DOI: 10.26565/2226-0994-2025-72-12
УДК: 1:004.8

Наталя Більчук, Ілля Костров

МЕЖІ ІМІТАЦІЇ: ШТУЧНИЙ ІНТЕЛЕКТ МІЖ «ЗНАННЯ-ЩО» ТА «ЗНАННЯ-ЩОДО»

У статті здійснено філософсько-методологічний аналіз проблеми можливості машинного мислення, яка була вперше системно поставлена Аланом Тюрінгом у його класичній праці «Обчислювальні машини й інтелект». Автори звертаються до тесту Тюрінга як інструменту перевірки здатності штучних систем імітувати людську поведінку та мислення, та наданих ним критичних зауважень щодо неспроможностей штучних інтелектуальних систем. Особливу увагу приділено критиці тесту Тюрінга, здійсненій Г'юбертом Дрейфусом і Джоном Серлем. Дрейфус зазначає, що теорії штучного інтелекту базуються на гіпотетичних припущеннях, які класифікує як біологічне, психологічне, епістемологічне та онтологічне, та які не мають достатнього підґрунтя. Спираючись на феноменологічну традицію (Е. Гуссерль, М. Гайдеґґер), Дрейфус вводить розрізнення між «Знанням-Що» (Knowing-That) та «Знанням-Щодо» (Knowing-How) – як двома різними формами пізнання, лише одна з яких підлягає формалізації та алгоритмізації. «Знання-Що» розглядається як аналітичне мислення, «фактичне» знання, здатне описати, прояснити предмет, в той час як «Знання-Щодо» є контекстуальним, «фоновим» знанням, що складає основу розуміння. Водночас Серль, через мисленнєвий експеримент «китайської кімнати», аргументує неможливість наділення комп'ютера справжнім розумінням, навіть якщо він здатен імітувати діяльність (мислення, поведінку) людини. У статті порушено фундаментальні епістемологічні, онтологічні та методологічні питання, пов'язані з природою мислення, досвіду та фонових знань. Робота ґрунтується на оригінальних роботах А. Тюрінга, Г. Дрейфуса, Д. Серля з використанням феноменологічних теорій Е. Гуссерля, М. Гайдеґґера та логічного аналізу формалізації мови Л. Віттенштайна. Зроблено висновок, що інтелект не може бути зведений до суто обчислювальних операцій, а моделювання мислення потребує врахування неконцептуальних, контекстуальних і екзистенційних складників людського буття.

Ключові слова: *штучний інтелект, тест Тюрінга, «Знання-Що», «Знання-Щодо», «фонове знання», розуміння.*

У сучасному світі майже неможливо знайти людину, яка, знаючи про існування комп'ютерних технологій, не чула б про поняття штучного інтелекту (ШІ). Початок систематичних досліджень у цій галузі поклав науковий семінар, що відбувся влітку 1956 року в Дартмутському коледжі (США). Ініціаторами цього заходу виступили Джон Маккарті, Марвін Мінскі, Клод Шеннон і Натаніель Рочестер, які об'єднали провідних американських фахівців, пов'язаних із теорією управління, автоматикою, нейронними мережами, теорією ігор та проблемами моделювання інтелектуальної діяльності. Саме в межах цієї події було вперше офіційно вжито термін «Штучний інтелект» («Artificial Intelligence»), який з того часу увійшов до наукового обігу. Відтоді точаться активні дискусії як у академічному (науково-філософському) середовищі, так і в художньому (літературі, кінематографі), що охоплюють низку фундаментальних питань: на що здатні машини та комп'ютерні технології; у чому вони краще за людину; чи можливо створити машину, здатну імітувати або відтворювати діяльність людини; якими є межі людської діяльності / свідомості та штучних інтелектуальних систем; у чому вони можуть допомагати та які можливі наслідки тощо. Усі вони врешті-решт загострюють перед дослідниками проблематичність і невирішеність фундаментальних питань про сутність, структури, принципи людської діяльності, мислення та свідомості.

Засновником наукового підходу до дослідження ШІ в рамках комп'ютерних наук зазвичай вважається Алан Тюрінг – видатний науковець, який зробив вагомий внесок у

створення перших цифрових програмованих обчислювальних систем, зокрема, відомої машини Тюрінга. У своїй фундаментальній праці «Обчислювальні машини й інтелект» («Computing Machinery and Intelligence») (1950) Тюрінг порушив ключове питання: чи може комп'ютер мислити у тому сенсі, в якому ми наділяємо розумом людину? Він приходить до висновку, що головним критерієм у визначенні здатності мислити певної комп'ютерної системи є здатність імітувати мислення, так звана «гра в імітацію». Запропонований ним відомий метод перевірки, нині знаний як тест Тюрінга, полягав у тому, чи здатна людина, отримуючи відповіді на свої питання, визначити, хто їй відповідає: людина чи комп'ютер, тобто – у можливості машини імітувати інтелектуальну поведінку людини. Сам Тюрінг усвідомлював, що такі міркування будуть сприйняті критично, тому у цій самій статті навів низку власних критичних зауважень та відповідей на них, які і на сьогодні викликають гострі дискусії серед науковців і філософів. Серед відомих критиків тесту Тюрінга та дослідників перспектив ШІ у філософській царині були Г'юберт Дрейфус, Денієл Деннет, Джон Серль, Пол та Патриція Черчленди, Джон Маккарті та багато інших, які зробили значні внески у розвиток філософії штучного інтелекту та запропонували власні теорії.

Метою статті є розглянути дискусійні питання, що були висунуті Тюрінгом щодо неспроможності комп'ютерних систем відтворювати / імітувати людську мисленнєву діяльність спираючись на запропонований у роботах Г. Дрейфуса критичний аналіз видів гіпотез (біологічної, психологічної, епістемологічної, онтологічної) та введене ним розмежування «Знання-Що» та «Знання-Щодо» з використанням феноменології Е. Гуссерля, М. Гайдеггера та логічного аналізу Л. Вітгенштайна.

У наш час проблематика ШІ чи інтелектуальних систем заполонила увагу дослідників з різних галузей наукового знання та активно обговорюється на теоретико-методологічному й емпірико-практичному рівнях. Провідними дослідниками та творцями штучних інтелектуальних систем були А. Тюрінг, Ч. Строс, Д. Чалмерс, Д. Маккарті, М. Мінські, А. Ньюелл, К. Шеннон, К. Ленгтон та багато інших. Питання, на що здатні комп'ютерні системи, окрім згаданих, ставили у своїх роботах Г. Дрейфус, Д. Деннет, Дж. Серль, С. Гарнад, Пол та Патриція Черчленди, Т. Нагель тощо. Серед вітчизняних дослідників тематики філософських засад ШІ слід зазначити В. Штанько, К. Райхерт, О. Добровольську, Ю. Мельник, О. Поліщук, О. Мороз, Н. Сморжевського, а також, дисертаційні роботи А. Синиці, О. Швиркова. Проте, тематиці цієї статті, а саме, критичному аналізу неспроможностей штучних інтелектуальних систем в працях А. Тюрінга, Г. Дрейфуса, Д. Серля, дослідники філософії ШІ приділяли недостатньо уваги. Особливо це стосується Г. Дрейфуса, роботи якого майже не висвітлюються в україномовному філософському дискурсі. Одним з таких винятків є стаття доцента Одеського національного університету ім. І. І. Мечникова К. Райхерта «Г'юберт Л. Дрейфус проти Деніела К. Деннета у фільмі «Робокоп» [Райхерт, 2017] та розділ у підручнику «Вступ до філософії штучного інтелекту» Андрія Синиці [Синиця, 2023]. З огляду на вищевикладене, при написанні статті автори передусім спиралися на оригінальні роботи А. Тюрінга, Г. Дрейфуса, Д. Серля. Автори вважають інноваційним впровадження у вітчизняний філософський дискурс ідей Г. Дрейфуса щодо аналізу гіпотетичних припущень теоретичних засад штучного інтелекту та введених ним понять «Знання-Що» та «Знання-Щодо».

У своїй праці Тюрінг наголошує на необхідності постановки базового питання: «Чи здатні машини мислити?» При цьому він зазначає, що аналіз варто розпочати з уточнення значень термінів «машина», «інтелект» і «мислення». На його думку, не дивлячись на те, що ці поняття можуть бути окреслені з урахуванням їх уживання в повсякденній мові, такий підхід є доволі проблематичним і може виявитися двозначним. Покладання лише на звичне мовне вживання може призвести до хибних висновків у межах наукового дослідження, оскільки природна мова часто не забезпечує достатньої точності для філософського й технічного аналізу [Turing, 1950].

Таким чином, Тюрінг звертає увагу на одну з ключових проблем філософії, яка протягом століть привертала увагу мислителів минулого (як от: Рене Декарт у «Роздумах про метод» (1637), Дені Дідро у «Філософських думках» (1746), Жульєн Офрі де Ламетрі (1748), Альфред Айєр у монографії «Мова, істина та логіка» (1936) та інші) і яка особливо актуалізувалася в сучасному світі у зв'язку з розповсюдженням комп'ютерних технологій, штучних інтелектуальних систем тощо. Філософи, від Платона до Канта, в різний спосіб прагнули осмислити сутність і винятковість людського (природнього) мислення та пізнання; від Арістотеля до Віттенштайна вишукували можливості відтворення їх структур і форм, що безпосередньо перегукується з проблемами формалізації мови та створення штучних (оцифрованих) інтелектуальних систем. Та попри це, не дивлячись на велетенський доробок (внесок) у цій царині мислителів минулого, М. Гайдеггер у середині XX ст. зазначав, що жодна доба не була настільки необізнаною у фундаментальних питаннях людської сутності та існування, як сучасна. Позитивістко-постмодерністський проєкт відмови від метафізичних засад не вирішив, а лише посилив / поглибив невизначеність ситуації, а комп'ютерна та інформаційно-технологічна революція ще більше загострила та актуалізувала ці питання, адже, «будь-яка спроба вирішити ту або іншу конкретно-наукову проблему, пов'язану з ІІІ, безпосереднім чином виводить на необхідність прийняття того або іншого однозначного рішення щодо самих спірних питань нашого світогляду» [Швирков, 2006, с.10].

Оригінальність підходу до розв'язання проблеми машинного мислення полягає у запропонованому ним експериментальному методі. Суть початкового експерименту Тюрінга, який отримав назву «гра в імітацію», полягала в тому, що «слідчий», на підставі відповідей, отриманих у письмовій формі, має встановити, хто з двох співрозмовників є чоловіком, а хто – жінкою. «Гра» ускладнюється тим, що один з «гравців» навмисно намагається ввести «слідчого» в оману, тоді як інший, навпаки, прагне допомогти правильно ідентифікувати статі. Тюрінг пропонує модифікувати цю ситуацію, підставивши замість одного з гравців машину. Метою експерименту було визначити, чи буде «слідчий» так само часто помилятися у розпізнаванні, як і в оригінальному варіанті з двома людьми? Таким чином, Тюрінг формулює нову версію початкової проблеми – питання про здатність машин імітувати людську поведінку та мислення [Turing, 1950]. Сьогодні ця концепція має практичне втілення: тест вважається успішно пройденим, якщо незалежні оцінювачі не можуть з достовірністю розрізнити відповіді машини від відповідей людини. Суттєво, що критерієм успіху в цьому випадку є не коректність чи точність відповідей, а їхня схожість з тими, які дав би реальний людський співрозмовник. Отже, тест Тюрінга слугує емпіричним індикатором здатності штучної системи до симуляції людської комунікативної поведінки.

Започаткована Тюрінгом проблема створення цифрової системи здатної імітувати інтелектуальну поведінку людини знайшла своє практичне втілення як у різноманітних модифікаціях самого тесту Тюрінга («ієрархія тестів Тюрінга» Стівена Гарнада), так і в створенні нових програм («Еліза» (англ. ELISA) Джозефа Вейценбаума (1966), «PARRY» Кеннета Колба (1972) тощо), а з 1990 року було впроваджено щорічну премію Льобнера за створення програми, що найліпше імітує людську мисленнєву діяльність.

Тест Тюрінга відіграв ключову роль у формуванні концептуальних засад ІІІ та став важливою віхою в історії його розвитку. Згодом запропонована модель зазнала значної критики з боку різних науковців, однак, попри це, вона й надалі зберігає значний теоретичний і практичний вплив. Аргументи проти тесту Тюрінга стали підґрунтям для осмислення нових підходів до розуміння природи інтелекту, а також допомогли окреслити межі застосування машинного імітування людської поведінки у сфері ІІІ. Не дивлячись на те, що тесту Тюрінга вже сімдесят п'ять років, його критичний аналіз залишається актуальним і на сьогодні.

Першим, хто почав дискусію щодо тесту, був сам Тюрінг у тій самій статті «Обчислювальні машини й інтелект». Він висунув низку критичних думок та аргументів, що відрізняються від його власних та навів свої відповіді на них [Turing, 1950]. Серед них:

- «теологічне заперечення», тобто, питання про метафізичну сутність людини, як то, «душа»;

- «заперечення «голів у піску»» про небезпечні фактори та ризики, до яких може призвести використання машин, комп'ютерних технологій та інтелектуальних систем (на що сам Тюрінг, іронічно зауважує: так, ризик є, проте науково-технічний прогрес не зупинити);

- «математичне заперечення» про принципову неповноту та обмеженість формалізації (теореми Геделя про неповноту формалізації), проте, і картезіанське гасло про безмежність людського розуму не доведене;

- «аргумент від свідомості», «від різноманітних неспроможностей», «від неформальної поведінки» стосуються неспроможності машин та інтелектуальних систем відчувати, переживати, оцінювати, нести відповідальність, тобто, поводитися «не за програмою», а на підставі інтуїтивних і емоційних імпульсів;

- «аргумент від неперервної нервової системи», тобто, що нервова система людини, на відміну від машин, не має дискретного характеру.

Зазначені Тюрінгом критичні зауваження започаткували цілу низку дискусій: від суто технічних (створення комп'ютерних машин та програм) до філософсько-антропологічних, в яких знов і знов повертаються до питання, наскільки недостатньо ми знаємо про сутність, природу, принципи та структури власного, людського мислення.

Одним з ранніх аналітиків припущень щодо можливості проведення чітких аналогій між роботою комп'ютерів і діяльністю людського мозку, був американський філософ Г'юберт Дрейфус, який закінчив Гарвардський університет, де отримав докторський ступень, працював у Массачусетському технологічному інституті, та вважається одним з провідних дослідників феноменології Е. Гуссерля, М. Гайдеґґера, філософії М. Фуко та інших. Він ретельно аналізує «аргументи» Тюрінга щодо «неспроможностей» штучних інтелектуальних систем та, спираючись на філософські теорії щодо сутності пізнання та мислення Платона, Г. Ляйбніца, Д. Г'юма, І. Канта та інших, приходить висновку про недосяжність уявлень про спроможність штучних інтелектуальних систем до відтворення людської поведінки, критикуючи досягнення своїх сучасників, зокрема Аллена Ньюелла, Герберта Саймона та інших.

У своїй книзі «Що не можуть комп'ютери» («What Computers Can`t Do», 1972) та статті «Алхімія і штучний інтелект» («Alchemy and AI», 1965) Дрейфус звертає увагу на наявність у людини несвідомих (інтуїтивних) процесів, які, через свою «неусвідомленість», не можуть бути відображені у формальний спосіб, а тому, не можуть бути імітовані. Він виокремлює чотири ключові науково-філософські припущення: біологічне, психологічне, епістемологічне та онтологічне, які, на його думку, стали підґрунтям упевненості дослідників ШІ в тому, що людський інтелект зводиться до маніпуляції символами. Дрейфус наголошує, що в кожному з цих випадків ідеться радше про *гіпотези*, які серед дослідників ШІ ілюзорно приймаються за аксіоми, тобто апріорні істини, що гарантують успішність розробок реальних прототипів. Він підкреслює необхідність критичного переосмислення цих положень через призму емпіричних результатів і філософського аналізу.

Згідно з біологічною гіпотезою [Dreyfus, 1972, p. 71-74], припускається, що людський мозок обробляє інформацію дискретними операціями за допомогою певного біологічного еквівалента «вмикання / вимикання». На початкових етапах розвитку нейронауки вчені звернули увагу на те, що нейрони людського мозку функціонують за принципом «усе або нічого», генеруючи імпульси, які мають чітко виражений дискретний характер. На підтримку цього припущення, американські нейропсихологи та нейрофізіологи Волтер Піттс і Воррен Маккалох створили математичну модель нейрона.

Ця математична модель ґрунтувалася на гіпотезі про те, що діяльність нейронів відбувається за аналогією з процесами між логічними елементами булевої алгебри, що дозволяло уявити нейронну активність мозку як сукупність логічних операцій, що потенційно може бути відтвореною у вигляді електронних схем. З поширенням цифрових комп'ютерів у 1950-х роках ця ідея трансформувалась у припущення, згідно з яким людський мозок є матеріальною схематичною системою, що оперує бінарними знаками – нулями та одиницями. Проте, Дрейфус ставить під сумнів біологічне припущення, що діяльність людського мозку носить виключно дискретний характер. Спираючись на результати нейрофізіологічних досліджень, зокрема, професора Массачусетського технологічного інституту Волтера А. Розенбліта, він вказував на наявність аналогових компонентів у часових і динамічних характеристиках нейронної активності. Вже на початку 1970-х років біологічна гіпотеза значною мірою втратила свою актуальність, тому, позиція Дрейфуса залишилась без заперечень з боку наукової спільноти. Отже, за висновками Дрейфуса, наявність аналогової, а не тільки дискретної, діяльності нейронів людського мозку спростовує аксіоматичність біологічного припущення.

Психологічна гіпотеза [Dreyfus, 1972, p. 75-100] заснована на тезі, що розум можна розглядати як пристрій, що поєднує окремі фрагменти інформації відповідно до формальних правил, тобто, що мислення є виключно обробкою інформації за допомогою формальних методів (аналіз, синтез, узагальнення тощо). Цей розділ Дрейфус розпочинає з питання-зауваження: «Чи працює мозок як цифровий комп'ютер – суто емпіричне питання, яке має вирішити нейрофізіологія; проте, не можна дати простої відповіді до спорідненого, але зовсім іншого питання: чи *функціонує* (виділено нами) розум як цифровий комп'ютер, тобто чи виправдано використовувати комп'ютерну модель у психології» [Dreyfus, 1972, p. 75]. Дрейфус наголошує на тому, що значна частина нашого знання про світ не зводиться до звичайного оперування певним, наявним знанням, а радше ґрунтується на складних установках, інтенціях і схильностях, які спонукають нас *інтерпретувати* ситуацію тим чи іншим чином. Він звертається до філософської традиції та пише: «Кант безпосередньо аналізував весь досвід, навіть сприйняття, з точки зору правил, а уявлення про те, що знання включає набір явних інструкцій, є ще давнішим…» [Dreyfus, 1972, p. 88]. Ще Платон у своїх діалогах, зокрема «Евтифрон» або «Про благочестя», порушував питання щодо аналізу структур нашого мислення у такій спосіб: не зважаючи на те, що наше мислення чи поведінка має раціональну структуру та складається з правил та «прагматичних накладань», втім, чи потребується їх усвідомлення та розуміння, щоб мислити / поводитися розумно? Надалі Дрейфус розглядає теорію «емпіричного суб'єкта» Д. Г'юма та концепт «трансцендентального суб'єкта» І. Канта як окремі гіпотези, з яких робить загальний (спільний для обох) висновок. Він стверджує, що навіть коли ми використовуємо певно визначенні, звичайні, повсякденні поняття та судження, вони функціонують у не завжди усвідомленому для нас епістемологічному контексті, у царині так званого «*фонового знання*», яке не оформлюється в щось конкретне (чітко визначений образ), проте є необхідною складовою для мислення. В наступній своїй роботі для виокремлення цих двох видів розумової діяльності, він буде використовувати терміни «Знання-Що» та «Знання-Щодо». Дрейфус був прихильником та провідним дослідником феноменології Гуссерля та Гайдеґґера, а тому, це «*фонове знання*» є радше *феноменологічною*, аніж простою редукцією. Людське знання не зводиться до простої сукупності фактів, як цей термін розуміє Віттенштайн, тобто, «наявний стан речей», а існує, передусім, як *контекстуальний*, образно-практичний досвід. Таким чином, Дрейфус спростовує психологічне припущення, адже, розумову діяльність не можна розглядати як звичайне обчислювання чи оперування фактами; на нього впливає багато неусвідомлених людиною чинників.

Епістемологічна гіпотеза [Dreyfus, 1972, p. 101-117] базується на припущені, що будь яке знання можна формалізувати. Дрейфус зазначає, що хоча і психологічне, і епістемологічне припущення виходять з платонівського уявлення про мислення як

формалізоване, проте, вони мають важливу різницю: психологічне припущення передбачає, що правила, якими користуються при формалізації діяльності (мислення, поведінки) людини – це ті самі правила, що скеровують саму діяльність; у той час як ті, хто робить епістемологічне припущення лише стверджують, що будь яка несвавільна поведінка може бути формалізована згідно з деякими правилами, і що ці правила, якими б вони не були, можуть потім використовуватися комп'ютером для відтворення поведінки. Навіть якщо психологічне припущення, тобто уявлення про те, що людський інтелект функціонує як система обробки символів, буде визнане хибним, дослідники ШІ (зокрема, Джон Маккарті, один із засновників галузі) можуть наполягати на тому, що формальна система, яка оперує символами, у принципі здатна репрезентувати будь-яке знання, незалежно від того, як саме це знання представлено в людському розумінні.

Онтологічне припущення [Dreyfus, 1972, p. 118-136] засноване на положенні, що уявлення про світ складаються з незалежних феноменів, які можна репрезентувати незалежними символами. Дрейфус виокремлює це як більш тонке, але не менш фундаментальне припущення, на яке спираються багато дослідників ШІ, а також футурологів та авторів наукової фантастики. Йдеться про переконання, що не існує обмежень для формалізованого, наукового пізнання, оскільки будь яке явище у Всесвіті, за цим припущенням, може бути описане (прояснене) мовою символів або в межах наукових теорій. Це припущення передбачає, що усе, що існує, може бути зведене до об'єктів, властивостей об'єктів, класів об'єктів, відношень між об'єктами – тобто до таких структур, які можуть бути виражені за допомогою логіки, мови або математики. Оскільки мова йде про структуру буття, Дрейфус класифікує це як онтологічне припущення. Якщо ж виявиться, що це припущення є хибним, тоді постає питання не лише про межі нашого пізнання, але й про обмеження в потенціалі машинного інтелекту допомагати людині в дослідженні і розумінні світу.

Критичні зауваження Дрейфуса, так само як і Тюрінга, щодо перспектив розвитку штучних інтелектуальних систем націлені не стільки на заперечення можливостей їх розвитку, скільки на порушення споконвічних дискусійних питань щодо світоустрою, структур і принципів людського пізнання та мислення. Розглядаючи різні теорії, від таких, які стверджують, що наше пізнання є віддзеркаленням світу, до тих, хто вважає, що саме наше мислення і конструює (створює) уявлення про світ, Дрейфус не відстоює жодну з них, а лише наголошує, що це не аксіоми, а гіпотези. Проте, розвиток наукового знання неможливий поза розуміння та усвідомлення ймовірності (а не догматичності) будь якого припущення. Таке розуміння поглиблює проблематику та стимулює пошук.

У своїй праці «Розум над машиною» («Mind Over Machine», 1986) написаній у період як найбільшого оптимістичного захоплення, так і розчарувань в галузі досягнень штучного інтелекту, Дрейфус продовжує працювати над питаннями, що були порушені ним у попередній роботі, зокрема, при розгляді психологічного та епістемологічного зауважень щодо контекстуальності, «*фоновості*» людського знання. Він ретельно аналізує розбіжності між людською аналітичністю та програмами, які претендували на її моделювання. Ця робота стала розвитком і поглибленням ідей, висловлених у книзі «Що не можуть комп'ютери», де він критикував підхід до досліджень ШІ, який отримав назву «когнітивне моделювання» й був представлений у 1960-х роках такими постатями, як Аллен Ньюелл і Герберт Саймон.

Дрейфус стверджував, що розв'язання задач людиною та її професійна майстерність ґрунтуються не на формальному аналізі усіх можливих комбінацій у пошуках потрібного рішення, а на тлі *контекстуального* розуміння ситуації, тобто, на підставі певного інтуїтивного відчуття, що є суттєвим або доречним у даному випадку. Він описував цю відмінність через розрізнення між «*Знанням-Що*» (Knowing-That) і «*Знанням-Щодо*» (Knowing-How), спираючись на багатозначний термін Гайдеггера «*Vorhandenheit*» у визначенні «*наявності*» (*present-at-hand*) та «*спідручності*» (*ready-to-hand*) [Dreyfus, 1986, p. 22].

«*Знання-Що*» (Knowing-That) – це знання про те, *що* собою являє предмет чи явище: які має властивості, ознаки, якості тощо. За визначенням І. Канта, вони складають основу аналітичних суджень; це наша свідома, покрокова здатність до розв'язання проблем шляхом розкладання на прості або визначення властивостей та ознак, що притаманні самому предмету / явищу. Ми використовуємо ці навички, коли стикаємось із складними завданнями, які вимагають зупинки, відступу та ретельного перегляду ідей одна за одною. Це, за визначенням Дрейфуса, *проста формальна розумова діяльність*, яка нагадує комбінування інформації за певним алгоритмом. Знання оформлюється у дуже чіткі та прості судження, перетворюється на символи та формули, позбавлені контексту, якими ми маніпулюємо за допомогою логіки та мови. Як зазначає Вітгенштайн, «усе, що взагалі можна подумати, можна подумати ясно» [Вітгенштайн, 1995: 4.116]. Це, так би мовити, існуючий стан речей, його «фактичність» (за Вітгенштайном) або «наявність» (за Гайдеґґером). Такі навички, як продемонстрували через свої психологічні експерименти Аллен Ньюелл і Герберт Саймон, можуть бути формалізовані та відтворені за допомогою комп'ютерних програм. Дрейфус погоджується, що штучні інтелектуальні системи здатні адекватно імітувати «Знання-Що»; це репрезентація функціонування дискретності людського мислення.

Натомість «*Знання-Щодо*» – це, передусім, *асоціативна розумова діяльність*; спосіб, яким ми зазвичай взаємодіємо з навколишнім світом. Таке знання не потребує чіткого вербального чи контекстуального вираження, може бути не оформленим у певні судження. «*Знання-Щодо*» – це звичні, повсякденні «*образи*», цілісні уявлення щодо того, що являє собою предмет чи явище для нас, тобто, «*спідручність*» (за Гайдеґґером). Ми діємо без свідомого символічного міркування, коли, наприклад, впізнаємо знайоме обличчя, виконуємо повсякденні справи або інтуїтивно знаходимо правильне слово чи відповідь. Ми ніби «переходимо» до відповідної реакції, не розглядаючи альтернативи, «схоплюємо» предмет у його контекстуальній, «*фоновій*» приналежності. Це, за словами Дрейфуса, є суттю та особливістю людського сприйняття: здатність усвідомлювати чим є предмет ще до того, як його пізнаєш, не коригуючись сформованими знаннями чи правилами, тобто, *розуміти*.

Отже, запропоноване Дрейфусом розподілення інтелектуальної діяльності людини на «Знання-Що» та «Знання-Щодо» виводить дискусії щодо можливостей / неспроможностей ШІ на новий рівень: виокремлення «знання-*пояснення* (прояснення) наявного стану речей від «знання-*розуміння*».

Згідно з Дрейфусом, людське розуміння речей чи відчуття ситуації ґрунтується не тільки на дискретному «Знанні-Що», але і на аналоговому «Знанні-Щодо», на несвідомих образах, інтуїціях, ставленнях, на формування яких впливають багато чинників, які неможливо однозначно визначити, а отже, і формалізувати. Цей «контекст» або «фонове знання» є, за термінологією Канта, «*трансцендентальними*» умовами будь якого знання, та перегукується з концепцією *Dasein* у Гайдеґґера; це форма знання-розуміння, яке не зберігається в нашому мисленні з сукупністю символів, а існує інтуїтивно, певним чином. Це «*фонове знання*» впливає на те, що ми щось помічаємо, а чогось – ні, чогось очікуємо, а якісь можливості не розглядаємо: ми проводимо розрізнення між тим, що є суттєвим, і тим, що не є таким у даній ситуації. Дрейфус не вірить, що програми ШІ, як вони були реалізовані в 1970-80-х роках, можуть відтворити це «*фонове знання*» або здійснювати той тип швидкого розв'язання проблем, який воно дозволяє. Він стверджував, що наше несвідоме знання ніколи не можна буде відтворити символічно, а отже, імітувати.

На проблему *розуміння* як фундаментальну у процесі прояснення на що неспроможні штучні інтелектуальні системи, звернув увагу Джон Роджерс Серль. У своїй праці «Чи є розум комп'ютерною програмою?» («Is the brain's mind a computer program?») він змістив акценти в постановці проблеми, порушуючи питання не лише про здатність машини імітувати мислення, тобто «Знання-Що», а й про можливість імітувати розуміння, «Знання-Щодо». Тому, акцент переноситься з моделювання когнітивної поведінки на питання імітації свідомості. Подібну проблему, хоча й у більш ранній формі, порушував

Дені Дідро, який у своїх філософських роздумах зазначав, що якщо папуга навчиться відповідати на запитання без вагань, її можна буде вважати розумною. Серль, своєю чергою, ставить принципове запитання: чи можна вважати свідомими відповіді на питання, які сформовані без розуміння самих питань? Цим він підкреслює відмінність між зовнішньою поведінкою, що здається розумною, та внутрішнім усвідомленням, яке є визначальною рисою людського мислення.

Запропонований Серлем мисленнєвий експеримент, що отримав назву «китайська кімната», став одним із найвідоміших аргументів проти функціоналістського підходу до ШІ. У своїй концепції Серль пропонує уявити ситуацію, в якій людина, яка не володіє китайською мовою, перебуває в кімнаті, заповненій кошиками з китайськими ієрогліфами. Цій особі надається докладна інструкція мовою, яку вона розуміє, яка дозволяє здійснювати маніпуляції з символами на основі їхньої форми, а не значення. Наприклад, інструкція може містити такі дії: «Візьміть символ N з кошика №1 і розмістить його поруч із символом M з кошика №2». Зовнішні спостерігачі, які володіють китайською, надсилають у кімнату послідовності символів, а людина, що виконує інструкції, формує у відповідь нові послідовності символів, не усвідомлюючи їхнього змісту [Searle, 1990]. Серль підкреслює, що за описаною схемою можливо генерувати відповіді на запитання, які не відрізняються від відповідей носія китайської мови, що зовні може створювати враження наявності справжнього знання мови. Така система, за його словами, здатна «успішно пройти» тест Тюрінга. Утім Серль наголошує: «Незважаючи на це, я все одно абсолютно не знаю китайської… Як і комп'ютер, я лише маніпулюю символами, не надаючи їм жодного значення» [Searle, 1990]. Цей приклад дає змогу йому сформулювати важливу тезу: тест Тюрінга, або ж «гра в імітацію», не є вичерпним критерієм визначення наявності інтелектуальних якостей. Він дозволяє лише встановити здатність до поведінкової симуляції – відтворення зовнішньої мовленнєвої чи когнітивної діяльності, яку, однак, можливо реалізувати як на основі осмисленої інтелектуальної активності, так і без справжнього розуміння чи усвідомлення.

Отже, Джон Серль акцентує увагу на тому, що процес розуміння для людини виходить за межі суто формального оперування символами. Натомість цифрові обчислювальні системи за своєю природою здатні лише до формального маніпулювання знаками без будь-якого доступу до їх змістовного наповнення. Цей підхід ліг в основу концепції так званого «слабкого інтелекту», згідно з якою штучна система може моделювати інтелектуальну поведінку, не володіючи при цьому справжнім розумінням чи свідомістю. І тест Тюрінга, і мисленнєвий експеримент Серля відкрили широкий спектр ключових філософських і методологічних питань, що стосуються меж можливостей ШІ.

Проведений у статті аналіз свідчить, що питання про здатність машин до мислення, окреслене ще Тюрінгом у середині XX ст., не зводиться лише до практичної реалізації, а радше загострює фундаментальні методологічні, епістемологічні та онтологічні проблеми, які стосуються самого розуміння людського мислення, свідомості та знання. Зокрема, критичний аналіз Дрейфуса демонструє неспроможність редукції людського мислення до формальних логіко-математичних структур, у яких знання розуміється виключно як маніпуляція символами. Розмежування ним понять «Знання-Що» та «Знання-Щодо» дає змогу глибше зрозуміти аспекти людської розумової діяльності та осмислити можливості сучасних інтелектуальних систем, які хоча й здатні імітувати раціональну поведінку, однак не спроможні відтворити інтуїтивні, контекстуальні й «фонові» аспекти людського пізнання. Водночас, залучення феноменологічної редукції (Гуссерль, Гайдеггер) дозволяє акцентувати увагу на «фоновому» знанні, яке, не може бути конкретизоване та формалізоване, проте складає трансцендентальні передумови розумової діяльності та стає підґрунтям його розуміння. Аналогічно, експеримент Серля з «китайською кімнатою» демонструє принципову відмінність між маніпулюванням знаками («Знання-Що», поясненням) та наявністю свідомого розуміння («Знання-Щодо», розумінням), наголошуючи, що останнє залишається поза межами формального обчислення. Таким

чином, філософський аналіз свідчить, що штучні інтелектуальні системи, як би досконало вони не імітували поведінкові аспекти мислення, залишаються неспроможними до відтворення його екзистенційно-феноменологічного виміру. Людське мислення – це не лише обчислення або репрезентація фактів, а насамперед спосіб бути-в-світі, вкорінений у довільному, але контекстуально насиченому досвіді, який не підлягає повній формалізації.

Отже, перспективи дослідження ШІ мають виходити за межі вузького інструментального підходу й включати розробку філософської онтології свідомості, здатної врахувати не лише структурні, а й смислові, цілісні та трансцендентальні засади людського пізнання.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

Антологія сучасної філософії науки, або усмішка ASIMO / наук. ред. В. Мельник, А. Синиця. Львів: ЛНУ імені Івана Франка, 2017. 568 с

Вітгенштайн Л. *Tractatus Logico-Philosophicus; Філософські дослідження* / пер. Євгена Поповича. Київ: Основи, 1995.

Добровольська О. В., Штанько В. І. Філософський аналіз еволюції штучного інтелекту. *Studies in history and philosophy of science and technology*. 2019. Vol. 28. № 1. С. 10-19. - Режим доступу: http://nbuv.gov.ua/UJRN/studhphst_2019_28_1_4.

Райхерт К. В. Хьюберт Л. Дрейфус против Дэниэла К. Деннетта в кинофильме *Robocop* (2014). *Філософія та політологія в контексті сучасної культури*. 2017. Вип. 3 (18). С. 56–64.

Синиця А. *Вступ до філософії штучного інтелекту : підручник*. Львів : ЛНУ імені Івана Франка, 2023. 292 с.

Швирков О. І. *Проблема штучного інтелекту і людиновимірність штучних інтелектуальних систем: дисертація на здобуття ступеня кандидата наук*. Житомир: Житомирський державний університет імені Івана Франка, 2006. Державний реєстраційний номер 0406U004726

Dreyfus H. L. *What Computers Can't Do: A Critique of Artificial Reason*. New York: Harper & Row, 1972. 231 p.

Dreyfus H. L., Dreyfus S. E. *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*. New York: Free Press, 1986. 240 p.

Searle, J. R. Is the brain's mind a computer program? *Scientific American*. 1990. 262. P. 26-31. https://www.cs.princeton.edu/courses/archive/spr06/cos116/Is_The_Brains_Mind_A_Computer_Program.pdf

Turing A. M. Computing Machinery and Intelligence. *Mind*. 1950. 49. P. 433–460 https://redirect.cs.umbc.edu/courses/471/papers/turing.pdf.

Більчук Наталя Леонідівна
кандидат філософських наук, доцент
кафедра філософії та суспільних наук
Національний аерокосмічний університет «Харківський авіаційний інститут»
61070. Харків, вул. В. Манька, 17, Україна
E-mail: n.bilchuk@khai.edu
ORCID: https://orcid.org/0000-0001-5505-5533

Костров Ілля Андрійович
аспірант кафедри філософії та суспільних наук
Національний аерокосмічний університет «Харківський авіаційний інститут»
61070. Харків, вул. В. Манька, 17, Україна
E-mail: i.kostrov@khai.edu
ORCID: https://orcid.org/0009-0009-4869-8392

Внесок авторів: всі автори зробили рівний внесок у цю роботу

Конфлікт інтересів: автори повідомляють про відсутність конфлікту інтересів

Стаття надійшла до редакції: 01.03.2025 Схвалено до друку: 28.04.2025

BOUNDARIES OF IMITATION: ARTIFICIAL INTELLIGENCE AMONG "KNOWING-THAT" AND "KNOWING-HOW"

Bilchuk Natalia L.
Associate Professor
Department of Philosophy and Social Sciences
National Aerospace University "Kharkiv Aviation Institute"
17, V. Manka st. Kharkov, 88000, Ukraine
E-mail: n.bilchuk@khai.edu
ORCID: https://orcid.org/0000-0001-5505-5533

Kostrov Illia A.
Student of postgraduate
Department of Philosophy and Social Sciences
National Aerospace University "Kharkiv Aviation Institute"
17, V. Manka st. Kharkov, 88000, Ukraine
E-mail: i.kostrov@khai.edu
ORCID: https://orcid.org/0009-0009-4869-8392

ABSTRACT

This article presents a philosophical and methodological analysis of the problem of the possibility of machine thinking, which was first systematically posed by Alan Turing in his classic work *Computing Machinery and Intelligence*. The authors refer to the Turing Test as a tool for evaluating the ability of artificial systems to imitate human behavior and thinking, while also engaging with Turing's own critical reflections on the inherent limitations of artificial intelligence. Special attention is devoted to the critiques advanced by Hubert Dreyfus and John Searle. Dreyfus contends that the foundational assumptions of artificial intelligence – classified by him as biological, psychological, epistemological, and ontological – are insufficiently substantiated. Drawing on the phenomenological tradition (E. Husserl, M. Heidegger), Dreyfus introduces a distinction between *"Knowledge-That"* and *"Knowledge-How"* as two different forms of knowledge, only one of which is subject to formalization and algorithmization. *Knowing-that* is conceptualized as analytical thinking, "factual" knowledge capable of describing and clarifying an object, whereas *knowing-how* is contextual or "background" knowledge that forms the basis of understanding. At the same time, John Searle, through his thought experiment of the "Chinese Room," argues that a computer cannot possess genuine understanding even if it can imitate human behavior and thought processes. The article addresses fundamental epistemological, ontological, and methodological issues related to the nature of thinking, experience and background knowledge. The work draws on primary sources by Turing, Dreyfus, and Searle, while also incorporating phenomenological insights from Husserl and Heidegger, as well as Wittgenstein's logical analysis of language formalization. It concludes that intelligence cannot be reduced to purely computational operations and that model of human thinking requires consideration of non-conceptual, contextual, and existential components of human existence.

***Keywords**: artificial intelligence, Turing test, "Knowledge-That", "Knowledge-How", "background knowledge", understanding.*

REFERENCES

Anthology of contemporary philosophy of science, or ASIMO smile (2017). / sc. ed. V. Melnyk, A. Synytsia. Lviv: Ivan Franko National University of Lviv. (In Ukrainian).

Wittgenstein, L. (1995). *Tractatus Logico-Philosophicus; Philosophical studies* (Ye. Popovych, Trans.). Kyiv: Osnovy. (In Ukrainian).

Dobrovolska, O. V., & Shtanko, V. I. (2019). Philosophical analysis of the evolution of artificial intelligence. *Studies in history and philosophy of science and technology*, 28, 1, 10-19. URL: http://nbuv.gov.ua/UJRN/studhphst_2019_28_1_4 (In Ukrainian).

Rayhert K. W. (2017). Hubert L. Dreyfus vs. Daniel C. Dennett in Robocop (2014). *Philosophy and political science in the context of modern culture*, 3(18), 56–64.

Synytsia, A. S. (2023). *Introduction to the philosophy of artificial intelligence: textbook*. Lviv: LNU named after I. Franko. (In Ukrainian).

Shvyrkov, O. I. (2006). *The problem of artificial intelligence and human measurements of artificial intellectual systems*: dissertation for the degree of Candidate of Sciences. Zhytomyr: Ivan Franko Zhytomyr State University. (In Ukrainian).

Dreyfus, H. L. (1972). *What Computers Can't Do: A Critique of Artificial Reason*. New York: Harper & Row.

Dreyfus, H. L., & Dreyfus, S. E. (1986). *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*. New York: Free Press.

Searle, J. R. (1990). Is the brain's mind a computer program? *Scientific American, 262*(1), 26–31. https://www.cs.princeton.edu/courses/archive/spr06/cos116/Is_The_Brains_Mind_A_Computer_Program.pdf

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460. https://redirect.cs.umbc.edu/courses/471/papers/turing.pdf.

Authors Contribution: All authors have contributed equally to this work
Conflict of Interest: The authors declare no conflict of interest

Article arrived: 01.03.2025 Accepted: 28.04.2025

Як цитувати: Більчук, Н., & Костров, І. (2025). МЕЖІ ІМІТАЦІЇ: ШТУЧНИЙ ІНТЕЛЕКТ МІЖ «ЗНАННЯ-ЩО» ТА «ЗНАННЯ-ЩОДО». *Вісник Харківського національного університету імені В. Н. Каразіна. Серія «Філософія. Філософські перипетії»*, (72), 128-138. https://doi.org/10.26565/2226-0994-2025-72-12
In cites: Bilchuk, N., & Kostrov, I. (2025). BOUNDARIES OF IMITATION: ARTIFICIAL INTELLIGENCE AMONG "KNOWING-THAT" AND "KNOWING-HOW". *The Journal of V. N. Karazin Kharkiv National University, Series Philosophy. Philosophical Peripeteias*, (72), 128-138. https://doi.org/10.26565/2226-0994-2025-72-12 [In Ukrainian]