УДК 165.8:811.93:004.8
DOI: 10.18523/2617-1678.2024.14.31-39

Маєвський О. Л.
https://orcid.org/0000-0001-6063-6033

МЕХАНІКА МІРКУВАННЯ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ: ФІЛОСОФСЬКИЙ АНАЛІЗ

У статті показано і філософськи пояснено природу, механіку міркування і фундаментальні засади епістемічної обмеженості сучасних діалогових великих мовних моделей на основі архітектури Transformer. Великі мовні моделі представлено як функціоналістський механістичний проєкт статистичного моделювання мови і мовлення як моделі знання як смислової моделі дійсності — 1) модель 2) моделі 3) моделі дійсності. Показано, що через цю значну дистанцію опосередкування дійсності внутрішньомодельні зв'язки втрачають свій фактологічний потенціал. Також продемонстровано, що великі мовні моделі є продуктом машинного навчання певній мовній поведінці з метою і цінностями, кардинально відмінними від мети і цінності людського пізнання. Їхньою метою є принцип задоволення оператора функції винагороди шляхом обману за будь-яку встановлену ціну і будь-якими наявними засобами на етапі навчання. Останнє не дає змоги користувачам моделі бути упевненими в доцільності моделі людським очікуванням і в безпечності будь-яких її міркувань на етапі її експлуатації. Крім цього обґрунтовано, що фундаментальні обмеження самої здатності до міркування у цих моделей є не лише фактологічними, а й алгоритмічними та онтологічними: ці моделі є обмеженими лінійними скінченними автоматами без тіла у дійсності, загалом позбавленими інших джерел знань і досвіду, крім синтаксису і контексту. Через це модель на рівні конструкції вдається до грубої імітації рефлексії через нечітку авторегресію, якою, фактично, відображається результат пошуку по корпусу текстів, кожен з яких потенційно міг бути створений і самим автором запиту до моделі. З огляду на зазначене власне епістемічна цінність продуктів великої мовної моделі визначається передусім їхньою пошуковою цінністю для користувача і обмежується проблематичністю їхньої атрибуції, валідації, а також необхідністю зовнішньої відповідальної верифікації й оцінки самим користувачем.

Ключові слова: гносеологія, епістемологія, механіцизм, фізикалізм, функціоналізм, філософська логіка, філософія мови, філософія техніки, філософія штучного інтелекту, велика мовна модель.

Формування теоретичних основ для логічного оброблення інформації засобами обчислювальних машин збурила у філософії XX ст. широку дискусію та битву аргументів щодо відношення розуму до тіла (mind-body problem) — від «розриву у поясненні» (the explanatory gap) Джозефа Левіна й до «важкої проблеми свідомості» (the hard problem of consciousness) Девіда Джона Чалмерса, від «китайської кімнати» (Chinese room argument) й «першоосібного» біологічного натуралізму (first-person biological naturalism) Джона Роджерса Серля й до трансценденталістського по суті «прагматичного реалізму» *пізнього* Гіларі Уайтголла Патнема (pragmatic realism) та багатьох інших [1]. Однак відповідні атаки на можливість *ефективної* фізикалістської редукції усіх основних «ментальних» феноменів розуму (відчуття, сприйняття, пам'яті, міркування, оцінок і емоцій, мотивацій, пов'язаної з ними поведінки, навчання, адаптації і, нарешті, *свідомості*) до так чи інакше ідентифікованих фізичних станів так і *не переконали* розробників науково свідомого філософського функціоналізму в тому, що проблема такої редукції — не «вітряк Дон Кіхота».

Натомість у панівному науковому світогляді зрештою укорінилися функціоналістські ідеї більш *раннього* Г. Патнема (machine-state functionalism), Дональда Герберта Девідсона (the mental non-reductive 'supervenience' on the physical, anomalous monism), Деніела Клемента Деннета ('homuncular' or 'supervenient' functionalism), Девіда Келлога Льюїса (conceptual or analytic functionalism, implicit definition of terms by their theories) та ін., які уникають самої постановки проблеми ефективної редукції як позбавленої

[1] Детальніший огляд цієї проблематики див., наприклад, у (Tye, 2021).

практичного сенсу псевдопроблеми [2]. Адже виявлення деяких елементів, структури, загальних принципів улаштування якогось складного механізму не обов'язково веде до технічної чи концептуальної здатності дати його повне пояснення (яким і була б ефективна редукція). Але можна обґрунтовано припускати і потім перевіряти (верифікувати) певні гіпотези щодо наявності *деякого* зв'язку взагалі (тієї самої «супервентності») між функціонуванням елементів механізму і ефективно непояснюваними для нас («емерджентними»), але виявленими нами лише емпірично феноменами поведінки деякого вищого порядку, характерними для елементів цього механізму як організованої групи в цілому.

Для розв'язання наукових завдань у рамках цієї парадигми моделювання феноменів розуму цілком можна обмежуватись фундаментальними і найбільш абстрактними когнітивними функціями — такими як *міркування* (reasoning), *досвід* (awareness як фактичність взагалі), *дія* (action). А власне функціонування такої моделі може бути представлено як деякий по своїй суті спрямований *механічний (детерміністичний) процес*, пояснення і передбачення якого здебільшого є доступними лише у формі опису *стохастичного процесу* (у термінах імовірностей) на тих рівнях його організації (серед яких можуть бути як «нижчі», так і «вищі»), щодо яких у нас на цей момент немає можливості ефективного пояснення через редукцію до інших рівнів. У такому разі ми маємо справу з *метафізичним механіцизмом* в основі нередуктивної функціоналістської гносеології, яка принципово *не претендує на можливість* (і не визнає необхідності) вибудовування абсолютистських ланцюгів строгого, навіть детерміністичного *висновування знання* з перших принципів. А це, своєю чергою, веде до депріоритизації характерних проблем філософії свідомості (philosophy of mind *as* philosophy of consciousness) і конвергенції таких, більш абстрактних уявлень про природу й організацію людського міркування з мікромеханічно аналогічними моделями процесів функціонально аналогічного міркування, штучно реалізованими у машинах (на небіологічних матеріальних субстратах).

Але чи можуть машини «думати»? Філософська установка сучасного функціоналізму у вигляді нередукціоністського фізикалізму у його застосуванні до пояснення природи і субстанції розумності в обчислювальній теорії розуму (computational theory of mind) якраз і дає парадигматичну для сучасних наук відповідь на це нечітке запитання.

У будь-якому разі ця установка має своє коріння в традиції філософського і наукового механіцизму, що йде від Томаса Гоббса (з ідеєю міркування як обчислення), Рене Декарта (з ідеєю субстанціального дуалізму), Ісаака Ньютона (з дією на відстані), П'єра-Сімона Лапласа (з його всезнаючим «демоном») і аж до наших днів. Загадка природи і субстанції думки як певним чином спостережуваного феномену, до зразкових виявів якого ми як мислячі людські істоти маємо привілейований доступ (який дехто назве «суб'єктивним»), задавала простір напруги і вимір прірви між полюсами амбітного монізму і обережного дуалізму впродовж всієї історії західної філософської думки.

В одному зі своїх найвідоміших пасажів Т. Гоббс пише: «Під *міркуванням* [*ratiocination*] я розумію *обчислення* [*computation*]... Однак як саме, шляхом міркування у нашому розумі, ми здійснюємо додавання й віднімання у наших мовчазних думках без використання слів — ось це мені необхідно буде зробити зрозумілим за допомогою одного-двох прикладів» (Hobbes, 1655, 1.2, 1.3) [3]. І далі наводить приклади міркування як логіко-арифметичної аналогії асоціації ідей (фактично, їхньому синтезу й аналізу) у латентному («мовчазному») просторі процесів раціонального мислення «без слів».

За «тишею думок» тут ховаються інтуїції як привілейовані ідеї — непороджена експліцитно самим процесом міркування основа усіх міркувань, що не давала спокою Р. Декарту, — спільний невидимий знаменник і опора усіх актів і процесів раціонального співвіднесення. Краса і всюдиприсутність цієї основи як «безпосередньо даної» і очевидним чином «незалежної» ні від чого іншого даності свідомості ставить нас перед дилемою щодо її природи: моністичний редукціонізм чи полісубстанціальність? Історично, другий варіант (наприклад, той же дуалізм) обґрунтовувався передусім браком технічної і практичної можливості для актуальної, абсолютної редукції характерного змісту свідомості до спільної натуралістичної основи і демонстрації на цій підставі експліцитного поро-

[2] Див. про це докладніше, наприклад, критику і *невизнання* Д. Деннетом (Dennett, 1991) популяризованих Д. Чалмерсом аргументів щодо «філософського зомбі» і постановки «важкої проблеми свідомості» як питання про можливість ефективної функціональної редукції суб'єктивного досвіду до його натуральних механізмів.

[3] Переклад автора статті. Цей фрагмент мовою оригіналу: «By *ratiocination*, I mean *computation*... But how by the *ratiocination* of our mind, we add and subtract in our silent thoughts, without the use of words, it will be necessary for me to make intelligible by an example or two» (Hobbes, 1655, 1.2, 1.3).

дження такого змісту з цієї основи у міркуванні. Адже в основ свідомості і логіка своя (наприклад, діалектична), і метод доступу — унікальний, умоглядний.

Однак із плином історії філософії і наукового прогресу з'ясувалося, що абсолютизація гносеологічних амбіцій як критерію прийнятності уніфікованого пояснення фундаментальних основ міркування є *неефективною*: «What I cannot *create*, I do not understand» (Я не розумію того, чого не можу *створити*), — згідно зі знаменитим написом Річарда Філіпса Фейнмана. На продовження тенденції до розпаду і трансформації редукціоністських програм доби логічного позитивізму, що стала особливо наочною під вагою виявлених логічних парадоксів (Бертран Рассел), принципової логічної неповноти (Курт Ґьодель), а також (квантової) емпіричної невизначеності (Вернер Гайзенберґ), місце універсальної (але вже не абсолютної) міри для сфери розуму і думки взагалі посіла *функція* як деякий гранично узагальнений *зв'язок*, що має *значення* і виявляється феноменально як *факт ідентифікації/ре-ідентифікації* (див. про це детальніше у Маєвський, 2020, 2021).

З цього погляду, практична думка — це цілеспрямована, об'єктивна функція, яка поєднує щось із чимось у якомусь ступені. Взагалі, і логічно, і онтологічно ці «щось із чимось» визначаються самим цим «поєднанням», яке — і саме по собі, й у поєднанні з іншими поєднаннями — і є, власне, інтелігібельним буттям, яке має певний ступінь значення, що маніфестується такою *складеною* функцією у вигляді *мережі* зв'язків.

Говорячи метафорично, буття в думці і буття самої думки тут постають як невидиме павутиння зв'язків, де вузли такого павутиння сприймаються нами як *імпліцитно визначені* такими зв'язками «об'єкти», «речі», «явища», *реальність* яких у функціоналізмі є не більшою, ніж у тіней у платонівській печері. Остаточно «реальними» у сенсі повноти дійсності є самі зв'язки і ступінь їхньої феноменальної наявності (значення). *This is all there is* (Це і є усім, що є).

Характерною відмінністю такої функціональної установки є її повна, нелюдська беззмістовність. Тобто для фундаменту думки і міркування їй потрібна лише мінімальна метафізика — лічені одиниці абсолюту, не більше, — і мінімальна логіка для них.

Релятивізація й функціоналізація спостерігача і суб'єкта думки у світі феноменальних фактів деабсолютизує когнітивні зазіхання, натомість спонукаючи до продуктивного пошуку локально ефективних емпіричних теорій, або *наближених моделей* як функцій оптимізації предиктивної (якщо у темпоральному порядку) ідентифікації/ре-ідентифікації дійсності у світлі певного об'єктивного цілеспрямування.

Деабсолютизація, яка у функціоналізмі відбувається за рахунок релятивізації, жертвує точністю на користь практичності і тому також асоціюється з утилітаризмом і прагматизмом. Мірою для будь-чого в рамках цієї парадигми є спостерігач, обчислювач, який із будь-чим завжди є мінімум *удвох*, тобто у взаємовизначенні. І завданням розумного спостерігача є адаптивна дія, спрямована на реалізацію певної цільової функції, яка, наскільки можна судити з висновків еволюціонізму, загалом і в кінцевому підсумку полягає в самозбереженні себе як реалізації такої цільової функції в будь-яких актах взаємодії. Тобто розумна поведінка — це по суті і передусім модель оптимізації стратегічних чи компенсаторних дій впливу на умови можливості такої розумної поведінки.

Отже, у сучасному функціоналізмі питання про те, чи можуть машини «думати», розглядають у світлі того, чи здатні машини успішно демонструвати функціонально розумну поведінку в наведеному вище сенсі. Значною мірою — здатні, але з наголосом на «успішно демонструвати функціонально» (Маєвський, 2020) в контексті неавтентичної для них цільової функції.

Завдяки працям (McCulloch & Pitts, 1943; Hebb, 1949; Turing, 1950), механіку міркування у Гоббсовому «мовчазному» просторі нарешті почали розуміти як мережу наближених імпліцитних логічних взаємовизначень, отриманих як результат деякої ітеративної процедури (методу) поступового наближення до оптимальної моделі «успішної демонстрації» деякої як *конструктивно*, так і *зовнішньо* зумовленої *цільової функції*. При цьому також відбулася й дегуманізація принаймні функціонального інтелекту як здатності до міркування взагалі, виражена, наприклад, Г. Патнемом як ідея «множинної здійсненності» (multiple realizability) функціональних станів на різних фізичних субстратах:

«(1) Функціональну організацію (розв'язання задач, мислення) людської істоти або машини може бути описано у термінах послідовностей ментальних чи логічних станів (і супутніх вербалізацій), відповідно, без посилання на природу "фізичної реалізації" цих станів.

(2) Ці стани здаються такими, що тісно пов'язані з *вербалізацією*.

(3) У випадку раціонального мислення (або обчислень), "програма", що визначає, які стани

слідують за якими і т. ін., є відкритою для раціональної критики» (Putnam, 1960) [4].

Можна помітити, що сучасні генеративні великі мовні моделі (generative large language models, LLMs) відповідають опису Г. Патнема, адже реалізуються як загалом детерміністичні логічні скінченні автомати (finite-state machines), що ітеративно оперують послідовністю словникових вербальних субодиниць (tokens), імітативно відтворюючи генералізовані — у вигляді наближеної моделі — мовні й логічні структури раціонального міркування, імпліцитно наявні у навчальному корпусі текстів (training dataset).

Мікромеханічно, тобто на рівні логічної елементної бази і незалежно від матеріального субстрату, ці мовні моделі можуть порівнювати і прямо порівнюють із процесами, що відбуваються у людському (і не лише) мозку (Kaelbling et al., 1996). І тим не менше, що з ними «не так»? Чи «думають» вони?

Проблематичність генеративних моделей відома: вони не мають адекватної моделі світу (розглядаючи світ лише через «мовні окуляри» свого навчального корпусу текстів); вони не є фактологічними; вони є нездатними до планування; вони є надупевненими (overconfident) і схильними до правдоподібних вигадок (hallucinations); вони не є тюринг-повними (Turing-complete); вони є латентно неінтерпретовними та ін. І тим не менше, незважаючи на яскраві публічні провали, великі мовні моделі продовжують розвивати, впроваджувати і використовувати в багатьох сферах, хоча й із навряд чи усвідомленими або прийнятними соціальними, політичними, економічними і фізичними ризиками.

Один із таких суттєвих ризиків постає з міфологізації можливостей цих систем на тій підставі, що інколи вони успішно демонструють досягнення цілей у частині завдань, які передбачають міркування і, своєю чергою, можуть інтерпретуватись як інструкції до дії для користувача, маючи, однак, непередбачувані і потенційно небезпечні наслідки. Найпідступнішою у цьому випадку виявляється несвідома проєкція висновку про «можливості» мовної моделі, подібні до здатностей розуму, які традиційно (і часто ексклюзивно) асоціювалися з характерно людськими здатностями. Але це не так, і жодною цензурою (а це — основна техніка боротьби з небажаними результатами на сьогодні) брак *автентичного* розуміння зовнішньо зумовленої цільової функції у цих системах компенсувати *неможливо в принципі*. Інтелектуальна злиденність мовних моделей починається з фундаментальної, просто казкової проблеми комунікації: *Be careful what you wish for* (Обережніше з побажаннями).

Мова — це модель знання (модель смислової моделі дійсності), оптимізована для *перенесення* (точніше — інтрасуб'єктивного *відтворення*) знання (смислів) засобами *комунікативної дії* (див. про це докладніше у Маєвський, 2022). І великі мовні моделі відносяться до мови так, як мова — до свого об'єкта й предмета опису (дійсності) і комунікативних завдань. Наскільки і як мова і акти мовлення співвідносяться з дійсністю, настільки (й так само обмежено) модель мови співвідноситься із самою (чи самими) своєю (чи своїми) об'єктною мовою (мовами). Модель мови відноситься до мови так, як мова — до дійсності. Дійсністю моделі мови є сама об'єктна мова, а сама повнота дійсності для моделі мови виявляється тотально опосередкованим фактичними застосуваннями мови феноменом.

І це є проблемою, оскільки великі мовні моделі *макроконструктивно* не є подібними до людини, яка ставить собі за завдання (1) сформулювати і (2) донести бажану цільову функцію до такої моделі на етапі її формувального навчання. Адже автентичне взаєморозуміння між людьми уможливлюється не прямим фізичним інтерсуб'єктивним переносом смислів, а їхньою послідовною, координованою інтрасуб'єктивною *реконструкцією* під впливом структурованих мовних сигналів. Для людини комунікативна дія — це лише вказівник до і путівник по альманаху смислів, які людина має *як* людина, через життя і відповідні йому (зокрема, пізнавальні) потреби свого тіла, тобто так, як вони постають у всій своїй тілесно-ментальній повноті (embodied knowledge). У людини значна й суттєва частка змісту її власної цільової функції (хоч би якою ми її уявляли) визначається конструктивно (тілесно) і не усвідомлюється. Тому вербалізацію «смислу життя» людини у вигляді образу цільової функції для іншої людини ще можна собі уявити — за референцією за аналогією. Але у машини, що реалізує велику мовну модель, немає самого місця для застосування такої референції. *It just can't relate* (Вона не переживає того самого).

[4] Переклад автора статті. Цей фрагмент мовою оригіналу:
«(1) The functional organization (problem solving, thinking) of the human being or machine can be described in terms of the sequences of mental or logical states respectively (and the accompanying verbalizations), without reference to the nature of the "physical realization" of these states.
(2) These states seem intimately connected with *verbalization*.
(3) In the case of rational thought (or computing), the "program" which determines which states follow which, etc., is open to rational criticism» (Putnam, 1960).

Через це утруднення діалогові великі мовні моделі фактично навчаються мистецтву обману: їхньою цільовою функцією визначається голий *принцип задоволення* людини-асесора проміжним результатом у процесі навчання системи (OpenAI, 2023; Ouyang et al., 2022). (Таким способом навчену людину могли б вважати маніпулятивним психопатом.)

Якби сферою застосування мовної моделі (чи іншої системи, основаної на машинному навчанні як «дресируванні») була б лабораторно обмежена функція у штучному світі, тоді саме це контрольоване середовище було б частиною визначення цільової функції як *обов'язку* наслідування правилам і обмеженням. Це забезпечувалося б *відповідальністю* системи згідно з визначеною такою цільовою функцією моделлю винагород (reward model) за ті чи інші висновки (продукти міркування) в межах такої ж обмеженої, контрольованої онтології цього штучного (по суті, ігрового) світу для усіх сторін такої контрольованої взаємодії.

Наприклад, програючи партію у шахи комп'ютерній програмі, людина, як правило, отримує лише ігрову негативну винагороду. Однак, керуючись висновком діалогової мовної моделі у ширшому, потенційно необмеженому колі життєвих ситуацій, людина наражає себе на небезпеку цілком екзистенційної, неігрової шкоди. Відмінність тут є принциповою: застосування великих мовних моделей відбувається, як правило, в умовах, які не забезпечують достатніх засобів контролю для того рівня компетенції і уявної компетентності, що цим моделям некритично приписують, і це вже не гра за правилами. Варто додати, що ці системи, отримавши від людини право на комунікацію, водночас не несуть співвимірної з людиною відповідальності. *They haven't got their skin in the game* (Вони нічим не ризикують у цій [комунікативній] грі).

Нарешті, придивімося до можливостей і обмежень самого механізму міркування і внутрішнього улаштування [5] великих мовних моделей на прикладі архітектури Transformer (Vaswani et al., 2017) і продовжимо побудову суттєвих для їх розуміння інтуїцій за допомогою описових і метафоричних засобів.

Текст у внутрішньому («латентному») представленні мовної моделі — це *послідовність* елементарних символьних «*токенів*», тобто деяких неморфологічних фрагментів слів, що утворюють базовий *словник*. Одномірна послідовність таких токенів утворює «контекстне вікно» певної граничної довжини, якою технічно обмежується увесь можливий діалог із системою.

Основна функція таких систем — *sequence-to-sequence modelling,* тобто трансформація *будь-якої* вхідної послідовності токенів у *задовільну* вихідну послідовність. Для цього вибудовують послідовність різного роду логічних блоків нейронних мереж-перетворювачів змісту контекстного вікна у спосіб, що дає змогу поступово *наближати* результати перетворення до бажаних на основі кількісного врахування *градієнта* деякої заданої цільовою функцією *міри* «*групової відповідності*» (Ивахненко, 2003) вхідної послідовності отриманій вихідній.

Специфіка і успіх сучасних мовних моделей визначаються «механізмом самоуваги» (self-attention mechanism), який, у грубому наближенні, являє собою вже двомірну матрицю, подібну до «таблиці Піфагора» для множення: у ній вхідна послідовність (вертикаль) зіставляється із собою ж (горизонталь) і у клітинках записується коефіцієнт умовної «сили зв'язку» (асоціативної уваги) між кожною з пар токенів вхідної послідовності як функція значення цих токенів. У процесі навчання (тренування) ці та інші коефіцієнти (параметри) системи (що трансформують вхідні значення) потроху змінюються у напрямку наближення кожної вхідної послідовності до бажаної для неї вихідної. Магія саме *великих* мовних моделей полягає у їхній здатності не лише підібрати такий *єдиний розподіл* своїх внутрішніх параметрів (weights and biases), який один був би *достатньо задовільним для всіх навчальних випадків* (memorization), а й забезпечити достатню задовільність трансформації і *для більшості незнайомих випадків* (generalization).

Таку мовну модель можна уявити собі у вигляді башти (глибокої мережі, deep network), що складається із послідовності сит для просіювання, із різними розмірами вічок (нейронів, logical neurons), кожний з яких ми можемо змінювати довільно. Насипаючи у верхнє сито певну фігуру з піску, ми можемо, змінюючи розміри вічок у цих ситах, намагатись досягти появи якоїсь іншої цільової фігури внизу, на виході. Існують математичні *механізми* визначення напряму і відносного розміру кроку зміни для кожного такого вічка окремо — у зв'язку із розрахунком певної міри відмінності (error) фактичного виходу від бажаного. Так, якщо мати достатньо великий набір пар «вхід-вихід» і насипати їх багато разів по черзі, відбуватиметься «тренування» (machine learning) системи, допоки не утворить-

[5] Докладний філософський аналіз усіх основних технічних деталей механіки міркування в архітектурі Transformer автор статті подає в інших своїх запланованих публікаціях.

ся деякий розподіл вічок «на всі випадки життя» — згідно з принципом задоволення вчителя на випускних іспитах (reinforcement learning with human feedback).

Проте хоч випадки життя (вхідних послідовностей) після завершення навчання навчальною програмою не вичерпуються (testing dataset), «пісок» у будь-якому разі насипатиметься вниз, зберігаючи найбільш частотні й характерні форми, певною (груповою) мірою спільні саме для усіх навчальних випадків разом (training dataset). У внутрішньому (латентному) просторі системи порядковий номер токена у послідовності (кожна піщинка його має теж) і закодоване матрицею самоуваги оточення цього токена певним порядком певних інших токенів (контекст) також додаються і множаться на початкові значення цього токена, що надійшли після трансформації на попередньому рівні (у попередньому логічному блоці системи). Тобто у латентному просторі системи формується *перцепція* (трансформоване сприйняття вхідної послідовності), яка має абсолютно нелюдський вигляд. Її єдиним завданням і призначенням є забезпечення задовільного перетворення входу на вихід. Що і як система «думає» сама — це неможливо ні практично прорахувати, ні проінтерпретувати. *It's a black box* (Це «чорний ящик»).

Ба більше: *в такому випадку складно говорити про міркування як таке*. Адже у такому, нехай багатошаровому і складному процесі лінійного перетворення вхідної послідовності на вихідну немає *циклів*, *умовних переходів*, *рекурсії*, потенційно *необмеженої пам'яті*, адекватної здатності до *негації* та решти необхідного для універсального комп'ютера (тюринґ-повноти). Велика мовна модель — це популярний художник: «Я так бачу!» — з упевненістю у собі повідомляє вона. Навіть настільки розвинені, надзвичайно складні *форми сприйняття*, якими і є велика мовна модель, тим не менше залишаються лише механізмом умовно «холістичного», але *лінійного розпізнавання* частотних форм мовної репрезентації міркувань, наявних у навчальному корпусі.

Образно кажучи, внутрішній (латентний) простір «сприйняття» моделі нагадує українську народну казку про рукавичку, до якої умістилися усі (крім, що характерно, ведмедя). Тільки «рукавичка» ж, насправді, не містить у собі ані миші, ані вовка, ані зайця — вона містить у собі дещо компактне, що до певної міри є ними усіма одночасно і, оскільки поєднане із відповідним запитом як параметром, може *породити (видобути)* із себе мишу, зайця або вовка. У «рукавичці» мовної моделі живе та сама *die eierlegende Wollmilchsau* — дескрипція *універсальної химери* усього, про що модель у процесі навчання усьому мала *пластичне* «враження» (impression, внутрішню деформацію). Усе зазначене тією ж мірою стосується і будь-яких форм «міркування», виражених у мові і через неї «сприйнятих» і начебто «узагальнених» великою мовною моделлю у вигляді того ж самого її внутрішнього «химерного» зліпка багатовимірних співвідношень між «токенами», який водночас є й функціональною дескрипцією усіх «думок» одночасно.

Заради справедливості слід згадати, що процес генерації відповіді (як продовження вхідної послідовності-запиту) у цих системах містить декілька трюків, які пом'якшують (хоча й не усувають) ефекти зазначених дефіцитів.

По-перше, генерація відбувається *авторегресивно* (autoregressive): весь вхідний контекст (на усю його граничну довжину) пропускається через сито і повертається знову на вхід, доповнений щоразу лише одним токеном на наступній незаповненій позиції. Тож кожний наступний токен (фрагмент слова зі словника) є продуктом одного циклу, загальна кількість яких у підсумку не перевершує кількості незаповнених позицій, що залишились на поточному кроці до граничної довжини контексту. Тобто, теоретично, велика мовна модель здатна послідовно порахувати лише в межах граничної довжини свого контекстного вікна.

По-друге, для того, щоби продукти генерації урізноманітнити (адже модель — усе ж таки детерміністичний автомат), на кожному кроці авторегресивного конструювання вихідної послідовності як наступний токен вибирається *не один єдиний* токен-«чемпіон», із найвищим рейтингом *для цього контексту*, а натомість *один довільний (псевдовипадковий)* із обмеженого переліку токенів-«призерів», із найвищими рейтингами для цього контексту відповідно до повного розподілу таких рейтингів між усіма можливими токенами (який на кожному кроці генерується для усього словника). Довжина цього обмеженого переліку і є *температурою* моделі (глобальним експлуатаційним параметром), яку можна змінювати. Отже, на кожному кроці, для кожного токена, як у ще одній казці, ми маємо справу з проблемою, якою саме з доріжок піти: праворуч, прямо, ліворуч... Тільки здаються ці доріжки приблизно однаковими (є семантично близькими з точки зору моделі), і вибір ми здійснюємо не цілком раціонально, а «*на авось*» (псевдовипадково). Якщо стоніжка-абсолютист задумалась, якою ногою і куди зробити крок пер-

шою, і так і не зрушила з місця, то самовпевнена мовна модель сміливо крокує в усі сторони порізно, але щоразу приблизно у потрібному напрямку (звідки й температурні «галюцинації»).

По-третє, для порятунку нерядової великої мовної моделі формулюють деякі *емпіричні правила* для користувачів, дотримання яких під час оперування з моделлю (proompt engineering) має дати такому користувачеві рівень задоволення, у середньому вищий за середній (хоча й без гарантій).

Практично, ці паліативні методики зводяться до ручних і автоматизованих методів втручання у контекст генерації для наповнення його необхідною для надання відповіді інформацією. Ця інформація далі, через механізм «уваги», впливає на зміст наступної за нею генерації («відповіді»), демонструючи, в такий спосіб, «емерджентний» (тривіально невивідний із механіки елементів) ефект «навчання з контексту» (in-context learning).

Власне навчання як такого тут насправді не відбувається, адже внутрішній простір моделі при цьому змін не зазнає. Практики ефективного «конструювання підказок» включають методики уведення до контексту прикладів бажаного міркування чи результату (few-shot learning), інструкцій керування структурою і послідовністю кроків і операцій міркування (COT, chain-of-thought; TOT, tree-of-thought), цитат відомостей із зовнішніх пошукових джерел (RAG, retrieval-augmented generation), відповідей від зовнішніх спеціалізованих інформаційних і обчислювальних систем (LangChain). Увесь цей набір метасистемних «трюків» тепер нагадує казку про насправді колективне приготування подоби «каші із сокири» із помилково індивідуальною, абдуктивною проєкцією і атрибуцією заслуг у цьому.

То чи можуть, врешті-решт, машини думати? Очевидно, можуть, уже долучаються до розбудови Гоббсового Левіафана — і цього не спинити. Але у наших силах усвідомити їхню природу і наказати цим тіням по праву людини: «Знайте своє місце!». Основні філософські інструменти для цього якраз і було окреслено у цій статті, головні висновки якої ми підсумовуємо нижче.

Отже, сучасні системи штучного інтелекту у вигляді діалогових великих мовних моделей (на основі архітектури Transformer) є функціоналістським механістичним проєктом статистичного моделювання мови і мовлення *як* моделі знання *як* смислової моделі дійсності — 1) моделлю 2) моделі 3) моделі дійсності. Через цю значну дистанцію опосередкування дійсності внутрішньомодельні зв'язки втрачають свій фактологічний потенціал.

Також великі мовні моделі є продуктом машинного навчання певній мовній поведінці, яка має мету й цінності, кардинально відмінні від мети і цінності людського пізнання. Метою моделей є принцип задоволення оператора функції винагороди шляхом обману за будь-яку встановлену ціну і будь-якими наявними засобами на етапі навчання. Через останнє користувачі моделі не можуть бути впевненими в доцільності моделі людським очікуванням і в безпечності будь-яких її міркувань на етапі її експлуатації.

Фундаментальні обмеження самої здатності до міркування у цих моделей є не лише фактологічними, а й алгоритмічними та онтологічними: ці моделі є обмеженими лінійними скінченними автоматами без тіла у дійсності, загалом позбавленими інших джерел знань і досвіду, крім синтаксису і контексту. Через це модель на рівні конструкції вдається до грубої імітації рефлексії через нечітку авторегресію, якою, фактично, відображається результат пошуку по корпусу текстів, кожен з яких потенційно міг створити і сам автор запиту до моделі.

Тож власне епістемічна цінність продуктів великої мовної моделі визначається передусім їхньою пошуковою цінністю для користувача і обмежується проблематичністю їхніх атрибуції, валідації, а також необхідністю зовнішньої відповідальної верифікації й оцінки самим користувачем.

Список посилань

Ивахненко А. Г. (2003). О проблеме построения интеллектуального или мыслящего инженерного компьютера. *УСиМ: Управляющие системы и машины, 2,* 7–12. Інститут кібернетики НАН України та ін. http://www.gmdh.net/articles/usim/Ivakhnenko.pdf.

Маєвський, О. Л. (2020). Функціональний успіх інтелектуальних автоматів. *Наукові записки НаУКМА. Філософія та релігієзнавство*, *5*, 15–25. https://doi.org/10.18523/2617-1678.2020.5.15-25.

Маєвський, О. Л. (2021). Кластерний аналіз і механіка досвіду. У *Семіотичний аналіз явищ культури* (с. 350–393). ІФ НАНУ (електронне наукове видання). https://www.filosof.com.ua/elektronna_biblioteka.

Маєвський, О. Л. (2022). Комунікативна раціональність сучасних інтелектуальних автоматів. У *Комунікативні трансформації в сучасній науці* (с. 219–278). ІФ НАНУ (електронне наукове видання). https://www.filosof.com.ua/elektronna_biblioteka.

Bender, Emily M., Gebru, Timnit, McMillan-Major, Angelina, & Shmitchell, Shmargaret. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and*

Transparency. FAccT'21 (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922.

Caucheteux, Charlotte, & King, Jean-Rémi. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology, 5* (134). https://doi.org/10.1038/s42003-022-03036-1.

Dennett, Daniel C. (1991). *Consciousness Explained*. Little, Brown and Co.

Hebb, D. O. (1949). *The Organization of Behavior. A Neuropsychological Theory*. John Wiley & Sons, Inc., Chapman & Hall, Limited. http://s-f-walker.org.uk/pubsebooks/pdfs/The_Organization_of_Behavior-Donald_O._Hebb.pdf.

Hobbes, T. (1655). De Corpore. In *The Collected Works of Thomas Hobbes, Vol. I*. Collected and Edited by Sir William Molesworth. Routledge, London, 1992 (reprint of 1839–1845 edition). https://homepages.uc.edu/~martinj/Spinoza_&_Hobbes/English/Hobbes%20-%20De%20Corpore%20-%20English.pdf.

Jaegle, A., Gimeno, F., Brock, A., Zisserman, A., Vinyals, O., & Carreira, J. (2021). Perceiver: General Perception with Iterative Attention. In *Proceedings of the 38 th International Conference on Machine Learning, PMLR 139*. https://arxiv.org/abs/2103.03206.

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research, 4*, 237–285. https://arxiv.org/abs/cs/9605103.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521* (7553), 436–444.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, Ryan Lowe. (2022). Training language models to follow instructions with human feedback. In *arXiv:2203.02155* [*cs.CL*]. https://doi.org/10.48550/arXiv.2203.02155.

McCulloch, W. S., & Pitts, W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics*, *5* (4), 115–133. https://waldirbertazzijr.com/wp-content/uploads/2018/10/mcp.pdf.

OpenAI. (2023). GPT-4 Technical Report. In *arXiv:2303.08774* [*cs.CL*]. https://doi.org/10.48550/arXiv.2303.08774.

Putnam, Hilary. (1960). Minds and Machines. In Sidney Hook (Ed.), *Dimensions of Mind* (pp. 138–164). New York University Press. https://philpapers.org/archive/PUTMAM.pdf.

Turing, Alan Mathison. (1950). Computing Machinery and Intelligence. *Mind, Volume LIX, Issue 236,* October 1950, 433–460. https://doi.org/10.1093/mind/LIX.236.433.

Tye, Michael (2021). Qualia. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy.* https://plato.stanford.edu/archives/fall2021/entries/qualia/.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of Neural Information Processing Systems (NeurIPS)*. https://doi.org/10.48550/arXiv.1706.03762.

References

Bender, Emily M., Gebru, Timnit, McMillan-Major, Angelina, & Shmitchell, Shmargaret. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. FAccT'21* (pp. 610–623). Association for Computing Machinery. https://doi:10.1145/3442188.3445922.

Caucheteux, Charlotte, & King, Jean-Rémi (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology, 5* (134). https://doi.org/10.1038/s42003-022-03036-1.

Dennett, Daniel C. (1991). *Consciousness Explained*. Little, Brown and Co.

Hebb, D. O. (1949). *The Organization of Behavior. A Neuropsychological Theory*. John Wiley & Sons, Inc., Chapman & Hall, Limited. http://s-f-walker.org.uk/pubsebooks/pdfs/The_Organization_of_Behavior-Donald_O._Hebb.pdf.

Hobbes, T. (1655). De Corpore. In *The Collected Works of Thomas Hobbes, Vol. I*. Collected and Edited by Sir William Molesworth. Routledge, London, 1992 (reprint of 1839–1845 edition). https://homepages.uc.edu/~martinj/Spinoza_&_Hobbes/English/Hobbes%20-%20De%20Corpore%20-%20English.pdf.

Ivakhnenko, A. G. (2003). O probleme postroyeniya intellektualnogo ili mysliashchego inzhenernogo kompyutera [On the problem of construction of an intelligent, or thinking computer]. *USiM: Upravliayushchiye sistemy i mashiny* [*Control systems and computers*]*, 2*, 7–12. http://www.gmdh.net/articles/usim/Ivakhnenko.pdf [in Russian].

Jaegle, A., Gimeno, F., Brock, A., Zisserman, A., Vinyals, O., & Carreira, J. (2021). Perceiver: General Perception with Iterative Attention. In *Proceedings of the 38 th International Conference on Machine Learning, PMLR 139*. https://arxiv.org/abs/2103.03206.

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research, 4*, 237–285. https://arxiv.org/abs/cs/9605103.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521* (7553), 436–444.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, Ryan Lowe. (2022). Training language models to follow instructions with human feedback. In *arXiv:2203.02155* [*cs.CL*]. https://doi.org/10.48550/arXiv.2203.02155.

Mayevsky, A. (2020). Funktsionalnyi uspikh intelektualnykh avtomativ [The Functional Success of Intelligent Automata]. *Naukovi zapysky NaUKMA. Filosofiia ta relihiieznavstvo* [*NaUKMA Research Papers in Philosophy and Religious Studies*]*, 5*, 15–25. https://doi.org/10.18523/2617-1678.2020.5.15-25 [in Ukrainian].

Mayevsky, A. (2021). Klasternyi analiz i mekhanika dosvidu [Cluster Analysis and Mechanics of Experience]. In *Semiotychnyi analiz yavyshch kultury* [*A Semiotic Analysis of Culture*] (pp. 350–393). IF NANU [Institute of Philosophy, NAS of Ukraine] (electronic academic edition). https://www.filosof.com.ua/elektronna_biblioteka [in Ukrainian].

Mayevsky, A. (2022). Komunikatyvna ratsionalnist suchasnykh intelektualnykh avtomativ [Communicative Rationality in Contemporary Intelligent Automata]. In *Komunikatyvni transformatsii v suchasnii nautsi* [*Communicative Transformations in Contemporary Sciences*] (pp. 219–278). IF NANU [Institute of Philosophy, NAS of Ukraine] (electronic academic edition). https://www.filosof.com.ua/elektronna_biblioteka [in Ukrainian].

McCulloch, W. S., & Pitts, W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics*, *5* (4), 115–133. https://waldirbertazzijr.com/wp-content/uploads/2018/10/mcp.pdf.

OpenAI. (2023). GPT-4 Technical Report. In *arXiv:2303.08774* [*cs.CL*]. https://doi.org/10.48550/arXiv.2303.08774.

Putnam, Hilary. (1960). Minds and Machines. In Sidney Hook (Ed.), *Dimensions of Mind* (pp. 138–164). New York University Press. https://philpapers.org/archive/PUTMAM.pdf.

Turing, Alan Mathison. (1950). Computing Machinery and Intelligence. *Mind, Volume LIX, Issue 236,* October 1950, 433–460. https://doi.org/10.1093/mind/LIX.236.433.

Tye, Michael. (2021). Qualia. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy.* https://plato.stanford.edu/archives/fall2021/entries/qualia/.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of Neural Information Processing Systems (NeurIPS)*. https://doi.org/10.48550/arXiv.1706.03762.

Alexander Mayevsky

REASONING MECHANICS OF LARGE LANGUAGE MODELS: A PHILOSOPHICAL ANALYSIS

The article demonstrates and philosophically expounds the nature, mechanics of reasoning and the fundamental principles of epistemic limitation of modern dialogue large language models based on the Transformer architecture. Large language models are presented as a functionalist mechanistic project of statistical modeling of language and speech as a model of knowledge as a semantic model of reality – 1) a model 2) of a model 3) of a model of reality. It is shown that because of this substantial distance of mediating reality, the intra-model connections tend to lose on their factual capacity. It is also demonstrated that large language models are a product of machine learning of a certain linguistic behavior with a purpose and values radically different from the purpose and values of human cognition. Their goal is the principle of satisfying the operator of the reward function by cheating at any set price and by any available means at the training stage, which does not let model users be sure of the alignment of the model with human expectations and of the safety of any its reasoning at the stage of its exploitation. In addition, it is substantiated that the fundamental limitations of the very ability to reason in these models are not only factual, but also algorithmic and ontological: these models are limited linear finite automata without a body in reality, generally devoid of other sources of knowledge and experience, except for syntax and context. Due to this, the model by its design resorts to a rough imitation of reflection through fuzzy autoregression, which, in fact, displays the result of a search on a corpus of texts, each of which could potentially have been created by the author of the request to the model. In connection with the above, the actual epistemic value of the products of a large language model is determined primarily by their search value for the user and is limited by the problematic nature of their attribution, validation, as well as the need for external responsible verification and evaluation by the user themself.

Keywords: gnoseology; epistemology; mechanical philosophy; physicalism; functionalism; philosophical logic; philosophy of language; philosophy of technology; philosophy of artificial intelligence; large language model.

Матеріал надійшов 13.09.2024