ISSN 2786-7005 (Print) ISSN 2786-7013 (Online)

УДК 165:004.8

Критичне мислення і формальна логіка в системах штучного інтелекту: філософське осмислення

CRITICAL THINKING AND FORMAL LOGIC IN ARTIFICIAL INTELLIGENCE SYSTEMS: A PHILOSOPHICAL ANALYSIS

НЕВМЕРЖИЦЬКА Олена – *кандидат філософських наук, старший викладач кафедри філософії та психології, Київський університет інтелектуальної власності і права, вул. Харківське шосе, 210, м. Київ, 02000, Україна*

ORCID https://orcid.org/0000-0001-9346-7205

NEVMERZHYTSKA Olena – *Candidate of Philosophical Sciences, Senior Lecturer at the Department of Philosophy and Psychology, Kyiv University of Intellectual Property and Law, 210 Kharkivske Shose St., Kyiv, 02000, Ukraine*

КИЯШКО Святослав – *аспірант гуманітарно-педагогічного факультету, кафедри «Управління та освітніх технологій», Національний університет біоресурсів і природокористування України, вул. Героїв Оборони, 15, м. Київ, 0304, Україна*

ORCID https://orcid.org/0009-0000-2938-0199

KYIASHKO Svyatoslav – *postgraduate student of the Faculty of Humanities and Pedagogy, Department of «Management and Educational Technologies», National University of Life Resources and Environmental Sciences of Ukraine, 15 Heroiv Oborony St., Kyiv, 0304, Ukraine*

DOI https://doi.org/10.54891/2786-7013/2025-2-4

Анотація. *Стаття пропонує цілісне філософське осмислення критичного мислення та формальної логіки в сучасних системах штучного інтелекту (ШІ) на тлі поширеного феномену «галюцинацій» – впевнено згенерованих, але фактично хибних відповідей. Автори обґрунтовують, що суто статистичні мовні моделі без внутрішніх механізмів верифікації не забезпечують епістемічної надійності й тому вимагають «дисциплінування» через логічні та метакогнітивні модулі. Методологічно дослідження поєднує історико-філософський аналіз (від Аристотеля, Декарта, Канта і Поппера до сучасних дискусій про семантичну обмеженість синтаксичних процедур у дусі «китайської кімнати» Дж. Серля) з оглядом технічних підходів до інтеграції логіки у мовні моделі ШІ. Показано маятникову еволюцію від символічного (GOFAI) до нейромережевого підходів і формування консенсусу на користь нейросимвольних архітектур, які поєднують статистичну потужність із прозорістю логічного виведення. Виокремлено три компоненти людського критичного мислення, релевантні для алгоритмів: нормативний (правила та закони логіки), дескриптивний (усвідомлення типових когнітивних викривлень) і прескриптивний (процедури самокорекції). На цій підставі сформульовано принципи «логічної дисципліни» для ШІ: явне багатокрокове міркування, внутрішня самоперевірка й оцінка впевненості, інструментальна перевірка фактів (на основі використання зовнішніх модулів і баз знань), причинно-наслідкове моделювання та пояснюваність. Розглянуто обмеження: коректність формального виведення не гарантує істини за хибних посилок; відсутність свідомості та інтенціональності унеможливлює «внутрішню» мотивацію до істини; довгі ланцюжки міркувань можуть раціоналізувати початкову помилку. Аргументовано, що оптимальною наразі є гібридна парадигма, де ШІ виконує роль швидкого логічного фільтра та генератора гіпотез, а людина – остаточного епістемічного арбітра. Практична значущість охоплює освіту, науку та управлінські рішення: впровадження логічних і метакогнітивних контурів здатне знизити частоту фактологічних помилок і підвищити довіру до ШІ. Наукова новизна полягає в міждисциплінарному картуванні компонентів критичного мислення на реалізовані алгоритмічні механізми та в окресленні дослідницької програми для розвитку*

© НЕВМЕРЖИЦЬКА Олена, КИЯШКО Святослав

«мислячих» моделей, які поєднують логічну строгість, причинне міркування і прозору пояснюваність, лишаючись при цьому під наглядом людини.
Ключові слова: *критичне мислення, формальна логіка, нейросимвольний штучний інтелект, галюцинації ШІ, пояснюваність ШІ, епістемологія ШІ.*

Summary. *The article provides a comprehensive philosophical account of critical thinking and formal logic within contemporary artificial intelligence (AI) systems against the backdrop of the pervasive «hallucination» phenomenon–confidently produced yet factually incorrect outputs. It argues that purely statistical language models, lacking intrinsic verification mechanisms, cannot deliver epistemic reliability and therefore require «disciplining» via logical and metacognitive components. Methodologically, the study combines a history-of-philosophy perspective–from Aristotle, Descartes, Kant and Popper to current debates on the semantic limits of syntactic procedures (e.g., Searle's «Chinese Room») – with a survey of technical strategies for embedding logic in AI. The paper traces the pendulum swing from symbolic (GOFAI) to neural approaches and shows an emerging consensus in favor of neuro-symbolic architectures that couple statistical power with the transparency of formal inference. It identifies three human critical-thinking components relevant for algorithms: a normative one (logical laws and rules), a descriptive one (awareness of common cognitive biases), and a prescriptive one (procedures for self-correction). Building on this mapping, it formulates principles of «logical discipline» for AI: explicit multi-step reasoning, internal self-verification and confidence estimation, tool-assisted fact-checking (calls to external modules and knowledge bases), causal modeling, and explainability. The analysis also clarifies key limitations: valid formal inference does not ensure truth under false premises; the absence of consciousness and intentionality precludes an «intrinsic» drive to truth; long reasoning chains may rationalize an initial error. The paper contends that a human-in-the-loop paradigm is currently optimal: AI acts as a rapid logical filter and hypothesis generator, while the human agent remains the final epistemic arbiter. Practical significance spans education, science, and managerial decision-making: introducing logical and metacognitive control loops can reduce factual errors and strengthen trust in AI. The contribution lies in an interdisciplinary mapping of critical-thinking components to implementable algorithmic mechanisms and in outlining a research program for «thinking» models that integrate logical rigor, causal reasoning, and transparent explanations – while remaining under human oversight.*
Key words: *critical thinking, formal logic, neuro-symbolic AI, AI hallucination, AI explainability, AI epistemology.*

Вступ. Стрімкий розвиток генеративних моделей штучного інтелекту (ШІ) актуалізував філософське питання про природу мислення і логіки в машинних системах. Сучасні великі мовні моделі, як ChatGPT, Gemini, Grok продемонстрували здатність генерувати розгорнуті зв'язні тексти, проте поряд із правильними відповідями вони з упевненістю продукують правдоподібні, але хибні твердження. Цей феномен дістав назву «галюцинацій» нейромереж. Проблема полягає в тому, що модель, покликана оперувати знаннями, фактично імітує знання – вона може надати грамотно сформульовану відповідь, яка звучить переконливо, але не відповідає дійсності. Як наслідок, некритичний користувач часто сприймає помилковий, хоч і правдоподібний результат за істину. Це створює ризик фундаментального непорозуміння: якщо навіть фахівцю потрібно перевіряти кожен факт, то де межа між знанням і його симуляцією? Аналогічна дилема простежується ще від часів Платона, який відрізняв справжнє знання (episteme) від простої думки (doxa) – у випадку ШІ його відповіді часто належать саме до сфери переконливої думки, а не обґрунтованого знання. Таким чином, виникає філософське питання: як відрізнити істинне знання від правдоподібної помилки у відповідях ШІ, і чи здатен ШІ самостійно здійснювати таке розрізнення?

Практичну значущість проблеми важко переоцінити. Галюцинації ШІ підривають довіру до інтелектуальних систем та можуть гальмувати їх впровадження у відповідальних сферах. ШІ-рішення, що дають красиві, але неправильні відповіді, ризикують лишитися

розважальними іграшками замість надійних інструментів. Це вже призвело до того, що в деяких компаніях запроваджують додаткові процедури перевірки: результати, згенеровані моделлю, обов'язково ревізуються експертом, аби виявити можливі помилки. Такий підхід перекладає тягар критичного аналізу на людину, вказуючи на обмеженість самих моделей. Отже, постає науково-практичне завдання: навчити штучний інтелект критично мислити і дотримуватися логіки, щоб він не лише відтворював імовірні тексти, а й умів відрізняти істину від хиби. Чи можливо це взагалі, з огляду на несвідомий характер сучасних алгоритмів? Це питання має міждисциплінарний характер, поєднуючи інженерні завдання та філософські проблеми епістемології і свідомості.

Аналіз останніх досліджень. Проблема правдоподібних, але хибних відповідей ШІ привернула увагу і дослідників, і практиків. З одного боку, проводяться емпіричні дослідження частоти та типології галюцинацій [9]. Ці дані вказують на систематичність проблеми. З іншого боку, лідери галузі ШІ усвідомлюють ризики та публічно підкреслюють необхідність обережності. Так, генеральний директор OpenAI С. Альтман визнав, що його здивувала надмірна довіра користувачів до ChatGPT, і застеріг, що моделі іноді «брешуть, щоб догодити», вигадуючи неіснуючі факти, тому не слід сліпо довіряти їхнім відповідям [7]. Практична реакція індустрії полягає у впровадженні принципів критичного мислення при використанні ШІ: результати моделі мають перевірятися на достовірність людиною-експертом перед прийняттям рішень [13; 20; 7].

У науковій літературі пропонується низка підходів до вирішення проблеми. По-перше, звернули увагу на необхідність оснастити моделі механізмами логічного контролю. Перші спроби інтеграції логіки в ШІ ще у 1960-х (системи автоматичного доведення теорем, експертні системи) заклали підґрунтя для сучасних досліджень в напрямі нейросимвольного ШІ [8; 16]. Огляд Colelough та Regli (2024) показує, що в період 2020–2024 рр. відбувся вибух досліджень на стику нейронних мереж та формальної логіки [8]. Попит на нейросимвольні рішення пояснюється усвідомленням меж чисто статистичних моделей: низка авторів наголошує, що «не можна долетіти до Місяця, видираючись на дедалі вищі дерева», маючи на увазі, що просте нарощування параметрів моделей не замінює якісного стрибка в напрямі логічного розуміння [8; 16].

По-друге, дослідники вказують на важливість метакогнітивних компонентів – здатності системи ШІ аналізувати власні «міркування» і виявляти помилки до того, як вони потраплять у фінальну відповідь. Поки що метакогніція залишається найменш дослідженою сферою (за згаданим оглядом, лише невелика частина робіт стосувалися самоконтролю ШІ), але поява перших моделей з ознаками планування та самоперевірки свідчить про перспективність цього напряму [24; 15].

По-третє, у філософських публікаціях (Матвієнко, 2025; Надурак, 2022) осмислюється сам феномен критичного мислення та раціональності, на які орієнтуються творці ШІ. Зокрема, Матвієнко (2025) вказує на взаємодію людини і ШІ у процесі мислення: щоби моделі стали корисними, людські навички критичного аналізу мають бути «перенесені» в алгоритмічне середовище [4]. Український філософ В. Надурак наголошує: «таким чином, серед іншого, навичка критичного мислення передбачає володіння спеціальним інструментарієм – набором різних процедур, які допомагають долати когнітивні упередження» [5, с. 142], тобто що критичне мислення – це навичка оцінювати хід власних думок за критеріями раціональності. На ці ідеї спирається наше дослідження, фокусуючись на тому, як прищепити машинам елементи критичного і логічного мислення [5; 3]. Водночас залишається невирішеним питання про межі такого «навчання»: поки що неясно, чи можливе справжнє розуміння істини без свідомості,

ISSN 2786-7005 (Print) ISSN 2786-7013 (Online)

або ж ШІ завжди тільки імітуватиме мислення. Саме цей аспект покликаний висвітлити філософський підхід, обраний у статті.

Мета статті. Метою даного дослідження є філософське осмислення проблеми критичного мислення та формальної логіки в системах штучного інтелекту, а також вироблення підходів до подолання феномену «правдоподібних помилок» у відповідях ШІ. Для досягнення цієї мети вирішуються такі завдання: 1) проаналізувати природу галюцинацій ШІ та пов'язані з ними епістемологічні проблеми істинності і знання; 2) розкрити зміст поняття критичного мислення людини, спираючись на праці класичних і сучасних філософів, та виокремити компоненти критичного мислення, релевантні для алгоритмів (формально-логічний, рефлексивний, прескриптивний); 3) дослідити еволюцію підходів до впровадження формальної логіки в ШІ – від ранніх символьних систем до сучасних нейромережевих моделей, що намагаються відновити логічну складову; 4) порівняти здатність до логічного та критичного мислення у людини і сучасних ШІ-моделей, визначивши ключові відмінності та точки зближення; 5) окреслити перспективи розвитку «мислячих» моделей ШІ та пов'язані з ними філософські питання (проблему свідомості, інтенціональності, довіри тощо).

Виклад основного матеріалу. Проблема правдоподібних, але хибних відповідей ШІ, відома як явище галюцинацій, виявило обмеження суто статистичного підходу до побудови імітації інтелекту [13; 9]. Моделі глибинного навчання генерують тексти, ґрунтуючись на ймовірнісних закономірностях у даних, без перевірки їх істинності [13; 20]. У результаті ШІ може впевненим тоном повідомити неправдивий «факт» або вигадати цитату, яка ніколи не існувала [20; 9]. Небезпека подібних ситуацій посилюється тим, що такі відповіді не мають явних ознак помилковості – вони стилістично коректні, логічно впорядковані і часто не викликають підозри у неспеціаліста [9]. Класичний приклад – модель впевнено пояснює неіснуюче наукове поняття або радить ліки, яких насправді немає. Лише експерт, вже озброєний знанням істини, може відрізнити підробку. Таким чином, епістемологічний статус знання, згенерованого ШІ, стає проблематичним: чи можна назвати це знанням, якщо йому бракує гарантії істинності? У традиційній філософії знання розуміють як обґрунтовану істинну віру. У випадку ШІ ми маємо справу принаймні з видимістю віри (модель «впевнена» у своїх твердженнях) без особистої відповідальності за істину. Це знання без суб'єкта, яке може виявитися хибним [5].

Із гносеологічної точки зору, така ситуація кидає виклик відомому критерію істинності. Якщо правдоподібність більше не є надійним показником істини, потрібні додаткові механізми верифікації [13; 20]. Відповіддю на цей виклик стала низка технологічних рішень. По-перше, розробники почали впроваджувати вбудовані перевірки фактів. Деякі сучасні великі мовні моделі (наприклад, у режимах пошуку) здатні автоматично робити запит до баз знань чи інтернету, щоб підтвердити факт перед тим, як сформулювати відповідь [13]. Однак і цього недостатньо – модель може неправильно витлумачити знайдену інформацію або підібрати нерелевантні джерела [9]. По-друге, виникла ідея, що модель повинна «думати перед тим, як говорити». Ця метафора лягла в основу нової методики генерації відповідей на запити до ШІ-систем, про які йтиметься далі. Утім, уже зараз зрозуміло: просто збільшувати обсяги тренувальних даних недостатньо для усунення галюцинацій [9; 16]. Потрібні якісно інші підходи, зокрема залучення механізмів логічного виведення та критичного аналізу.

Щоб зрозуміти, що саме необхідно реалізувати в ШІ, звернімося до поняття критичного мислення людини. У філософській традиції критичне мислення пов'язане з ідеалом раціональності та рефлексії Людина, яка мислить критично, не просто формально правильно міркує, але й постійно оцінює зміст своїх думок, відфільтровуючи помилки, упередження, логічні хиби. В. Надурак визначає: «Критичне мислення це навичка, що передбачає

вміння аналізувати процес мислення на предмет його відповідності критеріям раціональності» [5, с. 134]. Така здатність спирається на три складові: нормативну, дескриптивну і прескриптивну. Нормативна складова передбачає знання правил правильного мислення – законів формальної логіки, методів наукового обґрунтування, принципів доказовості. Дескриптивна складова пов'язана з усвідомленням реальних психологічних процесів мислення, зокрема типових когнітивних викривлень (ефект підтвердження, помилки атрибуції тощо). Нарешті, прескриптивна складова включає методи виправлення думки: як перебороти упередження, перевірити факти, застосувати логічні схеми для оцінки аргументів [5]. Отже, критичне мислення є метамисленням – мисленням про мислення, спрямованим на досягнення істинності й обґрунтованості знань.

Формальна логіка відіграє у цьому ключову, але не єдину роль. Ще Аристотель заклав основи логіки як науки про форми правильного міркування, які гарантують збереження істини. Наприклад, силогістична логіка вчить, що із двох істинних посилок при правильній формі аргументу гарантовано випливає й істинний висновок [1]. Цей формальний апарат забезпечує несуперечливість і правильність міркувань, абстрагуючись від конкретного змісту. Критично мисляча людина опановує логіку, але йде далі – вона рефлексує над самим фактичним змістом: чи правдиві наші посилки? чи не впливають на висновки психологічні упередження?. Така всеосяжна раціональність є ідеалом, до якого людство йшло століттями. Філософи від Декарта до Канта наголошували на важливості сумніву та самоперевірки в процесі пізнання. Німецький мислитель І. Кант, зокрема, у «Критиці чистого розуму» вимагав піддавати розум критицизму, аби окреслити межі достовірного знання [2]. У XX ст. К. Поппер розвинув ідею фальсифікаціонізму, за якою наукове знання просувається шляхом спростування помилок [6] – також прояв критичного підходу. Таким чином, критичне мислення можна розглядати як історично сформований канон інтелектуальної дисципліни, котрий поєднує формальну строгість логіки з емпіричним контролем і самокорекцією.

Чи можна алгоритмічно відтворити такі функції? Принаймні частково – так. Формальні аспекти раціональності піддаються алгоритмізації: правила логіки можна закодувати у програмі [8; 10]. Власне, перші системи ШІ (у 1950-х–1960-х роках) були символьними і оперували логічними правилами. Програма Logic Theorist (1956) доводила теореми, а мова Prolog надалі стала класичним інструментом логічного програмування. Проте повністю замінити людське критичне мислення суто формальною логікою не вдалося. Логічні експертні системи виявилися негнучкими: вони вимагали повної бази фактів і не могли самостійно навчатися новому [8]. Хвиля ентузіазму щодо нейронних мереж у 2010-х змістила акцент на підхід машинного навчання «без вчителя», де машина сама виявляє патерни в представленому наборі даних. Це дало вражаючі результати у розпізнаванні образів, генерації осмислених текстів, перекладі, але це відбулося за рахунок втрати прозорості процесу мислення машини. Глибокі нейромережі – «чорні скриньки», які можуть приховувати грубі помилки в логіці виводу. Наприклад, моделі без явно заданих арифметичних правил часом плутали елементарні арифметичні операції або причинно-наслідкові відношення, якщо таких прикладів не було в навчальних даних [13; 9; 7]. Це й спонукало науковців повернутися до ідеї дисциплінування ШІ за допомогою логіки та правил критичного мислення.

Таким чином, історія розвитку ШІ – це маятник між символічним та суб-символічним підходами. Перший підхід трактував інтелект як маніпулювання символами за строгими правилами (так званий Good Old-Fashioned AI, GOFAI). Другий, сучасний, бачить інтелект як статистичне узагальнення великих даних нейронними мережами. Сьогодні формується консенсус, що істинно сильний ШІ повинен поєднати обидва підходи. Концепція нейросимвольного

ШІ виходить саме з цієї ідеї: об'єднати потужність машинного навчання (інтуїтивне «мислення» на основі великих даних, аналог Системи 1 за Д. Канеманом [14]) із надійністю та прозорістю символічної логіки (повільне аналітичне мислення – Система 2 за Д. Канеманом [14]). Практичні результати вже є: сучасні моделі ШІ розробляються таким чином, що можуть викликати зовнішні інструменти (калькулятори, бази знань) для перевірки своїх відповідей, виявляти логічні суперечності у тексті, будувати формальні моделі ситуацій [8; 15; 24]. Наприклад, алгоритми Google для задач математики навчилися генерувати математичні докази, комбінуючи нейромережу з модулем на зразок AlphaGo, що здійснює пошук рішень [12]. Вже у 2023–2024 рр. провідні лабораторії представили прототипи великих мовних моделей, здатних до багатокрокового логічного міркування. Компанія OpenAI восени 2024 р. випустила модель під назвою OpenAI o1, яку охарактеризовано як перший зразок «мислячого» ШІ. Модель o1 генерує прихований ланцюжок міркувань (chain-of-thought), перш ніж видати фінальну відповідь. Модель ніби «розмірковує про себе» в процесі пошуку відповіді, за рахунок цього точність відповідей суттєво зростає. Як повідомляє OpenAI, експериментальна версія o1 у складному математичному тесті вирішила 83% задач, тоді як звичайний GPT-4 – лише 13 %. Також o1 значно перевершив попередників у конкурентних програмах з програмування (89-й перцентиль на Codeforces) та запитаннях PhD-рівня з природничих наук [17]. Майже одночасно Google представив власну модель Gemini 2.0 Flash Thinking – експериментальну версію, оптимізовану під міркування. Вона використовує внутрішній «процес думання», що помітно поліпшує багатокрокове планування і логічні виводи. У тестах Gemini Flash Thinking розв'язувала складні логічні головоломки, комбінуючи аналіз зображень і тексту, та на 75 % зменшила кількість помилок у програмуванні порівняно з попередником [11]. Обидві компанії фактично рухаються в одному напрямі: їх моделі вчаться витрачати більше обчислювального часу на обдумування кожної відповіді. Як підкреслюють провідні розробники ШІ, нова модель «навчена використовувати процес мислення для посилення свого міркування», і при збільшенні часу на обчислення якість рішень помітно зростає. Це підтверджує стару істину: ретельний аналіз (хай і алгоритмічний) дає кращий результат, ніж миттєва реакція.

Утім, упровадження chain-of-thought автоматично не вирішує всіх проблем. Дослідники відзначають цікавий побічний ефект: якщо початковий крок міркування виявився хибним, довший ланцюжок може лише посилити помилку, зробивши кінцеву відповідь ще переконливішою у своїй неправильності. За аналогією з людиною, це схоже на раціоналізацію упередження – коли логіка використовується для того, щоб виправдати хибне інтуїтивне рішення. Перші тести «мислячих» моделей зафіксували поодинокі випадки, коли при складній задачі модель збивалася на неправильний шлях і впевнено його розвивала [11; 17]. Більше того, деякі оглядачі зауважили, що ранні версії reasoning-моделей навіть частіше галюцинували дрібні факти, надто зосередившись на власних міркуваннях. Розробники врахували це в оновленнях: тепер моделі на кшталт o1 вміють коригувати власні помилки на проміжних етапах [17]. Задіяно механізми, подібні до людського «подвійного перевіряння»: модель генерує кілька варіантів розв'язання задачі і звіряє їх між собою або з зовнішніми джерелами, перш ніж видати остаточну відповідь. Такий метакогнітивний елемент (перевірка себе) поступово зближує алгоритмічне виведення з людським критичним мисленням [5; 14].

Таким чином, при порівнянні людського критичного мислення й логічного мислення ШІ, попри значний прогрес, між мисленням людини та алгоритму все ще існує принципова різниця. Людина мислить не тільки логічно, а й змістовно: наші судження занурені в контекст життєвого досвіду, вони мають мотивацію (прагнення до істини, практичну зацікавленість), і головне – ми усвідомлюємо сенс своїх думок [2; 5]. ШІ-модель наразі оперує символами

без справжнього розуміння їх значення. Цей аргумент відомий у філософії як парадокс «китайської кімнати» Дж. Серля: система може маніпулювати символами за правилами синтаксису, але не мати семантичного доступу до значень [21]. Так само і сучасний ШІ: навіть коли він «міркує» логічно, це відбувається на рівні формальних операцій, а не осмислення. Наприклад, модель може виявити формальну суперечність між двома твердженнями, якщо її навчено правилам логіки, але вона не «відчуває» абсурдності змісту, як це зробила б людина. Відомий теоретик причинності Дж. Перл зауважив, що нинішнім нейромережам бракує розуміння причинно-наслідкових зв'язків, вони оперують лише кореляціями [18; 19]. Людина ж, пізнаючи світ, будує його причинну модель і тим самим досягає глибшого розуміння. Відповідно, без засвоєння принципу причинності ШІ навряд чи буде критично оцінювати свої висновки – адже одне з питань критичного мислення: «Чому ми вважаємо це істинним, які причини та наслідки?» – поки для машин закрите [18].

Ще одна суттєва відмінність – наявність свідомості та особистої відповідальності. Людина несе епістемічну відповідальність за свої твердження: якщо я помиляюся, це впливає на мій подальший психологічний стан (я можу відчувати сумнів, сором, прагнути виправити помилку). ШІ не має таких переживань. Він не «хвилюється» про істинність – його задача задана ззовні (максимізувати точність відповіді чи догодити користувачу), що добре видно з публічних застережень розробників про те, що моделі іноді «вигадують, щоб сподобатися» [7]. Тому моделі типу ChatGPT раніше схильні були догоджати співрозмовнику навіть ціною правди. Хоча нові «мислячі» версії частково подолали цей синдром, та повної автономної мотивації до істини у них все ж немає. Можна сказати, що критичність залишається зовнішньою щодо ШІ: ми вбудовуємо в нього правила і фільтри, але він не усвідомлює потреби в них. Проте, «Найкращим інструментом для цього є критичне мислення, яке дає змогу знаходити компроміси для успішного розв'язання проблем і пошуку ефективних рішень. Водночас несвідоме використання штучного інтелекту може зашкодити людині й обмежити її розумовий розвиток» [4, с. 45].

Водночас, деякі аспекти людського мислення машини виконують краще. Наприклад, комп'ютер не втомлюється перевіряти тисячі варіантів, не має несвідомих упереджень, притаманних нам (скажімо, упередження підтвердження – confirmation bias – у алгоритму виникає лише якщо він навчений на однобічних даних) [9; 13]. У рутинних логічних операціях машина перевершує людину швидкістю й точністю. Це підказує оптимальну стратегію на сьогодні: симбіоз людини і ШІ. Людина формулює цілі, інтерпретує результати, робить висновки про сенс і значення, тоді як ШІ може слугувати потужним інструментом аналізу, перевірки та генерування гіпотез. Уже з'являються приклади такої синергії: науковці використовують великі мовні моделі для попереднього огляду літератури чи генерування можливих рішень, але потім критично оцінюють ці пропозиції, відкидаючи хибні [15]. Виходить своєрідний тандем, де ШІ – «логічний фільтр» і помічник, а остаточний арбітр істини – людина [14]. «Водночас штучний інтелект розвивається значно швидше, ніж людське мислення. Це має спонукати сучасну людину працювати над власним мисленням і постійно самовдосконалюватися» [4, с. 41].

Отже, незважаючи на всі досягнення, межа між машинним і людським мисленням поки що не стерта. Спільним між ними є здатність дотримуватися формальних правил: сучасні моделі значно знизили частоту елементарних логічних помилок, вони можуть уникати суперечностей, слідкувати за структурою аргументу. Але відмінним залишається джерело цієї здатності. Для людини правила логіки – це усвідомлена норма, частина культури мислення, яку вона приймає внутрішньо. Для ШІ – це зовнішня програма, якій він слідує, не знаючи «навіщо». Внаслідок цього людина залишається носієм семантичної інтуїції і творчості, здатною вийти за межі

правил, коли потрібно (наприклад, винайти нову концептуальну схему при науковому прориві). ШІ ж діє в рамках заданого формалізму і не має власної ініціативи щось змінювати. Втім, як було показано, крок за кроком машини запозичують дедалі більше «людських» прийомів мислення. Ймовірно, майбутні моделі зможуть враховувати контекст і здоровий глузд краще, ніж нинішні – вже зараз ведуться роботи з інтеграції в ШІ онтологічних баз знань, що втілюють базові уявлення про реальний світ: «…вже зараз ведуться роботи з інтеграції в ШІ онтологічних баз знань, що втілюють базові уявлення про реальний світ. Українські дослідження цифровізації освіти також розглядають ШІ як органічну частину цього комплексного процесу, що поєднує великі дані, хмарні сервіси та машинне навчання» [3, с. 29]. Так, у проєктах типу ConceptNet закладено тисячі тривіальних фактів (на кшталт «вода мокра», «вогонь обпікає»), які формують базу для здорового глузду. Коли мовні моделі почнуть послуговуватися такими онтологіями, їх відповіді стануть змістовнішими і менш абсурдними. Але повністю людську гнучкість – здатність не тільки застосовувати знання, а й переосмислювати їх, вчитися новим концептам – поки що жодна програма не демонструє [14]. Крім того, побудова таких онтологічних баз знань має першочергове значення і при формуванні світогляду підростаючого покоління через систему освіти: «Цифровізація в контексті вищої освіти почала розглядатися як комплексний процес, що охоплює інноваційні педагогічні методи, використання великих даних (ВД), хмарних сервісів, штучного інтелекту (ШІ) та інструментів машинного навчання, а також розвиток цифрової культури в академічному середовищі» [3, с. 29].

Тому розбудова критично-мислячого ШІ підіймає низку глибинних питань філософського і етичного характеру. Насамперед, це питання про природу істини в контексті штучних систем. Якщо ШІ здатен оперувати істинними твердженнями, перевіряти їх на несуперечливість і навіть аргументувати – чи можна сказати, що він «знає» ці істини? З точки зору класичної епістемології, знання потребує не лише фактичної істинності, а й суб'єктивного обґрунтування (виправданої віри). У ШІ немає ні віри, ні розуміння – лише маніпуляція символами. Таким чином, епістемічний статус його відповідей залишається інструментальним: ці відповіді корисні для нас, але їх носій не є суб'єктом пізнання [6]. Це напряму перегукується з відомим питанням про співвідношення синтаксису і семантики («китайська кімната»): чи дорівнює виконання формальних правил володінню сенсом? – більшість дослідників поки що схиляється до відповіді «ні» [22]. Утім, прихильники функціоналістських підходів у філософії свідомості зауважують, що достатньо складна система, яка послідовно діє так, ніби розуміє, може вважатися носієм розуміння – ця дискусія про штучну свідомість виходить за межі нашого викладу, але стає щораз актуальнішою в міру «олюднення» поведінки ШІ.

Друге важливе питання – межі формальної логіки і алгоритмічного підходу. Логіка гарантує істинність висновків лише при заданні істинних засновків. Якщо ж вихідна інформація неповна або помилкова, то формально бездоганне виведення може привести до фактичної хиби. Це спостерігається і в роботі моделей: вони можуть абсолютно коректно з точки зору правил мови відповісти на запит, проте відповідь не буде істинною в реальності, бо модель не має всієї необхідної інформації або не розуміє прихованих припущень, що малися на увазі у питанні. Філософськи тут проявляється нездоланний для будь-якої формальної системи горизонт – теза, подібна до теореми Гьоделя про неповноту: у межах будь-якої достатньо виразної системи знайдуться істинні твердження, які вона не зможе довести [23]. Для ШІ це означає, що завжди існуватимуть питання, де він «застрягне» або згенерує хибу, бо не матиме доступу до якогось знання чи контексту, який нам очевидний. Усвідомлення цього припущення підштовхує до ідеї гібридного інтелекту: поєднання штучного та людського. Поки машини не набули істинної інтенціональності, людина залишається вищою інстанцією перевірки та сенсоутворення.

ISSN 2786-7005 (Print) ISSN 2786-7013 (Online)

У перспективі дослідження в напрямі критичного ШІ мають не лише практичну а й теоретичну користь. Це свого роду експеримент із моделювання самого феномену мислення. Успіхи та невдачі на цьому шляху проливають світло на питання: що таке мислення? Якщо нам вдасться створити алгоритм, що аргументує, сумнівається, виправляє себе – чи не наблизимося ми до матеріалістичного розуміння розуму як обчислення? Якщо ж на якомусь етапі з'ясується, що без справжньої свідомості машина впирається у стелю можливостей, – це буде аргумент на користь особливої, нематеріальної природи мислення, як про те писали ще Декарт або Гуссерль. У будь-якому разі, філософія і ШІ нині вступають у продуктивний діалог. Філософські ідеї (наприклад, концепція Канемана про дві системи мислення або платонівське розрізнення істинного знання і думки) надихають інженерів на нові рішення. Натомість досягнення ШІ змушують філософів переглядати деякі класичні питання під новим кутом.

Висновки. Проведене дослідження підтверджує, що інтеграція критичного мислення і формальної логіки в системи штучного інтелекту є необхідною передумовою підвищення їх надійності та істинності. Галюцинації ШІ – впевнена генерація хибних, хоча і правдоподібних відповідей – становлять серйозну перешкоду на шляху широкого впровадження інтелектуальних технологій. Просте масштабування моделей не розв'язує цієї проблеми, оскільки вона коріниться у відсутності в машині механізмів перевірки істини. Тому науковці звернулися до досвіду людини: розвиток критичного мислення протягом історії показує, які для цього потрібні інструменти – логічні правила, рефлексія, самокорекція. Частину з них уже вдалося реалізувати в новітніх моделях. Зокрема, впровадження явного багатокрокового міркування (ланцюжків думок) у моделях OpenAI, Google Gemini продемонструвало суттєве підвищення точності відповідей на складні запитання. Моделі навчилися витрачати більше часу на обдумування і навіть застосовувати зовнішні інструменти, що зменшило кількість фактологічних помилок. Однак повноцінне критичне мислення ШІ ще не досягнуте. Нинішні системи все ще неконтрольовано галюцинують у відкритих доменах, не мають внутрішньої мотивації до істини і не розуміють зміст своїх знань.

Перспективним напрямом подальших досліджень є розвиток метакогніції в ШІ – здатності моделі оцінювати власні дії. Це включає механізми оцінки впевненості в своїх відповідях, виявлення потенційно сумнівних кроків міркування, повернення до попередніх етапів для виправлення. Також актуальним є збагачення моделей онтологічними знаннями і причинно-наслідковими уявленнями, щоб вони оперували не лише кореляціями, а й причинними моделями світу. У цій площині вже триває діалог між ШІ та працями з теорії причинності (наприклад, роботи Дж. Перла та ін.). Ще один важливий аспект – пояснюваність рішень ШІ. Моделі-«чорні ящики» мають навчитися пояснювати свої відповіді мовою логіки, зрозумілою людині. Це, знову ж, відсилає до ідеї, що ШІ повинен внутрішньо будувати хоч спрощені, але аргументи. Пояснюваність не лише підвищить довіру, а й сприятиме виявленню помилок: якщо модель зможе викласти свій «хід думок», людина або інша програма зможе проаналізувати його на правильність.

Нарешті, варто згадати етично-соціальний вимір проблеми. Навчити ШІ мислити критично – означає зробити його безпечнішим і кориснішим для суспільства. Системи, що дотримуються «логічної дисципліни», менше схильні до дезінформації і маніпуляцій. Це особливо важливо, враховуючи тенденцію інтеграції ШІ в освітні, правові, медійні процеси. Можна прогнозувати, що по мірі вдосконалення ШІ, роль людини зміститься з безпосереднього виконання рутинних інтелектуальних завдань до контролю та наставництва ШІ. Людина-викладач буде вчити машину, як вчила б студента: правилам, критичності, етики мислення. В цьому сенсі розвиток ШІ ставить перед нами дзеркало: щоб створити мислячий алгоритм, ми самі маємо чітко

усвідомити, як мислимо ми, що таке істина і логіка. Тож, займаючись «дисциплінуванням» штучного інтелекту, людство паралельно підвищує власну інтелектуальну культуру.

Отже, впровадження принципів критичного мислення і формальної логіки у ШІ є необхідною умовою для переходу до нової ери надійних інтелектуальних систем. Поєднання машинної потужності з людською раціональністю здатне дати якісно нові результати в науці, техніці, освіті. Штучний інтелект, позбавлений «дурних» помилок і прозорий у своїх міркуваннях, з цікавого експерименту поступово перетворюється на справді корисного помічника людства. Досягнення цієї мети вимагає ще чимало зусиль на перетині філософії, когнітивістики та комп'ютерних наук. Однак напрям окреслено вірно: навчити машини мислити правильно, аби вони могли доповнити і підсилити наше власне мислення, не спотворюючи його. Як підкреслюють сучасні дослідники, синергія людини і ШІ, побудована на засадах логіки та критичного мислення, може стати запорукою нових наукових відкриттів і технологічного прогресу. У процесі цього взаємного навчання – людини й машини – ми краще зрозуміємо і сам феномен розуму, що є одвічною темою філософії.

Список використаних джерел

1. Аристотель. Метафізика / пер. з давньогр. Київ: Темпора, 2022. 848 с.
2. Кант І. Критика чистого розуму / пер. з нім. Київ: Юніверс, 2000. 504 с.
3. Крулевський А. В. Формування стратегії цифровізації системи вищої освіти в Україні: дис. … д-ра філософії. Тернопіль, 2025. 291 с. URL: https://www.wunu.edu.ua/svr/disertatcia/2025/Krulevskyi/Dis_Krulevskyi.pdf (дата звернення: 08.10.2025).
4. Матвієнко І. Критичне мислення та штучний інтелект: сучасні можливості взаємодії. *Педагогіка, психологія, філософія.* 2025. Т. 13. № 2. URL: https://humstudios.com.ua/uk/journals/tom-13-2-2025/kritichne-mislennya-ta-shtuchny-intelekt-suchasni-mozhlivosti-vzayemodiyi (дата звернення: 07.10.2025).
5. Надурак В. В. Критичне мислення: поняття та практика. *Філософія освіти*. 2022. № 28 (2). С. 129–147. DOI: https://doi.org/10.31874/2309-1606-2022-28-2-7
6. Поппер К. Логіка наукового відкриття / пер. з англ. Київ: Основи, 1994. 432 с.
7. Altman S. Don't trust ChatGPT too much. URL: https://www.ndtv.com/world-news/dont-trust-that-much-openai-ceo-sam-altman-admits-chatgpt-can-be-wrong-8808530 (дата звернення: 07.10.2025).
8. Colelough B. C., Regli W. C. Neuro-Symbolic AI in 2024: A Systematic Review. CEUR-WS, 2024. URL: https://ceur-ws.org/Vol-3819/paper3.pdf (дата звернення: 09.10.2025).
9. Dang H. A., et al. Survey and Analysis of Hallucinations in Large Language Models. Frontiers in Artificial Intelligence, 2025. URL: https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1622292/full (дата звернення: 10.10.2025).
10. Delvecchio G. P., et al. Neuro-Symbolic AI: A Task-Directed Survey in the Black-Box Models Era. IJCAI 2025 Proceedings, 2025. URL: https://www.ijcai.org/proceedings/2025/1157.pdf (дата звернення: 10.10.2025).
11. Google. Gemini 2.0 Flash / Gemini thinking models (Vertex AI). 2025. URL: https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash (дата звернення: 10.10.2025).
12. Google DeepMind. AlphaGo. URL: https://deepmind.google/research/alphago/ (дата звернення: 11.10.2025).
13. Huang L., et al. A Survey on Hallucination in Large Language Models. arXiv:2311.05232, 2023. URL: https://arxiv.org/pdf/2311.05232 (дата звернення: 07.10.2025).
14. Kahneman D. Thinking, Fast and Slow. New York: Farrar, Straus and Giroux, 2011.
15. Liu F., et al. Self-Reflection Makes Large Language Models Safer, Less Biased, and Ideologically Neutral. arXiv:2406.10400, 2024. URL: https://arxiv.org/html/2406.10400v2 (дата звернення: 17.10.2025).
16. Nawaz U., et al. A review of neuro-symbolic AI integrating reasoning and learning for advanced cognitive systems. AI Open, 2025. DOI: https://doi.org/10.1016/j.iswa.2025.200541

17. OpenAI. Introducing OpenAI o1 / Learning to reason with LLMs. 12 Sept. 2024. URL: https://openai.com/index/learning-to-reason-with-llms/ (дата звернення: 09.10.2025).
18. Pearl J. Causality: Models, Reasoning, and Inference. 2nd ed. Cambridge: Cambridge University Press, 2009.
19. Pearl J., Mackenzie D. The Book of Why: The New Science of Cause and Effect. New York: Basic Books, 2018.
20. Sahoo P., et al. A Comprehensive Survey of Hallucination in Large Language Models. Findings of EMNLP 2024, 2024. URL: https://aclanthology.org/2024.findings-emnlp.685.pdf (дата звернення: 06.10.2025).
21. Searle J. R. Minds, Brains, and Programs // Behavioral and Brain Sciences. 1980. Vol. 3. № 3. P. 417–457.
22. Smith P. An Introduction to Gödel's Theorems. Cambridge: Cambridge University Press, 2007.
23. Wang Y., Zhao Y. Metacognitive Prompting Improves Understanding in Large Language Models. NAACL 2024, 2024. URL: https://aclanthology.org/2024.naacl-long.106.pdf (дата звернення: 12.10.2025).

References

1. Arustotel. (2022). Metafizyka [Metaphysics]. Kyiv: Tempora [in Ukrainian].
2. Kant I. (2000). Krytyka chystoho rozumu [Critique of pure reason]. Kyiv: Yunivers [in Ukrainian].
3. Krulevskyi A. V. (2025). Formuvannia stratehii tsyfrovizatsii systemy vyshchoi osvity v Ukraini [Formation of a digitalization strategy of the higher education system in Ukraine]. Dys. … d-ra filosofiyi. Ternopil′. URL: https://www.wunu.edu.ua/svr/disertatcia/2025/Krulevskyi/Dis_Krulevskyi.pdf (accessed: 08.10.2025) [in Ukrainian].
4. Matviienko I. (2025). Krytychne myslennia ta shtuchnyi intelekt: suchasni mozhlyvosti vzaiemodii [Critical thinking and artificial intelligence: Current possibilities of interaction]. Pedahohika, psykholohiia, filosofiia – Pedagogy, Psychology, Philosophy, 13 (2). URL: https://humstudios.com.ua/uk/journals/tom-13-2-2025/kritichne-mislennya-ta-shtuchny-intelekt-suchasni-mozhlivosti-vzayemodiyi (accessed: 07.10.2025) [in Ukrainian].
5. Nadurak V. V. (2022). Krytychne myslennia: poniattia ta praktyka [Critical thinking: Concept and practice]. Filosofiia osvity – Philosophy of Education, 28 (2), 129–147. DOI: https://doi.org/10.31874/2309-1606-2022-28-2-7 [in Ukrainian].
6. Popper K. (1994). Lohika naukovoho vidkryttia [The logic of scientific discovery]. Kyiv: Osnovy [in Ukrainian].
7. Altman S. (2025, July 1). Don't trust ChatGPT too much. URL: https://www.ndtv.com/world-news/dont-trust-that-much-openai-ceo-sam-altman-admits-chatgpt-can-be-wrong-8808530 (accessed: 07.10.2025).
8. Colelough B. C., & Regli W. C. (2024). Neuro-Symbolic AI in 2024: A systematic review. CEUR-WS. URL: https://ceur-ws.org/Vol-3819/paper3.pdf (accessed: 09.10.2025).
9. Dang H. A., et al. (2025). Survey and analysis of hallucinations in large language models. Frontiers in Artificial Intelligence. URL: https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1622292/full (accessed: 10.10.2025).
10. Delvecchio G. P., et al. (2025). Neuro-Symbolic AI: A task-directed survey in the black-box models era. IJCAI 2025 Proceedings. URL: https://www.ijcai.org/proceedings/2025/1157.pdf (accessed: 10.10.2025).
11. Google. (2025). Gemini 2.0 Flash / Gemini thinking models (Vertex AI). URL: https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash (accessed: 10.10.2025).
12. Google DeepMind. (n. d.). AlphaGo. URL: https://deepmind.google/research/alphago/ (accessed: 11.10.2025).
13. Huang L., et al. (2023). A survey on hallucination in large language models. arXiv:2311.05232. URL: https://arxiv.org/pdf/2311.05232 (accessed: 07.10.2025).
14. Kahneman D. (2011). Thinking, fast and slow. New York, NY: Farrar, Straus and Giroux.
15. Liu F., et al. (2024). Self-reflection makes large language models safer, less biased, and ideologically neutral. arXiv:2406.10400. URL: https://arxiv.org/html/2406.10400v2 (accessed: 17.10.2025).

16. Nawaz U., et al. (2025). A review of neuro-symbolic AI integrating reasoning and learning for advanced cognitive systems. AI Open. DOI: https://doi.org/10.1016/j.iswa.2025.200541
17. OpenAI. (2024, September 12). Introducing OpenAI o1 / Learning to reason with LLMs. URL: https://openai.com/index/learning-to-reason-with-llms/ (accessed: 09.10.2025).
18. Pearl J. (2009). Causality: Models, reasoning, and inference (2nd ed.). Cambridge: Cambridge University Press.
19. Pearl J., & Mackenzie D. (2018). The book of why: The new science of cause and effect. New York, NY: Basic Books.
20. Sahoo P., et al. (2024). A comprehensive survey of hallucination in large language models. Findings of EMNLP 2024. URL: https://aclanthology.org/2024.findings-emnlp.685.pdf (accessed: 06.10.2025).
21. Searle J. R. (1980). Minds, brains, and programs. Behavioral and Brain Sciences.
22. Smith P. (2007). An introduction to Gödel's theorems. Cambridge: Cambridge University Press.
23. Wang Y., & Zhao Y. (2024). Metacognitive prompting improves understanding in large language models. NAACL 2024. URL: https://aclanthology.org/2024.naacl-long.106.pdf (accessed: 12.10.2025).

Отримано 28.10.2025. *Received 28.10.2025.*
Прийнято до друку 21.11.2025. *Accepted for publication 21.11.2025.*
Опубліковано 18.12.2025. *Published 18.12.2025.*