МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Національний університет «Києво-Могилянська академія»
Факультет гуманітарних наук
Кафедра філософії та релігієзнавства

Кваліфікаційна робота

(освітній ступінь - бакалавр)

на тему **“Критика самосвідомості штучного інтелекту у філософії Деніела Деннета”**

Виконала: студентка 4-го року навчання,
Спеціальність: 033 Філософія
Мерзлікіна Софія Миколаївна

Керівник: Циба В’ячеслав Миколайович
кандидат філософських наук,
доцент кафедри філософії та релігієзнавства

Рецензент _____

(прізвище та ініціали)

Кваліфікаційна робота захищена
з оцінкою «_____»
Секретар ЕК _____
«_____»_____20___р.

Київ - 2020

ЗМІСТ

ВСТУП

Дискусії щодо можливості створення штучного інтелекту та змісту цього поняття тривають більше, ніж півстоліття, і за цей час суттєво поглибили наукові знання щодо можливостей як техніки, так і людського розуму. Від першого прототипу тесту Алана Тюрінга 1930-х років, де штучний інтелект постає як раціональна машина, що імітує математичні помилки людини, бере початок історія розвитку штучного інтелекту (далі - ШІ) у двох напрямах: філософському і математичному. ШІ як комп'ютерні системи, що мають спільні з людським розумом характеристики та здатні виконувати конкретні задачі, успішно реалізована у проектах персональних асистентів, автопілотів, продовжується вивчення побудови складних нейронних мереж. Однак реалізовані проекти ШІ суттєво обмежені виконанням однієї чи декількох функцій і не є відображенням ідеї ШІ як свідомого агента, що може брати участь у комунікації і щодо якого ми можемо застосовувати поняття моральності.

У ШІ як філософській проблемі, що полягає у можливості створення системи, здатної до самостійного (без втручання спостерігача і корегування) аналізу і вирішення проблем (на відміну від технічної імітації) такого прогресу не спостерігається. Причиною є тісний взаємозв'язок ШІ та філософії свідомості, у якій відсутній консенсус щодо природи свідомості та до якої змушені звертатися теоретики розумних машин. Питання, що постало для ШІ ("Чи може машина мислити") виявляє проблемні положення у філософії свідомості, такі як умовність поняття кваліа, співвідношення вродженого та набутого знання, значення інтроспекції та інші - як наслідок, неможливо навіть імітувати мислення людини, якщо достеменно невідомий його процес. Уявні експерименти, що були поставлені задля критики ШІ, виявили фундаментальні відмінності між поглядами теоретиків сучасної філософії свідомості, хоч і висвітлювали проблему ШІ досить точково (усвідомлення ШІ своєї дії не може

бути підтверджено чи спростовано протягом однієї перевірки, оскільки подібне може не виконуватися і з людиною).

Щоб досягти створення машини сильного ШІ, необхідно з'ясувати природу свідомості (біологічну, метафізичну чи комукативну), або віднести проблему свідомості і самосвідомості до ряду невирішених проблем у компетенції феноменології. Філософський підхід Д. Деннета користується здобутками нейробіології, фізики та комп'ютерних наук, що дають змогу розглянути ШІ у новій перспективі та з'ясувати, які проблеми наразі актуальні у філософії свідомості і обґрунтувати розвиток ШІ на рівні інтенційних систем.

Об'єктом дослідження є теорія самосвідомості штучного інтелекту. *Предметом дослідження* є критика самосвідомості ШІ у філософській концепції Деніела Деннета. *Метою дослідження* є виявлення основ біхевіористичного підходу до потенціалу ШІ у працях Д. Деннета. Сукупний аналіз філософської позиції Д. Деннета щодо свідомого ШІ, сучасних обґрунтувань можливостей ШІ як комп'ютерного алгоритму[1] та досягнень нейробіології потенційно доводить можливість розвитку ШІ до рівня інтенційної системи (за визначенням інтенційної системи Д. Деннета). У підсумку, має бути доведено евристичний потенціал ШІ як складної інтенційної системи завдяки використанню підходу функціоналізму, однак поставлено під сумнів результативність функціоналізму у з'ясуванні феномену людської свідомості на користь емерджентної теорії.

Завдання роботи полягають у :

- Визначенні актуальних проблем у сфері розробки ШІ і філософії свідомості (вступ та розділ І "Передумови дослідження свідомого ШІ).

[1] Winfield A. Intelligence is not one thing / A. Winfield // Journal of Artificial General Intelligence. - 2020. - Vol. 11 (2). - P. 97-100. – Режим доступу: https://content.sciendo.com/view/journals/jagi/11/2/article-p1.xml

- Доведенні приналежності ШІ до класу інтенційних систем та формулюванні поняття самосвідомості для застосування до ШІ (Розділ ІІ “Можливість самосвідомого ШІ”).
- Критичній оцінці функціоналістського тлумачення свідомості (Розділ ІІІ “Критичні зауваження до філософських аргументів Д. Деннета”).
- Обґрунтуванні перспектив розвитку ШІ (висновки роботи).

Реалізацію кожного з цих завдань відображено у структурі моєї роботи.

Джерельну базу роботи склали оригінальні праці Д. Деннета різних років: “Brainchildren", "Consciousness explained", "Content and Consciousness" та інші (див. Список використаних джерел). Для вказівки на існуючий напрям скептицизму у філософії свідомості та критики сильного ШІ (що є протилежними до методу Д. Деннета) згадані мисленнєві експерименти Дж. Сьорла, Г. Патнема та Т. Наґеля. Для визначення актуальності функціоналізму у роботі були використані сучасні дослідження нейробіології та програмування ШІ.

Для написання роботи були застосовані наступні *методи:* компаративне дослідження текстів та метод моделювання.

РОЗДІЛ І
ПЕРЕДУМОВИ ДОСЛІДЖЕННЯ СВІДОМОГО ШІ

1.1 Штучний інтелект і переосмислення проблем філософії свідомості

Проблема раціонального пояснення зв'язку свідомості людини з її фізичним тілом для філософії є класичною: перші спроби осмислити можливість існування несвідомих людей-автоматів з'являються вже у Рене Декарта[2] у "Медитаціях про першу філософію" 1641-го року. Р. Декарт проводить мисленнєвий експеримент з людиною, що мала б ті самі функції, що і людина свідома, рухалася рефлекторно, однак не мислила - відрізнити свідому і несвідому людину у такому разі буде напрочуд важко. Пізніше, у 20-му столітті тактика мисленнєвого експерименту набуває поширення у філософії свідомості, а людина-автомат Декарта еволюціонує до проблеми філософського "зомбі" Роберта Кірка. Мисленнєвий експеримент є способом не лише уявити, а і логічно обґрунтувати ситуацію, що неможлива в реальному житті або складно виконувана, однак її моделювання має суттєвий потенціал у вирішенні філософської проблеми. Для мисленнєвого експерименту застосовуються загальноприйняті символи, поведінкові зразки та середовища, що робить висновки з експерименту спільними для багатьох його спостерігачів. Такий тип експериментів застосовується і для аналізу ШІ: класичним можна вважати експеримент "Китайську кімната" Дж. Сьорла, що розпочинає філософську дискусію щодо ШІ - навіть якщо машина здатна виконувати поставлені для неї завдання, вона не володіє осмисленням того, що робить. Питання про природу свідомого стану залишається, однак спостерігаючи за процесом мислення ШІ і будучи його творцем, можна з упевненістю стверджувати, що свідомість у ньому не наявна. У такому трактуванні проблематика ШІ набуває технічного, а

[2] Леонов А. Роберт Кірк: засновник філософських зомбі/ А. Леонов // Філософська думка. - 2016. - №2. - С. 71

не філософського спрямування: майбутній розвиток математики та комп'ютерних наук може означати і створення досконалішої системи ШІ з структурно-схожими до людських характеристиками, однак можна поставити під сумнів відтворення біологічних особливостей людини, що уможливлюють специфічність мислення як суто людської здатності.

Протилежний підхід у питанні свідомості ШІ, а також значення ШІ у розумінні людського мислення демонструє Деніел Деннет у низках філософських робіт. Робота "Когнітивне колесо: проблема фреймінгу ШІ"("Cognitive wheels: The Frame Problem of AI") переформульовує питання "Чи може машина мислити?" на таке: чому ШІ мислити не вдається? Порівняймо дві кімнати: китайську кімнату[3] Дж. Сьорла і кімнату з роботом[4] Д. Деннета. У першому випадку ми спостерігаємо, як людина в кімнаті спілкується з тими, хто за дверима кімнати китайською мовою. Людина в кімнаті не знає значення символів китайської, однак правильно відповідає на поставлені їй запитання завдяки інструкціям англійською мовою, що не пояснюють значення символів. Комунікація між ШІ та людьми відтворювалася б так само: без розуміння значення символів ШІ і можливості їх зрозуміти після цієї комунікації. Ключовим є наше знання процесу обміну інформацією: ми знаємо, як саме людині в кімнаті вдається спілкуватися китайською.

У кімнаті Д. Деннета знаходиться батарея, що необхідна роботу для підзарядки та закріплена на ній бомба. На відміну від кімнати Сьорла, ми спостерігаємо невдачі робота у виконанні завдання: він забирає з кімнати батарею разом з бомбою, оскільки не може зробити правильних висновків про ситуацію в кімнаті без попереднього досвіду. Людина в китайській кімнаті виконує інструкцію, у той час як робот мусить виконати складнішу операцію

[3] Searle J. R. Minds, brains, and programs / J. R. Searle // Behavioral and Brain Sciences. - 1980. - Vol. 3(3). - P. 417-457.

[4] Dennett D. C. Cognitive wheels: the former problems of AI / D. C. Dennett // Brainchildren: Essays on Designing Minds. - Cambridge: MIT Press, 1998. - P. 181-205.

вибору. Навіть після перепрограмування робот буде помилятись: він змушений обирати серед безлічі засновків правильний, однак є суттєво обмеженим у часі. Неможливо наперед встановити дедуктивний механізм, за яким робот сортуватиме релевантні та нерелевантні засновки[5], не задаючи наперед всі умови експерименту, тим самим нівелюючи його подібність до реального. "Проблема фреймінгу" (The frame problem of AI), яку визначає Д. Деннет, є проблемою існування ШІ у незнайомому середовищі, для якого ШІ не має моделі, що звужуватиме його процеси мислення до виконання конкретної задачі. На думку Деннета, неможливість ШІ вирішувати ряд простих мисленнєвих задач вказує на епістемологічну проблему, що не порушувалася філософами минулих поколінь - очевидно, що людина має певні біхевіористичні диспозиції, але як саме вона їх застосовує і уникає помилок? Проблема вродженого знання у філософії може бути вирішена завдяки зверненню до свідомості як цілісного феномену - спроби об'єднати феноменологію та інтроспекцію зумовили розгляд часткових проявів свідомості, що є недостатнім (одними з перших застосування такого методу демонструють Д. Г'юм та Дж. Лок). Дуалізм суттєво сповільнив філософські і наукові здобутки в теорії мислення - Д. Деннет продовжує позитивістську ідею хибності інтроспекції як методу пізнання. Саморефлексія спричиняє хибне уявлення[6] про процеси мислення, таким чином і для створення ШІ застосовується хибна модель.

Можна зробити висновок, що для того, щоб аналізувати присутність свідомості у ШІ, необхідно першочергово з'ясувати як мислить людина - єдина жива істота у якої свідомість присутня. ШІ має повторювати механізми не лише мислення людини, а і його наслідків - поведінки, яку ми не завжди можемо пояснити набутим досвідом у процесі навчання чи результатом ретельного обмірковування. Мисленнєвим експериментом, що унаочнює проблемність

[5] Dennett D. C. Cognitive wheels: the former problems of AI / D. C. Dennett // Brainchildren: Essays on Designing Minds. - Cambridge: MIT Press, 1998. - P.182
[6] Там само. С. 186

імплементації в ШІ алгоритму навіть типових життєвих задач є "Нічний бутерброд[7]"(Midnight snack problem) Д. Деннета - для того, щоб зробити бутерброд вночі, необхідно знати, що тримаючи банку з майонезом у лівій руці, неможливо цією ж рукою намазувати хліб. Необхідно знати, що якщо наливати пиво у склянку, воно зникає з пляшки. Помилкою є програмування ШІ як такого, що знає вибіркові факти, необхідні для здійснення задачі - такий підхід ігнорує наявність великої кількості опорних знань, з яких формується алгоритм дії. Якщо створення бутерброда можна віднести до логічних стабільних процесів, що мають певну послідовність, протилежним прикладом буде уявний експеримент з дітьми і печивом. Якщо обидві дитини братимуть печиво з коробки без дозволу і при цьому одна з них отримуватиме за це фізичне покарання, а інша ні, пізніше ми спостерігатимемо очікувану реакцію: дитина, яку покарали, утримуватиметься від спроб взяти печиво ще раз. З побутової ситуації постає декілька проблем, які необхідно враховувати для конструювання ШІ.

По-перше, існує поведінкова схильність до конкретної реакції на стимул: 'ідея' болю означає, що дитина може плакати, кривитися, але вона не танцюватиме і не усміхатиметься. Конкретні реакції не мали б жодного сенсу, якщо не були б передвизначені біологічно та детермінували поведінку. Відповідно, схожий принцип співставлення стимулу з реакцією має бути втілений в ШІ для його існування в соціальному середовищі.

По-друге, подальші реакції дитини на подібну ситуацію можуть не відповідати новим умовам середовища - навіть за відсутності загрози вона може не вдаватися до спроб взяти печиво. Виникає ускладнення для ШІ: необхідно вибрати правильну комбінацію дій серед тисяч і більше можливих (його реакція має так само відповідати стимулу, як і людська), але у випадку точного розрахунку ця дія може виявитися невідповідною ситуації. Постає питання про

[7] Dennett D. C. Cognitive wheels: the former problems of AI / D. C. Dennett // Brainchildren: Essays on Designing Minds. - Cambridge: MIT Press, 1998. - P. 187

те, яким має агентом має бути ШІ: суто раціональним і перевершувати людину, чи керуватися тими самими “вродженими” ідеями та смислами, що і людина. Людина здійснює індуктивне прогнозування ситуації і приймає відповідне рішення, позитивне чи негативне. Д. Деннет стверджує, що “проблема фреймінгу” в ШІ є ширшою за проблему індукції і не вирішується виключно дослідженням алгоритму для суб’єктивної ймовірності та фіксації переконань[8]. Навіть якщо ШІ володіє емпіричним знанням про навколишній світ і може виконати точний прогноз, йому бракуватиме механізму, який визначить, який з цих прогнозів потрібно враховувати - проблеми виділення важливих засновків з-поміж інших неможливо уникнути. Можливості ШІ обмежуються теоремою неповноти Геделя[9], що визначає межу розвитку ШІ через співвідношення поставлених задач та внутрішнього знання. Існують такі комбінації задач, за яких дієвцю неможливо надати ідеальний інтелект, якщо він не здатен навчатися - таким чином, спеціальні алгоритми для конкретної задачі ШІ не завжди означають можливість її успішного виконання. Наразі здатність ШІ до самонавчання лише прогнозується.

Проблема фреймінгу - теорії вибору одного релевантного засновку з безлічі нерелевантних, є вступом до низки запитань, спільних для нейробіології, філософії свідомості та комп’ютерних наук, оскільки точково унаочнює, що наразі не існує структурованого викладу як теорії свідомого сприйняття, так і обґрунтування поведінкових особливостей людини. Новизна проблеми фреймінгу є сумнівною, оскільки схожий механізм концентрації свідомості на одному предметі з-поміж інших належить до феноменологічного визначення інтенційності, що хоча і не має на меті описувати цей свідомий процес інформаційно, однак виокремлює ту саму проблему. Для Д. Деннета, як

[8] Dennett D. C. Cognitive wheels: the former problems of AI / D. C. Dennett // Brainchildren: Essays on Designing Minds. - Cambridge: MIT Press, 1998. - P. 194

[9] Laiard J. Intelligence, Knowledge & Human-like Intelligence / J. Laiard // Journal of Artificial General Intelligence. - 2020. - Vol. 11 (2). - P. 41-44. - Режим доступу: https://content.sciendo.com/view/journals/jagi/11/2/article-p1.xml

функціоналіста, інтенційність визначається як здатність мати інтенції певного порядку та передує свідомості: виникає необхідність пояснення окремої проблеми фреймінгу та появи інтенцій вищого порядку у людини на противагу дослідження цілісного феномену свідомості у феноменології.

1.2 Скептицизм у філософії свідомості. Можливість функціонального стану для системи ШІ.

Інтерес до ШІ може вважатися раціональним для філософії тільки у випадку, якщо він вирішує низку проблем - саме таку функцію, на думку Д. Деннета, пропонують комп'ютерні науки. Неможливо уявити істоту, що має тотожний людському статус переживань, біологічного часу, рівень інтелекту, тому про ШІ не варто говорити - саме з відсутністю інтенції зазирнути за межі неможливого Деннет пов'язує позиції Томаса Наґеля та Коліна МакҐіна[10] (неможливо дослідити суб'єктивну феноменологію через об'єктивну фізіологію). Простір для функціонального дизайну[11], яким є ШІ, виступає посередником між оманливою інтроспекцією (основою нашого уявлення про феноменологію) та функціональними процесами мозку, які досліджує наука. Комп'ютерні науки дають змогу відслідковувати процеси, що складають комплексні механізми роботи мозку та пояснюють квазі-розуміння[12] - аналогічним способом ми можемо уявити, як завдяки когнітивістському опису мисленнєвих процесів можна пояснити розуміння людини .

Філософія свідомості теоретизує свідомість як феномен, що описує взаємодію свідомості з індивідуальними значеннями, які мають речі (узагальнено - символами). Для Д. Деннета таке тлумачення свідомості є вузько семантичним і проблемним для пояснення свідомих процесів: якщо значення ,

[10] Dennett D. C. Consciousness explained / D. C. Dennett. - Boston: Little, Brown, 1991. - P. 433
[11] Dennett D. C. Brainstorms: Philosophical Essays on Mind and Psychology / D. C. Dennett. - Cambridge: MIT Press, 1981. - P. 3
[12] Dennett D. C. Consciousness explained / D. C. Dennett. - Boston: Little, Brown, 1991. - P. 439

притаманні речам, є незмінними, достатньо описувати лише символічний зв'язок між свідомістю та річчю, на яку вона спрямована, однак не пояснювати процес утворення цього зв'язку. Альтернативним є гетерофеноменологічний[13] підхід Деннета, що має на меті опис ментальних станів суб'єкта лише інформаційно - виділяти необхідно лише ту частину інформації, що важлива для виконання конкретної задачі. Реалізацією гетерофеноменологічного підходу можна назвати ШІ як критичний інструмент, що унаочнює хибні висновки щодо феноменологічного розуміння свідомості: Д. Деннет зазначає, що подібно до розголошення секрету фокусу ми дізнаємося про процеси мозку, які послаблюють позиції феноменології. Завдяки моделюванню процесів мозку ми дізнаємося про відсутність розрізнення між основними та другорядними процесами, суперечливість кваліа та уявлення про концентрацію значень і реакцій на них у межах однієї структури - свідомості. Критичною спробою спростування "фокусу" є мисленнєві експерименти Неда Блока ("Китайська нація[14]") та Дж. Сьорла ("Китайська кімната"), що спрямовані на критику ШІ і показують його нездатність до оперування на феноменологічному рівні. Якщо "Китайська кімната" вказує на неможливість ШІ оперувати символами, у "Китайській нації" Блока демонструється неможливість пояснення феномену ментального стану у термінах функціоналізму. "Китайська нація" зображає наступне: протягом святкування 60-го дня народження Мао Цзедуна мільярд китайців (що є тотожним мільярду нейронів, які уможливлюють роботу мозку) посилають один одному сигнали за такою ж схемою, за якою відбувається передача сигналів у мозку Мао Цзедуна впродовж чотирьох годин. Упродовж цих чотирьох годин Мао Цзедун відчуває себе щасливим, однак пізніше страждає від головного болю. Чи можна стверджувати, що мільярд китайців буде у такому ж ментальному стані, як Мао Цзедун? Навіть якщо функціоналіст

[13] Dennett D. C. Cognitive wheels: the former problems of AI / D. C. Dennett // Brainchildren: Essays on Designing Minds. - Cambridge: MIT Press, 1998. - P. 188

[14] Block N. Troubles with Functionalism / N. Block // Minnesota Studies in the Philosophy of Science. - 1978. - №9. - P. 261–325.

відповідає, що так, за Блоком теорія функціоналізму не пояснює феномен ментального стану. Д. Деннет визначає протистояння функціоналізму та емерджентної теорії як наслідок різного підходу до пояснення процесів свідомості: функціоналізм описує етапи “фокусу”, тоді як емерджентизм вказує на прихованість і магічність знання. У мисленнєвих експериментах Деннета та Сьорла роль реципієнта, або спостерігача експерименту є різною: у першому випадку він просто спостерігає процеси і описує можливими засобами, у другому випадку не виконується спроба розкрити феномен, однак для реципієнта унаочнюється його наявність.

У той час як відтворення спільного уявлення про речі на феноменологічному рівні у людини та ШІ є проблемним питанням, спільна думка про об’єкт та його дескрипцію є цілком реалізованою через поняття “функціонального стану”. Для спільності ментальної події, визнання чогось як таке, що має певний колір, запах, призначення достатньо мати спільну логічну детермінацію в конкретній ситуації, одну загальну функціональну мову - так, неначе комп’ютери з різним устаткуванням працюють за однаковою програмою [15]. Можливість спільного функціонального стану вказує на необхідність ШІ бути операційною системою особливого класу - інтенційною.

Інтенційність є невід’ємною характеристикою свідомості та визначає її спрямованість на предмет, його покладання у мисленні - у феноменології інтенційність є умовою можливого досвіду свідомості та переживань свідомості у реальному часі. Інтенційність як здатність відповідати взаємністю іншим суб’єктам та мати мету, відмінну від базових біологічних, притаманна функціоналізму - переживання та існування серед інших об’єктів та суб’єктів все ще важливі, однак постають як функції найвищого порядку, а не феномени. Таким чином, інтенційність є необхідною умовою індукційного функціонального типу формування свідомості, на противагу емерджентній

[15] Dennett D. C. Brainstorms: Philosophical Essays on Mind and Psychology / D. C. Dennett. - Cambridge: MIT Press, 1981. - P. xvi

теорії свідомості Дж. Сьорла, де інтенційність є результатом природної активності людської свідомості, а в інших об’єктів може бути лише набутою (приписаною). Інтенційні системи Д. Деннет визначає як концепти[16], що можуть бути пояснені та спрогнозовані завдяки умовному наділенню переживаннями та переконаннями. Можливість передбачення поведінки інтенційної системи є ключовою характеристикою можливості її подальшого розвитку до свідомого суб’єкта.

Можна розглянути дві підстави оцінки поведінки інтенційної системи: її перспектива дизайну та фізичний стан. Дизайн характеризується як сукупність функціональних дизайнів системи, тоді як фізичний стан означає фактичні складові системи об’єкта. Для ШІ це будуть програмне забезпечення та технічне устаткування. Удосконалення техніки та комп'ютерної науки призводить до того, що поведінку системи ШІ стає складно або неможливо спрогнозувати - єдиною можливістю є припущення найбільш раціонального варіанту поведінки, яким система керуватиметься. Виконання штучним інтелектом певних задач на високому рівні, що перевершує можливості людини не пов’язане з теорією свідомої дії, однак є втіленням інтенційності. Належність ШІ до інтенційних систем є умовою порівняння можливостей людини і машини та спроб відтворення біологічних процесів мозку у вигляді цифрових процесів програмного забезпечення. “Необхідно захистити різновид інтенційності як загальну теорію приписування ментальних станів від переконливого, проте проблематичного підходу інженерії, де складні інтенційні системи мусять мати спільну характеристику, а саме внутрішню систему чи мову ментальної репрезентації[17]” - розширення Д. Деннетом класу інтенційних систем залучає до них і ШІ.

[16] Dennett D. C. Brainstorms: Philosophical Essays on Mind and Psychology / D. C. Dennett. - Cambridge: MIT Press, 1981. - P. 3
[17] Там само. - P. xxi

Коли складна інтенційна система демонструє правильне виконання задачі, ми необхідно приписуємо їй риси, що притаманні людині: окрім раціональності це почуття та переконання, що є корисними для прогнозування. Штучність раціональних дій ШІ не заперечує того, що ШІ належить, як і людина до класу інтенційних систем і володіє логічними категоріями істини[18]. Визначивши, що інтенційність притаманна людині незалежно від того, чим визначаються критерії її особистості можна підсумувати, що все, що постає з факту належності до інтенційних систем буде характеризувати особистість.

Людина належить до підкласу інтенційних систем, де передумовою свідомості є наявність мови і здатність до комунікації. Деніел Деннет вказує на те, що хоча мова людини є найбільшим досягненням еволюції і сприяє подальшому розвитку навіть більш ускладнених раціональних систем[19], мову все ще можна розглядати як наслідок адаптації до навколишнього середовища (біхевіористичний талант). Для ШІ виведення мови з феноменологічного поля до наукового означає можливість імплементації мови в програмне забезпечення ШІ і детальне дослідження механізму комунікації людини.

Питання, що необхідно постає з розвитку складних нейронних систем ШІ і об'єднує мисленнєві експерименти "Китайської кімнати" Дж. Сьорла, "Мізків у бочці" Г. Патнема є таким: який статус для нас має реальність та чи можна вважати ШІ симулякром? Питання є неоднозначним, оскільки багатьом істотам, у тому числі і людям, раціональність приписується або за спостереженнями, або апріорно (всі, хто належать до людського роду є свідомими істотами). Якщо образ ШІ застосовується для філософського глузування над його можливостями, створюється штучна ситуація, яку неможливо назвати ґрунтовним експериментом (подібно до проблеми фреймінгу з роботом ШІ, для пацюків у

[18] Dennett D. C. Brainstorms: Philosophical Essays on Mind and Psychology / D. C. Dennett. - Cambridge: MIT Press, 1981. - P. 9
[19] Там само. - P. 17

лабораторії створюються експерименти, де немає простору для вчинення незапланованої дії).

Питання реальності чи симулякру може бути відтермінованим: Д. Деннет зазначає, що використання концепції інтенційної системи має потенціал лише у тому випадку, коли на певному етапі розробки ми перестаємо перевіряти припущення раціональності системи[20] і вважаємо, що системі можна приписувати бажання та переконання. Завдання інтенційної системи постає у створенні проміжної ланки між символічним світом суб'єктів і їх мовних ігр (поведінкові механізми людини, які вона застосовує) та суто фізіологічними особливостями організму.

1.3 Структура свідомості. Субперсональний рівень доступу до свідомого.

ШІ є приводом для нового дослідження питань, що виникли в філософії свідомості століттями тому: досі актуальною є "Проблема Г'юма[21]" - якщо ми відмовляємося від самості, що здатна оперувати ідеями та враженнями (які складають основу внутрішньої репрезентації), ми необхідно маємо визнати, що ідеї та переживання мають опрацьовувати самі себе. Якщо так, то самість постає як сукупність почуттів та ідей, що перебувають у постійному русі завдяки законам асоціації і не потребують спостерігача, наділеного інтелектом. Так само проблемним є і психологічний підхід до свідомості, що стверджує необхідність "гомункулуса" у ролі інтерпретатора: не існує нічого, що мало б внутрішньо втілену репрезентацію будь-чого, для сприйняття ідеї необхідний зовнішній для неї механізм. Обидва питання, а також дилеми щодо проекції відчуття і наявності внутрішнього спостерігача (зовнішню картинку потрібно

[20] Dennett D. C. Brainstorms: Philosophical Essays on Mind and Psychology / D. C. Dennett. - Cambridge: MIT Press, 1981. - P. 19
[21] Там само. - P. 122

переформатовувати для внутрішнього сприйняття - якщо у нас є механізм для перцепції, то завдяки чому він працює?) ШІ може вирішити у випадку досягнення найвищого ступеню технологічного розвитку. Індуктивна логіка дозволяє редукувати припущення гомункулуса, що відповідає за сприйняття, до найпростішого механізму, що відтворюється схематично - за такого підходу можна теоретизувати процес сприйняття інформації людиною.

Деніел Деннет пропонує визначити схему, за якою можна виокремити характеристики персональності людини як того самого гомункулуса, що роз'яснює функціонування мислення. Умовно є внутрішні та зовнішні мисленнєві процеси людини, що нею не усвідомлюються і подібні процеси, що усвідомлюються. Структурувати процеси мислення можна завдяки трьом[22] типам доступу до наявного знання: комп'ютативному, публічному та доступу персональної свідомості. Доступ персональної свідомості відокремлюється як усе, суб'єктом чого є особистість, але не є жодна з частин її тіла. Таким чином, можна досліджувати комп'ютативний та публічний доступ як важливі ланки у когнітивному процесі, які не є частиною феноменологічного рівня оперування свідомості і можуть бути перенесені на рівень моделі неорганічного походження.

Комп'ютативний процес відокремлений від доступу персональної свідомості та означує всі процеси, якими керує нервова система людини: для обробки зображень необхідний доступ до інформації внутрішнього вуха та руху очей, що відбувається як несвідомий процес. Публічний доступ є також відокремленим від персонального доступу свідомості (нашого "Я") та постає як виклад історії про проведені операції в системі. Якщо ототожнити мозок з комп'ютером, публічний доступ є рівнем під-операцій, проміжним етапом, що надає змогу проаналізувати інформацію з комп'ютативного рівня. Оголошеному допоміжному механізму надається комп'ютативний доступ до такої інформації,

[22] Dennett D. C. Brainstorms: Philosophical Essays on Mind and Psychology / D. C. Dennett. - Cambridge: MIT Press, 1981. - P. 150

до якої ми прагнемо отримати публічний доступ[23]. Д. Деннет зазначає, що публічний доступ є необхідно окремим процесом, оскільки спостерігається різниця між основним виконавчим процесом і його доступу до допоміжних процесів та доступу, який має процес публічного доступу до основного виконавчого процесу. Утворена ієрархічна структура розкладається на прості елементи, що задовольняє встановлену на початку вимогу уникати необхідності зведення механічних процесів до появи гомункулуса, що їх спостерігає і осмислює. Все ж, сукупність публічного доступу, що повідомляє нас про витяг з подій в системі, та комп'ютативного доступу до всієї інформаційної бази ніяк не описують появу суб'єкта, жоден процес не стає "Я[24]", що має досвід до свідомості. Система публічного доступу є підготовчою стадією або передумовою свідомості, яка описана феноменологічно і постає як наслідок, який можна спостерігати. Система публічного доступу фактично є викладом вибіркових важливих процесів з комп'ютативного рівня, цю інформацію може використовувати лише свідоме "Я", що і спостерігається у процесі людської комунікації. Хоча рівень публічного доступу не характеризує суб'єкта (ми не можемо визначити, як саме виникає "Я"), наявність суб'єкта легітимує існування публічного доступу та має з ним зв'язок. Д. Деннет вказує на те, що істоти без ознак мови не мають систем публічного доступу (або вони існують як рудименти) і отже не можуть бути свідомими - така точка зору є суперечливою і оскаржувалася іншими філософами теорії свідомості. Тварини, як такі, що не мають інтенцій та не аналізують власні комп'ютативні процеси (для цього потрібен рівень публічного доступу), функціонують лише на комп'ютативному рівні.

Пропозиція Д. Деннета ґрунтується на припущенні широкого функціоналу запропонованих ним двох рівнів мисленнєвого процесу:

[23] Dennett D. C. Brainstorms: Philosophical Essays on Mind and Psychology / D. C. Dennett. - Cambridge: MIT Press, 1981. - P. 151

[24] Там само. - P. 152

субперсональні поняття доступу уможливлюють подальше додавання ускладнень, оскільки система не є завершеною. Будь-яка теорія, як психологічна, когнітивістська, так і функціоналістська, яку представляє Д. Деннет є теорією субособистісного рівня свідомості і не деталізує процес її утворення. Відмінністю пропозиції Деннета є надання передумов для свідомого мислення - поступово звужується коло задач, яке може належати виключно персонально свідомому доступу. Наприклад, явище сомнамбулізму демонструє, що процеси орієнтації у просторі, балансування, звернення уваги на певний предмет можуть відбуватися без безпосередньої участі свідомості. З дослідження уваги у несвідомому стані постає питання, як звернення уваги на об'єкт стає свідомим.

Таким чином, трирівнева структура свідомості Д. Деннета (публічний, комп'ютативний і персональний рівень) є способом побудови свідомості як індуктивного процесу та його інформаційного опису - успішне визначення публічного та персонального доступу означає можливість відтворення свідомості в об'єктах неприродного походження. Класичний для філософії свідомості метод інтроспекції зазнає модифікації: Д. Деннет пропонує визначити інтроспекцію як таку, що є наслідком біологічного стану переживання: ми виявляємо себе у стані, коли необхідно описати все те, що відбувається, і опис відбувається завдяки здатності до комунікації[25].

[25] Dennett D. C. Brainstorms: Philosophical Essays on Mind and Psychology / D. C. Dennett. - Cambridge: MIT Press, 1981. - P. 159

РОЗДІЛ ІІ. МОЖЛИВІСТЬ САМОСВІДОМОГО ШТУЧНОГО ІНТЕЛЕКТУ

2.1 Застосування когнітивістського методу для верифікації свідомості

Як це - відчувати свою належність до певного виду істот, мати внутрішні процеси та “бути таким, як”? Д. Деннет укотре ставить питання, що найбільш відоме з експерименту Т. Наґеля “Як воно - бути кажаном” - завдяки чому можна стверджувати належність чи неналежність до групи свідомих істот? Окрім вимог бути істотою з достатнім рівнем інтелекту та соціалізованості, існує неоголошений обов’язок задовольняти ряд інших характеристик - особливістю підходу біхевіоризму і когнітивізму є те, що здійснюється спроба віднайти ці характеристики завдяки аналітичному підходу.

Не виключаючи можливість створення системи ШІ[26], що повторювала б функціонал біологічної істоти та не поступалася їй у здатності виконувати мисленнєві операції, Д. Деннет стверджує, що приписування такій системі свідомості (і підстави цього приписування) ще належить з’ясувати. Можливість існування свідомої істоти, що функціонує на суттєво нижчому рівні за людський і відтворює найпростіші рефлексії на зовнішні подразники (наприклад, плач від болю) не є достатньою для ствердження свідомості цієї істоти - визнавши її свідомість, ми тим самим заперечуємо унікальність свідомості людини.

Для постановки питання про ствердження свідомості людини або істоти, яку ми хочемо віднести до групи свідомих, необхідно з’ясувати належність критерію свідомості до внутрішнього або зовнішнього. Д. Деннет зазначає, що звичним і раціональним видається спосіб зовнішнього спостерігача, коли ми намагаємося дослідити наявність чи відсутність гомункулуса як увімкненого

[26] Dennett D. C. Brainstorms: Philosophical Essays on Mind and Psychology / D. C. Dennett. - Cambridge: MIT Press, 1981. - P. 171

світла[27], що вирізняє свідомість і ускладнює поведінковий механізм. Метод зовнішнього спостерігача є проблемним, коли ми звертаємося до поведінки істот, що мають свідомість: людина може повертатися у часі для згадки про події і має сумніви щодо того, чи була вона свідомою у певний момент. Лише порівняння здібностей на реальний момент часу і поведінки в минулому дає змогу стверджувати власну свідомість, подібний спосіб верифікації можна вважати дійсним і для інших свідомих істот. Переведення критерію свідомості з зовнішнього (спостерігач - "інший") до внутрішнього ("Я" є спостерігачем) є спробою вийти з логічної пастки підтвердження існування свідомості заздалегідь встановленими критеріями особистості, що можуть встановити видимість свідомості, але не підтвердити існування внутрішньої активності свідомості. Д. Деннет захищає позицію можливості свідомості у сутності, описаної виключно когнітивістськими термінами - якщо цю тезу буде спростовано, когнітивізм позбавлений будь-якого сенсу.

Питання для верифікації свідомості є наступними:

- чи може трирівнева конструкція, запропонована Деннетом (рівні комп'ютативного, публічного та персонального доступу) виглядати як така, що має свідомість?
- якщо певна сутність задовольняє всі критерії верифікації, чи можемо ми стверджувати, що така сутність дійсно має свідомість?

Створення системи, що здатна справити враження свідомої і перевершити здібності людини є реалізованою задачею, що підтверджувалось у шахових турнірах з ШІ та спробами пройти тест Тюрінга (машини Eliza, Parry, Eugene Goostman та ін.). Аналіз мисленнєвих експериментів з ШІ підтверджує, що оскаржується не потенційна можливість ШІ мати свідомість, а спосіб, яким досягається імітація свідомості - гра з психологічними особливостями людини. Для верифікації свідомості обов'язково необхідний "інший", завдяки якому

[27] Dennett D. C. Brainstorms: Philosophical Essays on Mind and Psychology / D. C. Dennett. - Cambridge: MIT Press, 1981. - P. 172

постає можливість порівняння та інтеракції. Однак таким чином не уникнути дилеми зі створенням заздалегідь програшної ситуації для ШІ: 1) він не є суто зовнішнім спостерігачем, якщо знає дизайн і програмне забезпечення ШІ (дані експерименту не будуть достовірними) 2) “Інший” має досвід лише власної свідомості і не знає не лише “як воно - бути кажаном”, а і не має критеріїв для верифікації свідомості іншої людини. Здатність до постановки правильних запитань визначає точність відповідей: якщо ми запитуємо про специфічність мислення кажана, як пропонує Т. Наґель, нас очікує невдача. Якщо ми намагаємось з’ясувати усі можливі риси свідомості кажана (за умови їх існування, розглядається свідомість як відокремлена структура, що може належати різним істотам (наприклад, прибульцям[28]) і не втрачає своєї суті, має певні критерії.

Деніел Деннет висуває когнітивістський підхід для огляду питання “проблеми свідомості іншого” через призму “Я” свідомості. Прогрес може відбутися завдяки застосуванню досягнень психології до власного переживання свідомості: Д. Деннет визначає когнітивізм як спосіб детального дослідження персональності. Дослідження “персонального” означає опис деталей життя та визначення походження того, як саме людина ознайомлюється з персональним[29] - необхідно виокремити підстави для ствердження не квазі-свідомого, а дійсно свідомого “Я”.

2.2 Вимоги до самосвідомого штучного інтелекту. Свідомість і особистість.

Визначивши методи дослідження мозкової діяльності та виокремивши три рівні доступу до свідомості, ми робимо припущення щодо створення

[28] Dennett D. C. Brainstorms: Philosophical Essays on Mind and Psychology / D. C. Dennett. - Cambridge: MIT Press, 1981. - P. 197
[29] Там само. - P. 173

свідомої системи ШІ. Припущення, що свідомість можна відокремити і віднести до субперсонального рівня (а отже, відтворити її і визнавати існування свідомих істот, що не є людьми) змушує нас продовжити дослідження і обґрунтувати потенціал самосвідомого ШІ. Як ми визначаємо самосвідомість і чого необхідно набути свідомому ШІ?

Визначення самосвідомості тісно пов'язане з поняттям особистості, оскільки ми наділяємо самосвідомістю усіх, хто належить до людської спільноти. Щоб наділити когось і щось особистістю, необхідно займати певну позицію[30] щодо нього: ми не визначаємо особистість за відомими нам критеріями, а надаємо статус особистості беззаперечно, особистість конституюється нашим відношенням. Для ШІ особистість означатиме або її беззаперечне ствердження (у такому разі ми не лише надаємо ШІ певні якості, а і задаємо вимоги моральної відповідності), або необхідність застосування критеріїв особистості, яких наразі не існує. Особистість обов'язково наділена мовою - Д. Деннет визначає вербальну комунікацію умовою повноти особистості. Асоціативність мови і особистості чітко простежується у виборі тестування і експериментів, які проходить ШІ - вони виключно лінгвістичні і спрямовані не лише на раціональність мислення і відповідне мовлення, а безпосередню схожість з мовленням людини. Відсутність вербальної комунікації заперечує повноту особистості[31], наприклад, у тварин, однак мовне питання стосується скоріше самосвідомості. У "Поясненні свідомості"("Consciousness explained[32]") Д. Деннет зазначає, що ствердження відсутності свідомості у тварин за критерієм браку вербального апарату комунікації означало б припущення свідомості як радикального відмежування між свідомими і несвідомими істотами. Неможливо повністю відокремити свідомі і несвідомі акти людини один від одного, тому можна не ставити

[30] Деннет Д. Условия присутствия личности/ Д. Деннет ; [пер. с англ. Г. Хасина] // Логос. - 2003. - №2 (37). - С. 138
[31] Там само. - С. 138
[32] Dennett D. C. Consciousness explained / D. C. Dennett. - Boston: Little, Brown, 1991. - P. 447

питання про свідомих чи несвідомих тварин. Для ШІ можна припустити наступне: за загальним принципом ми не заперечуємо можливість його свідомості (як і у тварин), однак його самосвідомість вимагатиме більше за побудову апарату вербальної комунікації. Відсутність загальних критеріїв визначення особистості та свідомості означає, що необхідно встановити мінімальні вимоги, які має задовольняти ШІ. Свідомість людини є особливою вищою формою свідомості, що не присутня в інших біологічних видах, тому детальний розгляд самосвідомості ШІ логічно проводити у межах визначення самосвідомості людини, враховуючи ознаки персонального. Субперсональне стосується рівня публічного та комп'ютативного доступу, а персональний доступ ще належить охарактеризувати, застосовуючи звернення до контексту навколишнього середовища і використання "Я" звернення (що саме з моєї поведінки характеризує не квазі-свідомість). Оскільки людина переважно визначає себе через комунікацію, і свідомість, і самосвідомість мають мовний критерій як достатній для ствердження свідомого.

Д. Деннет виділяє шість[33] критеріїв особистості різного ступеню важливості, одні з яких будуть необхідними передумовами для свідомої істоти, але не достатніми - інші можна застосовувати як підтвердження належності до свідомих істот. Критерії є наступними:

- Раціональність. Усі свідомі істоти є необхідно раціональними, що цілком реалізується у ШІ - алгоритм здатний до розрахунку найбільш успішної комбінації дій серед можливих для досягнення результату.
- Інтенційність. Агент вважається раціональним, якщо він є і інтенційним - системі можуть приписуватись уявлення та бажання задля прогнозування її подальшої поведінки або пояснення вже виконаного. Інтенції об'єднують страхи, очікування, враження, що ефективно описують об'єкт, що діє. Інтенційні системи можуть бути високого та низького рівня,

[33] Деннет Д. Условия присутствия личности/ Д. Деннет ; [пер. с англ. Г. Хасина] // Логос. - 2003. - №2 (37). - С. 137

залежно від ефективності надання їм інтенційного статусу. Таким чином, системами нижчого рівня інтенції можна вважати навіть рослини, оскільки можна надавати їм “бажання” сонця, тепла, однак різноманіття їх поведінки суттєво обмежене. До складніших інтенційних систем належать тварини, яким приписується більша кількість інтенцій - хоча їх поведінку можна обґрунтувати як сталу і раціональну, інтуїтивно ми надаємо їм переконання і бажання, які спрацьовують у прогнозуванні швидко та ефективно. Відсутність оригінальної поведінки тварин характеризується як її механічність, що спирається на інстинкти та рефлекси. Можна виокремити інтенції другого порядку, або вищі інтенції, якими володіє людина, оскільки вона не лише має власні інтенції, а і уявлення щодо них. Без прийняття інтенційної позиції щодо об’єкту його не можна вважати особистістю.

- Позиція щодо об’єкта. Ми сприймаємо об’єкт і називаємо його, або вважаємо його належним до певної категорії, що визначає його роль і конституює його до певної межі. Спочатку відбувається визнання особистості, відповідно до цього особистості надаються критерії.
- Взаємність. Той, на кого спрямована інтенція і щодо кого існує позиція, має певним чином відповідати взаємністю - вступати в комунікацію. Ми можемо не розрізняти складні і прості інтенційні системи за їх справжнім володінням інтенціями чи їх приписуванням, оскільки заперечуючи можливість інтенцій у комп’ютерів та тварин ми заперечуємо власній інтуїції[34] - можна виокремити інші критерії, що вирізняють інтенційні системи другого класу. Саме таким критерієм є взаємність, інтенційна система займає інтенційну позицію у відношенні інших об’єктів - якщо розуміти її як здатність до володіння інтенціями високого порядку, взаємність буде обумовлена раціональністю, інтенційністю та

[34] Деннет Д. Условия присутствия личности/ Д. Деннет ; [пер. с англ. Г. Хасина] // Логос. - 2003. - №2 (37). - С. 141

встановленням позиції щодо об’єкта інтенції. Система може бути взаємною, проте не мати здібностей до вербальної комунікації та не бути самосвідомою: це пояснюється тим, що і людина як істота інтенційних систем другого порядку має невисловлювані інтенції (які не усвідомлює і не може описати), отже, таке відмежування було б надто радикальним. У межах функціонуючої схеми і тварин, і людину можна відносити до системи інтенційності другого порядку, якщо займати щодо неї безумовну позицію. Незалежно від дійсної репрезентації інтенції у об’єкта чи її приписування, ми отримуємо однаковий результат прогнозування: Д. Деннет зазначає, що сам факт мислення не означає осмислення і тим паче не вказує на свідомість. Таким чином, до систем, що проявляють взаємність можна відносити і ШІ.

- Здатність до вербальної комунікації. Вербальна комунікація розширює здатність системи до вияву взаємності і передбачає наявність наміру, що проявляється у розмові чи повідомленні. Через символ передається взаємність, яку ми передбачаємо: виконати А, щоб об’єкт виконав В, щоб я зміг виконати С. Комунікація приховує в собі намір, що є передумовою цієї комунікації і може бути як свідомим, так і несвідомим - вербальна комунікація часто побудована на натяках та грі сенсів, які інтенційна система другого порядку може застосовувати. Реципієнт інтенції може не повторювати логічного циклу розмислів, які передбачає у ньому ініціатор комунікації - тим не менш, ініціатор може досягати своєї мети. Відмінністю системи другого порядку є те, що свідомо чи несвідомо вона прогнозує взаємність іншої системи, і такі дії не є ні суто формальними, ні суто автоматичними. Намір, комунікація і усвідомлення дії є тріадою, що уможливлюють гру[35] запитань та відповідей між тими, хто бере участь в комунікації, оскільки лише усвідомлюючи свою дію можна

[35] Деннет Д. Условия присутствия личности/ Д. Деннет ; [пер. с англ. Г. Хасина] // Логос. - 2003. - №2 (37). - С. 150

стверджувати її намір. Якщо система є інтенційною, належить до другого порядку і використовує принцип взаємності, здатність до вербальної комунікації уможливлює її здатність до особистісної взаємодії, де всі особистості є під впливом аргументації та переконання, а їх інтереси взаєморегулюються (усі учасники використовують принцип раціональності). Комунікації недостатньо для визначення свідомості, оскільки ми можемо уявити філософського "зімбі[36]", як складну форму зомбі: він наділений інтенціями, квазі-переконаннями, відповідає взаємністю іншим об'єктам та має здатність до моніторингу власних процесів (аналізує поведінку), однак такий зімбі не є свідомим.

- Свідомість особливого типу - самосвідомість. Здатність до вербальної комунікації є передумовою для самосвідомості - рефлексивної самооцінки, що досягається завдяки позиції щодо самого себе за участі в комунікації і запитуванню. Окрім бажань щось виконувати або не виконувати, особистість має здатність схвалювати або не схвалювати власні дії, що визначає її позицію щодо самої себе. Саме завдяки рефлексії самосвідомі істоти можуть змінюватися і розвиватися, надалі брати участь у комунікації.

У підсумку, з шести критеріїв особистісності, три є спільними для багатьох живих істот (раціональність, інтенційність, позиція), взаємність є критерієм розмежування між системами першого і другого порядку інтенційності, а критерії вербальної комунікації та самосвідомості притаманні виключно людині. ШІ проходить демаркацію трьома першими критеріями та потенційно може задовольняти критерій взаємності, оскільки інтуїтивно ми надаємо ознак взаємності живим істотам, що не належать до людей лише задля обґрунтування їх поведінки (технічно). Проблема виникає на етапі втілення в ШІ вербальної комунікації та та саморефлексії. Навіть з вирішенням проблеми

[36] Dennett D. Intuition Pumps and Other Tools for thinking / D. Dennett. - New York: W. W. Norton & Company, 2013. - P.290

фреймінгу, що дасть змогу ШІ влучно оперувати психологічними прийомами та повноцінно брати участь у комунікації, ШІ не вистачатиме волінь та намірів - для цього необхідна ґрунтовна теорія субперсонального рівня (комп'ютативний та публічний доступ). Припустивши, що теорія субперсонального рівня створена і ШІ здатен до комунікації, ми зіштовхуємося з проблемою імплементації самосвідомості, оскільки критеріїв щодо її ствердження чи заперечення не існує: п'ять здійснених характеристик особистості не гарантують, що до них буде долучена шоста. Деніел Деннет розмежовує особистість метафізичну і моральну[37], де метафізична складає сукупність всіх шести критеріїв і є оболонкою особистості моральної - метафізичну особистість можна приписувати подібно до інтенційності. Постає питання щодо приписування самосвідомості і самосвідомості як "мети в собі", яким відповідають метафізичний і моральний вимір особистості. Метафізичний вимір особистості є необхідною умовою присутності особистості у сенсі моральному - у такому трактуванні здається, що ШІ може технічно відповідати метафізичному виміру особистості без охоплення морального, що є більш ніж достатнім для участі ШІ у мовних іграх та комунікаційних взаємодіях. Метафізичність є необхідною, але недостатньою для ствердження моральності дієвого агента, що знову підтверджує можливість ШІ зайняти свою позицію у класах інтенційних систем без необхідності вирішення морального питання. Суттєві труднощі виникають, якщо погодитись з тезою Д. Деннета про належність морального та метафізичного поняття особистості до одного континууму: ці поняття є нестабільними, однак не окремими або не пов'язаними. Таким чином, можна припустити, що в ШІ у будь-якому разі буде необхідно врахувати моральну сторону особистості, для якої немає повних критеріїв. Деннет зазначає, що поняття особистості є нормативним - ми можемо потрапити у пастку, встановивши критерії демаркації для особистостей і тих,

[37] Деннет Д. Условия присутствия личности/ Д. Деннет ; [пер. с англ. Г. Хасина] // Логос. - 2003. - №2 (37). - С. 137

хто ними не є (подібно до нормативного поняття свідомості, де запитання про специфічність мислення кажана створює ряд оціночних суджень, що не сприяють дослідженню свідомості). Неможливо встановити бал для проходження тесту на наділення особистістю, що не був би мимовільним - як результат, у разі сумніву щодо наявності у людини ознак моральної особистості ми не маємо критерію оцінки. Так само неможливо обґрунтувати наявність особистості у свого “Я”.

Для самосвідомості ШІ поняття, що можуть мати статус нормативності (свідомість, особистість) складають здавалось би неподоланні труднощі, оскільки маючи доступний зразок інтенційної системи для наслідування - людини, ми не маємо даних щодо можливостей і характеристик її рівня персонального доступу (за трирівневою структурою Д. Деннета). Однак, перспективним є вивчення субперсонального рівня свідомості, на якому вимоги до ШІ є досить прозорими. Нормативності не уникнути, якщо скептично ставитися до можливості створення свідомого ШІ та створювати численні перевірки на його відповідність поведінки людини, однак будь-які критерії будуть суперечити нашій інтуїтивній логіці: ми не перевіряємо присутність особистості у людей, а безумовно приписуємо її, так само відбувається і з моральним виміром особистості. Якщо свідомість та особистість за деяких умов можна надавати безумовно, не враховуючи належність ШІ до мети “в собі”, ШІ цілком може зайняти свою нішу серед складних інтенційних систем другого порядку.

РОЗДІЛ ІІІ
КРИТИЧНІ ЗАУВАЖЕННЯ ДО ФІЛОСОФСЬКИХ АРГУМЕНТІВ Д. ДЕННЕТА

Повертаючись до ситуації “Нічний бутерброд” ("The midnight snack problem[38]", варто зазначити, що аргументи Д. Деннета щодо можливої вродженості знання не видаються переконливими. Умовний розподіл знання на вроджене та набуте є необхідністю для аналізу людського мислення і виокремлення алгоритму поведінки, що не залежить від біологічних характеристик людини та може бути змодельований в ШІ. Деніел Деннет слушно зазначає, що велика кількість звичних для нас механізмів поведінки належать до набутого знання: знання з фізики, як то усвідомлення гравітації і відповідне управління предметами у просторі, з’являються ще у немовлят[39], однак не є вродженими. Припускається, що вродженими є фундаментальні для існування знання - наприклад, усвідомлення того, що дві речі не можуть знаходитись одночасно в одному місці. Якщо з емпіричним знанням можна погодитись, то питання вродженого знання є неоднозначним. Так, усвідомлення зникнення рідини з пляшки під час її переливанні у чашку може вказувати не на фундаментальне знання про умову зникнення чогось одного для появи чогось іншого, а на те, що людина неодноразово за допомогою зорового сприйняття підтверджувала цей факт. Так само я не лише знаю, що вода тепер знаходиться у склянці, а і відчуваю, що пляшка полегшала.

Тим не менш, ідея вродженого знання у людини є цікавою для реалізації і може бути розглянутою під дещо іншим кутом та поглибити наші знання щодо природи інтелекту. Наприклад, Алан Вінфілд зазначає[40], що інтелект не є прерогативою виключно тварин та не осмислюється у межах лінійної шкали від

[38] Dennett D. C. Cognitive wheels: the former problems of AI / D. C. Dennett // Brainchildren: Essays on Designing Minds. - Cambridge: MIT Press, 1998. - P. 187
[39] Там само. - P. 187
[40] Winfield A. Intelligence is not one thing / A. Winfield // Journal of Artificial General Intelligence. - 2020. - Vol. 11 (2). - P. 97-100. – Режим доступу: https://content.sciendo.com/view/journals/jagi/11/2/article-p1.xml

нульового інтелекту до суперінтелекту. Інтелект є складною структурою, до комплексу якого можуть входити морфологічний, індивідуальний, соціальний та інтелект групи (swarm intelligence). Про вроджене знання нам повідомляє морфологічний інтелект, що визначається як використання організмами їх природних властивостей, переваг морфології для успішної взаємодії з навколишнім середовищем. Індивідуальний інтелект є здатністю організму інстинктивно реагувати на стимул та здобувати нові рефлекси (або адаптуватися) завдяки навчанню. Важливим є зауваження щодо механізму навчання: він не має значення, якщо навчання відбувається і процес є самостійним. Таким чином можна припустити, що в експерименті Деннета йдеться про комбінацію індивідуального та морфологічного інтелекту для успішного виконання завдання.

Суперечливим може видатись і вже згаданий приклад з дітьми і печивом, де фізичне покарання викликає біль і формує подальший досвід уникнення болю завдяки відмові від печива. Д. Деннет вважає сталі емоційні реакції біхевіористично обґрунтованими, а отже вродженими - для того, щоб ідея печива або болю означала саме те, що означає, необхідна біологічна детермінація. Чи означає це, що окремі фрагменти знання про умови функціонування простору (переміщення предмету в одне місце означає його зникнення з іншого тощо) і визначені типи поведінки згідно з 'ідеєю', на яку вони спрямовані, є однаково обумовлені вроджено? Перший аргумент не доводиться і не спростовується емпірично, а лише вказує на можливу схему побудови алгоритму ШІ. Другий аргумент вказує на об’єктивну біхевіористичну обумовленість станів стимул-реакція, яким притаманна відстежувана логіка. Проблемою “дизайну ідей[41]” є варіативність людської поведінки залежно від обставин середовища, у якому відбувається дія - біль не завжди означатиме, що людина його уникатиме. В умовах однієї кімнати, 15 хвилин часу, людини

[41] Dennett D. C. Cognitive wheels: the former problems of AI / D. C. Dennett // Brainchildren: Essays on Designing Minds. - Cambridge: MIT Press, 1998. - P. 183.

наодинці і кнопки з електричним струмом ми отримаємо неочікувані результати. 67% чоловіків[42] і 25 % жінок, що брали участь в експерименті завдали собі болю струмом, оскільки перебувати наодинці у розмислах було нудно і нестерпно. Експеримент не спростовує тезу Д. Деннета про детермінований зв'язок ідеї-реакції, однак поглиблює його проблематику - навіть уникаючи складнощів семантики ситуацій, які ми розглядаємо (що однак буде необхідним для подальшої інтеграції ШІ в соціальні взаємодії), ми зіштовхуємося з варіативністю людської поведінки стосовно визначеної "ідеї" (біль є тільки тим, чим він насправді є).

"Проблема фреймінгу" ШІ та складність моделювання і вирішення побутових ситуацій (велика кількість комбінації дій для їх вирішення) вказують на специфічність людського мислення, яке необхідно дослідити. У роботі Д. Деннета "Когнітивне колесо: проблема фреймінгу ШІ"("Cognitive wheels: The Frame Problem of AI") згадується визначення *ceteris paribus*[43], як необхідна умова прийняття рішення людиною. Мислення у межах *ceteris paribus* означає побудову картини світу, умови якої є константою, що сприяє швидкому ухваленню рішення. *Ceteris paribus* є методом вирішення проблеми фреймінгу ШІ, якби він відтворював людське мислення: принципи визначення вихідних положень невідомі, однак велика кількість помилкових засновків відсіюється на етапі вибору стратегії. Логічно пов'язуються причини і їх наслідки: якщо підкинути предмет у повітря, він впаде внаслідок дії гравітації (так відбувається завжди, а отже умови залишаються незмінними). Аномалії можуть порушувати звичний стан речей, додаючи до задачі нових умов, при цьому *ceteris paribus* залишається чинним для інших випадків. Предмет, який ми підкинули у повітря, може впасти пізніше, ніж мав би за законами фізики і розрахунками, які ми виконали - він зіштовхнувся з іншим предметом. Подібним чином можуть не

[42] Just Think: The challenges of disengaged mind/ [T. D. Wilson, D. A. Reinhard, E. C. Westgate and others] // Science. - 2014. - Vol. 345(6192). - P. 76

[43] *Ceteris paribus* - з лат. "за рівних однакових умов"

спрацьовувати рішення для задач, оскільки з'являються нові аксіоматичні умови, що заперечують попередні рішення і змінюють *ceteris paribus*. Здатність ухвалювати правильні рішення в умовах аномалії є вагомою ознакою мислення людини і умовою, необхідною для розробки питання свідомості ШІ - виключно ситуації *ceteris paribus* недостатньо, щоб розв'язати проблему фреймінгу ШІ.

Підсумком невирішеної проблеми фреймінгу ШІ у Д. Деннета постає ідея "когнітивного колеса". Когнітивне колесо є ідеєю дизайну влаштування мозку людини, завдяки якому вона здатна обирати серед великої кількості деталей найбільш релевантні і ухвалювати рішення на їх основі. Мікроскопічні деталі у вигляді колес, що були біологічно обумовлені у найпростіших організмів (бактерії та одноклітинні еукаріоти[44]) і пришвидшували їх рух завдяки обертанню, виконали роль зразка для припущення існування подібного механізму у високорозвинених істот. Когнітивне колесо як ідея є певним спрощенням, однак системою високого рівня, що покликана пояснювати взаємодію високорозумної істоти (у тому числі ШІ) на рівні нижчому за феноменологічний. "Замість спроб змоделювати здатність до відслідковування речей у межах їх розташування у крос-секційних тимчасових зонах знання, що виражається у визначеннях вже визначеного (назвах того, чим речі є у своїй суті) та предикатів, ми могли б змоделювати відслідковування речей більш прямолінійно, залишаючи всю інформацію з перехресних зон знання про ситуацію 'тепер' переважно імпліцитною і важко доступною для вилучення з її формату (так само, як і для людини[45])" - Деніел Деннет визначає виокремлення рівня категорійного відбору у пізнанні людини як основну задачу "когнітивного колеса". Звертаючись до мисленнєвих експериментів з роботом, що мусив знешкодити вибухівку, проблеми "Нічного бутерброду" ("The Midnight Snack

[44] Dennett D. C. Cognitive wheels: the former problems of AI / D. C. Dennett // Brainchildren: Essays on Designing Minds. - Cambridge: MIT Press, 1998. - P. 200.
[45] Там само. - С. 203

Problem") та визначеного ланцюжку дій для уникнення болю, можна визначити такі проблеми ШІ (і їх вирішення завдяки “когнітивному колесу”):

1. Робот з ШІ не може відокремити релевантні засновки від нерелевантних і пріоритизувати їх, відповідно, не може знешкодити вибухівку - йому не вистачає часу на виконання завдання. Можливе обґрунтування існування механізму “когнітивного колеса” (вже у його біологічних процесах, а не як ідея дизайну) надасть зразок для втілення подібної моделі на всіх рівнях моделювання ШІ: дизайну, програмного та апаратного забезпечення. У підсумку, робот з ШІ функціонуватиме з залученням ситуації *ceteris paribus*: закони гравітації, правила пріоритетності дій (спочатку знешкодити вибухівку, потім ізолювати, якщо обмеження у часі є критичними), механізми орієнтації у часі і просторі (визначення 'мало' і 'багато', співвідношення розмірів об’єктів) надають можливість ШІ не лише правильно виконувати завдання. "Когнітивне колесо" як ідея дизайну дає змогу ШІ функціонувати без “дизайнера”, а саме мати у заданих параметрах лише задачу без переліку дій для її виконання. Наявність такого механізму є обов’язковою передумовою дискусії щодо можливої самосвідомості ШІ.
2. "Нічний бутерброд" (The Midnight Snack Problem) знову підтверджує необхідність орієнтації у ситуації з заданими параметрами, якими людина людина користується без безпосереднього усвідомлення. “Когнітивне колесо” як і в експерименті з роботом створює *ceteris paribus* (більш обмежений) для конкретної задачі. Колесо як пропозиція дизайну, а не результат біологічної еволюції людини тут особливо важливе, оскільки воно нівелює суперечливу тезу Д. Деннета про вроджене і набуте знання, що обов’язково потребує роз’яснення для розуміння механізму

роботи колеса. Хоча розглядається теорія створення ШІ як автономного суб'єкта, що зможе у подальшому виконувати певні дії без втручання в основний алгоритм (на противагу покроковій задачі людини у "китайській кімнаті" Сьорла), ШІ матиме перелік заданих функцій, що відповідатимуть поняттю "вродженого" знання. Очевидно, що навіть при визначенні повного переліку знань, що належать до вроджених (за умови, що такі знання підтверджуються), ШІ вимагає інакшого від людського співвідношення надане/набуте внаслідок різної природи. Важливо зазначити, що "когнітивне колесо" у даному експерименті порушує проблему аномалії у ситуації *ceteris paribus*: людина має набуті знання про навколишній світ і застосовує їх для передбачення своїх дій (у ситуації "Нічний бутерброд" ("The Midnight Snack Problem") людина здивується, якщо звичні їй речі будуть приклеєні до полиці). ШІ, натомість, навіть маючи попередні знання, може аналізувати ситуацію як зміну середовища і сприймати її як таку після адаптивної поведінки. "Когнітивне колесо" уможливлює вибудовування взаємодії ситуацій нормального і ненормального (аномалій), де зміна звичного c*eteris paribus* вимагатиме суттєвого аналізу зі сторони ШІ.

3. Реакція дітей на покарання продемонструвала, що особливістю людського мислення є формування сталих моделей ідея-значення, що складають основу поведінки людини. Подібна модель може бути реалізована в ШІ завдяки когнітивному колесу, що відтворюватиме подібно до експерименту "Нічний бутерброд" ситуацію *ceteris paribus*. Когнітивне колесо не обробляє інформацію про сприйняття ситуації тут і зараз (такі дані мають бути імпліцитними і входити до ширшого інформаційного блоку). Застосування ідеї речей як таких

("Things in essence") є можливим рішенням програмувати коректні дії ШІ на початковому етапі розробки без прагнення втілити здатність до аналізу семантичних рівнів спілкування, де біль може набувати інших сенсів. Варіативність поведінки людини, що показує зміну зв'язки ідея-значення (біль стає розвагою) стає аномалією, як і в попередньому експерименті. *Ceteris paribus* як ключова характеристика когнітивного колеса лише вказує на можливість аномалій, однак не роз'яснює можливості їх розв'язання.

Проблема уявної ситуації з імплементацією зв'язку ідея-значення (такі емоційні константи для людини, як біль) у дизайн ШІ заслуговує більш детального розгляду, оскільки вирішується "когнітивним колесом" лише технічно. Реакція на зовнішній подразник, що є аналогічною людській задовольняє потребу у функціональній відповідності ШІ людині, однак повертає нас до питання "симуляції" свідомості: якщо робот відтворює біль, чи може він дійсно його відчувати? Як і у випадку з тестом Тюрінга, "тремтіння" та "плач" ШІ, вказівка на розташування больової точки можуть переконати у досконалості системи ШІ, однак не вказують на філософське обґрунтування істинності переживань. "Чи є штучний інтелект справжнім інтелектом?"[46] - питання без очевидної відповіді, оскільки людина має досвід використання штучних замінників природних об'єктів, що посутньо не відрізняються від їхнього зразка.

Штучний інтелект займає позицію поміж штучними барвниками, що є повним аналогом натуральних, та штучними квітами, щодо яких неможливо ствердити ні повноцінне відтворення ідеї, ні функціоналу - результат застосування ШІ залежить від інтерпретації його спроможності до заміни. Д. Деннет стверджує, що ШІ, як належний до категорії штучного, теоретично може

[46] Dennett D. C. Brainstorms: Philosophical Essays on Mind and Psychology / D. C. Dennett. - Cambridge: MIT Press, 1981. - P. 196

бути ідеєю продуктивною для розвитку та замінити справжній інтелект[47]. Якщо припустити створення ШІ, що не поступається людському функціоналу (йдеться не про одноразову перевірку), інтелект стає функцією - Деннет порівнює її з диханням. Незалежно від природи суб'єкта, до якого ця функція долучається, і способу отримання функції втрата істинності не відбувається. Як висновок, ми можемо припускати різні типи функціонування інтелекту у різних істот, однак інтелект існуватиме у межах єдиної інтерпретації. Нівелювання різниці між штучним та "справжнім" інтелектом у контексті проблеми болю у неживих істот важлива для демонстрації того, що подібна аналогія не може бути проведена для переживання болю - фізичного і психологічного. Якщо біль є сутнісно біологічним феноменом і пов'язаний з відтворенням видів, циклом народження і смерті, його неможливо відтворити, оскільки він не є виключно функцією. Психологічний вимір не пояснює широкого обсягу значень, яких біль набуває протягом життя людини. Біль у трактуванні Д. Деннета є комплексним феноменом, що нерозривно пов'язаний не лише з біологічною функцією та його свідомим переживанням, а і уявленням про страждання, обов'язок, зло. Біль як етичний, соціальний та спеціальний принцип означає, що окрім алгоритму розпізнавання емоційної ситуації та органічної реакції ШІ (проблема, яку вирішує механізм "когнітивного колеса"), необхідно дослідити можливість імплементації етичного світогляду - без цього робот, що відчуває біль є безглуздим, однак саме завдання постає невиконанним[48].

Аналогічно до труднощів з уявленням свідомого ШІ через сприйняття свідомості як феномену персонального і спричиненого сукупністю факторів, ШІ, що здатний відчувати біль вважається нездійсненною ідеєю - робот не має соціального статусу, історії життя, особливостей біологічного походження. Для того, щоб додати до ШІ бажану функцію, що має за зразок нейробіологічні

[47] Dennett D. C. Brainstorms: Philosophical Essays on Mind and Psychology / D. C. Dennett. - Cambridge: MIT Press, 1981. - P. 197
[48] Там само. - P. 198.

особливості людини, необхідно визначити походження і дію цих особливостей - перешкодою є відсутність єдиного визначення болю як об'єкту перцепції. Біль може бути як об'єктом інтенції, на який звернена наша свідомість, так і об'єктом з незалежним статусом. Підтвердити чи спростувати жодну з позицій неможливо на даному етапі розвитку нейробіології, а філософські позиції можуть бути протилежними і тим не менш обґрунтованими, оскільки лише припускають статус об'єкту болю через інтуїтивне сприйняття. Філософською позицією Д. Деннета є уникнення фундаментального питання про належність болю до певного виду об'єкту перцепції на користь питань, які пропонують продуктивні рішення для конструювання ШІ або поглиблюють розуміння проблеми. Як відчуває біль повний клон людини, її біохімічна копія та технологічна? Чи були б всі три копії несвідомими роботами без органів чуттів? Невідповідність робота зі ШІ інтуїтивним вимогам щодо критеріїв переживання болю може означати не містицизм феномену болю, а неузгоджене трактування болю у звичному його розумінні[49] - за розвитку ґрунтовної субперсональної теорії болю він може бути реалізованим в ШІ. Наукове підтвердження походження больової реакції означає розширення інтуїтивного розуміння болю до чітких критеріїв, що зробить закиди до штучності болю в ШІ безпідставними. Вирішення проблеми фреймінгу не анулює питання щодо феноменологічного сприйняття ситуації людиною в процесі вирішення задачі, однак окремі її складові (імплементації механізму реакції ідея-значення) наштовхують на дослідження найбільш перспективних реакцій організму людини, які можливо відстежити або прогнозується, що це можна зробити. Якщо рівень персонального доступу до свідомості відстежується завдяки його виокремленню від публічного та комп'ютативного доступу (ми поступово звужуємо обсяг задач, які належать персональному доступу через виокремлення ознак персональності), то подібним чином можна локалізувати інтерпретацію

[49] Dennett D. C. Brainstorms: Philosophical Essays on Mind and Psychology / D. C. Dennett. - Cambridge: MIT Press, 1981. - P. 228

болю у свідомості людини. Д. Деннет пропонує розглядати біль як такий, що ідентифікується через події пост-інтерпретації[50] - відповідно, ми можемо відстежувати на якому з трьох рівнів персональної організації біль ідентифікується, або що з фінальної інтерпретації ми вирізняємо як біль.

Д. Деннет зазначає, що інтерпретація подій у трирівневій системі є не суто атомарним явищем, а комбінацією складних процесів - “когнітивне колесо” в інтерпретації філософа потребує або суттєвого уточнення, або ускладнення, оскільки вирішує лише частину проблеми свідомого ШІ. Сучасне дослідження роботи нейромереж якщо не ставить під сумнів існування 'когнітивного колеса', то принаймні ускладнює його завдання й функції, що мають бути відтворені у ШІ для можливості розмови про його самосвідомість. Експеримент щодо дослідження переходу мозку від усвідомлюваного сприйняття до підсвідомого[51] (авторства Ф. А. Лучіні, Дж. Дель Фераро та ін.) демонструє, що у процесах мислення досі є багато невизначеного і нейробіології ще належить виконати багато завдань. Підтверджується хибність інтроспекції: під час виконання певної задачі у мозку активуються дві зони, одна з яких є підсвідомою і є неконтрольованою для людини, що заважає їй повноцінно відтворювати послідовність рішень для виконання задачі. Особливо цікавим є взаємозв’язок між свідомим і підсвідомим мисленням: експеримент суперечить тезі про знаходження центру функціональної мережі мозку у його частині, що відповідає за свідомі процеси. Головним здобутком експерименту є дослідження процесів мислення, що є підсвідомими: визначено, що мозкова активність розповсюджується від зони підсвідомого до свідомого, при цьому ядро діяльності підсвідомого залишається ядром у структурі функціональної мережі мозку. Початок процесу мисленнєвої діяльності знаходиться у локальних

[50] Dennett D. C. Brainstorms: Philosophical Essays on Mind and Psychology / D. C. Dennett. - Cambridge: MIT Press, 1981. - P. 225

[51] Lucini F. How the brain transitions from conscious to subliminal perception / F. Lucini, G. Ferraro, M. Sigman, A. Hernan. - 2020. Режим доступу: https://arxiv.org/pdf/1903.09630v2.pdf

.

ланцюжках мозку, що спричиняють подальше кодування інформації: процес може переходити з несвідомого у свідомий саме завдяки кодуванню.

Використання сучасних досягнень нейробіології допомагає вдосконалювати філософські припущення щодо можливого операційного механізму для ШІ. Ідея “когнітивного колеса” полягає в альтернативній заміні нескінченного відбору релевантних засновків механізмом, що застосовує ситуацію *ceteris paribus* і аналізує інформацію поверхнево - вона зашифрована в інформаційних блоках і представлена загальними термінами (у Д. Деннета - ідеями як такими) подібно до того, як оперує певною інформацією людина. Припущення когнітивного колеса є робочою теорією, оскільки в біологічних процесах мозку підтверджується кодування інформації і подальша обробка вже спрощеного інформаційного обсягу. Однак когнітивне колесо має два суттєвих недоліки, один з яких зауважує сам Деніел Деннет: “...уможливлення вагомих посередницьких рівнів не означає обґрунтування їхнього існування. У конкретному випадку може відбуватися сходження від упізнаваного феноменологічно рівня психологічної дескрипції до втілення когнітивного колеса без пояснення того, як людським істотам вдається існувати з цією феноменологією[52]”. Когнітивне колесо не торкається проблеми людської свідомості, а лише симулює діяльність людського мозку, однак навіть така теорія суттєво змінює підхід до ШІ. Стала модель когнітивного колеса означає, що існування суб’єкта ШІ як активного свідомого агента є неможливим - реальна діяльність свідомого суб’єкта ніколи не є завершеною.

Другою і вагомішою проблемою є те, що когнітивне колесо постає як єднальний механізм свідомого і підсвідомого рівня оперування інформацією без роз’яснення щодо порядку відтворення послідовності процесів та переваги рівнів одним над одним. Експеримент Лучіні, Феррaро та ін. з дослідження процесів активності мозку висуває цікаве припущення: оболонка ядра активації

[52] Dennett D. C. Cognitive wheels: the former problems of AI / D. C. Dennett // Brainchildren: Essays on Designing Minds. - Cambridge: MIT Press, 1998. - P. 202.

свідомого сприйняття може бути не сталим набором нейронів, а різнитися залежно від функціональних вимог специфічного об'єкту перцепції у різні моменти часу[53]. Якщо припущення виявиться релевантним, необхідно відзначити емерджентну, а не функціоналістську теорію свідомості більш перспективною: емерджентний порядок виникає внаслідок взаємодії структурних частин свідомості (де ціле є більшим, ніж сума частин[54]), те саме відбувається і в переході нейронної активності від несвідомих до свідомих процесів. Оскільки когнітивне колесо має повторювати механізм опрацювання інформації мозком людини, його теоретичний потенціал видається досить слабким після припущення свідомого процесу як наслідку роботи активної структури нейронів, що постійно перебудовується. Навіть без розгляду свідомого процесу як феноменологічно обумовленого, когнітивне колесо постає сталим механізмом, що потенційно не зможе опрацьовувати великі обсяги інформації, якщо не матиме підструктур з додатковими функціями - усі проблеми описані у "проблемі фреймінгу" будуть повторені. Варто зазначити, що Д. Деннет не ставить за мету створити "свідомого" робота, задача полягає у зміні постановки питання про мислення: що насправді відбувається, коли ми виконуємо ту чи іншу дію, і як це можливо відтворити?

Оскільки "когнітивне колесо" спрямоване на вирішення конкретної задачі, доцільно порівняти дослідження мозку Лучіні та Ферраро з іншою теорією Д. Деннета - трирівневою системою комп'ютативного, публічного та персонального доступу. Рівень персонального доступу, що визначає самосвідомість у Деннета і лише передбачається, без пояснень його структури, жодним чином не відображається в активності мозку, тому ми можемо порівнювати ступені однакового рівня - субперсонального. Комп'ютативний рівень відповідає несвідомим процесам мозку, що відбуваються у

[53] Lucini F. How the brain transitions from conscious to subliminal perception / F. Lucini, G. Ferraro, M. Sigman, A. Hernan. - 2020. Режим доступу: https://arxiv.org/pdf/1903.09630v2.pdf

[54] Комар Л. В. Емерджентний підхід як основа еволюційного пояснення свідомості/ Л. В. Комар // Мультиверсум. Філософський альманах. - 2009. - №77. - С. 197

функціональному ядрі нейронної мережі і не припиняють активності навіть під час завершення свідомого процесу. Проблема постає з відповідністю публічного доступу свідомому процесові з експерименту, оскільки перший є лише підготовчим етапом до ствердження свідомості, сама свідомість є ознакою рівня персонального доступу. Так само проблемним є співвіднесення наукового розуміння свідомого процесу і філософського:

- Якщо припущення експерименту з нейробіології виявиться правильним, і для свідомого процесу набір нейронів є постійно змінним (залежить від часу і об’єкту перцепції), підсвідомий рівень є основним, оскільки з нього активізуються свідомі процеси. Філософський зміст поняття свідомості розуміється скоріше як пасивна здатність, а не активна структура.
- Обґрунтування підсвідомого і свідомого процесів не вирішують питання персональності та самосвідомості (навіть якщо припустити об’єднання рівнів публічного та персонального доступу для порівняння зі свідомим процесом) - явища сомнамбулізму, бажання, які ми не усвідомлюємо, але вони проявляються вказують на необхідність розмежування свідомості і самосвідомості і детального дослідження процесів, у яких простежується різниця перцепції.

Для розробки ШІ необхідно порівнювати філософські і наукові теорії для пошуку спільних точок зіткнення, що уможливлюють прогрес дослідження. Суттєві невідповідності можуть вказувати як на недостатньо розвинутий технічний потенціал науки (філософська теорія може бути підтверджена, але в майбутньому), так і на хибність філософської теорії. Перспективним для ШІ видається субперсональний підхід, що також не є ґрунтовно дослідженим, однак відповідає етапу розробки ШІ та досягненням нейробіології. За відсутності критеріїв свідомості та самосвідомості важливим є розв'язання проблем, які є необхідними умовами для їх появи: саме такими є “проблема фреймінгу” та інтенційність системи ШІ.

ВИСНОВКИ

Сьогодні ШІ є об'єктом міждисциплінарної розробки, що визначає перспективу його розвитку та невирішеність питання свідомих процесів у філософії та нейробіології. Окремі приклади комп'ютерної розробки ШІ спрямовані на вирішення конкретної задачі: наприклад, ШІ, розроблений у Стенфордському університеті (дослідниками Хе Хе, Ньян Пенгом та Персі Ліангом) запрограмований створювати панчі[55] та жартувати, і це його єдина функція. Поразка таких розробок виявляє суттєві недоліки комп'ютерної теорії, водночас детально досліджується психологія людини: що саме ми вважаємо смішним та на чому базується акт імпровізації. Натомість філософія свідомості та нейробіологія мають на меті виробити комплексний підхід, що має визначити статус свідомості як феномену, функції чи можливо їх сукупності. Деніел Деннет описує низку проблем, на які вказує ШІ і пропонує для їх вирішення індуктивний метод: дослідження можливості існування людини у ситуації *ceteris paribus*, субперсональні і персональний рівень свідомості людини, зв'язок ідей та їх значень у мисленні людини, критерії визначення свідомості та самосвідомості.

Індуктивний метод конструювання свідомості (і її трирівнева структура з трьома видами доступу, як наслідок) Деннета не суперечить сучасному дослідженню з нейробіології[56], де центром початку свідомого процесу є підсвідома частина функціональної мережі мозку. Утім, дослідження вказує і на ймовірність емерджентності свідомості, що формується безпосередньо у даний момент часу (відповідно, є залежною від середовища та сукупності елементів, що її утворюють). Питання оперування ШІ значенням символів, а не їхнім кодом (що заперечувалось у "Китайській кімнаті" Сьорла) здається неактуальним, оскільки нейрофізіологічна природа абстрактних (поняття,

---

[55] He He, Peng, N., Liang, P. Pun generation with surprise / He He, P. Liang. - 2019. - Режим доступу: https://arxiv.org/pdf/1904.06828.pdf

[56] Lucini F. How the brain transitions from conscious to subliminal perception / F. Lucini, G. Ferraro, M. Sigman, A. Hernan. - 2020. Режим доступу: https://arxiv.org/pdf/1903.09630v2.pdf

символи) та чуттєвих образів є спільною[57], утім свідомість не тотожна інтелекту, а значить її функціоналізм мусить бути додатково доведений. Метод інтроспекції для дослідження свідомості у Д. Деннета втрачає свою актуальність та визначається як наслідок біологічного переживання, однак сучасні дослідження нейробіології підтверджують важливість власної оцінки ситуації як основної для визначення відмінності феноменальної свідомості від механізму прийняття рішень та категоризації. Так, суб'єктивний свідомий опис власних дій може не відповідати об’єктивній реакції на стимули та навпаки[58], однак переживання свідомості є реальними процесами нейробіології, їх спрощення до функції не є очевидним.

Намагаючись уникнути появи “гомункулуса” на певному етапі формування свідомості, Д. Деннет як представник функціоналістського напряму філософії позбавляє свідомість статусу феномену, який тим не менш присутній: “ціле” свідомого цілком може бути більшим за суму частин, що створює ціле і залежати від умов нейробіологічного середовища, себто бути унікальним. У той час як основні процеси нервової системи характеризуються паралельним рухом, теорія глобального операційного простору (“Global workspace theory[59]”) пропонує визначити характерним для процесів свідомості серійний рух. Це вкотре підтверджує суттєву відмінність свідомих процесів від загальних нейробіологічних та ставить питання про доцільність створення свідомого ШІ: свідомі процеси значно повільніші та менш продуктивні за несвідомі паралельні процеси, отже несвідомий ШІ матиме значно більший потенціал у виконанні задач. Так, ШІ цілком може належати до інтенційних систем складного типу, що було аргументовано в роботі, проте не бути свідомим активним агентом. Утім,

---

[57] Звенігородський О. С., Качур І. В. Модель структури свідомості як множини процесів. // Штучний інтелект. - 2018. - №1. - С. 9

[58] Morales J. The Neural Substrates of Conscious Perception without Performance Confounds / J. Morales, B. Odegaard, B. Maniscalco. - 2019. Режим доступу: https://doi.org/10.31234/osf.io/8zhy3

[59] Baars J. B. How Deliberate, Spontaneous and Unwanted Memories Emerge in a Computational Model of Consciousness/ J. B. Baars, U. Ramamurthy, S. Franklin // Involuntary memory. - Singapore: Black Publishing Ltd, 2007. - P. 180

для успішного виконання задач ШІ спільне поле проблеми фреймінгу Д. Деннета та інтенційності свідомості (інтенційність є похідною свідомості, чи окремим процесом, що їй передує) має бути з’ясованим.

Поняття самосвідомості ШІ мало би означати здатність цієї автоматичної системи до рефлексії над власними свідомими станами, однак цей процес не має бути суто механічним (як алгоритм аналізу власної поведінки). Роз’яснення рівня персонального доступу свідомості може показати не лише можливу межу розвитку ШІ, а і підтвердити нездатність теорій емерджентизму та функціоналізму повно описати свідомі процеси на користь нової або доповненої філософської теорії - це перспективно для усіх наукових галузей, що беруть участь у дослідженні ШІ. Як наслідок, використання сучасних досягнень нейробіології, як науки, що ще не досягла кризової межі знань, у філософії свідомості може суттєво змінити наше уявлення про самосвідомість та її природу.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Андрощук Г. О. Штучний інтелект: тенденції розвитку технологій / Г. О. Андрощук // Матеріали XVIII Міжнародної науково-практичної конференції: “Побудова інформаційного суспільства: ресурси і технології” 19-20 вересня 2019 р., м. Київ. - Київ: УкрІНТЕІ. - 2019. - С. 189-196.
2. Деннет Д. Условия присутствия личности/ Д. Деннет ; [пер. с англ. Г. Хасина] // Логос. - 2003. - №2 (37). - С. 135-153.
3. Звенігородський О. С., Качур І. В. Модель структури свідомості як множини процесів. // Штучний інтелект. - 2018. - №1. - С. 7-14.
4. Кебуладзе В. Інтенційність як характеристика емпіричних психічних актів і як трансцендентальна умова можливості досвіду/ В. Кебуладзе // Філософська думка. - 2009. - №4. - С. 84-91.
5. Комар Л. В. Емерджентний підхід як основа еволюційного пояснення свідомості/ Л. В. Комар // Мультиверсум. Філософський альманах. - 2009. - №77. - С. 196-296.
6. Леонов А. Роберт Кірк: засновник філософських зомбі/ А. Леонов // Філософська думка. - 2016. - №2. - С. 71-77.
7. Лютий Т. Моделі свідомості: від “гомункуляризму” до “метатеорій”/ Т. Лютий // Мультиверсум.Філософський альманах. - 2007. - №62. - С. 63-70.
8. Baars J. B. How Deliberate, Spontaneous and Unwanted Memories Emerge in a Computational Model of Consciousness/ J. B. Baars, U. Ramamurthy, S. Franklin // Involuntary memory. - Singapore: Black Publishing Ltd, 2007. - P. 177-207.
9. Block N. Troubles with Functionalism / N. Block // Minnesota Studies in the Philosophy of Science. - 1978. - №9. - P. 261–325.
10. Dennett D. C. Brainstorms: Philosophical Essays on Mind and Psychology / D. C. Dennett. - Cambridge: MIT Press, 1981. - 424 p.

11. Dennett D. C. Brainchildren: Essays on Designing Minds / D. C. Dennett. - Cambridge: MIT Press, 1998. - 418 p.
12. Dennett D. C. Darwin's Dangerous Idea: Evolution and the Meanings of Life / D. C. Dennett. - London: Simon & Schuster, 1996. - 586 p.
13. Dennett D. C. Kinds of Minds: Towards an Understanding of Consciousness / D. C. Dennett. - New York: Basic Books, 1996. - 184 p.
14. Dennett D. C. Consciousness explained / D. C. Dennett. - Boston: Little, Brown, 1991. - 500 p.
15. Dennett D. C. Cognitive wheels: the former problems of AI / D. C. Dennett // Brainchildren: Essays on Designing Minds. - Cambridge: MIT Press, 1998. - P. 181-205.
16. Dennett D. C. Philosophy as Naive Anthropology: Comment on Bennett and Hacker / D. C. Dennett // Neuroscience and Philosophy: Brain, Mind and Language. - New York: Columbia University Press, 2007. - P. 73-96.
17. Dennett D. C. Content and Consciousness / D. C. Dennett. - London: Routledge & Kegan Paul, 1969. - 198 p.
18. Dennett D. Intuition Pumps and Other Tools for thinking / D. Dennett. - New York: W. W. Norton & Company, 2013. - 512 p.
19. Dennett D. Kinds Of Minds: Towards An Understanding Of Consciousness / D. Dennett. - New York: Basic Books, 1996. - 184 p.
20. He He, Peng, N., Liang, P. Pun generation with surprise / He He, P. Liang. - 2019. - Режим доступу: https://arxiv.org/pdf/1904.06828.pdf
21. Just Think: The challenges of disengaged mind/ [T. D. Wilson, D. A. Reinhard, E. C. Westgate and others] // Science. - 2014. - Vol. 345(6192). - P. 75–77.
22. Laiard J. Intelligence, Knowledge & Human-like Intelligence / J. Laiard // Journal of Artificial General Intelligence. - 2020. - Vol. 11 (2). - P. 41-44. -

Режим доступу: https://content.sciendo.com/view/journals/jagi/11/2/article-p1.xml

23. Lucini F. How the brain transitions from conscious to subliminal perception / F. Lucini, G. Ferraro, M. Sigman, A. Hernan. - 2020. Режим доступу: https://arxiv.org/pdf/1903.09630v2.pdf
24. Morales J. The Neural Substrates of Conscious Perception without Performance Confounds / J. Morales, B. Odegaard, B. Maniscalco. - 2019. Режим доступу: https://doi.org/10.31234/osf.io/8zhy3
25. Nagel T. What is it like to be a bat? / Nagel T. // Philosophical Review. - 1974. - Vol. 83 (4). - P. 435-450.
26. Putnam H. Brain in a vat / H. Putnam. - 1981. Режим доступу: http://ieas.unideb.hu/admin/file_2908.pdf
27. Searle J. R. Minds, brains, and programs / J. R. Searle // Behavioral and Brain Sciences. - 1980. - Vol. 3(3). - P. 417-457.
28. Searle J. R. The rediscovery of mind/ J. R. Searle. - Cambridge: MIT Press, 1992. - 286 p.
29. Tamietto M. Neural bases of the non-conscious perception of emotional signals / M. Tamietto, B. de Gelder // Nature Reviews Neuroscience. - 2010. - Режим доступу: https://doi.org/10.1038/nrn2889.
30. Winfield A. Intelligence is not one thing / A. Winfield // Journal of Artificial General Intelligence. - 2020. - Vol. 11 (2). - P. 97-100. – Режим доступу: https://content.sciendo.com/view/journals/jagi/11/2/article-p1.xml