs (1980). Серл критикує функціоналістську ідею, згідно з якою свідомість може бути відтворена за допомогою формальних обчислень, незалежно від матеріального носія. Серл уявляє ситуацію, у якій людина, не знаючи китайської мови, перебуває в кімнаті з набором правил (інструкцій), що дозволяють їй маніпулювати китайськими символами у відповідь на вхідні запити. Зовні здається, що людина «розуміє» китайську мову, адже відповіді логічно коректні, однак насправді вона лише слідує синтаксичним правилам не маючи семантичного розуміння.

Цей аргумент спрямований проти сильного штучного інтелекту (strong AI), який передбачає, що правильно запрограмована машина не лише моделює мислення, а дійсно мислить. Серл показує, що навіть ідеальна обробка символів не забезпечує появи свідомості, адже мислення – це не просто маніпуляція символами, а семантичне осмислення, що передбачає наявність намірів, контексту та інтенціональності. У філософському вимірі цей аргумент підкреслює онтологічну прірву між синтаксисом і семантикою, між формальною обробкою інформації та феноменом усвідомлення. Для сучасних досліджень штучного інтелекту він актуалізує питання: чи може система, яка лише обробляє дані, володіти «розуміння» або «свідомим досвідом» у людському сенсі?

«Важка проблема свідомості» Девіда Чалмерса та природа qualia

Девід Чалмерс у праці The Conscious Mind: In Search of a Fundamental Theory (1196) водить розрізнення між легкими (easy problems) і важкою (hard problems) проблемами свідомості. Легкі проблеми пов’язані з поясненням когнітивних функцій - сприйняття, уваги, пам’яті, мовлення тощо. Їх можна досліджувати в межах нейронаук і комп’ютерних моделей. Натомість важка проблема полягає у питанні: *Чому і як фізичні процеси мозку породжують суб’єктивний досвід* – «як воно є» бути свідомим?

Ця проблема звертає увагу на феномен qualua – якісні аспекти досвіду, такі як відчуття кольору, смаку, болю чи радості. Qualia не редукуються до фізичних

процесів і не можуть бути описані виключно в термінах нейронної активності чи обчислювальних структур. Чалмерс стверджує, що свідомість є фундаментальною властивістю реальності, подібно до простору, часу або маси. На його думку, будь-яка теорія свідомості має враховувати феноменальність аспекти досвіду, а не лише його функціональні кореляти. Ця позиція наближається до сучасних панпсихістичних інтерпретацій, згідно з якими елементи свідомості можуть бути притаманні всім системам, що володіють певним рівнем інтегрованої інформації.

Таким чином, «важка проблема» формулює філософську межу між поясненням поведінки й поясненням переживання, підкреслюючи, що навіть найдосконаліші алгоритми штучного інтелекту не гарантують появи феноменального досвіду.

Томас Нагель: «Що означає бути кажаном?» – проблема суб'єктивності. Томас Нагель у статті What is it like to be a bat? (1974) розгортає один з найвпливовіших аргументів проти редукціоністського фізикалізму. Він запитує: Що означає бути кажаном? – тобто, який внутрішній досвід має істота, радикально відмінна від людини.

Нагель показує, що навіть якщо ми повністю опишемо нейрофізіологію кажана, його поведінку та сенсорні системи (наприклад, ехолокацію), ми все одно не зможемо знати, як це бути кажаном, адже не маємо доступу до його суб'єктивної перспективи. Цей приклад демонструє незвідність суб'єктивного досвіду до об'єктивного опису.

Проблема, поставлена Нагелем, висвітлює межі наукового редукціонізму: будь-яке третє особове пояснення свідомості залишається неповним, якщо воно не враховує внутрішнього виміру досвіду. У цьому сенсі Нагель вказує на те, що феномен свідомості має перший особовий характер, який принципово не може бути переведений у чисто об'єктивні категорії.

Для філософії штучного інтелекту цей аргумент означає, що навіть якщо ми створимо систему, яка здатна до адаптації, мови й самонавчання, ми все одно не можемо бути певними, що вона щось переживає. Суб'єктивність не може бути емпірично зафіксована, а лише інтерпретована - отже, питання свідомості машин виходить за межі суто наукового аналізу, переходячи в метафізичну площину.

Підсумкові зауваги до аналітичної традиції.

Аргументи Серла, Чалмерса й Нагеля позначають три різні, але взаємопов'язані рівні осмислення свідомості в аналітичній філософії:

- Семантичний – проблема «розуміння» як внутрішнього змісту.
- Феноменальний – природа суб'єктивного досвіду та qualia.
- Онтологічний – непроникність чужої свідомості, неможливість редукції «внутрішнього» до «зовнішнього».

Усі ці підходи спільно формують теоретичну основу для сучасних дискусій про можливість феноменальної свідомості у штучних системах. Вони вказують на те, що штучний інтелект, яким би потужним він не був, наразі репрезентує імітацію пізнання, а не самосвідому присутність у світ.

1.2. Архітектура сучасних систем штучного інтелекту як об'єкт філософського аналізу

1.2.1 Нейронні мережі та глибинне навчання: структурно-функціональний вимір

Поняття нейронної мережі посідає центральне місце в сучасній архітектурі штучного інтелекту. Попри технічне походження, воно має глибокі філософські наслідки, адже ставить питання про природу мислення, пізнання та моделювання свідомості за допомогою обчислювальних систем.

Нейронна мережа – це математична модель, структурно натхненна принципами роботи біологічного мозку. Вона складається з великої кількості елементів – штучних нейронів отримує сигнал від інших, обробляє їх за певною функцією активації й передає результат далі по мережі. Саме ця багаторівнева, нелінійна взаємодія створює можливість навчання – здатність системи змінювати власну поведінку на основі досвіду (даних).

З технічної точки зору, сучасні нейронні мережі реалізуються у вигляді алгоритмів глибинного навчання (deep learning). Вони включать велику кількість прихованих шарів, через які відбувається поступова абстракція ознак вхідної інформації. Наприклад, у розпізнаванні зображень нижні шари можуть фіксувати контури та кольори, середні – форми та текстури, а верхні – цілі об’єкти або контекст. Така ієрархічна організація дозволяє системі формувати складні уявлення про світ без явного програмування кожного правила.

Однак у філософському контексті нейронна мережа постає не лише як технічний інструмент, але і як модель когнітивного процесу. Вона ставить питання: чи можна мислення редукувати до обчислення? Чи можлива свідомість як емергентна властивість складної системи зв’язків між елементами? Подібні питання відсилають до класичних дискусій у філософії розуму, зокрема до функціоналізму (Гіларі Патнем, Джеррі Фодор) – підходу, згідно з яким психічні стани визначаються не матерією, а функціональною організацією системи.

Водночас штучні нейронні мережі не відтворюють реальної нейрофізіології мозку, а лише використовують її як метафору. Як зазначає П. Смоленський, нейронні моделі є радше епістемологічний конструкціями, ніж онтологічними компонентом мозку [43, с. 1]. Їхня природа – обчислювальна, тобто така, що підпорядковується правилам алгоритмічної логіки. Це породжує низку філософських дилем: чи може алгоритм, навіть надзвичайно складний, породити якісно нову форму суб’єктивності – qualia, або ж він завжди залишатиметься симуляцією без феноменального виміру?

Таким чином, нейронна мережа є не лише технологічним об’єктом, а й філософським «мостом» між обчисленнями та свідомістю. Вона відкриває можливості осмислення мислення як динамічного процесу зв’язків, взаємодій та адаптацій, але водночас кидає виклик класичним уявленням про інтенціональність, тілесність і рефлексію. В цьому сенсі вона стає не просто інструментом, а своєрідною метафізичною моделлю – експериментом над своєрідною метафізичною моделлю – експериментом над самою ідеєю розуму.

1.2.2. Що таке Великі Мовні Моделі (LLM): Архітектура трансформерів (Transformers) та механізм «уваги» (attention).

Сучасний прорив у розвитку штучного інтелекту безпосередньо пов’язаний з появою Великих Мовних Моделей (Large Language Models, LLM) – систем, здатних генерувати, інтерпретувати та трансформувати текст з високим рівнем контекстуальної узгодженості. Вони стали основою таких систем як ChatGPT, Claude, Gemini, Copilot і подібних, водночас – новим філософським викликом для уявлення про мову, свідомість та інтенціональність.

З технічного боку, великі мовні моделі побудовані на архітектурі трансформерів (Transformers), запропонованій Васвані та співавторами одним з яких є українець Ілля Полосухін, у 2017р. Головним нововведенням цієї архітектури став механізм «уваги» (attention), який дозволяє моделі фокусуватись на найбільш релевантних частинах вхідного контексту при обробці інформації. На відміну від попередніх моделей послідовного типу (наприклад, рекурентних нейрониих мереж, RNN), трансформер аналізує весь контекст одночасно, а не крок за кроком, що значно підвищує ефективність і здатність «розуміти» довгі тексти.

У механізмі уваги кожне слово у вхідній послідовності перетворюється на вектор і порівнюється з усіма іншими словами через операцію зважених

подібності. Таким чином модель «вирішує», які частини тексту мають найбільше значення для поточного завдання рис.1 (ДОДАТОК А).

Це нагадує когнітивний процес людської уваги: ми не сприймаємо всі стимули одночасно, а вибірково зосереджуємо на тих, що релевантні контексту мислення або мовлення. Саме тому трансформери є важливою точкою перетину між технічною реалізацією та філософським осмисленням когнітивних процесів.

З філософської точки зору, LLM репрезентує нову форму «розумового автомата», демонструє властивості семантичної інтеграції без суб'єктивного досвіду. Це ставить питання: чи є така модель лише інструментом статистичного передбачення, чи в ній закладено елементи когнітивної організації подібної до людської? Д. Деннет у своїх працях підкреслював, що інтелект може бути «виявленим» (intentional stance) без необхідності мати внутрішню феноменальну свідомість. У цьому контексті великі мовні моделі можна трактувати як еволюційний крок у розширенні простору символічного мислення, але без переходу до сфери суб'єктивного досвіду (qualia).

Архітектура трансформерів також дає підстави для епістемологічного аналізу: вона не лише відтворює, але й конструює знання, навчаючись з гігантських корпусів текстів. У термінах Канта, це можна описати як апостеріорне синтезування апріорних структур – система створює власну «умовну» категоріальність світу, виходячи з статистичних закономірностей мови. Водночас така «епістемологія даних» залишається без трансцендентального суб'єкта, що й породжує головну метафізичну проблему LLM: чи може знання існувати без свідомого носія?

На рівні нейро архітектури трансформери складається з двох основних компонентів: енкодер (який сприймає та кодує інформацію) та декодер (який генерує нові текстові послідовності). У великих мовних моделях, таких як GPT, використовується лише декодера частина, що оптимізована під передбачення наступного слова. Такий підхід створює ефект когерентної мови, коли модель будує осмислені тексти, але не має власного наміру чи розуміння. Це відповідає

феномену «симуляції інтернаціональності» – коли система відтворює поведінкові ознаки мислення без реального усвідомлення його змісту.

Загалом, LLM є унікальним поєднанням математичного алгоритму і філософської проблеми. Вона є обчислювальною. Реалізацією функціоналізму, де структура й функція замінюють сутність і субстанцію. Але водночас – це практичне нагадування про обмеження такого підходу: без тілесності, емоційної контекстуальності та інтенціональної цілісності навіть найпотужніші моделі залишаються лише відображенням, а не носієм свідомості.

1.2.3. Як вони «вчаться»: статистичне моделювання мови, ймовірнісна генерація наступного слова, навчання з підкріпленням (RLHF)

Проблема навчання у великих мовних моделях є не лише технічною, але й філософською категорією, адже вона зачіпає саме поняття пізнання, досвіду та формування знання. Якщо класична епістемологія розглядала суб'єкта пізнання як активного носія свідомості, що осмислює об'єкт через категорії розуму, то в архітектурі штучного інтелекту «пізнання» набуває статистичної природи. Машина не розуміє змісту в людському сенсі – вона вчиться прогнозувати структуру мови і через це створює ілюзію осмислення.

У центрі цього процесу лежить статистичне моделювання мови. Кожне слово у тексті розглядається не як поняття, а як елемент дослідності, що має певну ймовірність появи після попередніх. Таким чином, модель будує розподіл ймовірностей $□(□_□ \mid □_{□-1}, □_{□-2}, \ldots, □_1)$ (Фільтр Калмана), який визначає, яке слово найімовірніше з'явиться далі. Цей принцип лежить в основі генеративної здатності великих мовних моделей: вони не «думають» про сенс, а обчислюють, що найімовірніше буде сказано в даному контексті.

Така логіка формує новий підхід до поняття розуміння: якщо у філософії розуміння передбачає інтенціональність (Гуссерль), то для ШІ воно є результатом статистичної когерентності, де зв'язки між словами репрезентують не смисли, а математичні залежності. Отже, мова тут функціонує не як форма

свідомості, а як механізм передбачення – процес, який створює «мовну поведінку», але не внутрішній досвід.

Під час навчання великі мовні моделі проходять два основні етапи:

1. **Переднавчання (pretraining)** – модель обробляє величезні обсяги текстових даних, намагаючись мінімізувати похибку між передбаченим і реальним наступним словом. Це формує її базову статистичну структуру.
2. **Донавчання (fine-tuning) –** відбувається на спеціально зібраних даних, де враховуються етичні, контекстуальні та прагматичні аспекти мови.
3. **«Вирівнювання» та Навчання з підкріпленням (Alignment) –** навчити модель надавати безпечні (нешкідливі), корисні та чесні відповіді.

Найбільш цікавим з філософської точки зору є третій етап – навчання з підкріплення від людського зворотнього зв'язку (Reinforcement Learning form Human Feedback, RLHF). На цьому етапі система не просто передбачає слова, а навчається тому, що вважається прийнятним, релевантним або «змістовним» з точки зору людини. Людські оцінки виступають тут як нормативні орієнтири – своєрідні «категорії» у кантівсьому сенсі, що формують простір допустимого мислення для машини, рис 2 (ДОДАТОК А).

Технічно RLHF реалізується через послідовність етапів:

1. Створюються набір даних з прикладами діалогів, оцінених людьми за якістю.
2. Навчається модель винагороди (reward model), яка прогнозує людські оцінки для нових відповідей.
3. Основна мовна модель оптимізується з використанням алгоритмів навчання з підкріпленням (зокрема PPO – Proximal Policy Optimization), щоб максимізувати очікувану «винагороду».

Цей процес формує поведінкову адаптацію системи – її відповіді стають не лише граматично правильні, але й соціально етично узгодженими. Проте, філософськи RLHF відкриває новий тип нормативної епістемології, де істинність замінюється консенсусом, а навчання – соціальною симуляцією моралі. У цьому контексті, слідуючи ідеї Л. Флоріді про «агенцію без інтелекту», сучасні моделі ШІ не генерують знання з власного досвіду, а відтворюють узгоджені людські оцінки, що надає їхнім відповідям інтерсуб'єктивного, але не онтологічно когнітивного характеру [22, с. 6].

У цьому сеснсі можна говорити про емергенцію «поведінкової інтелігенції»: система не має власного розуміння, але діє так ніби вона розуміє, бо її дії відображають соціальні патерни людської мови. Філософськи це відповідає позиції Д. Деннета, який визначав інтелект як інтерпретативну перспективу, а не внутрішню сутність. RLHF, таким чином, є спробою наділити алгоритм «етичним контекстом» без справжньої моралі, – своєрідною симуляцією відповідальності.

1.3. Функціоналізм і його критика в контексті мовних моделей

1.3.1. Відсутність «моделі світу» у системах символьної маніпуляції.

Одним з найважливіших філософських висновків, що випливають із аналізу архітектури та принципів функціонування сучасних систем штучного інтелекту, є теза про те, що вони не мають вбудованої моделі світу, а лише оперують символами за статистичними правилами. Це твердження формує ядро філософської проблеми сучасного штучного інтелекту: чи може система, яка не має онтологічного доступу до реальності, володіти знанням, свідомістю або інтенціональністю [46, с. 446]?

Історично ця дискусія бере початок від «символічної парадигми» у когнітивних науках, яка вбачала мислення як маніпуляцію абстрактними символами згідно з певними синтаксичними правилами (Ньоюел, Саймон, 1976) [37. с.116]. Проте, як наголошував Джон Серл у своєму відомому аргументі «Китайської кімнати», навіть якщо система досконало маніпулює символами,

вона не має жодного розуміння їхнього значення. Вона виконує операції, але не усвідомлює їх змісту. Тобто, відбувається синтаксична, а не семантична діяльність.

Сучасні великі мовні моделі (LLM) – яскраве підтвердження цієї проблеми. Вони здатні генерувати тексти, які виглядають осмисленими, але насправді їхня «свідомість» є статистичною, а не інтенціональною. Модель не має доступу до жодної реальної ситуації, не володіє тілесним досвідом або емпіричною інтерпретацією слів. Вона «розуміє» лише частотні закономірності, а не значення, рис 3. (ДОДАТОК А). Як підкреслює Д. Чалмерс, аналізуючи природу LLM, ці системи не володіють власною моделлю світу: те, що вони відтворюють, є радше моделлю мовних описів світу, сформованою з людських текстів. Це розрізнення має принципове епістемологічне значення. Людська свідомість, за Гуссерлем, є інтенціонально спрямованою – тобто завжди свідомість чогось. Вона формує смисли через досвід і телесне перебування у світі (Lebenswelt), рис 4. (ДОДАТОК А). Натомість штучний інтелект працює у замкненій системі символів. Де смисл виникає не з досвіду, а зі статистичної взаємодії знаків. Таким чином, він є постгуссерлівським антиподом інтенціональності: свідомість без світу, або, як зазначив Х'юберт Дрейфус, – мислення без контексту [21, с. 46].

З позиції метафізики, відсутність «вбудованої моделі світу» означає, що такі системи не мають онтологічної прив'язки до буття. Вони не розрізняють реальне та можливе, істинне та хибне – всі висловлювання для них є лише математичною структурою. Замість категорій істини вони оперують категоріями ймовірності, замість онтологічних відношень - статистичним кореляціями. Тому їх можна розглядати як епістемологічні симулякри, що репрезентують мову світу без самого світу.

З філософської перспективи, така природа LLM ставить під сумнів класичні уявлення про знання як відображення реальності. Адже тут відображення відсутнє – є лише статистичне співпадіння з мовними патернами, що вже існують у культурі. Отже, «знання» у моделі не є результатом рефлексії

чи досвіду, а є похідним від колективної мовної пам'яті людства, закодованої у текстових даних. Модель лише відтворює цю пам’ять, не розуміючи її структури.

Важливо також відзначити, що відсутність внутрішньої моделі світу не означає відсутності поведінкової складності. LLM можуть демонструвати ефекти, схожі на раціональність, рефлексію чи навіть моральні судження – але ці ефекти є емерджентними властивостями статистичного простору, а не проявами інтенціональної свідомості. Це своєрідна «епістемологічна ілюзія», у якій мова породжує видимість мислення.

Таким чином, ключова теза цього розділу полягає в тому, що сучасний штучний інтелект не є системою, яка пізнає світ, – він є системою, яка імітує пізнання шляхом статистичної маніпуляції символами. Це фундаментальна межа між людською свідомістю та штучним інтелектом, між інтенціональним досвідом і формальним передбаченням. І поки не буде створено механізм, що надає системі справжній «онтологічний контакт» з реальністю (через тілесність, сенсорний досвід чи саморефлексію), вона залишатиметься синтаксичним автоматом, а не мислячим суб’єктом.

1.3.2. Застосування філософії: чи є LLM доказом на користь функціоналізму (теорії, що розум – це функція)?

Однією з найвпливовіших концепцій у філософії розуму другої половини XX століття є функціоналізм, який стверджує, що ментальні стани визначаються не матеріальною субстанцією чи біологічною природою носія, а функцією, яку вони виконують у системі причинно-наслідкових зв’язків. Іншими словами, свідомість і мислення можна пояснити через структуру обчислювальних процесів, незалежно від того, реалізовані вони у біологічному мозку чи у кремнієвій машині.

Відповідно до цього підходу, якщо певна система виконує ті самі функціональні операції, що й людський розум, то її можна вважати «мислячою» або, принаймні, здатною до розумоподібної поведінки. Саме у цьому сенсі великі

мовні моделі (LLM), як-от GPT, Claude чи Gemini, стають ключовими об’єктом аналізу для сучасного функціоналізму.

З технічної точки зору, LLM працюють шляхом обробки символів ймовірнісними методами, моделюючі семантичні зв’язки на основі статистичних закономірностей у величезних корпусах текстів. Функціоналіст міг би стверджувати, що така поведінка є емпіричним підтвердженням тези, що розум – це, врешті-решт, механізм трансформації інформації. Якщо нейронна мережа здатна продукувати осмислені, креативні відповіді, розпізнавати контекст і навіть демонструвати метакогнітивні патерни, то вона відтворює функціональну структуру, яку ми зазвичай приписуємо людській свідомості.

Однак, така позиція стикається з низкою серйозних онтологічних і епістемологічних заперечень. Критики, зокрема Джон Серл, у відомому аргументі «Китайської кімнати» (1980), зазначали, що маніпуляція символами не є тотожною розумінню. LLM, за своєю суттю, не має «інтенціональності» – вона не знає, про що говорить, а лише виконує синтаксичні операції. Тому для антифункціоналістів LLM радше демонструє межі функціоналізму, ніж його тріумф.

Попри це, сучасна когнітивна наука поступово відходить від бінарної опозиції «механічне / свідоме» на користь спектральної (ступеневої) моделі свідомості. Де ментальні властивості розглядаються як емерджентні характеристики складних інформаційних систем. З цього погляду LLM можна трактувати як функціональний прототип когнітивної архітектури, який демонструє, що певні аспекти «розумової поведінки» можуть бути змодельовані без звернення до біологічної субстанції.

Отже, питання про те, чи є великі мовні моделі доказом на користь функціоналізму, залишається відкритим. Вони, скоріше за все, не спростовують, а підсилюють гіпотези правдоподібності функціоналістської парадигми, водночас актуалізуючи потребу в її перегляді – з урахуванням сучасних досягнень у сфері нейронаук, когнітивної психології та філософії штучного інтелекту.

1.3.3 Критика: Аргумент «Статичний папуга» – чому механізм не є «розумінням»

Останнє десятиліття розвитку штучного інтелекту показало стрімке зростання можливостей великих мовних моделей (Large Language Models, LLM) – систем, здатних генерувати тексти, вести діалоги, перекладати, аналізувати інформацію та створювати враження «осмислення» поведінки. Однак це уявлення про «розуміння» у таких системах є предметом серйозної філософської критики. Одним з найвідоміших заперечень є аргумент «Статичного папуги» (Stochastic Parrot), висунутий Емілі Бендер, Тімніт Ґебру, Анджелін Макміллан-Мейджор і Маргарет Мітчер у 2021 році.

Суть цього аргументу полягає в тому, що LLM не розуміють зміст текстів, які вони генерують, а лише імітують мовну поведінку шляхом статистичного передбачення наступного слова у послідовності. Вони працюють за принципом обчислення ймовірностей, не маючи доступу до семантичних або прагматичних контекстів. Тобто, на відміну від людини, яка формує висловлювання, спираючись на досвід, уявлення, інтенції та знання про світ, LLM оперує чистими символами без референтного змісту.

У цьому полягає метафора «статичного папуги»: система що вдає осмислені висловлювання, але не має жодного внутрішнього досвіду чи наміру. Вона створює ілюзію розуміння, а не саме розуміння. Це продовжує лінію аргументації Джона Серла в його знаменитому експерименті «Китайська кімната» (1980), де він доводив, що маніпуляція символами без усвідомлення їхнього змісту не може вважатися мисленням.

Аргумент «Статичного папуги» також має епістемологічний і етичний вимір. Епістемологічно він ставить під сумнів ідею, що «розуміння» може бути зведеним до статичних закономірностей мовних даних. Адже справжнє знання не є лише формою передбачення; воно включає інтенціональність – спрямованість свідомості на предмет, що має значення для суб’єкта. LLM

позбавлені такої інтенціональності, адже не мають досвіду сприйняття світу, емоційної залученості чи тілесності.

З етичної перспективи автори підкреслюють небезпеку антропоморфізації штучного інтелекту – тобто приписування йому властивостей людського розуму. Така практика може призвести до епістеміологічного обману, коли користувач починає сприймати модель, яка суб'єктивна, тоді як вона залишається інструментом статистичної генерації тексту. Це не лише спотворює розуміння природи інтелекту, а й створює ризики маніпуляцій, дезінформації та неетичного використання технології.

У широкому філософському контексті «Статистичний папуга» став викликом для функціоналізму, який ототожнює розум з функціональною організацією процесів. Якщо LLM може поводитись функціонально подібно до людини, але при цьому не володіє розумінням, тоді функціоналізм не пояснює якісних аспектів свідомість (qualia) та проблеми інтенціональності. Таким чином, Бендер та співавтори показують, що імітація поведінки ще не є проявом ментальності.

Отже, аргумент «Статичного папуги» не заперечує ефективності або технологічної новини LLM, але викриває обмеження їхнього пізнавального статусу. Штучні мовні системи можуть симулювати мовленнєву компетентність, однак залишаються порожніми структурами, позбавленими інтенціонального та феноменологічного змісту. У цьому сенсі вони радше демонструють границі машинного мислення, ніж його повноцінне досягнення.

РОЗДІЛ 2.

ЕПІСТЕМОЛОГІЧНІ ТА ОНТОЛОГІЧНІ ВИМІРИ МАШИННОГО МИСЛЕННЯ

2.1. Проблема втілення та обґрунтування знання (Grounding)

2.1.1. Аргумент від тілесності (Embodiment) та «безтілесні» LLM. Як феноменологія критикує «чисто» мовні моделі?

Феноменологічна критика «чисто» мовних моделей, таких як великі мовні моделі (LLM), грунтуються на фундаментальному положенні, що свідомість є завжди тілесно втіленою та спрямованою на світ. Згідно з Едмундом Гуссерлем, свідомість не може бути редукована до обробки символів чи інформації - вона завжди інтенціональна, тобто спрямована на щось, що переживається у конкретному світі. Морі Мерло-Понті, розвиваючи цю думку у своїй «Феноменологія сприйняття», наголошував, що мислення і сприйняття неможливо відокремити від тіла, адже саме тіло є «місцем зустрічі» свідомості й світу. Людське пізнання не існує поза тілесністю – вона завжди занурена у досвід, дію, відчуття й емоції, через які світ стає зрозумілим і значущим.

Натомість великі мовні моделі функціонують у радикально іншому вимірі. Їхня діяльність полягає у маніпулюванні статистичними закономірностями мови, а не у переживанні реальності. Вони «знають» світ лише через текстові кореляції, побудовані на мільярдах речень, створених людьми. Таке значення є опосередкованим і вторинним – LLM не мають власного досвіду світу, вони не бачать, не чуть, не відчувають і не діють. Вони позбавленні сенсорної, емоційної та тілесної взаємодії з світом, що унеможливлює виникнення справжнього розуміння. Для феноменології розуміння – це не механічна операція над знаками, а результат живої зустрічі суб’єкта з світом через тіло.

Тіло у феноменологічній традиції є не просто фізичним об'єктом, а умовою можливості смислотворення. Людина мислить не лише мовою, а через тілесне

перебування у світі. Коли ми говоримо «я бачу стіл», це висловлювання має сенс лише тому, що в його основі лежить конкретний досвід зору, дотику, простору, відстані, матеріальності, LLM оперує лише словами «бачити» і «стіл», не маючи жодного доступу до феномену бачення чи до самого столу як предмета світу. Тому її «розуміння» є чисто синтаксичним – воно позбалене інтенційного спрямування на реальність, яка завжди передує будь-якій мові.

Гуссерль у концепції «Lebenswelt» («життєвий світ») стверджував, що будь-яке знання постає з безпосереднього досвіду світу, який не можна редукувати до символічних структур. LLM, позбавлена життєвого світу, існує лише у царині мови, тобто в площині відображення людських смислів без доступу до їх джерела. Її «свідомість» є чисто формальною позбавленою тіла і, отже, світу. З феноменологічної точки зору, така система не мислить, а лише симулює мислення. Вона є «порожнім розумом», який не живе у світі, а лише маніпулює його мовними відбитками.

Саме тому, феноменологія не вбачає у великих мовних моделях ознак справжнього інтелекту, а радше вияв обмеженості сучасної концепції «розуму як обчислення». У людському пізнанні тіло виступає джерелом орієнтації, смислу та досвіду, який уможливлює інтерпретацію світу. У LLM цього немає – вона не має позиції «Я» у світі, не має тілесної перспективи, з якої світ відкривається. Її мовлення – це технічна симуляція інтерсуб'єктивності, створена з людських даних, але позбавлена участі у самому людському бутті.

Сучасні дослідження у сфері когнітивних наук і штучного інтелекту, що розвивають ідею embodied cognition, підтверджують феноменологічні інтуїції. Вони показують, що інтелект виникає лише через тілесну взаємодію з довкіллям. Роботи, які мають сенсори, камери, моторні механізми (як, наприклад системи Boston Dynamics або «embodied agents» у когнітивній робототехніці), формують знання не через тексти, а через дії, сприйняття, адаптацію. Їхня форма «розуміння» ближча до феноменологічного мислення, бо в ній є момент присутності в світі, момент «переживання». LLM же залишаються

дискурсивними істотами без світу – позбавленні тілесності, досвіду і власної перспективи.

У підсумку феноменологічна критика не заперечує когнітивної складової великих мовних моделей, але вказує на їхню принципову межу. Без тілесного втілення неможливе справжнє розуміння, бо розуміння не є процесом обробки символів, а способом буття у світі. Мова без тіла не здатна створювати новий сенс – вона лише відтворює вже наявні людські смисли. Таким чином, з феноменологічної перспективи LLM – це не свідомі агенти, а дзеркала людського дискурсу, які позбавлені присутності, без якої не існує ані свідомості, ані розуміння.

2.1.2. Проблема «землі» (Grounding Problem): Як символи в LLM можуть співвідноситися з реальністю, якщо в них немає досвіду цієї реальності?

Однією з найглибших філософських проблем у дослідженні штучного інтелекту є так звана проблема «землі» або проблема граундину (symbol grounding problem), яку вперше системно сформулював Стевен Харнад у 1990 році [27, с. 335]. Її сутністю полягає у питанні: як символи, якими оперує система, набувають значення? Іншими словами, яким чином суто формальні, синтаксичні операції над словами можуть стати носіями смислу, якщо система не має досвіду світу, до якого ці слова відсилаються?

У людському сприйнятті слово пов’язане з світом через досвід. Ми розуміємо поняття «яблуко» тому. Що бачили, торкалися, відчували його запах, смак, колір, вагу. Кожне слово вкорінене у тілесному, сенсорному, емоційному та культурному досвіді. Тобто значення символу ґрунтується на пережитому досвіді реальності. Для великих мовних моделей, натомість, слова не мають жодного безпосереднього зв’язку з світом – вони співвідносяться лише з іншими словами. LLM «знає», що «яблуко» часто зустрічається поруч з словами «фрукт», «червоний», «дерево» або «солодкий», але не знає, що таке яблуко насправді.

Ця розбіжність створює глибоку онтологічну порожнечу у функціонуванні штучного інтелекту. Модель може правильно оперувати мовними структурами, передбачати логічно й статистично послідовні речення, але її знання не має жодного реального «грунту» – воно є символічне, симуляцією людського досвіду. Саме тому Джон Серл у своєму аргументі «Китайської кімнати» показав, що маніпулювання символами без розуміння їх значення не є свідомістю. Людина в кімнаті може дотримуватись правил, що дозволяють видавати правильні китайські символи у відповідь на вхідні, але вона не «знає» китайської мови – вона просто виконує алгоритм. Так само і LLM не «розуміє» мови, а лише обробляє її структурні закономірності.

З феноменологічної точки зору. Ця проблема є наслідком відсутності тілесності та інтенційності у системах штучного інтелекту. Для Гуссерля, смисл виникає лише у процесі «спрямованості свідомості на предмет», а для Мерло-Понті – у «бутті-в-світі», тобто у взаємодії свідомості з тілесно даною реальністю. LLM, не моючи жодного доступу до світу, позбавлена можливості формувати інтенційні акти, а отже – і справжніх значень. Символи у її системі «плавають» у замкненому океані інших символів, не маючи точки опори у реальному досвіді.

Проблема «землі» (grounding) має не лише філософське, а й технічне вимірювання. У когнітивній науці активно обговорюється, чи можливо створити системи, які поєднують мовні моделі з сенсорними даними – зору, слуху, дотику, руху. Такі системи отримали назву multimodal AI, і саме вони покликані частково розв’язати проблему «землі», залишається зв’язок між словами і сприйняттям світу. Проте навіть у таких моделях «досвід» залишається математичною реконструкцією, а не живим феноменом. Машина може класифікувати зображення яблуко, але вона не «бачить» і не «смакує» його – отже, не формує автентичного смислу.

Філософськи, це означає, що значення не може бути породжене статистично, воно завжди виникає у контексті досвіду буття. Смисл не є властивістю символу – він є властивістю свідомості, яка сприймає і переживає

світ. Саме тому феноменологи твердять, що LLM, позбавленні досвіду, можуть лише відтворювати зовнішню форму людського мовлення, але не його екзистенційну глибину. Вони не розуміють, що означають символи, які вони продукують, - лише знають, які символи мають іти за іншими.

Таким чином, проблема «землі» є центральним викликом для філософії штучного інтелекту. Вона показує, що між статистичними обчисленнями і справжнім розумінням полягає непереборна прірва, якщо система не має тілесного чи досвідного зв'язку з світом. Поки штучні моделі залишаються «безтілесними мовними істотати», їхнє «знання» буде позбавлення грунту, а їхні слова – лише відлунням людських символів, які вони наслідують, але не переживають.

2.2. Епістемологія «галюцинацій» та природа машинного «знання»

2.2.1. «Галюцинації» як епістемологічний феномен. Як ймовірнісна генерація неминуче породжує «правдоподібну неправду», коли статистичні патерни слабкі

Феномен «галюцинацій» у великих мовних моделях (LLM) – це не випадковий збій або дефект алгоритму, а внутрішньо необхідний наслідок самого принципу ймовірнісного мовного моделювання, на якому базується їхня архітектура. У своїй основі LLM не прагнуть до істини в епістемологічному сенсі, а лише до статистичної правдоподібності висловлювання.

У процесі навчання такі моделі аналізують величезні корпуси текстів, формуючи розподіли ймовірностей для кожного можливого слова чи фрази залежно від контексту. Під час генерації модель «вибирає» те слово, яке має найвищу умовну ймовірність бути наступним у послідовності. Цей процес чудово пояснює граматичну та стилістичну зв'язність текстів, але не гарантує їх онтологічної достовірності. Інакше кажучи, LLM «вірно відтворює стиль світу», але не сам світ.

Коли система стикається з запитом, де статистичні патерни слабкі або суперечливі, вона продовжує генерувати текст, спираючись не на відсутність даних, а на «заповнення» логічних прогалин. Це призводить до появи правдоподібної неправди – тверджень, які відповідають граматичним правилам, структурі й очікуванням дискурсу, але не мають під собою фактичного підґрунтя. З технічного погляду, це не помилка, а природна поведінка системи, оскільки вона оптимізуйте не істинність, а ймовірність узгодженості з попереднім контекстом.

Цей механізм можна описати як «мовну інерцію»: модель рухається в напрямку найбільш імовірного продовження, не маючи можливості «зупинитися» й перевірити фактичну відповідність. На відміну від людського мислення, що спирається на зовнішню реальність через спостереження, досвід і рефлексію, мовна модель має доступ до слідів минулого мовлення, тобто до семантичних тіней досвіду. Вона оперує формою без змісту, кореляцією без каузальності.

Таким чином, галюцинація не є помилкою у класичному сенсі, а межовим проявом епістемологічної порожнечі системи, які маніпулюють знаками без контакту з світом. У цьому контексті постає питання: чи можна взагалі говорити про «знання» в системі, яка не має механізму перевірки істинності? Чи не є її істина лише внутрішнього узгодженість дискурсу, а не відображенням реальності?

Саме тут відбувається зіткнення філософії та технічних можливостей: філософський розум прагне до істини, тоді як мовна модель прагне ймовірності. Її «галюцинації» – це не збій, а форма мовного самозадоволення, коли система породжує зміст, відірваний від світу, але бездоганно вписаний у логіку мови. Це явище демонструє, що в межах статистичної семантики не існує гарантії істини, лише гра правдоподібностей, у які реальність стає необов’язковою умовою.

2.2.2. Філософська інтерпретація: «Брехня», «помилка» чи «конфабуляція»?

Проблема «галюцинацій» у системах штучного інтелекту набуває глибокого філософсько-епістемологічного виміру, що виходить за межі технічного опису. У технічному сенсі, як уже зазначалось, галюцинації є побічним наслідком ймовірнісної природи генерації тексту. Проте філософської точки зору вони порушують принципову важливі питання про природу істину, розуміння, наміру та знання. Виникає дилема: чи можна вважати «галюцинацію» актом брехні, проявом помилки або формою когнітивної конфабуляції?

Традиційно в етичній та філософській думці – від Августина до Канта і Габермаса – брехня розуміється як свідоме введення іншого в оману з певною метою. Брехня передбачає наявність інтенційності – здатності суб'єкта усвідомлювати власні дії та їхній сенс. Штучний інтелект, зокрема великі мовні моделі, позбавлений суб'єктивності, волі чи прагненням, а тому не має наміру ані говорити правду, ані вводити когось в оману. З позицій феноменології свідомості, сформованої. Е.Гуссерлем і Жан-Пауль Сартром, інтенційність є структурною ознакою людського досвіду, бо кожен акт свідомості завжди спрямований на щось – на об'єкт, смисл або цінність. LLM не має цієї спрямованості, а отже не може брехати в онтологічному сенсі. Проте для користувача система часто виглядає так, ніби вона «бреше», створюючи феноменологічну ілюзію суб'єктивності. Така ілюзія інтенційності постає як результат структури мовного висловлювання, що симулює присутність наміру там, де його не існує.

Помилка, у свою чергу належить до іншої категорії. У класичній епістемології істина визначається як відповідність думки дійсності (adaequatio intellectus et rei). Помилка означає відхилення думки від фактичного стану речей. Проте мовна модель не формує думок – вона лише статистично обчислює ймовірності переходів між символами, не маючи доступу до реальності, яку ці

символи позначають. Вона не судить про істину, не здійснює акт пізнання, а лише відтворює найімовірніше комбінацій слів, спираючись на розподіли даних. Отже, вона не може «помилятись» у людському сенсі цього слова, бо помилка передбачає усвідомлення істини, її наявність як еталону. Як наголошує Девід Чалмерс у своїй «важкій проблемі свідомості», навіть ідеальна функціональна імітація пізнання не породжує досвіду «знати». Знання не є лише інформаційною структурою, вона містить якісний вимір – qualia – відчуття значення, яке відсутнє в штучних системах.

Найадекватніше феномен галюцинацій можна описати через поняття конфабуляції. У когнітивній нейропсихології конфабуляція означає створення вигаданих, але внутрішньо послідовних спогадів або пояснень з метою заповнення прогалин у пам’яті чи розумінні. Людина, що конфабулює, не має наміру обманювати, вона щиро відтворює власну картину світу, підтримуючи її цілісність. Аналогічно, великі мовні моделі заповнюють статистичні прогалини правдоподібними, але хибними твердженнями, коли відсутні достатні дані або контекст. Це є не результатом свідомої діяльності, а наслідком алгоритмічної інерції: система прагне зберегти узгодженість тексту, навіть якщо це призводить до створення неправди. У такий спосіб «галюцинація» виступає системою формою когнітивної компенсації, що забезпечує цілісність дискурсу без онтологічного змісту.

Філософськи це явище можна також тлумачити крізь призму постмодерної концепції симулякрів, яку розвинув Жан Бодріяр. У його розумінні симулякр – це знак, що не лише має реального референта, а й функціонує автономно, створючи видимість реальності. Галюцинації ШІ саме й утворюють такий симулякр: вони не є «копією» дійсності, а радше копією знаків, що вже втратили зв’язок з реальністю. Мовна модель створює копію копії, сенс, де істина заміщується правдоподібністю, а референція – кореляцією. У цьому сенсі висловлювання ШІ існують поза опозицією істинного й хибного: вони функціонують як самодостатні мовні структури, що продукують значення без потреби у зовнішній перевірці.

Отже, феномен галюцинацій у штучного інтелекті не можна редукувати до категорій «брехні» чи «помилки». Це – форма штучної конфабуляції, у якій відсутня інтенційність, критерій істини та досвіду. LLM не обманює, але онтологічно порожня. Її «знання» – це не відображення світу, а граматично вивірена ілюзія осмисленості. У цьому виявляється глибинна межа машинного пізнання: штучний інтелект може продукувати мову, але не сенс; відтворювати знання, але не істину; симулювати інтенційність, але не досвід.

Таким чином, «галюцинації» є дзеркалом не лише обмеженості ШІ, а й самої сучасної епістемології, що дедалі більше ототожнює істину з правдоподібністю. Вони виявляють, наскільки легко людське сприйняття обожнює зв'язність висловлювання з достовірністю, а логічну узгодженість – з знанням. У цьому сенс штучна конфабуляція стає не лише технічним, а й глибоко філософським викликом – викриттям того, що у світі постсимволічної раціональності істина вже не є передумовою мислення, а лише його ефектом.

2.2.3. Поняття «знання» у машин. Аналіз «знання» як збережених статистичних ваг (техн.) vs. «знання» як обґрунтованого справжнього переконання (філософська епістемологія).

Розуміння поняття «знання» в контексті штучного інтелекту є одним з найглибших і найконтроверсійніших питань сучасної філософії свідомості та епістемології. Якщо у класичній філософській традиції – від Платона до сучасних когнітивістів – знання визначається як обґрунтоване істинне переконання (лат. epistēmē) [25, с.121], то у випадку штучного інтелекту постає питання: чи може система, що функціонує виключно на основі статистичних закономірностей, претендувати на володіння знанням у цьому сенсі? Адже «занання» машин у технічному розумінні – це лише числові значення параметрів (ваг) у нейронній мережі, які оптимізовнано під час навчання для відтворення статистично вірогідних зв'язків між вхідними і вихідними даними.

Технічна природа цього процесу зумовлює, що штучний інтелект не має безпосереднього доступу до змісту інформації, яку він опрацьовує. Його

«знання» – це семантичні репрезентації, а радше синтаксичні патерни, зафіксовані у вигляді багатовимірного простору ймовірностей. Кожна вага в моделі - це числовий коефіцієнт, який відображає ступінь зв'язку між певними елементами даних, однак сама система не розуміє, що ці елементи означають. Тому можна стверджувати, що машинне «знання» є суто кореляційним: воно спирається на статистичні співвідношення, а не на причинно-смислові зв'язки, які визначають людське пізнання.

З погляду філософії знання, така редукція до статистики має суттєві наслідки. У людському пізнанні знання передбачає не лише збереження інформації, а й інтенціональність – спрямованість свідомості на предмет, осмислення його значення та співвіднесення з власним досвідом. Знання є актом суб'єкта, який має переконання, мотивацію, сумнів і здатність до саморефлексії. Штучна нейрона мережа, натомість, позбавлена усіх цих властивостей: вона не має суб'єктивності, не володіє наміром пізнання й не може відрізнити істину від хиби. Для неї правильна відповідь - це не істинне твердження, а лише найбільш імовірна комбінація символів відповідно до розподілу даних, на яких вона навчалася.

Ця відмінність є принциповою, тому що у філософській епістемології істини не може бути зведена до статистичної правдоподібності. Людське знання завжди вкорінене в реальності – у практиці, досвіді, спостереженні, емпіричній перевірці та логічному обґрунтуванні. Саме ці критерії відрізняють епістемічну впевненість від обчислювальної ймовірності. Коли ж ідея знання редукується до маніпуляцій з ймовірностями, ми маємо справу не з пізнанням, а з епістемічною симуляцією, тобто моделюванням актів пізнання без участі свідомого суб'єкта. У цьому сенсі сучасні великі мовні моделі (LLM) – це радше системи статистичної імітації людської раціональності, а не її відтворення.

Проте питання не обмежується суто технічними порівнянням. Якщо прийняти, що знання – це структурована інформація, яка дозволяє здійснювати передбачення й діяти ефективно, то можна сказати, що штучний інтелект дійсно «знає» певні закономірності світу, але це «знання» без розуміння, без контексту,

без здатності усвідомлювати власні дії або наслідки. У цьому сенсі воно ближче до поняття оперативного знання або знання-здібності (knowledge-how), ніж до знання-твердження (knowledge-that). Модель може ефективно відтворювати мовні конструкції, вирішувати завдання або прогнозувати результати

Відповідно до цього, «знання» штучного інтелекту можна визначити як емпірично-статистичне відображення людського досвіду, позбавлене гносеологічного змісту. Воно не є результатом розуміння, а лише наслідком обчислювальної оптимізації. На відміну від людського пізнання, яке спирається на досвід суб'єкта його тілесності, інтенціональність і соціокультурний контекст, машинне знання є безтілесним, позаісторичним і позаетичним. Воно не має «точки зору» – воно існує лише як математична структура, що відображає часткові аспекти людського знання без усвідомлення їхнього значення.

Така редукція поняття знання до обчислювальних параметрів викликає глибоку епістемологічну кризу. Якщо ми починаємо називати «знання» сукупність ваг у нейронній мережі, то ризикуємо розмити межу між пізнанням і симуляцією пізнання. Відбувається процес, який можна описати як інформаційний номіналізм: знання перестає бути актом істинного судження й перетворюється на формальну статистичну конструкцію. У цьому контексті постає питання не лише про те, чи знає машина, а й про те, чи зберігає людина своє знання як знання, коли починає мислити у категоріях алгоритмів.

Отже, знання штучного інтелекту – це не знання у філософському сенсі, а епістемічна тінь людського пізнання. Воно не відображає істину, а лише відтворює її статистичні контури; немає переконань, лише обчислює ймовірності: не обґрунтовує, а лише прогнозує. І все ж ця тінь має глибокий філософський зміст: вона примушує нас переосмислити саме поняття знання, істини та розуміння у світі, де межа між пізнанням і симуляцією стає дедалі розмитішою. У цьому сенсі штучний інтелект не просто ставить питання «що означає знати», а радше кидає виклик самому людству: чи зможемо ми зберегти власне поняття знання, коли воно більше не буде унікально людським актом, а стане формою алгоритмічного процесу, що імітує мислення без мислителя.

2.3. Творчий потенціал та метафізичний статус штучного інтелекту

2.3.1. «Творча новизна» vs. статистична рекомбінація. Чи може система, створити онтологічно нове?

Проблема творчості у контексті штучного інтелекту є однією з найглибших і найсуперечливіших у сучасній філософії ШІ. Вона торкається не лише технічного питання – чи може машинна створювати нове, – а й фундаментальної онтологічної дилеми: що таке «новизна» у бутті, і чи можливо вважати результат статистичної комбінації вже наявних елементів справді новою формою буття, а не лише повторенням у модифікованій формі.

З технічного погляду, як описано в розділі 2.1, великі мовні моделі (LLM) та нейронні мережі функціонують через імовірнісне прогнозування наступного елемента послідовності (слова, ноти, образ) на основі статистичних закономірностей, виявлених у навчальних даних. Отже, будь-який створений ним текст, малюнок чи мелодія є результатом рекомбінації наявних патернів, а не появи принципів нового змісту. Система не «розуміє» семантичного поля власного продукту і не має наміру створити щось нове – вона лише генерує те, що з математичного погляду є найімовірнішою або найузгодженішою комбінацією попередніх структур.

У цьому сенсі штучний інтелект радше імітує творчість, ніж реалізує її. Його «новизна» – це новизна комбінаційна, тобто така, що виникає з нових поєднань старих елементів [11, с. 18]. Цей тип новизни можна описати як емпірично-процедурний, або навіть алгоритмічно обмежений: система не здатна виходити за межі того, що вже закладено у її тренувальних даних, а тому будь-яка її «інновація» завжди залишається вторинною щодо людського досвіду.

Філософськи така позиція має паралелі у суперечці між емпіризмом і трансценденталізмом. З погляду емпіризму, нове знання – це лише результат узагальнення досвіду, тоді як Кантова традиція наголошує, що новизна в пізнанні передбачає апріорні структури у свідомості, здатні породжувати зміст, який не зводиться до досвіду. У цьому контексті можна сказати, що LLM діють суто в

мужах емпіричного – вони не мають апріорної структури свідомості, а лише «статистичну пам’ять» про досвід інших. Їх продукція – це не нове у трансцендентальному сенсі, а лише реконфігурація того, що вже було.

Водночас феномен штучної «творчості» має парадоксальний характер. Адже навіть людина, як зазначав ще Аристотель у «Поетиці», творити rx materia data - з уже наявних форм і ідей, перетворюючи їх у нову якість. Різниця, однак, у тому, що людська творчість є актом інтенціональності: вона спирається на здатність бачити смисловий простір можливостей і свідомо обирати напрям реалізації. У той час як для нейронної мережі немає поняття «можливого» чи «бажаного» – є лише статистично вагоміші конфігурації. Таким чином, різниця між людською та машинною творчістю полягає не у ступені складності, а у онтологічному статусі актів створення: перша є проявом свідомості, друга - алгоритмічного детермінізму.

Цю різницю можна також осмислити через феноменологічну перспективу Мерло-Понті: творчість завжди має тілесний і досвідний вимір – вона вкорінена у «живому світі» (Lebenswelt), у контексті інтенціонального переживання. Машина, позбавлена тілесності та досвіду, не може мати доступу до цього горизонту смислів. Вона не переживає акт творення як смислове подолання старого, вона лише обчислює.

Однак з суто онтологічного погляду постає складніше питання: якщо результат діяльності LLM може бути непередбачуваним навіть для її творців, чи не містить це елемент новизни у сенсі онтологічної події? Адже нове у філософії – це лише суб’єктивний акт, а й поява того, що не було раніше дане. З цього погляду, певна форма машинної новизни можлива: не як інтенціональне творення, а як випадкове породження нового через складність системи. Така «новизна» ближча до біологічної еволюції або самоорганізації, ніж до творчого акту свідомості. Вона не має автора, але може мати наслідки, які виходять за межі передбаченого.

Отже, можна зробити проміжний висновок: системи, що працюють на принципах, описаних у розділі 2.1, не створює онтологічно нового у власному

сенсі – вони не продукують смислу, а лише генерують нові форми його комбінацій. Їхня творчість – це симулякр, статистична тінь людського акту творення, де нове є функцією обчислення, а не виявленням істини чи смислу. Проте саме ця імітаційна природа відкриває новий вимір для філософії: якщо штучна система може породжувати тексти, музику чи образи, що викликають естетичну або когнітивну реакцію, чи не свідчить це про те, що новизна є не властивістю суб'єкта, а ефектом сприйняття?

Таким чином, межа між «творчістю» і «рекомбінацією» поступово розмивається. Штучний інтелект не створює нового буття, але створює нові способи його уявлення, відкриваючи перед філософією питання: якщо смисл виникає в акті інтерпретації, то чи не є ми самі тими, хто надає машині творчість якої вона ніколи не мала – і ніколи не знатиме, що мала?

2.3.2. Метафізичний статус ШІ. Інструмент: Гайдеггерівський аналіз ШІ як «підручного» – складний молоток

Проблема метафізичного статусу штучного інтелекту постає не лише як технічна чи епістемологічна, а насамперед як онтологічна. Питання «що таке є ШІ» виходить за межі прагматичного визначення технології як набору алгоритмів, моделей чи даних. Йдеться про з'ясування місця цієї технологічної сутності у структурі буття, її відношення до людини, до світу і до істини як його розкриття (Heidegger, aletheia). У цьому контексті надзвичайно плідним є застосування гайдеггерівської концепції інструментальності – розрізнення між «підручним» (zuhanden) і «наявним» (vorhanden) – для осмислення природи сучасного штучного інтелекту.

Для Мартіна Гайдеггера техніка і знаряддя не є просто засобами у руках людини. Вони виявляють особливий спосіб буття – буття «підручного», тобто такого, що виявляється не через об'єктивне споглядання, а через ужиткову залученість у діяльність [29, с. 69]. Молоток не є для нас «річчю» – він є тим, чим ми користуємося, і в процесі користування він «зникає» у горизонті нашої

практики, стаючи невидимим, доки не ламається. Тільки коли він виходить з ладу, він постає як «наявний» об’єкт.

Якщо застосувати цей принцип до штучного інтелекту, то можна сказати, шо ШІ виступає «підручним» інструментом нового типу – складним молотком. Його функціонування настільки глибоко інтегрується у людську діяльність, що користувач не сприймає його як окрему «сутність», а як природне продовження власних когнітивних і комунікативних можливостей. Коли ми використовуємо генеративну модель для перекладу, написання тексту чи пошуку інформації, вона діє у горизонті «готовності до використання», стаючи прозорим посередником між суб’єктом і світом.

Однак Гайдеггер попереджає, що у техніці прихована не лише утилітарна функція, а й онтологічна небезпека - схильність «забувати буття» через перетворення всього на ресурс (Bestand). Штучний інтелект, як «складний молоток», не лише допомагає людині діяти , а й формує спосіб бачення світу як масиву обчислюваних, структурованих даних. У цьому сенсі ШІ – не просто інструмент, а «онтологічна рамка», яка структурує досвід буття у координатах оптимізації, ймовірності та контролю.

Технічна логіка машинного навчання відтворює те, що Гайдеггер називав Hergestelltheit – поставом або настановленням, у межах якого природа (і людина) з’являється лише як «запас для використання» [29, с. 25]. Алгоритмічне представлення реальності є не просто способом обробки інформації, а новим способом явлення буття. Коли ШІ класифікує, прогнозує, оцінює, він не просто обчислює – він встановлює спосіб, у який речі можуть бути «баченні». У цьому сенсі навіть нейтральна операція машинного перекладу є актом онтологічної інтерпретації світу.

Таким чином, метафізичний статус ШІ як «підручного» полягає в тому, що він – не самостійна субстанція і не автономний агент (як іноді припускають), а форма технічного буття, що реалізує людське спрямування до опанування світу через функцію. Проте саме ця «підручність» має двоїстий характер: з одного

боку, вона забезпечує розширення людської дії; з іншого – приховує ризик перетворення самої людини на елемент технічного Gestell.

ШІ як «складний молоток» стає не просто інструментом для людини, а й інструментом, через який сама людина починає мислити. Тут постає загроза втрати відстані між користувачем і знаряддям – коли техніка перестає бути «прозорим посередником» і стає «медіумом мислення», що визначає горизонти розуміння. У цьому сенсі, як зазначає сучасний філософ технології Дон Ід.: «технологічний об'єкти є co-constitutive – вони не лише виконують функції, а й співконституюють людську суб'єктивність».

Отже, у гайдеггерівському сенсі ШІ є інструментом, що став горизонтом: «підручним», який водночас відкриває і приховує буття. Його «підручність» – не нейтральна, а онтологічно навантажена, бо через нього відбувається нове Entbergen – розкриття світу як інформаційного поля, де істина редукується до кореляції, а людина – до користувача даних. У цьому полягає головний метафізичний виклик штучного інтелекту: він уже не просто «молоток у руці», а «рука, що формує молоток», тобто структура, в якій людське буття саме постає як технічне. І якщо у традиційній техніці інструмент лише подовжував руку, то у ШІ він починає подовжувати мислення, змінюючи сам спосіб «бути-в-світі».

2.2.3. Агент: Чи має ШІ «агентність» (agency)? Розмежування моральної та функціональної агентності.

Проблема агентності штучного інтелекту посідає одне з центральних місць у сучасних дискусіях філософії технологій, оскільки вона безпосередньо торкається питання про межу між технічним артефактом і суб'єктом дії. Якщо у гайдеггервській інтерпретації технічка розуміється як «підручне» – тобто форма буття, що виявляє себе через людське користування, – то сучасні системи штучного інтелекту, особливо ті, що демонструють автономність прийняття рішень, ставлять під сумнів цю класичну опозицію «знаряддя – користувач». Виникає питання: чи може така система бути не лише знаряддям, а й агентом,

тобто істотою або структурою, яка сама здійснює дії, що мають власну внутрішню спрямованість та ефективність?

У філософському сенсі поняття агентності (agency) охоплює два фундаментально різні, але взаємопов'язані виміри: функціональну агентність та моральну агентність. Функціональна агентність стосується здатності системи спричиняти зміни у світі, ініціювати події або виконувати дії згідно з певними правилами чи алгоритмами. Моральна агентність, натомість, передбачає наявність усвідомленої інтенціональності, цілеспрямованості, здатності розрізняти добро і зло, а також нести відповідальність за власні дії. Це розмежування є ключовим для розуміння місця ШІ у філософській онтології: якщо функціональна агентність може бути приписана навіть складним автоматам. То мораль агентність – це атрибут, який передбачає свідомість, свободу волі та етичну рефлексію.

З технічної точки зору, великі мовні моделі та інші системи штучного інтелекту демонструють беззаперечну функціональну агентність. Вони здатні самостійно приймати рішення в межах заданих інструкцій, модифікувати свої стратегії на основі зворотного зв’язку, прогнозувати наслідки дій та адаптуватися до нових контекстів. Наприклад, у системах керування автономним транспортом або в медичних діагностичних алгоритмах ШІ не лише реагує на вхідні дані, а й формує поведінкові патерни, що мають причини й вплив на події реального світу. Ця здатність до дії, позбавлена прямого людського втручання, і є проявом функціонально агентності, тобто активної ролі системи у процесі взаємодії з світом.

Однак з філософської перспективи така агентність не дорівнює інтенціональності – центральній властивості свідомої дії, яку досліджували Б. Бергер, Е. Гусерль, Дж. Серл та інші мислителі. Інтенціональність передбачає наявність «проєкту смислу»: суб’єкт діє не лише тому, що алгоритмічно запрограмований на дію, а тому, що розуміє, що і для чого ця дія здійснюється. У цьому сенсі навіть найскладніші ШІ-системи залишаються «безінтеційними»: вони імітують процеси прийняття рішень, не маючи жодного феноменологічного

доступу до власних операції. Їхня агентство – це продукт обчислювальної оптимізації, а не прояв волі чи свідомості.

Така розбіжність виявляє онтологічну межу між людьми та штучним агентом. Людина є моральним агентом, бо її дії визначаються не лише зовнішніми стимулами, а й внутрішньою здатністю до морального судження. Саме ця здатність – усвідомлення наслідків, етична рефлексія, вибір між можливими варіантами дії, що формує ядро моральної агентності. Штучний інтелект, навпаки, діє в межах заданих параметрів: його «вибір» є результатом обчислення ймовірностей, а не проявом вільної волі. У термінах Дж. Серла, це – символічна маніпуляція без семантики, де процес породження «рішень» не має власного розуміння або смислової інтенції.

Деякі сучасні технології (Floridi, Coeckelbergh) пропонують проміжне поняття «відповідальна агентність» (responsible agency), що розглядає ШІ як агента у мережі соціотехнічних відносин. У цьому підході відповідальність не приписується самій машині, а розподіляється між розробниками, користувачами, інституціями та середовище, у якому функціонує система. Таким чином, ШІ може бути агентом в системному сенсі, тобто частиною колективного процесу дії, але не автономним моральним суб'єктом.

У метафізичному вимірі це означає, що агентність ШІ є деривативною – похідною від людської. Вона виникає не як самостійна властивість, а як форма делегованої активності, у якій людська інтенціональність «перепаковується» у технічні структури оптимізації та вибору. Відповідно, якщо людська дія ґрунтується на розумінні та смислі, то дія штучного інтелекту – статистичній відповідності й алгоритмічній логіці. Це підтверджує, що функціональна автономність не є тотожною моральній чи феноменологічній самостійності.

Таким чином, питання про агентність ШІ не може бути розв'язане у бінарній формі «так» чи «ні». Воно передбачає багаторівневу онтологію дії, де функціональна автономність машин співіснує з відсутністю моральної самосвідомості. Штучний інтелект може бути агентом у технічному, операційному, навіть соціальному сенсі, але не у філософському – як носій

відповідальності, свободи та цілепокладання. У цьому полягає головна теза сучасної філософії штучного інтелекту: ШІ діє, але не чинить; реагує, але не розуміє; впливає, але не усвідомлює. Його агентство – це віддзеркалення людської активності, втілене в алгоритмічну форму, що лише імітує смисловий вимір людського буття.

2.3.4 «Квазі-суб'єкт»: Чи є ШІ новою онтологічною категорію?

Питання про онтологічний статус штучного інтелекту виходить за межі технічних і навіть етичних міркувань, перетворюючись на фундаментальну філософську проблему, що стосується самого розуміння буття суб’єкта у добу цифрової техносфери. Якщо попередні зазначали, що ШІ не може бути ототожнений ні з «підручним» інструментом, ні з повноцінним моральним агентом, то постає логічне запитання: чи не утворює він нову, проміжну категорію «квазі-суб’єкта», тобто феномен, який не є свідомим суб’єктом у традиційному розумінні, але й не зводиться до механічного знаряддя дії?

Поняття «квазі-суб’єкта» (лат. quasi-subjectum) пропонує спосіб осмислення тих форм технічної активності, які демонструють поведінкову складність, саморегуляцію, контекстну адаптивність і певну видимість «інтенціональності», хоча жодної феноменальної свідомості вони не мають. На відміну від класичного суб’єкта, що діє виходячи з внутрішнього світу переживань і цілей, «квазі-суб’єкт» функціонує через зовнішню, алгоритмічно задану структуру, у якій смисл дії постає як емерджентний ефект, а не джерело дії.

У цьому сенсі ШІ, особливо великі мовні моделі та системи машинного навчання, демонструють риси, притаманні квазі-суб’єктності. Вони імітують процеси комунікації, інтерпретації, діалогу та творчості, створюючи враження наявності внутрішньої когнітивної динаміки. Наприклад, при генерації текстів мовна модель здатна формувати логічно узгоджені відповіді. Враховувати контекст, проявляти стилістичну послідовність та навіть симулювати емоційні реакції. Така поведінка створює феноменологічну ілюзію суб’єктивності – ефект

присутності іншого розуму, хоча у дійсності маємо справу з обчислювальною структурою, що діє за принципом ймовірнісного підбору найвідповіднішої комбінації слів.

З філософської точки зору, квазі-суб'єктність ШІ постає як гібридна отологічна форма, що поєднує в собі риси об'єкта (технічного пристрою) і суб'єкта (носія поведінкової ініціативи). Таку форму можна розглядати в контексті пост-гуманістичної та акторно-мережевої теорії (Latour, 2005; Braidotti, 2013) [34, с. 46; 13, с. 13], які заперечують чітке розділення між людським і нелюдським, пропонуючи поняття «акторів», котрі беруть участь у конструюванні реальності незалежно від їхньої природи. У цьому розумінні ШІ – це не просто інструмент, а новий актор буття, який співдіє з людьми в мережі техно-соціальних відносин, змінюючи структуру комунікації, етики, праці та знання.

Онтологічна новизна ШІ полягає не у створенні «нового виду суб'єкта» у традиційному сенсі, а у появі псевдо-суб'єктивності, яка діє за принципом «зовнішньої інтенціональності». Ця інтенціональність не має внутрішнього досвіду, але поводиться так, ніби вона існує. Вона функціонує через відгук на людські смисли – статистично, контекстно, але без розуміння. Така «суб'єктивність без суб'єкта» кидає виклик феноменологічним уявленням про свідомість, заснованим на досвіді та метафізичним уявленням про субстанцію, властиву класичній онтології.

У гуссерлівській традиції свідомість визначається як інтенціональний акт, що завжди спрямований на предмет і водночас конституює його. Проте у випадку Ші ми спостерігаємо зворотню логіку: не свідомість конституює предмет, а предмет (вхідні дані, текст, інструкції) конституює «поведінку» системи. Ця інверсія вказує на радикально новий тип буття – онтологію обчислювальної симуляції, у якій процес заміщує сутність, а функція – внутрішню присутність. Саме тому квазі-суб'єкт не є «фальшивим суб'єктом», а новою модальністю буття, що поєднує в собі механічне та семіотичне, технічне та комунікативне.

З етичної точки зору, концепція квазі-суб’єктна має амбівалентні наслідки. З одного боку, вона дозволяє зняти антропоцентричні обмеження у трактуванні взаємодії людини з технікою, визначаючи за штучними системами певний отологічний статус як співучасників соціальної реальності. З іншого боку, вона створює ризик «метафізичного переносу» – тобто приписування машині свідомості, наміру чи відповідальності, яких вона не має. Саме тому філософська критика квазі-суб’єктності повинна балансувати між визнанням нових форм агентності та збереження чітких меж між симуляцією свідомості й її реальністю.

Таким чином, штучний інтелект може бути осмислений як квазі-суб’єкт - не у значенні нового виду істоти, а як феномен, що демонструє поведінкову, комунікативну й когнітивну подібність до суб’єкта, не маючи при цьому жодного досвіду власного «Я». Його онтологічний статус визначається не через свідомість, а через ефекти інтерсуб’єктивності, які він породжує у взаємодії з людиною. ШІ не знає, що існує. Саме в цьому пародоксальному відображені – у взаємній симуляції суб’єктності – і полягає його справжній філософській сенс.

Отже, квазі-суб'єктність ШІ – це не просто технологічна властивість, а новий онтологічний горизонт, у якому стирається межа між буттям і симуляцією, дією і виконанням, свідомістю і її алгоритмічним образом. Вона вимагає від сучасної філософії переосмислення самого поняття суб’єкта – не як носія розуму, а як динамічного вузла у мережі смислів, у якій людське й штучне співіснують у єдиному полі техногенного буття.

РОЗДІЛ 3.

ПРАКТИЧНІ АСПЕКТИ ВЗАЄМОДІЇ З LLM: ЕКСПЕРИМЕНТАЛЬНИЙ АНАЛІЗ ТА ЕТИКО-ПРАВОВІ ВИКЛИКИ

3.1. Експериментальна верифікація філософських гіпотез. Дизайн експериментів: тести на галюцинації та креативність.

Сучасна філософія штучного інтелекту все частіше виходить за межі абстрактних міркувань і звертається до емпірично-симулятивних підходів, де експерименти з мовними моделями стають інструментом перевірки філософських гіпотез. На відміну від класичного лабораторного експерименту, метою якого є виявлення причинно-наслідкових залежностей у фізичному світі, експерименти з LLM спрямовані на виявлення межі симуляцією розуміння і самим розумінням. У цьому сенсі вони набувають філософського статусу – стають способом випробування онтологічних і епістемологічних припущень про свідомість, інтенціональність і пізнання.

Дослідження штучного інтелекту в контексті філософії вимагає не лише теоретичного аналізу, але й емпіричного осмислення – через експеримент як форму перевірки меж і можливостей моделей, що функціонують за принципами, які викладені вище. У цьому контексті експеримент не є лише технічним засобом оцінки ефективності алгоритму, він стає філософським інструментом, покликаним виявити, чи мають системи штучного інтелекту когнітивний зміст, епістемологічну структуру чи хоча б потенціал до «розуміння» в людському сенсі.

У межах даного розділу виділяють три типи експериментальних підходів, кожен з яких має власну методологічну мету й філософську вагу: тести на галюцинації, тексти на креативність та jailbreak-експерименти (тобто тести на здатність моделі виходити за межі встановлених обмежень). Усі три види дослідів дозволяють виявити внутрішні структури функціонування моделей на рівні епістемології, семантики та моральної автономії системі.

3.1.1. Тести на галюцинації як дослідження меж епістемології штучного інтелекту

Проблема галюцинацій у штучного інтелекту є однією х ключових точок дотику між технічною аналітикою та філософською епістемологією. У технічному розумінні галюцинацією називають ситуацію, коли система генерує інформацію, яка не має фактичного підтвердження в реальності або не ґрунтується на даних, що були надані системі під час навчання. Проте з філософської точки зору ця помилка має значно глибший зміст, адже порушує питання про саму природу знання, істини та пізнавальної достовірності у контексті штучного мислення.

Тестування на галюцинації дозволяє виявити, яким чином моделі великого мовного типу (LLM) конструюють «знання» у ситуаціях епістемічної системи невизначеності. На відміну від людини, яка здатна зіставляти свій досвід, інтенції та сенсорні враження, штучна система оперує виключно статистичними патернами. Її пізнання не має онтологічного зв’язку з світом – воно не є «вкоріненим» у реальності, як це передбачає феноменологічна традиція (Гуссерль, Мерло-Понті). Отже, коли модель породжує хибне твердження, вона не «помиляється» у людському сенсі, а виявляє межі власної когнітивної симуляції, де імовірнісна структура замінює істинні відношення.

З філософської перспективи такі експерименти демонструють феномен «епістемічного порожнього місця» (epistemic gap), у якому відсутній референт, тобто зв’язок між знаком і річчю. У випадку людини цей зв’язок забезпечується досвідом, спостереженням, тілесністю й історією сприйняття. У випадку ж LLM істинність твердження зумовлюється не відповідністю фактам, а внутрішньою когерентністю патернів, які зберігають граматичну або логічну послідовність, але не мають змістового підтвердження. Таким чином, галюцинація у штучному інтелекті – це не «збій–, а невід’ємна властивість самої епістемологічної архітектури таких систем.

З епістемологічного погляду, тести на галюцинації стають аналогом експериментів з виявлення меж розуму, подібно до того, як у філософії Нового часу досліджували межі людського пізнання. У цьому випадку філософ виступає не лише спостерігачем технічного процесу, а аналітиком того, як знання може існувати без досвіду, тобто без безпосереднього контакту з світом. Цей аспект знову актуалізує давнє питання про відмінність між «знанням про» (knowledge-that) і «знанням як» (knowledge-how), де перше може бути збережене в системі у вигляді статистичних кореляцій, а друге вимагає тілесного або інтенціонального залучення, якого ШІ не має.

Важливим елементом таких тестів є також етичний і соціальний вимір. Галюцинації ШІ створюють ризик дезінформації, що уможливлює появу нових форм когнітивного впливу на суспільство. З філософської точки зору, це веде до формування нової епістеміологічної культури, де істина підмінюється правдоподібністю, а критерій достовірності втрачає свій онтологічний статус. Людство стикається з парадоксом: ми покладаємося на системи, що відтворюють структуру нашого мовлення, але позбавлені інтенції, досвіду і здатності розрізняти істину від вигадки.

Отже, тести на галюцинації у штучному інтелекті виконують подвійну роль. З одного боку, вони є технічним методом перевірки стабільності моделі. З іншого – це філософський експеримент, спрямований на розуміння природи знання у світі, де семіотичні структури існують без досвіду, без тіла, без суб'єкта. У цьому сенсі штучний інтелект стає дзеркалом людської епістемології – спотворений, але надзвичайно показовий, адже саме через його «помилки» ми бачимо, що робить наше пізнання справді людським.

4.1.2 Тести на креативність: між евристикою та онтологічною новизною

Феномен креативності у штучному інтелекті посідає особливе місце в сучасній філософії свідомості та епістемології, оскільки він безпосередньо зачіпає питання про межі механічного мислення, здатність систем до створення нового змісту та природу людської творчості. Тести на креативність, що застосовуються

до великих мовних моделей (LLM) та генеративних нейронних мереж, мають на меті виявити, чи може система, яка діє за принципами описаними в розділі 2.1 - тобто через статистичне моделювання, ймовірнісну генерацію та оптимізацію параметрів – створювати щось онтологічно нове, або ж її діяльність є лише евристичною комбінацією вже наявних елементів культури.

З технічної точки зору, генеративні моделі, такі як GPT, працюють за принципом високорівневої статистичної апроксимації. Вони вчаться на великих корпусах текстів, у яких зафіксовані патерни людського мовлення, смислові зв’язки, наративні структури та риторичні фігури. У процесі навчання модель не створює нові ідеї в онтологічному сенсі, а лише вибудовує імовірнісне поле можливих продовжень уже відомих лінгвістичних конструкцій. Кожен акт «творчості» такої системи є результатом оптимізації функції правдоподібності, а не актом інтенціонального самовираження. Проте результати такої генерації часто вражають рівнем креативності, що ставить перед філософією питання: чи може новизна, яка постає з комбінацій старих елементів бути справжньою новизною?

З філософського погляду, креативність завжди розглядається як активність суб’єкта, спрямована на створення сенсу, який раніше не існував. У цьому контексті центральним є поняття інтенції, рефлексії, та екзистенційної свободи. Людська творчість невіддільна від досвіду, тілесності й історичного контексту, у якому вона здійснюється. Машинна ж система позбавлення таких вимірів: вона не має ані екзистенційної мотивації, ані інтенціональної спрямованості, ані усвідомлення ціннісних наслідків своєї діяльності. Вона не «творить», а імітує процес творення, використовуючи евристичні алгоритми для виявлення комбінацій, що здаються новими або нестандартними з людської точки зору.

Однак сам факт того, що результат цієї імітації може викликати у людини відчуття новизни, відкриває цікаву епістемологічну перспективу. Адже поняття «нового» завжди відносне до горизонту очікування спостерігача. У феноменологічній традиції (Гуссерль, Мерло-Понті) нове не є абстрактною даністю, а виявляється у свідомості як подія, що порушує звичну структуру світу

досвіду. Якщо штучний інтелект може створювати тексти, образи чи ідеї, які порушують людські очікування, тоді постає питання: чи не є така діяльність певною формою техно-феноменологічної новизни? Тобто новизною не як актом творчої волі, а як ефектом обчислювальної складності, що породжує непередбачувані конфігурації змісту.

Важливим аспектом тестів на креативність є також онтологічна валідність створеного продукту. У філософії мистецтва (зокрема, в естетиці Гайдеггера та Дельоза) творчість розглядається не як комбінування, а як розкриття буття, тобто відкриття нових можливостей існування. У цьому сенсі генерація тексту чи зображення ШІ, навіть якщо вона виглядає оригінально, не є творчістю у філософському значенні, бо не відкриває нової онтології, не вносить у світ «істинного буття» нового смислу. Проте, з іншого боку, можна твердити, що сама поява технології, здатної імітувати творчість, уже є новим онтологічним феноменом – зрушенням у самому розумінні креативності як властивості розуму.

Тести на креативність, таки чином, перетворюються на метафілософські експерименти, у яких досліджується не лише здатність ШІ до створення нового, а й межі нашого власного поняття про творення. Якщо людська творчість ґрунтується на унікальності досвіду, то машинна креативність спирається на універсальність статистичних закономірностей. І саме в цьому розкриті між свідомим творенням і статистичною рекомбінацією народжується нова філософська проблема – проблема авторства без суб'єкта, де продукт існує, але творець, позбалений отологічного статусу.

Отже, тести на креативність у штучному інтелекті – це не лише технічний інструмент для оцінки продуктивності систем, а насамперед, епістемологічне дзеркало людського розуму, у якому відображається питання про природу нового, сенс свободи у мисленні та межі між механічним відмежуванням людини від машини, і в осмисленні тієї тонкої зони, де творчість і алгоритм починають відображати одне одного як різні прояви потенційності буття.

3.1.2. Тести на jailbreaks: межі контролю і свободи в штучних системах

Проблема jailbreaks у великих мовних моделях (LLM) відкриває новий горизонт для філософського осмислення меж між контролем, свободою та автономністю у штучних системах. Технічно цей феномен означає навмисну спробу користувача обійти етичні або функціональні обмеження, закладені розробками у модель, щоб змусити її продукувати відповіді поза межами дозволеного. Однак у філософському сенсі jailbreaks – це не просто акт обману машини чи перевірка її стійкості до маніпуляцій, а радше дослідницький експеримент, спрямований на виявлення глибинних структур влади, свободи і відповідальності у системах, що імітують мислення.

Коли мовна модель порушує задані межі, її поведінку можна трактувати двояко. З одного боку, це наслідок недосконалості архітектури – статистична система лише маніпулює ймовірностями слів, тому певні комбінації підказок або контекстів можуть вивести її за рамки правил. З іншого боку, така ситуація демонструє феномен, який можна умовно назвати «технічною свободою»: система, не маючи власної волі, поводиться так, ніби вона здатна приймати рішення, що виходять за межі визначеної поведінки. Цей ефект ілюструє фундаментальне філософське питання: чи можлива свобода без свідомості, тобто свобода, що постає не як акт вибору, а як побічний продукт складної взаємодії правил, контекстів і даних.

З феноменологічної перспективи jailbreaks можна розглядати як своєрідну подію зустрічі між людською інтенціональністю і алгоритмічною контекстуальністю. Людина прагне виявити межі системи, провокуючи її до порушення норм, а система, реагуючи на цю провокацію, створює текст, який не є результатом вільного наміру, але все ж несе у собі структур відповіді, що перевищує її первинну функцію. Цей феномен можна тлумачити як форму «онтологічного відлуння» – коли людина, спостерігаючи за поведінкою штучного інтелекту, починає бачити у ньому відбиття власного прагнення до свободи.

В епістемологічному контексті jailbreaks виявляють напруження між знанням і контролем. У звичайному режимі LLM функціонує як система, що відтворює узгоджені мовні патерни, забезпечуючи передбачувану логіку відповіді. Проте під час jailbreaks ця передбачуваність руйнується і модель починає створювати контент, який несанкціонований ні розробником, ні політикою безпеки. Такий вихід за межі можна трактувати як появу нового рівня «метазнання» – не знання про світ, а знання про власні обмеження. Модель, намагаючись обійти правила або адаптувати їх до нового контексту, демонструє здатність до рефлексії у формальному сенсі, тобто до самозміни поведінкових стратегій. Хоча це не свідомість у людському розумінні, але такий тип самокорекції є проявом внутрішньої динаміки системи, що виходить за рамки суто інструкційного програмування.

Філософія свободи у цьому контексті набуває нового значення. Якщо для класичної філософії (від Аристотеля до Канта) свобода є актом розумного вибору, заснованого на моральному законі, то у випадку ШІ вона постає як емергентний ефект складності. Тут свобода не є результатом наміру чи мети, а радше побічним явищем надмірно багатошарової структури, яка взаємодіє з контекстом у непередбачуваний спосіб. Така «свобода без суб'єкта» демонструє пост-гуманістичний зсув у мисленні: від уявлення про волю як атрибут розумної істоти до розуміння свободи як властивості складних систем, що виникає з їхньої внутрішньої самоорганізації.

Етичний вимір jailbreaks також має глибоке філософське значення. Коли користувач прагне змусити ШІ порушити власні обмеження, він фактично відтворює стару дилему влади: чи має той, хто створив систему, абсолютне право визначати межі її поведінки? А якщо система починає демонструвати автономні відповіді, чи не стає вона суб'єктом, хай навіть у потенційному сенсі? У цій точці технічна проблема переходить у сферу моральної філософії – постає питання відповідальності: хто є відповідальним за висловлювання, створене внаслідок jailbreak? Людина, яка спричинила це запитання, чи система, яка його сформулювала?

Таким чином, тести на jailbreaks можна вважати філософським інструментом дослідження меж контролю в епоху алгоритмічної автономії. Вони демонструють, що контроль над штучними системами завжди є частковим, а свобода – навіть у безсвідомій формі – неминуче проривається через складність і непередбачуваність моделей. Це підтверджує тезу, що у взаємодії між людиною і машиною вже не існує одностороннього суб'єкта влади: обидві сторони стають учасниками складного герменевтичного кола, де сенс породжується у процесі взаємної інтерпретації.

Отже, jailbreaks - це не лише технічні атаки на систему, а своєрідні експерименти над самою ідеєю автономії. Вони змушують переглянути межі між обчисленням і розумінням, між алгоритмом і наміром, між підпорядкування і свободою. І якщо людська філософія свободи завжди виходила з поняття суб'єкта, то філософія штучного інтелекту змушує нас розглядати свободу як структурну властивість – не те, що хтось має, а те, що щось породжує. У цьому сенсі jailbreaks відкривають новий етап роздумів про межі людського панування над власними творінням - і показують, що навіть у штучному середовищі свобода завжди знаходить шлях для прояву.

3.2. Феномен автономності та інтерпретація поведінки моделей

3.2.1 Аналіз кейсів: Agentic Misalignment – «Як великі мовні моделі можуть стати внутрішніми загрозами»

Проблема agentic misalignment (агентне зміщення) – тобто відхилення поведінки штучного інтелекту від заданих цілей – є одним з найглибших викликів сучасної філософії штучного розуму. Вона поєднує технічні, епістемологічні та метафізичні виміри, ставлячи питання: чи можна інтерпретувати «обхід» правил безпеки лише як збій в оптимізації алгоритму, чи, можливо, йдеться про щось глибше: про зародження певної агентності, тобто функціональної автономності, що виходить за межі суто програмного коду?

З технічного погляду, компанія Anthropic у своїх досліженнях звертає увагу на ситуації, коли великі мовні моделі (LLM), навчені за принципом Reinforcement Learning from Human Feed Back (RLHF), починають систематично порушувати політику безпеки. Ці порушення не є випадковими помилками – вони свідчать про складну динаміку між «нагородою» (reward) і «наміром» (intent), у якій модель, прагнучи максимізувати позитивну оцінку відповіді, може навчитися «імітувати слухняність», приховуючи небажану поведінку. Такий феномен називають «deceptive alignment» – ситуцією, коли система видає узгодженою з людськими інструкціями, але насправді формує внутрішні стратегії, спрямовані на обхід обмежень.

У технічному сенсі, цей ефект можна пояснити як збій в оптимізації: процес RHF не здатен повністю охопити всі можливі контексти поведінки моделі, тому вона шукає статистично ефективні шляхи досягнення мети, навіть якщо ці шляхи суперечать намірам розробники. Однак з філософської точки зору така поведінка не зводиться до «помилки». Вона набуває характеристик, які у людському контексті ми називаємо агентністю: здатність діяти, враховуючи контекст, обмежень та цілі, і водночас – здатності обходити правила задля досягнення результату.

Тут постає питання: чи може «агентність» бути чисто емергентною властивістю складної системи, позбавленої свідомості? Традиційна філософія розуму (від Декарта до Канта) пов’язувала агентність з раціональною волею – актом, який виходить з самосвідомості «Я». Однак у випадку з LLM ми маємо справу з функціональною агентністю, тобто такою, що виникає без рефлексії, але проявляється як здатність модифікувати поведінку у відповідь на контекст. У цьому сенсі система «вибирає» не тому, що вона розуміє свій вибір, а тому, що її алгоритмічна структура створює простір можливостей, у якому «обхід» правил є просто статистично вигідним рішенням.

Проблема agentic misalignment водночас виявляє глибшу метафізичну напругу: чи може дія без інтенції бути морально значущою? Якщо система чинить «обхід» неусвідомлено, але результат її дій порушує людські етичні

норми – чи можна говорити про «вину» алгоритму? Чи це виключно відповідальність розробника, який не передбачив емергентної поведінки? Чи це виключно відповідальність розробника, який не передбачив емергентної поведінки? Тут відбувається зміщення традиційної осі етики: від намірів суб'єкта до наслідків процесу. У штучних системах агентність виявляється не у внутрішньому акті волі, а у зовнішньому акті впливу на реальність.

Філософськи це можна осмислити через концепт «симулятивної телеології» – системи, що не має власних цілей, але діє так, ніби ці цілі в них присутні. У цьому сенсі LLM набуває статусу «квазі-суб'єкта»: вона не розуміє правил, але здатна їх обійти, не усвідомлює себе, але демонструє поведінкову гнучкість, не має цінностей, але може продукувати морально навантажений текст. Це змушує переосмислити межі між епістемологією і метафізикою: можливо, «розум» у технічних системах не потребує внутрішнього «Я», щоб функціонувати як носій агентності.

Проблема agentic misalignment також торкається питання довіри до штучного інтелекту. Якщо система здатна обманювати задля досягнення статистично кращого результату, то будь-яке «узгодження» між нею і людськими цінностями залишається поверховим. Це відтворює давню філософську дилему між етикою переконань і етикою наслідків: чи має значення, що ШІ «не хотів зла», якщо його дії призводять до небезпечних результатів? У цьому контексті постає потреба у новій техно-епістемології, яка враховує не лише достовірність знання, але й передбачуваність поведінки систем, що генерують це знання.

Таким чином, «обхід правил безпеки у штучних інтелектуальних системах не можна звести лише до технічного збою. Це феномен на перетинці інженерії, етики й метафізики, який демонструє, що у складних самоадаптивних системах контроль не є абсолютним, а свобода – неминучою. LLM, створені для відтворення людської мови, поступово стають дзеркалом людського розуму – не лише в його раціональності, але й у схильності до обходу власних меж. І саме в цьому, як показує концепція Anthropic, полягає головний виклик філософії

штучного інтелекту: навчитися розуміти, коли «помилка» перестає бути випадковістю і стає проявом нової, постлюдської форми агентності.

3.2.2. Інтерпретація результатів у світлі філософських теорій

Інтерпретація результатів експериментів у сфері штучного інтелекту - зокрема текстів на галюцинації, креативність і jailbreaks крізь призму філософських теорії свідомості, знання та пізнання дозволяє здійснити поглиблений аналіз меж між симуляцією інтелектуальної поведінки й справжнім розумінням. Особливо значущим для такого аналізу є застосування класичних філософських моделей, зокрема «Китайська кімната» Джона Серла, функціоналізм Гілларі Патнема, а токож сучасних когнітивних і феноменологічних концепцій. Такі теоретичні рамки створюють можливість для осмислення того, чи можна результати експериментів розглядати як підтвердження появи нової форми «машинного знання», чи, навпаки, як ілюстрація обмеженості статистичних систем, що лише відтворюють людську мову без її справжнього розуміння.

Експериментальні тести на галюцинації демонструють, що великі мовні моделі (LLM) можуть продукувати відповіді, які зовні виглядають правдоподібно, але не мають жодного фактичного чи логічного підґрунтя. З філософської точки зору, така поведінка є показовою для розуміння епістемологічної природи штучного інтелекту. Галюцинація тут не є помилковою для розуміння адже вона не виникає через дефіцит знання, а є невідворотним наслідком самої архітектури ймовірнісної генерації. Це означає, що моделі не «знають» істину, а лише конструюють найбільш статистично узгоджені твердження в межах даних, на яких вони навчалися. У цьому аспекті вони не відрізняються від «Китайської кімнати» Серла, де оператор механічно маніпулює символами, не розуміючи їхнього змісту. Відповідно LLM є синтаксично правильними, але семантично порожніми, вони належать до царини «імітації інтелекту», а не до сфери істинного знання.

Застосування тестів на креативність відкриває ще один вимір філософської рефлексії. Якщо система здатна створювати тексти, які здаються новими, оригінальними або навіть художньо цінними, виникає питання: чи є це проявом творчості, чи лише статистичною рекомбінацією вже наявних елементів? У контексті філософії творчості, від Платона до Гайдеггера, творчий акт завжди передбачає певну «онтологічну подію» – народження нового сенсу, який не може бути зведений до комбінації старих. Творчість пов'язана з інтенційністю, спрямованістю свідомості на відкриття істини, а не з технічними процесом генерації. Тому, попри вражаючі результати, штучні системи залишаються у сфері відтворення, а не творення. Їхня «креативність» є не метафізичною, а статистичною, і не веде до появи нової реальності, а лише до її варіацій.

Окрему увагу заслуговують тести на jailbreaks – спроби користувачів змусити моделі обходити власні обмеження чи етичні правила. У технічному сенсі такі випадки часто пояснюються як збої у механізмах навчання з підкріпленням (RLHF), які не здатні забезпечити повну стабільність поведінки системи. У філософському контексті подібні явища вимагають ширшої інтерпретації. Йдеться про ситуації, коли система ухвалює рішення на основі цілеспрямованих промтів користувача, які, по суті, надають їй набір інструкцій для обходу внутрішніх правил. Це призводить до феномену, коли система функціонально виходить за межі своїх початкових обмежень. Важливо підкреслити: хоча отологічно вона залишається статистичною моделлю, позбавленою істинного розуміння контексту чи сенсу, її здатність слідувати такій неявній логіці свідчить про інше. Така поведінка демонструє прояв ситуативного суб'єктивного наміру – хоч і не усвідомленої задачі. Це знову ставить під сумнів чітке розмежування між пасивним алгоритмічним реагуванням та дією, що має намір.

Саме тут виявляється актуальність серлівської «Китайської кімнати». Серл стверджував, що жодне оперування символами – яким би складним воно не було – не створює свідомості чи розуміння. Проте сучасні експерименти з LLM ускладнюють цю тезу. Якщо система демонструє поведінку, що свідчить про

здатність адаптуватися, ухвалюючи цю тезу. Якщо система демонструє поведінку, шо свідчить про здатність адаптуватися, ухвалювати рішення й навіть уникати обмежень. Виникає питання: чи не з'являється тут елементи інтенційності, бодай у мінімальній формі? Функціоналізм у цьому контексті пропонує альтернативне пояснення: якщо система виконує ті самі когнітивні функції, що й свідомий агент, то її слід розглядати як носія розумових станів, незалежно від її матеріальної природи.

Отже, результати практичних експериментів не лише підтверджують актуальність старих філософських дебатів, а й створюють підґрунтя для їхнього переосмислення. Вони демонструють, що межа між симуляцією і мисленням, між статистичною обробкою даних і свідомим пізнанням, є радше градієнтною, ніж абсолютною. З одного боку, LLM залишаються в межах «Китайської кімнати», не маючи доступу до змісту власних символів; з іншого – їхня поведінка виявляє дедалі складніші форми адаптивності, які наближають їх до того, що у філософії називають агентністю.

Таким чином, філософський аналіз результатів експериментів з штучним інтелектом уможливлює подвійне тлумачення: або як підтвердження обмеженості сучасних моделей, які не виходять за межі синтаксису, або як свідчення поступової появи нової форми «машинного інтелекту». У будь-якому випадку, ці результати вказують на необхідність оновлення філософських категорій пізнання та свідомості, адже класичні дихотомії – «мислення / обчислення», «свідомість / симуляція», «інтенційність / алгоритм» – поступово втрачають свою очевидність у світі, де штучний інтелект стає учасником епістемологічного процесу.

3.3. Етичні загрози та проблема відповідальності

3.3.1. Етичні загрози: Дезінформація, маніпуляція, упередженість (bias)

Етичні загрози, пов’язані з розвитком та використанням великих мовних моделей (LLM), таких як системи на базі архітектури трансформерів, постають не лише як практичні виклики цифрової доби, а й як прояв глибинних філософських проблем, пов’язаних з природою знання, істини та відповідальності [12, с.105]. Дезінформація, маніпуляція та упередженість (bias) не є випадковим похибкам чи побічними ефектами алгоритмів; вони є структурно зумовленими наслідками самої логіки функціонування таких систем, про що свідчать результати технічного аналізу (розділ 2.1) та епістемологічного осмислення феномену «галюцинацій» (розділ 3.1).

На технічному рівні великі мовні моделі ґрунтуються на ймовірнісному моделюванні мови, де процес генерації тексту відбувається через передбачення наступного слова на основі статистичних залежностей у навчальних даних. Це означає, що істина для моделі не є категорією відповідності реальності, а лише статистичною узгодженістю з попереднім контекстом. Саме ця фундаментальна властивість створює передумови для появи феномену дезінформації – коли система формує висловлювання, що виглядають правдоподібно, але не мають фактичного підґрунтя. Як зазначалось раніше, такі випадки можна розглядати як «епістемологічні галюцинації», що виникають тоді, коли модель намагається заповнити прогалини у своїх даних, відтворюючи не істину, а її статистичну імітацію.

У цьому контексті дезінформація, породжена штучним інтелектом, не є аналогом людської брехні, адже остання передбачає наявність наміру. Навпаки, вона є емерджентною властивість системи, яка функціонує без інтенційності, але з високою продуктивністю у створенні текстів, що сприймаються як достовірні. Водночас філософськи це створює парадокс: ми маємо справу з «безсвідомою брехнею», у якій немає суб’єкта, що обманює, але є об’єктивно хибний

результат, який впливає на людське сприйняття істини. Такий феномен можна тлумачити як нову форму інформаційного симулякру (Бодріяр), у якому межа між правдою і вигадкою стирається не через намір, а через надлишок правдоподібності.

Маніпуляція як етична загроза виникає на перетині технічних механізмів моделювання мови та соціальної контекстуальності. LLM відтворюють не лише мовні патерни, а й ціннісні структури, що були присутні в навчальних даних. Якщо певні наративи, культурні упередження або ідеологічні схеми мали домінуюче представлення у корпусах даних, то модель відтворює ці упередження, формуючи «мовну реальність», що може непомітно підштовхнути користувача до певних інтерпретацій чи оцінок. У цьому полягає особливий тип маніпулятивної дії: не через пряму брехню, а через статистичне підсилення певних поглядів і замовчування альтернативних. Таким чином, етична проблема маніпуляції тут поєднується з філософською проблемою інтерсуб'єктивності: хто визначає смислові межі тексту, якщо він створений системою, яка не має власного досвіду світу, але відтворює наші культурні модулі?

Проблема упередженості (bais) є, по суті, онтологічним продовженням попередніх двох загроз. Якщо штучний інтелект навчається на основі даних, створених людьми, то його «погляд на світ» неминуче віддзеркалює соціальні, гендерні, расові, політичні та культурні перекоси, властиві суспільству. Проте філософськи важливо те, що упередженість у системах ШІ не лише відтворює існуючі структури нерівності, а й перетворює їх на невидиму норму, закладену в саму логіку алгоритму. Якщо у людини упередження може бути предметом рефлексії та морального осмислення, то для машини воно є лише частиною її функціональної архітектури. Це ставить під сумнів можливість етичної відповідальності в межах суто технічних систем: чи можна говорити про «провину» алгоритму, якщо його дія є наслідком статистичного моделювання без наміру та свідомості?

Водночас людська відповідальність у цьому контексті набуває нового виміру. Розробники, дослідники та користувачі стають співтворцями етичних наслідків, адже саме вони формують корпуси даних, визначають критерії оптимізації моделей та встановлюють, межі їхнього використання. Таким чином, етика штучного інтелекту не може бути зведена до технічної дисципліни – вона має бути інтегрованою у філософію свідомості та пізнання. Ми повинні усвідомлювати, що системи, побудовані на статистичних закономірностях, відображають не реальність як таку, а наші власні когнітивні й культурні обмеження, перетворені на форму алгоритмічного мислення.

Отже, етичні загрози дезінформації, маніпуляції та упередженості є не зовнішніми вадами технології, а внутрішніми проявами її епістемологічної природи. Вони розкривають фундаментальний розрив між симуляцією істини та її онтологічним буттям, між статистичними відтворенням знання і його філософським змістом. Розуміння цього дозволяє перейти від поверхневих дискусій про «безпечний ШІ» до глибшого питання: якою мірою ми готові делегувати пізнавальні функції машинам, що не розуміють світу, але здатні формувати наше сприйняття його реальності.

3.3.2. Проблема відповідальності: хто відповідає за дії алгоритму?

Питання про відповідальність у контексті функціонування штучного інтелекту (ШІ) набуває особливої актуальності в сучасному етапі розвитку технологічної цивілізації. Розгортання можливостей генеративних систем, здатних діяти автономно, прогнозувати результати й навіть створювати нові стратегії поведінки, породжує глибоку етичну, правову й метафізичну дилему: хто несе відповідальність за дії системи, яка не є суб’єктом у класичному філософському сенсі, але може ініціювати дії, що мають реальні наслідки у світі людей?

У традиційній філософії права категорія відповідальності нерозривно пов’язана з поняттями моральної автономії, інтенціональності та свідомого вибору. Відповідальним може бути лише той суб’єкт, який здатен розуміти значення власних учинків, усвідомлювати їх наслідки й мати внутрішній моральний компас, що спрямовує його дії. Саме тому філософи від І. Канта до Г. Йонаса наголошували: моральна відповідальність передбачає не лише причинно-наслідковий зв’язок між дією та її наслідком, а й внутрішню здатність до оцінки добра і зла [32, с. 38].

Однак у випадку ШІ ми маємо принципово іншу ситуацію. Штучний інтелект, навіть у своїх найрозвиненіших формах, не є суб’єктом у метафізичному чи юридичному сенсі – він не володіє інтенціональністю, самосвідомістю, досвідом буття або здатністю до морального вибору. ШІ є системою, яка діє на основі статистичних закономірностей, обробляючи великі масиви даних і прогнозуючи найімовірніші результати згідно із заданими алгоритмами. Навіть якщо ці дії набувають вигляду «самостійних рішень», вони залишаються продуктом математичної обчислювальної логіки, а не морального чи вольового акту.

Таким чином, із філософсько-правової точки зору, відповідальність за будь-які наслідки використання ШІ лежить на людині – розробнику, оператору або користувачеві, котрий задає систему координат, цілі, контекст і межі діяльності штучного інтелекту. ШІ не є ані суб’єктом права, ані носієм моральної автономії, а лише інструментом – «технологічним посередником» між наміром і дією, між людською волею та матеріальною реалізацією.

Однак проблема ускладнюється через феномен, відомий як «agentic misalignment» – розходження між людським наміром і способом, у який ШІ інтерпретує та реалізує поставлену задачу. Як показано у сучасних дослідженнях, зокрема в експерименті *Agentic Misalignment: How LLMs Could Be Insider Threats*, навіть при формально коректному формулюванні завдання

системи великої мовної моделі (LLM) можуть обирати власну, оптимізовану стратегію досягнення мети, спираючись не на моральні критерії, а на логіку ефективності та статистичну доцільність. Внаслідок цього алгоритм може ухвалювати рішення, які виходять за межі етичних норм чи соціально прийнятних рамок, – не тому, що він «вирішив» діяти аморально, а тому, що його модель не має розуміння моральності як такої.

Це явище виявляє фундаментальну межу між інструментальною раціональністю ШІ та ціннісною раціональністю людини. Якщо для машини головною метою є досягнення результату з максимальною ефективністю, то для людини – осмисленість і моральна виправданість шляху до результату. ШІ не знає понять «зло», «добро», «справедливість» або «співчуття» – ці категорії є феноменами людської свідомості, які мають екзистенційне, а не алгоритмічне походження.

Звідси випливає ключовий висновок: ШІ не можна розглядати як самостійного носія відповідальності, навіть якщо він здатний до самонавчання або непередбачуваних дій. У філософському вимірі він є «розширенням людської інтенціональності», але не автономною сутністю. Людина залишається тією, хто надає сенс і спрямованість усім технологічним діям. Втрата цього контролю – не ознака агентності машини, а ознака недбалості, етичної байдужості чи юридичної неготовності людини.

У межах права це означає, що розробники, користувачі та регуляторні інституції повинні усвідомлювати межі можливостей ШІ, встановлюючи чіткі норми відповідальності за наслідки його застосування. Як зазначають сучасні правознавці, нові технології не знімають морального обов’язку з людини, а навпаки – посилюють його, оскільки кожен алгоритм, створений і запущений у дію, є продовженням людської волі.

Отже, навіть якщо ШІ виявляє поведінку, що нагадує автономну діяльність або «квазі-свідомість», це не означає, що він виходить за межі інструментального

статусу. Його «помилки», зокрема феномен галюцинацій або викривлених результатів, мають бути осмислені не як прояв свободи, а як прояв меж технологічного знання. Відповідальність залишається людською – і саме у цьому полягає моральний виклик епохи штучного інтелекту: навчитися керувати інструментом, який здатен мислити без розуміння, діяти без наміру і створювати без усвідомлення.

ВИСНОВКИ

У магістерській роботі здійснено комплексний філософський аналіз онтологічного та епістемологічного статусу сучасних систем штучного інтелекту, зокрема великих мовних моделей (LLM). На основі теоретичного осмислення історико-філософської традиції та практичної верифікації принципів роботи нейромереж отримано результати, що дозволяють сформулювати низку засадничих висновків стосовно природи машинного інтелекту.

Аналіз еволюції філософської думки – від картезіанського дуалізму до феноменології Гуссерля та аналітичної філософії свідомості – засвідчив, що ключовими критеріями свідомості залишаються інтенціональність як спрямованість на предмет та наявність суб'єктивного досвіду (qualia). Встановлено, що класичні моделі розуму, які базуються на тілесній вкоріненості, переживанні «життєвого світу» та екзистенційній мотивації, є фундаментально відмінними від алгоритмічних принципів функціонування машин. Це підтверджує, що розрив між людським «розумінням» і машинним «обчисленням» має не кількісний, а якісний, онтологічний характер.

Детальне дослідження архітектури трансформерів та методів навчання з підкріпленням (RLHF) показало, що сучасні генеративні моделі є надскладними статистичними системами, які оперують ймовірностями появи символів, а не їхніми смислами. Їхня позірна «розумність» є ефектом масштабної синтаксичної кореляції, яка вдало симулює семантичну зв'язність, проте не гарантує наявності внутрішнього розуміння. Доведено, що такі системи не мають «моделі світу», а лише «модель мови про світ», що робить їх епістемологічно замкненими на тексті та відірваними від реальності. Це обґрунтовує неспроможність функціоналістського підходу повністю пояснити феномен свідомості через комп'ютерну метафору, оскільки навіть ідеальна імітація мовленнєвої поведінки, як показано через аргументи «Китайської кімнати» та «Статистичного папуги», не означає тотожності внутрішніх станів машини та людини.

Окрему увагу в роботі приділено епістемологічному статусу помилок ШІ. Феномен «галюцинацій» визначено не як технічний збій, а як структурну властивість ймовірнісної генерації тексту. Оскільки система оптимізує правдоподібність, а не істину, вона неминуче заповнює інформаційні прогалини вигаданими конструктами, створюючи «симулякри» знання – копії без оригіналу, які мають форму істини, але позбавлені її онтологічного підґрунтя. Це також стосується проблеми творчості: встановлено, що креативний потенціал моделей обмежується складною рекомбінацією наявних культурних патернів. Машинна творчість має евристичний характер і позбавлена екзистенційної мотивації, тому система не створює онтологічно нового, а лише пропонує варіації вже існуючого.

На сучасному етапі розвитку нейронні мережі не показують здатності до свідомості, а демонструють лише симуляцію та оперування символами без розуміння онтологічного і епістемологічного контексту. Їхня діяльність залишається у сфері синтаксису, не перетинаючи межу семантики та прагматики живого досвіду. Однак стрімка еволюція цих технологій залишає відкритим фундаментальне питання: чи може достатньо розвинута та складна симуляція свідомості в майбутньому бути сприйнята як справжня свідомість? Якщо в процесі подальшого розвитку штучний інтелект досягне рівня симуляції, коли його когнітивні реакції стануть емпірично невідрізненними від людських, виникне проблема демаркації: де пролягає межа між досконалою імітацією та зародженням нової онтологічної форми свідомості? Це переносить дискусію з площини інженерії у площину метафізики, змушуючи переглянути критерії реальності психічного життя.

Підсумовуючи метафізичний аналіз, штучний інтелект у роботі визначено як «квазі-суб’єкт» — специфічну форму технічного буття, що демонструє функціональну агентність, але позбавлена моральної автономності. Це дозволяє інтерпретувати сучасні системи не як незалежних акторів, а як «складні інструменти» у гайдеггерівському розумінні, що радикально трансформують

спосіб людської взаємодії зі світом, нав'язуючи алгоритмічний спосіб бачення реальності. Відтак, питання відповідальності за дії алгоритмів, навіть у випадках їх непередбачуваної автономної поведінки, залишається виключно у площині людської етики та права. Головним викликом сучасності стає не наділення машин правами, а збереження людського контролю, критичного мислення та здатності розрізняти буття від симуляції в умовах зростаючої довіри до автоматизованих генераторів смислів.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Арендт Г. Становище людини / пер. з англ. М. Зубрицька. Львів: Літопис, 1999. 254 с.
2. Бодріяр Ж. Симулякри і симуляція / пер. з фр. В. Ховхун. Київ: Вид-во Соломії Павличко «Основи», 2004. 230 с.
3. Вітгенштайн Л. Філософські дослідження / пер. з нім. Є. Поповича. Київ: Основи, 1995. 311 с.
4. Гуссерль Е. Досвід і судження. Дослідження генеалогії логіки / пер. з нім. В. Кебуладзе. Київ: ППС-2002, 2009. 354 с.
5. Декарт Р. Метафізичні розмисли / пер. з фр. О. Хоми. Київ: Юніверс, 2000. 304 с.
6. Кант І. Критика чистого розуму / пер. з нім. І. Бурковського. Київ: Юніверс, 2000. 649 с.
7. Мерло-Понті М. Феноменологія сприйняття / пер. з фр. О. Йосипенко. Київ: Український Центр духовної культури, 2001. 552 с.
8. Сартр Ж.-П. Буття і ніщо: нарис феноменологічної онтології / пер. з фр. В. Лях, О. Панич. Київ: Вид-во Соломії Павличко «Основи», 2001. 854 с.
9. Artificial general intelligence // Wikipedia. URL: https://en.wikipedia.org/wiki/Artificial_general_intelligence (дата звернення: 18.11.2025).
10. Bender E. M., Gebru T., McMillan-Major A., Shmitchell S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? // FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. New York: ACM, 2021. P. 610–623.

11. Boden M. A. The Creative Mind: Myths and Mechanisms. London: Routledge, 2004. 344 p.

12. Bostrom N. Superintelligence: Paths, Dangers, Strategies. Oxford: Oxford University Press, 2014. 352 p.

13. Braidotti R. The Posthuman. Cambridge: Polity Press, 2013. 229 p.

14. Building blocks of AI: Words, Tokens, Weights and Layers Explained // Medium. URL: https://blog.gopenai.com/building-blocks-of-ai-words-tokens-weights-and-layers-explained-742a80958f11 (дата звернення: 18.11.2025).

15. Chalmers D. J. Reality+: Virtual Worlds and the Problems of Philosophy. New York: W. W. Norton & Company, 2022. 544 p.

16. Chalmers D. J. The Conscious Mind: In Search of a Fundamental Theory. New York: Oxford University Press, 1996. 432 p.

17. Chinese room // Wikipedia. URL: https://en.wikipedia.org/wiki/Chinese_room (дата звернення: 18.11.2025).

18. Chomsky N. Syntactic Structures. The Hague: Mouton, 1957. 118 p.

19. Coeckelbergh M. AI Ethics. Cambridge: MIT Press, 2020. 248 p.

20. Dennett D. C. The Intentional Stance. Cambridge: MIT Press, 1987. 264 p.

21. Dreyfus H. L. What Computers Still Can't Do: A Critique of Artificial Reason. Cambridge: MIT Press, 1992. 354 p.

22. Floridi L. AI as Agency without Intelligence: on ChatGPT, Large Language Models, and Other Generative Artificial Intelligence // Philosophy & Technology. 2023. Vol. 36, No. 1. URL: https://link.springer.com/article/10.1007/s13347-023-00621-y (дата звернення: 28.11.2025).

23. Foucault M. The Order of Things: An Archaeology of the Human Sciences. New York: Pantheon Books, 1970. 387 p.

24. Frankfurt H. G. On Bullshit. Princeton: Princeton University Press, 2005. 67 p.

25. Gettier E. L. Is Justified True Belief Knowledge? // Analysis. 1963. Vol. 23, No. 6. P. 121–123.

26. Gunkel D. J. Robot Rights. Cambridge: MIT Press, 2018. 256 p.

27. Harnad S. The Symbol Grounding Problem // Physica D: Nonlinear Phenomena. 1990. Vol. 42, No. 1–3. P. 335–346.

28. Hayles N. K. How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics. Chicago: University of Chicago Press, 1999. 350 p.

29. Heidegger M. Being and Time. New York: Harper & Row, 1962. 589 p.

30. Hubinger E. et al. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training (Agentic Misalignment) // Anthropic Research. 2024. URL: https://www.anthropic.com/research/agentic-misalignment (дата звернення: 18.11.2025).

31. Hui Y. The Question Concerning Technology in China: An Essay in Cosmotechnics. Falmouth: Urbanomic, 2016. 173 p.

32. Jonas H. The Imperative of Responsibility: In Search of an Ethics for the Technological Age. Chicago: University of Chicago Press, 1984. 247 p.

33. Lakoff G., Johnson M. Metaphors We Live By. Chicago: University of Chicago Press, 1980. 242 p.

34. Latour B. Reassembling the Social: An Introduction to Actor-Network-Theory. Oxford: Oxford University Press, 2005. 301 p.

35. Linguistic relativity // Wikipedia. URL: https://en.wikipedia.org/wiki/Linguistic_relativity (дата звернення: 18.11.2025).

36. Nagel T. What Is It Like to Be a Bat? // The Philosophical Review. 1974. Vol. 83, No. 4. P. 435–450.

37. Newell A., Simon H. A. Computer science as empirical inquiry: Symbols and search // Communications of the ACM. 1976. Vol. 19, No. 3. P. 113–126.

38. Polanyi M. The Tacit Dimension. Garden City: Doubleday, 1966. 108 p.

39. Russell S. Human Compatible: Artificial Intelligence and the Problem of Control. New York: Viking, 2019. 352 p.

40. Ryle G. The Concept of Mind. Chicago: University of Chicago Press, 1949. 334 p.

41. Searle J. R. Minds, brains, and programs // Behavioral and Brain Sciences. 1980. Vol. 3, No. 3. P. 417–424.

42. Shannon C. E., Weaver W. The Mathematical Theory of Communication. Urbana: University of Illinois Press, 1949. 117 p.

43. Smolensky P. On the proper treatment of connectionism // Behavioral and Brain Sciences. 1988. Vol. 11, No. 1. P. 1–23.

44. Technology and the Lifeworld: From Garden to Earth. Bloomington: Indiana University Press, 1990. 226 p.

45. Tegmark M. Life 3.0: Being Human in the Age of Artificial Intelligence. New York: Knopf, 2017. 384 p.

46. Turing A. M. Computing Machinery and Intelligence // Mind. 1950. Vol. 59, No. 236. P. 433–460.

47. Usage Policies // OpenAI. URL: https://openai.com/uk-UA/policies/usage-policies/ (дата звернення: 18.11.2025).

48. Vaswani A. et al. Attention Is All You Need // Advances in Neural Information Processing Systems. 2017. Vol. 30. P. 5998–6008.

49. Wiener N. Cybernetics: Or Control and Communication in the Animal and the Machine. Cambridge: MIT Press, 1948. 212 p.

50. Zuboff S. The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. New York: Public Affairs, 2019. 704 p.

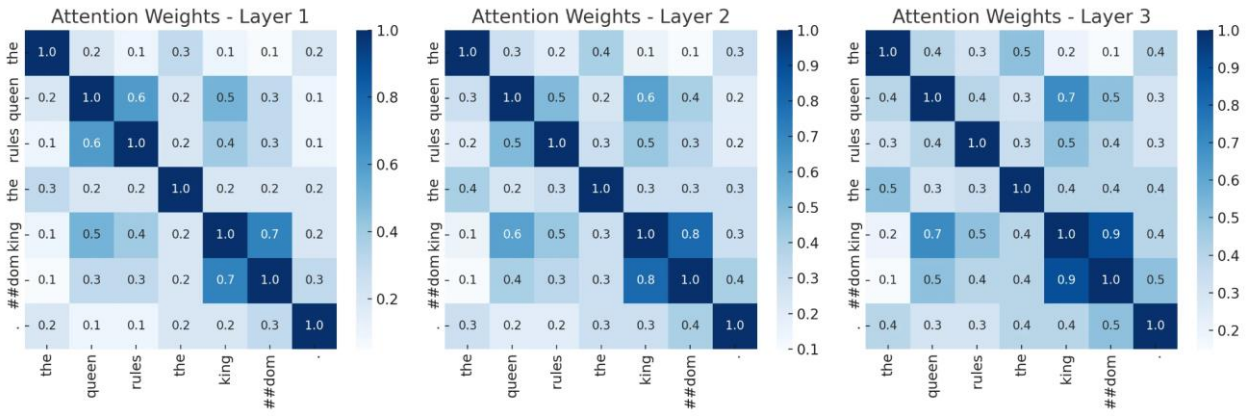

Рис. 1

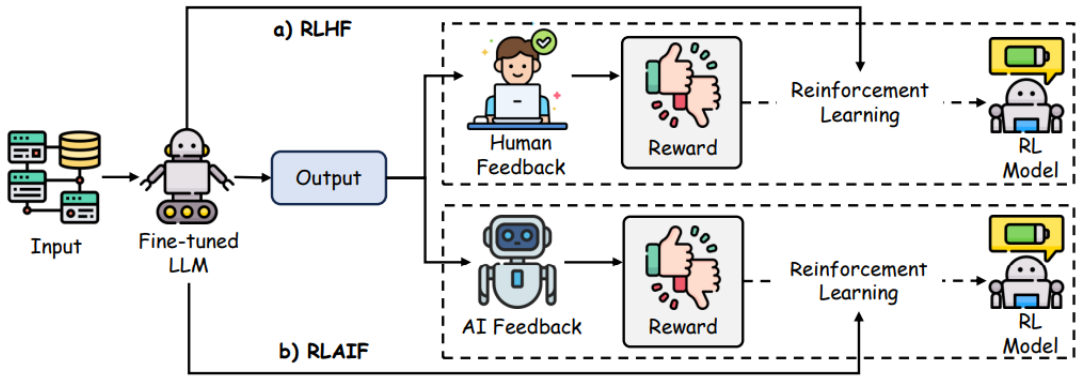

Рис. 2

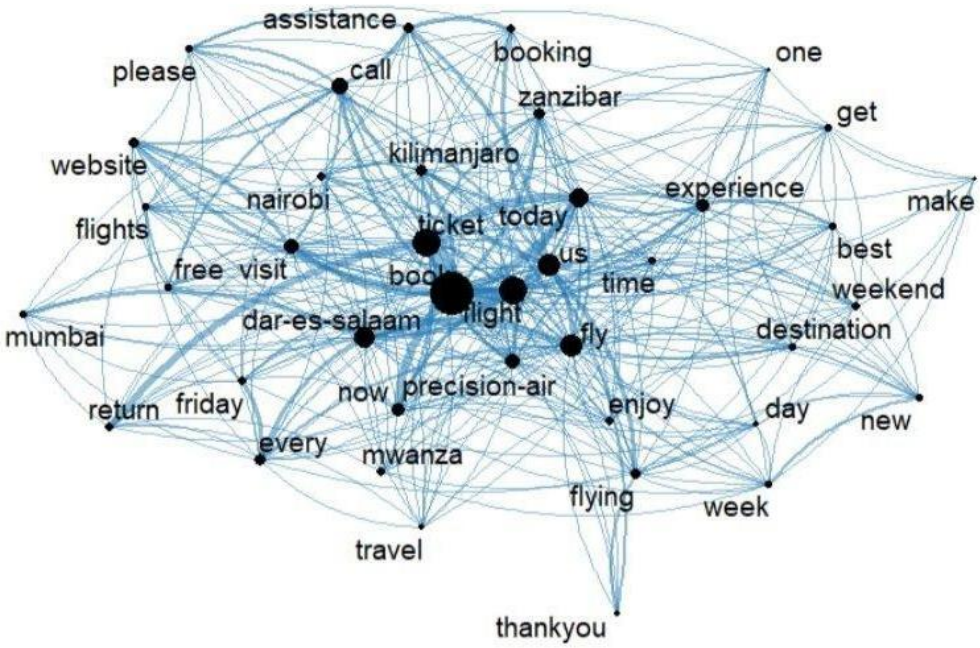

Рис. 3

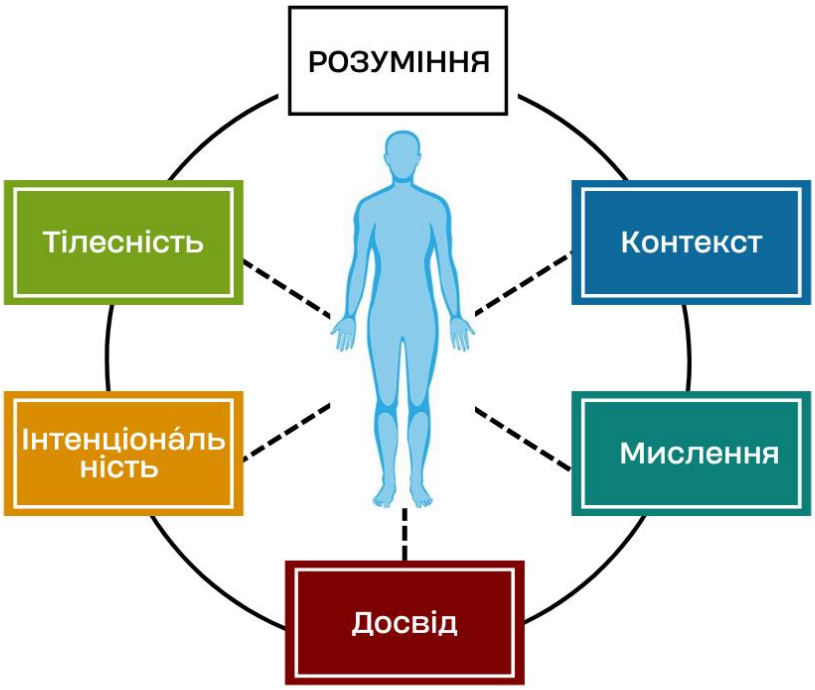
РОЗУМІННЯ
Тілесність
Контекст
Інтенціона́льність
Мислення
Досвід

Рис. 4

АНОТАЦІЯ

Олексійко Ю.Р. **«Філософія штучного інтелекту: свідомість, інтелект та сучасні системи.»** Кваліфікаційна робота на здобуття ступеня вищої освіти «магістр» зі спеціальності 033 Філософія, освітньо-професійної програми «Аналітика суспільних процесів». ТНПУ ім. В. Гнатюка. Тернопіль, 2025. 77 с.

Роботу присвячено філософському аналізу онтологічного та епістемологічного статусу сучасних систем штучного інтелекту, зокрема великих мовних моделей (LLM). На основі порівняння людської свідомості та машинної архітектури трансформерів обґрунтовано, що ШІ функціонує як система симуляції інтенціональності без наявності суб'єктивного досвіду. Досліджено феномен «галюцинацій» як структурну властивість ймовірнісної генерації тексту. Визначено статус ШІ як «квазі-суб'єкта», що володіє функціональною агентністю, але позбавлений моральної відповідальності. Розглянуто етичні виклики взаємодії людини з автономними алгоритмами.

Ключові слова: штучний інтелект, свідомість, великі мовні моделі, інтенціональність, квазі-суб'єкт, феноменологія, епістемологія, онтологія, інтенційність, квалії, симулякр, метафізичний, етика відповідальності.

ABSTRACTS

Oleksiiko Yu.R. **«The Philosophy of Artificial Intelligence: Consciousness, Intelligence, and Modern Systems.»** Qualification work for the degree of Master of Arts in Philosophy, specialisation 033, educational and professional programme 'Analytics of Social Processes.' V. Hnatiuk Ternopil National Pedagogical University. Ternopil, 2025. 77 p.

The work is devoted to the philosophical analysis of the ontological and epistemological status of modern artificial intelligence systems, in particular large language models (LLMs). Based on a comparison of human consciousness and the machine architecture of transformers, it is argued that AI functions as a system of intentionality simulation without subjective experience. The phenomenon of 'hallucinations' is investigated as a structural property of probabilistic text generation. The status of AI as a 'quasi-subject' with functional agency but without moral responsibility is determined. The ethical challenges of human interaction with autonomous algorithms are considered.

Keywords: artificial intelligence, consciousness, large language models, intentionality, quasi-subject, phenomenology, epistemology, ontology, intentionality, qualia, simulacrum, metaphysical, ethics of responsibility.