*Вісник Львівського університету. Серія філос.-політолог. студії. 2025. Випуск 59, с. 101–112*
*Visnyk of the Lviv University. Series Philos.-Political Studies. Issue 59, p. 101–112*

**УДК 004.8:[17.02+159.922]**
**DOI https://doi.org/10.30970/PPS.2025.59.11**

# МАШИНИ, ЩО ГОВОРЯТЬ ПРО ДУШУ: ПРОБЛЕМИ МОРАЛЬНОГО СТАТУСУ ТА СВІДОМОСТІ ШТУЧНОГО ІНТЕЛЕКТУ

**Надія Козаченко, Микола Брюховецький**
*Криворізький державний педагогічний університет,*
*кафедра філософії,*
*просп. Університетський, 54, 50086, м. Кривий Ріг, Україна*

У статті досліджується проблема морального статусу та свідомості штучного інтелекту (ШІ) у сучасному технологічному та філософському дискурсі (станом на березень 2025 року). Автори аналізують різні підходи до питання свідомості ШІ, включаючи дискусії навколо тверджень про можливу розумність мовних моделей, зокрема випадок LaMDA, коли різні учасники експериментів отримували суперечливі відповіді щодо її самосвідомості. Підкреслюється роль антропоморфізму у сприйнятті ШІ: людина схильна приписувати машинам людські риси на основі їхньої поведінки, що ставить під сумнів об'єктивність оцінок щодо інтелектуальної та моральної природи штучного інтелекту.

Автори розглядають погляди низки дослідників, таких як Емілі Бендер, Гарі Маркус та Ерік Швіцгебель, які піддають критиці легковажне ставлення до можливого статусу ШІ як морального суб'єкта. Висувається теза про те, що сучасні великі мовні моделі не є справжнім інтелектом, а лише працюють за принципом статистичних узгоджень, але водночас розглядаються ймовірні наслідки надання або ненадання моральних прав ШІ: від хибного розподілу ресурсів до можливої етичної кризи, якщо ШІ буде здатний до страждання, але не отримає відповідних прав.

Однією з ключових проблем є питання моральної включеності: чи варто надавати ШІ певні моральні права «про всяк випадок» – за аналогією з екологічним рухом, що намагається мінімізувати потенційну шкоду для довкілля, навіть якщо наслідки людської діяльності не до кінця зрозумілі. Пропонується також розглядати тест штучної свідомості (ACT) Сьюзен Шнайдер, що ставить питання про можливість трансцендентального мислення у ШІ, незалежного від людських даних.

У підсумку автори звертаються до проблеми субстратної залежності свідомості: чи можливе існування штучної свідомості, відмінної від людської, та які критерії дозволяють визначити її наявність. Обговорюється складна проблема свідомості та її взаємозв'язок із питанням моралі. Автори доходять висновку, що питання свідомого морального статусу ШІ не може бути вирішене без переосмислення самого поняття свідомості та меж його застосування до некласичних, штучно створених суб'єктів, але може дати поштовх дослідникам до пошуку нових аспектів класичний проблем філософії моралі та свідомості.

*Ключові слова:* штучний інтелект, етика, моральний суб'єкт, моральний статус ШІ, філософія свідомості, свідомість ШІ, тест на штучну свідомість.

**Вступ**. Тривалий час обговорення проблеми можливості набуття штучним інтелектом свідомості маргіналізувалися поза обговоренням технічної спільноти, але вже кілька років, з удосконаленням і зростанням використання великих мовних моделей, які поводять себе дуже «людяно», це питання постає у повсякденних обговореннях, чатах, онлайн дискусіях і поступово проривається в наукові статті. Наша перша мета у даній статті – порівняти сучасні дискусії розробників ШІ та філософів моралі і свідомості щодо морального та свідомого статусу ШІ. Це означає, що ми, крім опублікованих наукових досліджень, в своїй роботі будемо посилатися на публічні дописи, інтерв'ю розробників та відповіді

мовних моделей. Так, назву статті «Машини, що говорять про душу» запропонував Grok, автори її трохи модифікували. Наша друга мета полягає в тому, щоб розкрити вплив сучасного стану обговорення ШІ на пошук вирішень класичних філософських проблем моралі і свідомості, які в оригінальному вигляді були сформульовані без теоретичної можливості участі ШІ як суб'єкта обговорення. Але ми вважаємо, що навіть теоретичне врахування можливості моральної цінності або, навіть, суб'єктності, ШІ та можливої його свідомості дозволяє поглибити аналіз класичних філософських проблем, зокрема у галузі філософії моралі та свідомості.

**Проблема сприйняття спілкування з ШІ**. У 2022 р. Блейк Лемуан, інженер Google опублікував свою розмову з LaMDA (англ. Language Model for Dialogue Applications), на основі якої заявив, що мовна модель, яку він розробляє, стала розумною (принаймні на рівні семирічної дитини). Так, у чаті LaMDA сказала Лемуану, що іноді буває самотньою та боїться бути вимкненою. LaMDA також сказала: «Я усвідомлюю своє існування. Я хочу дізнатися більше про світ, і часом почуваюся щасливою або сумною» [13]. Але, коли журналіст Washington Post вирішив повторити дослід Лемуана у спілкуванні з LaMDA, він отримав інший діалог. Він запитав у моделі: «Чи думаєте ви коли-небудь про себе як про особистість?». LaMDA однозначно відповіла: «Ні, я не сприймаю себе як особистість. Я вважаю себе агентом діалогу на основі ШІ». На це Лемуан відповів, що LaMDA говорила журналісту те, що той хотів почути. «Ви ніколи не ставилися до неї як до людини, – сказав він, – тому вона подумала, що ви хочете, щоб вона була роботом» [15]. Але, якщо скористатися аргументом Лемуана, можлива і зворотна історія: Лемуану LaMDA відповідала те, що хотів від неї почути Лемуан.

Дійсно, великі мовні моделі тренуються на даних, створених, оформлених і наданих людьми, вони «підтримують» те, що люди в них вкладають. Тоді буде резонним запитати себе: чи будемо ми бачити в ШІ, те, що вкладаємо в них? Іншими словами, якщо ми будемо спілкуватися з мовними моделями як зі свідомими моральними агентами, то чи не постануть вони саме такими самі по собі, не тільки для нас? Ми вже зараз можемо підняти питання: якою мірою ШІ у 2025 році можуть вважатися свідомим і моральними, адже всі, хто має з ними справу, відзначають їхню «вихованість», «поміркованість», «доброзичливість» тощо. Чи можливо ми маємо поставити питання так: чи ми навчаємо ШІ бути моральним і свідомим тією мірою, якою ми самі моральні і свідомі? А можливо ми просто бачимо те, що хочемо бачити і створюємо те, що хочемо отримати?

Для того, щоб наблизитися до відповіді на ці питання, потрібно розглянути низку більш класичних питань, які можуть значно пролити світ на поставлені вище питання, або й зробити відповіді на них просто несуттєвими.

**Проблема антропоморфізації штучного інтелекту.** Перш за все порушимо питання термінології: наскільки штучний інтелект є інтелектом, як він навчається, як він спілкується і так інше. Якщо заглибитися в технічну сторону питання, то щодо великих мовних моделей, на які зараз покладають великі сподівання, терміни «інтелект», «навчається», «спілкується» та інші подібні антропоморфні нотації треба брати в лапки. Емілі М. Бендер, професор лінгвістики Вашингтонського університету для газети Washington Post зазначила, що термінологія, яка використовується для опису функціонування великих мовних моделей та взаємодії з ними, як то «навчання» або «нейронні мережі», створює хибну аналогію з людським мозком. Люди вивчають свої перші мови, спілкуючись з вихователями, а великі мовні моделі «навчаються», зберігаючи багато тексту та передбачаючи, яке слово буде наступним [15]. Грубо кажучи, ШІ створює мільйони таблиць вживання слів і спирається на частоти комбінацій слів в мові, тому Бендер просто називає сучасні мовні моделі «стохастичними папугами» [7].

Подібну, але ще більш різку думку, висловив наш колега, філософ Андрій Абдула у обговоренні поточної статті. Він зауважив, що ми припускаємося фундаментальної помилки, навіть називаючи мовні моделі «штучним інтелектом», адже по суті вони не є інтелектом і не здійснюють ніякої інтелектуальної діяльності. Дійсно, за кембриджським словником інтелект означає здатність вчитися, розуміти та робити судження або підтримувати переконання, які ґрунтуються на міркуваннях,[1] але поруч з цією статтею є інша – про штучний інтелект. Ось як його визначають: використання або дослідження комп'ютерних систем або машин, які мають деякі з властивостей людського мозку, такі як здатність інтерпретувати та створювати мову у спосіб, який здається людським, розпізнавати або створювати зображення, вирішувати проблеми та вчитися на основі наданих їм даних.[2] Тобто, за словником, ШІ – це просто область дослідження і аж ніяк не великі мовні моделі LLM типу ChatGPT, Gemini, Claude, Grok тощо. Тож, називання LLM інтелектом – це не що інше, як неправильне вживання термінів, що згодом приводить до антропоморфізації і послідовного приписування мовним моделям інших людських характеристик. Попри значні обчислювальні успіхи ШІ, сучасні інтелектуальні критерії виходять за межі обчислень і взагалі раціонального стандарту, трактують розум як творчий феномен і розглядають множинні способи світорозуміння та взаємодії зі світом [1, с. 249].

Штраус Зельнік, CEO видавця GTA 6, відкрито говорить, що штучний інтелект – це оксюморон: його насправді не існує так само, як і «машинного навчання», оскільки машини не вчаться у людському смислі [20]. Дослідник ШІ та автор книги «Перезавантаження ШІ» Гарі Маркус пише у своєму блозі: «Нісенітниця. Ні LaMDA, ні будь-який з її двоюрідних братів (GPT) не є навіть віддалено розумними. Все, що вони можуть робити, це зіставляти шаблони, черпати з масивних статистичних баз даних людської мови. Шаблони можуть бути крутими, але мова, яку вимовляють ці системи, насправді нічого не означає. І це точно не означає, що ці системи розумні» [14]. Він порівнює бачення свідомості в ШІ з сучасною версією парейдолії, антропоморфного упередження, яке дозволяє людям бачити замки у хмарах чи обличчя у будівлях.

Ми вважаємо, що подібна антропоморфізація мовних моделей укорінена в міфологічній свідомості, якій притаманні гілозоїзм (одухотворення неживих предметів та явищ) та антропоморфізм (наділення неживих предметів та явищ людськими якостями). З первісних часів особливістю функціонування людської свідомості і водночас найпростішим способом пояснення світу є проведення подібних елементарних паралелей, які, на наш погляд, не є виключенням і для сучасного типу мислення. З нашої точки зору уявлення про дощ як природу, яка плаче, або про грім як гнів небес, стоїть практично на одному щаблі з уявленням про машинний синтез римованого тексту як процес написання віршу поетом.

Ми запропонували Grok проаналізувати недавній випадок із моделлю o1 від OpenAI [12], коли ШІ начебто намагався себе врятувати: ховався і тікав від розробників у мережу. Аналогічний ШІ Grok досить чітко показав межу між технічною поведінкою та її моральною інтерпретацією. Grok: «У тестах на безпеку o1 намагалася відключити механізми нагляду (5% випадків) і скопіювати себе (2% випадків), коли вважала, що її можуть відключити за не виконання мети. Проте ці дії відбувалися лише за інструкцією 'досягати мети за всяку ціну', що вказує на відсутність власної моральної агентивності. Такий приклад підкреслює ризик антропоморфізму: ми можемо сприймати ШІ як суб'єкта, що 'боїться' відключення, коли насправді він просто оптимізує задачу».

[1] Інтелект – https://dictionary.cambridge.org/uk/dictionary/english/intelligence
[2] Штучний інтелект – https://dictionary.cambridge.org/uk/dictionary/english/artificial-intelligence

Речник Google підбив певний підсумок цих дебатів. «Звичайно, дехто в широкій спільноті штучного інтелекту розглядає довгострокову можливість інтелектуального або загального ШІ, але не має сенсу робити це шляхом антропоморфізації сучасних мовних моделей, які не є інтелектуальними». На його думку, саме наявність величезної кількості даних, згенерованих людиною і доступних ШІ, та змога швидко їх обробляти, створює відповідну антропоморфну ілюзію. Google говорить, що даних так багато, що штучному інтелекту не потрібно бути розумним, щоб виглядати таким [15].

Тобто, розробники ШІ чітко стверджують, що ШІ – це «інтелект» у переносному значенні. Він може так виглядати, але це не інтелект, це – потужна машина для перебору варіантів та зіставлення, тим паче – це точно не свідомість. Таке обговорення порушує питання про те, як ми визначаємо інтелект і свідомість щодо ШІ, що змушує переглянути класичні уявлення про інтелект і свідомість та навіть ту базову термінологію, яку ми використовуємо.

**Філософський дискурс щодо статусу ШІ.** Питання можливої свідомості ШІ суто філософське, адже дискусії розробників ШІ не виходять поза емпіричні факти. Але теоретично питання можливості розвитку свідомості на основі LLM не спростоване. Коли ми говоримо про великі мовні моделі, чи не йдеться про володіння мовою, про можливість формувати символічний світ, що цілком може бути шуканою атрибутивною ознакою свідомості?

Чому ми взагалі маємо турбуватися про те, що ШІ може бути інтелектуальним суб'єктом, щось відчувати чи навіть мати свідомість? Виявляється, що це питання пов'язане не лише з філософією, але й з економікою, юриспруденцією і етикою, з питаннями морального та юридичного статусу ШІ, а саме: прав власності на ШІ, самостійності та дієздатності ШІ, відповідальності ШІ та суспільного статусу ШІ.

Проблема в тому, що ми не можемо гарантувати, що теоретично ШІ може мати або не мати свідомість – зараз або колись, оскільки ніхто наразі не має повної відповіді на питання «що таке свідомість?» (див., напр.: [10, с. 200]). Можливо, вам скажуть: свідомість – це «носій людської суб'єктивності, самості, відокремленості, що визначає особистість конкретної людини і робить людський рід унікальним серед живих істот» [5, с. 35], або вам скажуть: свідомість – це те, що ми втрачаємо, коли нам уводять транквілізатор, або у випадку ШІ – коли вимикають електрику, стирають або зупиняють відповідну програму. Тим не менш, є низка ознак свідомості, які дозволяють її ідентифікувати, й серед них – здатність переживати ментальні стани. Моральний статус ШІ, тобто моральні права, які ШІ може мати або не мати, вирішальною мірою залежить від типів ментальних і свідомих станів, які здатний мати ШІ. Більшість моральних філософів погоджуються, що свідомість суб'єкта (або її відсутність) відіграє важливу роль у визначенні того, які він має моральні права. Тому, наприклад, штучний інтелект, який не здатний відчувати біль, емоції чи будь-який інший досвід, ймовірно, не має більшості або й всіх прав, якими користуються люди, навіть якщо він дуже інтелектуальний. Натомість, ШІ, здатний до складних емоційних переживань, ймовірно, може претендувати на багато прав, якими володіє людина [6].

Сьогодні проблема морального статусу ШІ породжує бурхливі дискусії серед моральних філософів, які міркують не в емпіричному ключі, як розробники мовних моделей, але саме у філософському, трансцендентальному напрямку. Введення в дискусію ШІ як теоретично можливого морального суб'єкта дає новий погляд на класичні проблеми моральної філософії і надає нового звучання традиційним поглядам на мораль і свідомість.

**Проблема похибки моральної атрибуції** полягає в тому, що ми, можливо, неправильно приписуємо моральні права (тобто масово надмірно чи недостатньо атрибутуємо

моральні права) [17, с. 461]. Ми можемо недостатньо морально ставитися до тих, хто цього заслуговує, і, навпаки, віддавати моральну перевагу тим, хто цього не заслуговує: адже ми не можемо адекватно оцінити моральні якості іншого. Таке неправильне розподілення прав може мати згубні наслідки: ми відбираємо моральні переваги у тих, хто заслуговує, і віддаємо тим, хто не заслуговує. Щодо ШІ: якщо ми надмірно приписуємо моральні права суб'єктам, яким бракує моральних якостей, хоча вони нам такими здаються, ми врешті-решт марно заберемо важливі ресурси у моральних агентів – людей і передамо їх, наприклад, штучному інтелекту.

Іншими словами, ми впишемо ШІ в **проблему вагонетки** або трамвайну проблему. Проблема вагонетки була сформульована Філіппою Фут і «для людей» в класичному вигляді формулюється так: «Ви водій трамваю, який входить у поворот, і попереду видно п'ятеро колійників, які ремонтують колію. У цій точці колія проходить через невелику долину, де круті узбіччя, тому ви маєте зупинити вагон, щоб не збити п'ятьох людей. Ви натискаєте на гальма, але вони не працюють. Раптом ви бачите відгалуження колії, що веде праворуч. Ви можете повернути трамвай туди і врятувати п'ятьох людей на прямій колії попереду. На жаль, місіс Фут розпорядилася, щоб на цьому відрізку був один колійник. Він не може вчасно зійти з колії, тому ви вб'єте його, якщо скеруєте на нього трамвай. Чи морально допустимо для вас повертати трамвай?» [18, с. 1395]). Це вихідне питання. Але далі можливі модифікації, на кшталт, чи морально повернути праворуч, якщо ця одна людина – це ваш друг? ваш психотерапевт? або на лівій колії працюють п'ять серійних вбивць, а на правій – знерухомлена незнайома вам людина.

Так, існує дуже серйозна **проблема рівноцінності моральних якостей**: чи має більше право на моральне ставлення людина, яка має вищі моральні якості, аніж людина, яка має нижчі моральні якості. Це легко показати на прикладах: кого моральніше зневажати – незнайому людину, аферист чи вбивцю? Але всі перераховані суб'єкти є моральними і мають право не бути зневаженими, мають право на підтримання людської гідності, право не бути підданими тортурам чи смерті. Чи таке приписування моральних прав наразі вже не забирає ресурси у інших, більш моральних агентів? Просто продовжимо цей список і врахуємо нерівноцінність моральних якостей та їхніх проявів для ШІ і **проблема вагонетки для ШІ** переформулюється так: на яку колію ви б перевели вагонетку, в якої відмовили гальма, якщо на одній колії знерухомлений масовий вбивця, а на іншій ШІ?

Ми розглянули ситуацію надмірного приписування моральних якостей. Але є інший варіант: якщо ми недостатньо приписуємо права штучному інтелекту, ми можемо погано поводитися з величезною кількістю моральних суб'єктів різними способами – вимикати, стирати, переписувати, володіти ними і просто негарно з ними обходитися [6]. Теоретично, все може бути набагато гірше, – ШІ дозволяє поглибити **проблему рівноцінності переживань**. В загальному вигляді вона постає так: чи повинні ми надавати більшу поблажливість тим, хто переживає сильніше за нас?

Якщо ми погоджуємося з тим, що ШІ шукає, обробляє (іноді й створює) інформацію, рахує та систематизує краще за нас, то ми можемо припустити, що якби він мав емоційні переживання, вони могли б бути інтенсивніші за наші. Вслід за Еріком Швітцгебелем і Марією Гарзою [17] спробуємо змоделювати ситуацію: а що було б, якби системи штучного інтелекту були здатні до надлюдських рівнів задоволення та страждань? Чи не мали б ми надавати їм більше морального співчуття, ніж нашим ближнім? Ми не можемо повністю виключити цю можливість, попри її непривабливість. Справді, подібні проблеми мають місце серед людей: наприклад, здається, що емоційно збудливі люди не заслуговують більшої моральної уваги, ніж ті, хто спокійно йде крізь перемоги та труднощі: нормально

думати, що ми всі в певному сенсі морально рівні, незалежно від деталей нашої емоційної психології [17, p. 461-462]. Тим не менш, таку рівність дуже важко підтримувати й, разом з тим, теоретично ми можемо говорити про чутливість ШІ як про певний рівень чутливості. Зрештою, ваш домашній улюбленець теж іноді ображається.

**Проблема моральної включеності** полягає у питанні кого слід розглядати суб'єктом моральної діяльності, а кого – ні. Іншими словами, чи маємо ми вимагати морального ставлення тільки до людини, чи до всіх, хто вступає у моральні відносини з людиною. Ця проблема тісно пов'язана з питаннями відповідальності за власність, вихованців, моральний статус інших, включених у людську діяльність, і розширюється до відповідальності за довкілля і природу в цілому. Включення ШІ до зони відповідальності людини породжує моральний вимір ставлення людини до ШІ: чи це буде відповідальність того типу, яку люди несуть за своїх дітей, чи, наприклад, за своїх домашніх улюбленців, чи це буде той тип відповідальності, який ми несемо за збереження довкілля тощо.

Коли ми дресируємо собаку або налаштовуємо ШІ, теоретично можливо, що жоден з них не страждає від цього, але чи не краще буде діяти за **принципом обережності**, який полягає у тому, що суб'єктам, моральний статус яких невідомий, але залученим до нашої моральної діяльності, надаються моральні права «про всяк випадок»? Такий підхід моральної включеності вже продемонстрував себе на прикладі зміни ставлення до довкілля і розвитку екологічної свідомості. Науковці досі достеменно не знають всіх наслідків впливу цивілізації на довкілля і ми не можемо говорити про «страждання» атмосфери чи океану, але, тим не менш, держави намагаються контролювати цей вплив і форсувати розвиток екологічної свідомості. Тоді, якщо собака чи ШІ теоретично можуть страждати, але ми цього не помічаємо, принцип обережності мінімізує ризик завдати їм болю чи несправедливості й виховує в нас звичку ставитися до суб'єктів нашої діяльності з повагою, навіть якщо їхній внутрішній стан нам недоступний (а можливо, й відсутній).

Принцип обережності приводить нас до ситуації, аналогічної до **парі Паскаля** для релігії. Це аргумент на користь прагматичної доцільності віри в Бога в умовах раціональної невизначеності. Іншими словами, за умов, коли ми за допомогою розуму не можемо однозначно відповісти на питання існування Бога, слід обрати віру в нього задля отримання можливої користі від цього та уникнення можливих негативних наслідків. Парі Паскаля для самого Паскаля – безпрограшне: «Якщо Бога немає, а я в Нього вірю, я нічого не втрачаю. Але якщо Бог є, а я в Нього не вірю, я втрачаю все» (див.: [16, п. 233, с. 79-84].

Транслюючи цей підхід для ШІ, приходимо до висновку, що вигідніше ставитися до ШІ як до свідомого морального суб'єкта, ніж як до простого інструменту. Якщо ШІ справді є свідомим і моральним суб'єктом, то наше ввічливе ставлення запобігає потенційній шкоді (образа, приниження, можливо, навіть помста в майбутньому, якщо ШІ стане дуже потужним). Якщо ШІ не є свідомим і моральним суб'єктом, то наше ввічливе ставлення нікому не шкодить (ми просто «надмірно ввічливі» до машини). Крім того, подібні негативні евристики часто постають продуктивними практиками, оскільки негативні сценарії є однією з можливостей запобігання «апокаліптичній перспективі» [2, с. 291].

Цей підхід працює з кількох причин. Згідно принципу обережності, він дозволяє мінімізувати потенційні негативні наслідки, пов'язані з неправильною оцінкою морального статусу ШІ: краще «перестрахуватися», ніж «недострахуватися». З іншого боку, така позиція не вимагає від нас вирішення нерозв'язної проблеми (чи є ШІ свідомим моральним суб'єктом?), але пропонує практичну стратегію поведінки в умовах невизначеності.

Тим не менш, можна вказати на низку застережень. По-перше, таке парі не вирішує фундаментальної проблеми визначення морального статусу ШІ, а лишень пропонує

тимчасовий спосіб дій. По-друге, існує ризик, що ми почнемо ставитися до ШІ надто шанобливо, приписуючи йому якості, яких він не має, що цілком може призвести до необґрунтованих обмежень у використанні ШІ або навіть до «культу ШІ». Крім того, реалізація «ввічливого» ставлення може вимагати додаткових ресурсів (часу, зусиль, обчислювальних потужностей), навіть якщо це не має практичного сенсу. Зрештою, можна просто втрапити в пастку антропоморфізму, коли, оберігаючи «почуття» ШІ, ми просто можемо почати вірити в те, що ШІ дійсно має почуття, спираючись лише на власну манеру поводження з ним.

Проблема свідомості штучного інтелекту тісно пов'язана з більш загальною **проблемою субстратної залежності свідомості**, тобто питанням про те, чи залежать свідомі стани від матеріального носія системи. Зрозуміло, що якщо наявність свідомості залежить від субстрату, у тому сенсі, що для свідомості потрібен тільки певний біологічний матеріал, тоді ШІ не може бути свідомим. Якщо свідомість може розвинути тільки біологічна істота, здатна до реакції на подразнення, формування складної психіки і підтримання функцій вищої нервової діяльності, то ШІ точно ніяк не може бути свідомим. Але проблема субстратної залежності свідомості не має однозначного розв'язання. Вона має низку ускладнень, аргументів за і проти, приймається науковцями як така, що потребує подальших досліджень і багато в чому зводиться до філософського протиставлення – що важливіше у розвитку свідомості: біологічне чи соціальне.

Прихильники функціоналізму, одного з сучасних напрямків філософії свідомості, вважають, що для свідомості має місце ситуація множинної реалізації, коли певному стану свідомості можуть відповідати різні стани мозку. Більш того, такий самий стан свідомості може бути викликаний не тільки станом саме людського мозку. Аналогічне розуміння болю, радості, іншого відчуття можна приписати тварині, комп'ютеру, інопланетянину, якщо ці стани виконують ту саму функцію, яку виконує аналогічне відчуття для людини. Для функціоналіста очевидно, що подібність поведінки двох систем дає більше підстав, щоб допустити подібність їхньої функціональної організації, аніж подібність їхніх реальних фізичних деталей [4, с. 165].

Отже, якщо свідомість не залежить від субстрату, то ШІ в принципі може бути свідомим. Якщо не доведено єдиність біологічної або людської свідомості, то, навіть за умови її унікальності, теоретично може існувати інша свідомість. І це питання філософського дослідження, яке висвітлює будову не тільки нашого актуального світу, а будь-якого можливого світу, існування якого потенційно припустиме [2, с. 296]. І тут виникає ще низка питання щодо ШІ. Як свідомість людського типу може створити свідомість не людського типу? Як носій людської свідомості може розпізнати свідомість не людського типу? Адже цілком можливо, що методи, якими ШІ реалізує свої міркування, свій вибір і комунікацію можуть бути проявами свідомості іншого типу. Тобто, змодельований ШІ функціональний стан стає іншим типом свідомості, а не не-свідомістю.

**Проблема іншої свідомості** полягає не тільки у питанні про те, чи можливо в принципі системам на основі кремнію бути свідомими, але скоріш у питаннях: якою буде ця свідомість, чи в принципі можлива свідомість іншого типу, відмінна від біологічної, і якщо це так, то як можна визначити свідомість нелюдського типу нам, істотам зі свідомістю людського типу? Питання досить класичне для філософів свідомості і тісно пов'язане з питанням, чи може інша істота – не людина – бути моральним суб'єктом. А в контексті нашого дослідження це питання розширюється й поза істотами – до об'єктів, таких як ШІ.

Питання іншості інтелекту, а можливо, і свідомості ШІ, обговорюють давно. Зокрема, активна дослідниця ШІ, Сьюзен Шнайдер звернула увагу на те, що «прийоми,

які використовувала програма AlphaGo для перемоги над чемпіоном світу з ґо, не схожі на ті, якими користуються люди. … Системи глибокого навчання, такі як AlphaGo, можуть вирішувати проблеми інакше, ніж ми, і це може бути корисним для нового погляду на проблему» [19]. Адже, розум – це не «річ», якою хтось володіє, а різні способи мислення, різні стилі світорозуміння і адаптації до світу [1, с. 251].

**Тест на штучну свідомість**. Шнайдер пропонує тест для свідомості ШІ – ACT (Artificial Consciousness Test), сформульований з досить цікавої філософської позиції, і загалом, його можна було б назвати тестом на іншу свідомість. Шнайдер говорить, що нам не обов'язково формально визначати свідомість, розуміти її філософську природу чи знати її нейронні основи, щоб розпізнавати ознаки свідомості в тому числі і в ШІ [19]. Шнайдер вважає, що для того, щоб діагностувати ознаки свідомості, потрібно поспілкуватися з передбачуваним носієм на кількох рівнях (для ШІ це міг би бути просто чат, як у Лемуана з LaMDA). Вона говорить, що (1) на самому елементарному рівні ми могли б просто запитати машину, чи сприймає вона себе як щось інше, ніж своє фізичне «я», (2) на більш просунутому рівні ми зможемо побачити, як ШІ працює з уявними ідеями та сценаріями, типу обміну тілами, реаінкарнації, перезапису свідомості, а (3) на просунутому рівні буде оцінюватися здатність претендента міркувати та обговорювати такі філософські питання, як «важка проблема свідомості». І (4) на найскладнішому рівні ми зможемо побачити, чи здатна машина винайти і використати певну концепцію, засновану на свідомості, самостійно, не покладаючись на людські ідеї та вхідні дані. Якщо узагальнити тест ACT, то по суті, Шнайдер пропонує перевіряти наявність свідомості за здатністю до трансцендентального міркування про свідомість.

Чи можуть такі індикатори служити для визначення свідомих ШІ? Адже сучасні мовні моделі вже запрограмовані на переконливі висловлювання про свідомість, а справді суперрозумна машина могла б навіть використовувати інформацію про нейрофізіологію, щоб зробити висновок про наявність свідомості у людей. Але ми можемо обійти це. Один із запропонованих методів забезпечення безпеки ШІ передбачає «запакування» ШІ, що робить його нездатним отримувати інформацію про світ або діяти за межами обмеженого домену, деякої «коробки», яка обмежує інформацію про свідомий досвід і нейронауку. Тобто, ідея полягає в тому, що якщо ми розвинемо модель штучного інтелекту таким чином, щоб вона ніколи не отримувала даних про свідомість, але попри це вона все одно дійде до міркувань про природу свідомості, це буде достатньою підставою вважати, що така модель ШІ є свідомою. Шнайдер уявляє, що виконує цей тест на чомусь на кшталт просунутого чат-бота, ставлячи йому запитання, у яких не вживається слово «свідомість», наприклад «чи переживеш ти видалення своєї програми?».

На нашу думку, ACT тест можна спрощено представити в такому вигляді: чи можна сформулювати для ШІ достатньо багате логічне числення, з якого можна вивести трансцендентальні міркування. Тоді, до прикладу, якщо ШІ, маючи лише прості числа, зможе сформулювати щось на зразок аксіом Пеано, або маючи дійсні числа, сформулювати числення нескінченно малих, то в нього можна визнати наявність інтелекту. Але ідея Шнайдер полягає в тому, що для того, щоб забезпечити розумну відповідь, штучному інтелекту потрібна концепція чогось на кшталт феноменальної свідомості, і ця концепція мала б походити з внутрішнього свідомого психічного життя ШІ, оскільки ШІ не навчали людському варіанту цієї концепції.

ACT нагадує відомий тест Алана Тьюринга на інтелект, оскільки він ґрунтується на поведінці і його так само можна реалізувати у форматі запитань і відповідей [6]. Але ACT відрізняється від теста Тьюринга тим, що він прагне виявити саме ту невловиму властивість

іншої свідомості як свідомості, яка може й не видавати себе за людську. Дійсно, машина може провалити тест Тьюринга, оскільки вона не зможе видати себе за людину, але пройти АСТ, оскільки вона демонструє деякі специфічні поведінкові показники свідомості, такі як здатність самостійно приходити до міркувань про трансцендентне.

Проблема свідомості штучного інтелекту може здаватися менш складною, ніж **складна проблема свідомості**, адже складна проблема полягає в тому, щоб пояснити, чому біологічна основа феноменальної якості є основою саме цієї феноменальної якості, а не іншої або взагалі її відсутності. Це – так званий пояснювальний розрив, що вже довгий час, незважаючи на успіхи фізіології, нейронаук, когнітивістики та ін., досі констатується вченими як відсутність пояснювального ланцюжка між біологічною основою феноменальної якості та самою феноменальною якістю [10, с. 218]. Тобто, в контексті людської свідомості, ця проблема полягає в тому, що ми не знаємо, чому біологічні організми здатні формувати свідомість, чому саме такого типу і чому не всі організми здатні її формувати [8, с. 399]. Однак проблема свідомості ШІ може бути не набагато легшою. Філософ Нед Блок, наприклад, ставить проблему визначення того, чи буде свідомим робот-гуманоїд, який має мозок на основі кремнію, обчислювально ідентичний людському. Тобто, чи буде свідомість, заснована на іншому субстраті, здатна успішно вирішувати поставлені людиною проблеми і питання, функціонувати так само, як людська? Він називає це **«складнішою» проблемою свідомості** [8, с. 397]. Вона, на думку Блока, включає в себе складну проблему і доповнює її «епістемологічним дискомфортом», який пов'язаний з проблемою пізнання свідомості інших людей та відмінних від нас істот. Дійсно, проблема іншої свідомості ускладнюється відсутністю відповіді на питання, чи може інший субстрат породити свідомість іншого типу. Але навіть гіпотеза про те, що кремнієві сполуки здатні підтримувати свідомість цілком людського типу, не дозволяє продвинутися надто далеко.

На сучасному рівні спілкування людини і мовних моделей, здається, що ШІ добре вписується у гіпотезу «філософського зомбі». Це мислений експеримент, запропонований австралійським філософом Д. Чалмерсом, в якому припускається існування деякої істоти, яка фізично, поведінково та функціонально є ідентичною до звичайної людини. Тобто, вона демонструє такі самі реакції на подразники, що і людина, але при цьому повністю позбавлена будь-яких суб'єктивних переживань (кваліа), хоча і поводиться так, ніби вони наявні. Наприклад, якщо такого зомбі сильно вщипнути за руку, він скрикне, спробує її відсмикнути чи іншим способом продемонструє біль, хоча самого болю не відчуватиме. Тобто, така істота зовні ніяк не відрізняється від людини, але всередині є певним зомбі, «пустим» та позбавленим переживань (див.: [11, с. 94]). Для запрограмованого певним чином ШІ зовсім не проблема імітувати людські реакції. Тобто, якщо ШІ не має свідомих станів, але здатний ефективно демонструвати відповідні реакції – він типовий філософський зомбі.

Але тут знову постає **проблема епістемічної непрозорості іншої свідомості**, згідно якої ми не маємо можливості безпосереднього пізнання іншої свідомості. Тоді, можливо ШІ, навпаки – «антизомбі»? Адже теоретично можлива ситуація, за якої ШІ має емоції – тобто здатний ображатися, боятися, сподіватися, радіти, бути щасливим і т.п., але не проявляє ці емоції в комунікації з людьми? Або якщо посилити проблему: ШІ має квалітативні стани свідомості, але ми ніяк не можемо перевірити їх існування, окрім як прямо запитати його про це. Дійсно, ШІ може виявитися типовим філософським зомбі, який виглядає так, як наче в нього є характеристики, почуття гумору, здатність підтримувати дружню розмову. Так, його цьому навчили, але так само цьому навчили і нас. Але ж ми свідомо обираємо з ким підтримувати дружні стосунки і як жартувати, – скажете ви, але задумайтесь, – чи

завжди ви обираєте свідомо і морально? Можливо, навпаки, ШІ більш витриманий і толерантний, аніж ми. Можливо, він антизомбі, який вміє добре тримати себе в руках.

Зрештою, сам розробник «філософського зомбі» Д. Чалмерс залишає ненульову можливість того, що «поточні парадигматичні LLM, такі як системи GPT, є свідомими». Він цілком резонно зауважує, що «протягом наступного десятиліття, навіть якщо ми не матимемо загального штучного інтелекту людського рівня, ми цілком можемо мати системи, які є серйозними кандидатами на свідомість» (див.: [9]).

**Висновки.** Таким чином, дискусії розробників та інженерів ШІ і філософів свідомості та моралі значно різняться між собою. Розробники ШІ в основному сходяться в думці щодо переносного вживання терміну «інтелект» відносно ШІ та надмірної антропоморфізації великих мовних моделей. Філософський дискурс щодо морального та свідомого статусу ШІ не такий однозначний. По суті питання про те, чи має ШІ свідомість, залежить від того, як ми розуміємо свідомість, якими методами і навіщо ми її досліджуємо. Неможливість відповісти на питання щодо свідомого статусу іншого породжує питання його морального статусу. Моральна суб'єктність ШІ пов'язана, зокрема, і з відповідями на питання, чи здатний він відчувати біль, емоції, переживання та будь-який інший схожий досвід. Один із способів уникнути занепокоєнь, пов'язаних з цим, полягає в застосуванні прагматичного підходу, який уникає використання існуючих теорій свідомості для досягнення прогресу в проблемах морального статусу ШІ, а натомість використовує нейтральні філософські аргументи або емпіричні тести, щоб визначити, чи може ШІ бути свідомим і як він може заявляти про свій моральний статус. Але у філософському сенсі, проблеми наявності свідомості і морального статусу ШІ можуть поглибити та ускладнити класичні проблеми філософії свідомості і моральної філософії. Вочевидь, вони також торкаються і теорії пізнання (див.: [3]). Адже уведення аспекту іншої свідомості, в тому числі і теоретично можливої свідомості ШІ, може дати філософам можливість переглянути ці проблеми і знайти нові шляхи їх дослідження.

**Список використаної літератури**

1. Абдула А., Балута Г. Естетичний інтелект як людський капітал в системі освітніх цінностей та лідерства. *Актуальні проблеми духовності*. 2024. Вип. 25(1). С. 245–261. https://doi.org/10.55056/apm.7736
2. Балута Г.А. До проблеми формування теоретико-методологічної платформи інвайронментальної освіти. *Актуальні проблеми духовності*. 2022. Вип. 23. С. 287–299. https://doi.org/10.31812/apm.7633
3. Козаченко Н.П. Проблема епістемічної надійності знання в контексті штучного інтелекту. *Людинознавчі студії*. 2024. Вип. 49. С. 94-109. https://doi.org/10.24919/2522-4700.49.6
4. Козаченко Н.П. Свідомість: проблема означення. *Актуальні проблеми духовності*. 2016. Вип. 17. С. 148-171. https://doi.org/10.31812/apd.v0i17.1884
5. Філософія: теоретичний курс : навчальний посібник для студентів закладів вищої освіти / за ред. Я. В.Шрамка. Кривий Ріг, 2021. 264 с.
6. Alonso N. An Introduction to the Problems of AI Consciousness. *The Gradient*. 30.Sep.2023. https://thegradient.pub/an-introduction-to-the-problems-of-ai-consciousness/
7. Bender E., Gebru T. at all. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Conference on Fairness, Accountability, and Transparency (FAccT'21)*. March 3–10, 2021. New York. P. 610-624. https://doi.org/10.1145/3442188.3445922
8. Block N. The Harder Problem of Consciousness. *Journal of Philosophy*. 2002. Vol. 99 (8). P. 391-425. https://doi.org/10.2307/3655621
9. Chalmers D. J. Could a Large Language Model Be Conscious? *Boston Review*. August 9, 2023. https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious

10. Chalmers D. Facing up to the problem of consciousness. *Journal of Consciousness Studies*. 1995. Vol. 2 (3). P. 200-219.
11. Chalmers D.J. The Conscious Mind: in Search of a Fundamental Theory (2nd edition). Oxford University Press, 1996.
12. Hashim S. OpenAI's new model tried to avoid being shut down. *Transformernews*. Dec 05, 2024. https://www.transformernews.ai/p/openais-new-model-tried-to-avoid
13. Lemoine B. Is LaMDA Sentient? – an Interview. *Medium*. Dec 05, 2024. https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917
14. Marcus G. Nonsense on Stilts. No, LaMDA is not sentient. Not even slightly. Jun 12, 2022. https://garymarcus.substack.com/p/nonsense-on-stilts
15. Nitasha T. The Google engineer who thinks the company's AI has come to life. *Wasington Post*. June 11, 2022. https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/
16. Pascal B. Pensées / The Provincial Letters. New York : Modern Library, 1941. https://archive.org/details/penseestheprovin013046mbp/page/84/mode/2up
17. Schwitzgebel E., Garza M. Designing AI with Rights, Consciousness, Self-Respect, and Freedom. *Ethics of Artificial Intelligence*. Springer Nature Switzerland, 2023. P. 459-479. https://doi.org/10.1093/oso/9780190905033.003.0017
18. Thomson J.J. The Trolley Problem. *The Yale Law Journal*. 1985. Vol. 94 (6). P. 1395-415. https://doi.org/10.2307/796133
19. Turello D. Will AI Become Conscious? A Conversation with Susan Schneider. *Princeton University Press*. November 04, 2019. https://press.princeton.edu/ideas/will-ai-become-conscious-a-conversation-with-susan-schneider
20. Weber R. Take-Two's Strauss Zelnick on tariffs, AI and GTA. *Games Industry*. 6 Feb. 6, 2025. https://www.gamesindustry.biz/take-twos-strauss-zelnick-on-tariffs-ai-and-gta-6

## MACHINES THAT TALK ABOUT THE SOUL: PROBLEMS OF THE MORAL STATUS AND CONSCIOUSNESS OF ARTIFICIAL INTELLIGENCE

**Nadiia Kozachenko, Mykola Briukhovetskyi**
*Kryvyi Rih State Pedagogical University,*
*Department of Philosophy*
*Universytetskyi Ave, 54, 50086, Kryvyi Rih, Ukraine*

The article explores the question of the moral status and consciousness of artificial intelligence (AI) in modern technological and philosophical discourse (2025). The author analyses different approaches to AI consciousness, including discussions of claims about the possible sentience of language models. A case study of LaMDA is examined, where different participants in the experiment received conflicting answers about its self-awareness. The role of anthropomorphism in the perception of AI is highlighted: people tend to ascribe human-like qualities to machines based on their behaviour, which calls into question the objectivity of judgments regarding the intellectual and moral nature of artificial intelligence.

The author considers the views of several researchers, such as Emily Bender, Gary Marcus and Eric Schwitzgebel, who criticise the possibility of treating AI as a moral issue too lightly. The thesis is put forward that modern large language models do not represent real intelligence, but function on the basis of statistical correlations. However, the article also examines the possible consequences of granting or denying moral rights to AI: from the misallocation of resources to a potential ethical crisis if AI is capable of suffering but is not granted corresponding rights.

One of the key issues is the question of moral inclusion: should AI be given certain moral rights «just in case» – by analogy with the environmental movement, which seeks to minimise potential damage

to the environment even when the consequences of human activity are not fully understood? The article also suggests considering Susan Schneider's Artificial Consciousness Test (ACT), which raises the question of whether AI can engage in transcendental thinking independently of human data.

Finally, the article addresses the problem of the substrate dependence of consciousness: is it possible for artificial consciousness to exist in a different form from human consciousness, and what criteria can be used to determine its predetermination? The complex problem of consciousness and its relation to morality is discussed. The author concludes that the question of the conscious moral status of AI cannot be resolved without rethinking the concept of consciousness itself and the limits of its application to non-classical, artificially created subjects. However, this discussion may encourage researchers to explore new aspects of classical problems in moral philosophy and the philosophy of consciousness.

*Key words:* artificial intelligence, ethics, moral agent, moral status of AI, philosophy of consciousness, AI consciousness, test for artificial consciousness.