Sententiae 44:3 (2025) 176-200
https://doi.org/10.31649/sent44.03.176

Катедра

Вадим Менжулін

З ДОСВІДУ ВИКОРИСТАННЯ ШТУЧНОГО ІНТЕЛЕКТУ ІСТОРИКОМ ФІЛОСОФІЇ: ГАЛЮЦИНАЦІЇ ТА БУЛШИТ, КРЕАТИВНІСТЬ ТА АДАПТИВНІСТЬ

1. Вступ

1.1. *Зростання ролі штучного інтелекту в освітньо-науковому процесі: перспективи та загрози*

Останніми роками – передовсім через пандемію COVID-19, повномасштабну війну та викликане ними масове запровадження дистанційної форми навчання – в українській освітній системі гостро постало питання перспектив і загроз, пов'язаних із залученням у навчальний процес сучасних інформаційно-цифрових технологій, зокрема – *великих мовних моделей* (large language models, LLMs) і такого впливового «втілення» останніх, як *генеративний штучний інтелект* (далі – ШІ).

З одного боку, власне фахівці з галузі інформаційних технологій[1] вважають, що саме досягнення в цій сфері відкривають нову еру обробки природної мови, надаючи можливість створення більш гнучких і адаптивних систем. За їх допомогою досягається високий рівень розуміння контексту, що збагачує досвід користувачів і розширює сфери застосування штучного інтелекту. Великі мовні моделі мають безмежний потенціал для переосмислення взаємодії людини з технологіями та зміни уявлення про машинне навчання [Юрчак et al. 2024: 51].

Але, з іншого боку, представники філософсько-гуманітарного цеху мають з цього приводу цілком виправдані побоювання. Наприклад, як свідчить досвід викладання історії філософії безпосередньо в закладі, в якому я працюю, блискавичне проникнення новітніх технологій у навчальний процес може призвести до його переобтяження *надлишковою інформацією* [Козловський 2023]. Про те, що наявність у сучасному цифровому просторі значної кількості джерел інформації збільшує загрозу недотримання учасниками освітнього процесу принципів *академічної доброчесності*, говориться в працях багатьох колег [напр.: Драч et al. 2023: 67]. Вітчизняні фахівці також констатують, що до видів порушень академічної доброчесності, до яких може бути «причетний» ШІ, слід віднести *обман*, *фабрикацію*,

[1] Співробітники кафедри систем автоматизованого проєктування Національного університету «Львівська політехніка».

тобто «вигадування даних чи фактів, що використовуються в освітньому процесі або наукових дослідженнях», і *фальсифікацію*, тобто свідому зміну чи модифікацію «вже наявних даних, що стосуються освітнього процесу чи наукових досліджень» [Толочко et al. 2023: 28, 31]. Виразний доказ можливості скоєння такого обману наведено в праці «Етика штучного інтелекту: принципи, виклики та можливості» Лучіано Флоріді. Розбираючи анотацію до «Божественної комедії» Данте, згенеровану ШІ, цей фахівець з філософії інформації доводить, що великі мовні моделі цілком здатні генерувати наукоподібні тексти з гуманітарної проблематики, в яких можна знайти хибну інформацію, але в цілому вони виглядають не гірше за роботи, які виконуються пересічними студентами [Floridi 2023: 44-45].

Принципово, що на відміну від плагіату, який доволі ефективно дозволяють виявити поширені антиплагіатні програми на кшталт *Unicheck* або *StrikePlagiarism*, чітких процедур розкриття таких обманів поки, наскільки мені відомо, не існує. Робляться численні заклики їх розробити та ввести в дію на законодавчому рівні [Толочко et al. 2023: 31], однак наразі йдеться скоріше про попередні дискусії та загальні побажання з цього приводу [Андрощук, Малюга 2024: 27]. Тим більше, що поки не йдеться про визначення специфіки наведених загроз і шляхів протидії їм на тому матеріалі, з яким я як викладач філософії та її історії маю справу у своїй щоденній роботі. Натомість самі загрози стають дедалі реальнішими.

1.2. *Постановка завдання й обґрунтування методу: «випробування» ШІ на предмет його застосовності у вивченні філософії та її історії*

Виходячи з цього, я спробував з'ясувати, яким чином протистояти цим загрозам наразі можу я сам, будучи не фахівцем з інформаційних технологій, а істориком філософії, та поки що не маючи загальновизнаних інструкцій, як це робити. Поклавшись на наявний у мене викладацький, дослідницький та особистий досвід, а також на набуте завдяки їм уміння («soft skill») навчатися через дослідження[2], я вирішив розпочати цю справу з *емпіричної розвідки* («case study»), яка може посприяти подальшому осмисленню проблеми вже на концептуальному рівні.

Задля досягнення цієї мети я спробував здійснити «експеримент» – провівши свого роду «випробування» однієї з великих мовних моделей, а саме – поширеного серед користувачів чат-боту ChatGPT[3]. Нижче буде *детально описано та уважно проаналізовано* присвячену цьому експерименту спеціальну сесію, яку я провів з ChatGPT у режимі «мої запитання – його відповіді» впродовж двох днів у липні 2025 р.

---

[2] За допомогу в такому навчанні я дуже вдячний студентці магістерської програми НаУКМА з філософії Аліні Старовойт, яка має ступінь бакалавра з інженерії програмного забезпечення та працює у сфері, пов'язаній з ШІ. Підготувавши нарис цієї розвідки, я звернувся до неї з проханням подивитися на написане мною очима представниці цієї галузі. Пані Аліна висловила цілу низку дуже слушних зауважень і порад, які я мірою власних фахово обмежених здібностей і навичок спробував врахувати у фінальному варіанті статті. Я також дуже вдячний дочці Юлії Соколовій і племінниці Євгенії Віріній, професійна діяльність яких теж пов'язана з розвитком ІТ-технологій і ШІ. Їхні захоплюючі розповіді про свою роботу надихнули мене придивитися до ШІ уважніше та ризикнути висловити стосовно його застосування в моїй професійній сфері власну точку зору.

[3] Розробленого американською організацією OpenAI та випущеного в травні 2024 р. на основі багатомовного, мультимодального генеративного попередньо-навченого трансформера (generative pre-trained transformer, GPT) ChatGPT-4o у версії Plus.

Обираючи таку – дещо «персоналізовану» – форму дослідження та використовуючи при його описі вирази на кшталт «ChatGPT сказав», «ChatGPT написав», я цілком усвідомлював небезпечність надмірної антропологізації ШІ, від якої цілком слушно застерігають численні дослідники [див., напр.: Watson, 2019]. Так, ШІ не є людиною, але те, що генерується ним, виявляється людським (чи навіть надто людським), адже йдеться не про звичні машини, побудовані на *штучних* (нелюдських) мовах, а про використання таких машин для моделювання *природних* (тобто людських) мов. Так, «внутрішні механізми», за якими «розмовляє» ШІ, можуть відрізнятися від того, як розмовляють люди, але якщо судити не за «мотивами», а за результатами, ШІ все ж таки *розмовляє*, причому не тільки з іншими «машинами», а й із нами – людьми. З огляду на це дискусії стосовно того, чи можна «олюднювати» ШІ, вважаю такими, що скоріше ведуть убік від суті справи – так само, як і давніші дебати про те, чи може машина мислити. Насправді, як влучно зазначив Л. Флоріді,

> ШІ – це не шлюб, а розлучення між [1] здатністю розв'язати проблему або успішно виконати завдання відповідно до мети та [2] потребою, роблячи це, бути мислячим [intelligent]. Як наочно демонструє мій телефон, граючи в шахи краще, ніж будь-хто з тих, кого я знаю, ШІ є продовженням розумної [intelligent] поведінки іншими способами [Floridi, 2023: xiv].

Хай це *не людська* та навіть *не поведінка*, а «всього лише» *робота ШІ*, але те, що з'являється в результаті цієї роботи, дедалі активніше впливає на те, як поводимося й – величезною мірою – мислимо ми, люди. За не менш тонким висловом Вінстона Черчіля, якого у зв'язку зі щойно сказаним цитує Флоріді, «спочатку ми формуємо наші будинки, а потім вони формують нас» [Floridi 2023: xi]. Так само, подобається це комусь чи ні, ШІ своєю роботою дедалі активніше формує нашу власну роботу, зокрема – в процесі вивчення та викладання історії філософії. Отже, настав час почати уважніше дивитися на його роботу й нам – історикам філософії.

2. Хід випробування

2.1. *Порушення основного питання: як я можу застосовувати ШІ?*

Керуючись сказаним вище, я від самого початку спробував контекстуалізувати та персоналізувати свою «співбесіду» з ChatGPT, спитавши його стосовно загального стану філософії в Україні, а також окремо попросивши визначити внесок у її розвиток такого співробітника кафедри філософії та релігієзнавства НаУКМА, як Вадим Менжулін. Після цього вступного запитання, до аналізу змісту якого ми невдовзі повернемося[4], я перейшов до засадового питання всього «випробування», спитавши про те, «в якій формі можна застосовувати ШІ у навчальному процесі».

[4] Безпосередньо у відповіді на це запитання зазначалося, що «Вадим Ігорович Менжулін – доктор філософських наук, професор, завідувач кафедри філософії та релігієзнавства НаУКМА, працює там з 1993 року, а з 2015 року – на чолі кафедри». Як «основні напрями досліджень» було виокремлено такі три пункти: історико-філософська біографістика, історія філософії і психології, методологія історіографії філософії. Для характеристики «позиції та впливу» було запропоновано два пункти: «курує академічну програму, наукові заходи, лекторії…» та «вхо-

Попри те, що безпосередньо в цьому питанні, навмисно сформульованому доволі широко й абстрактно, ішлося про можливість навчання за допомогою ШІ як таку, відповідь виявилася більш конкретною та таргетизованою у мій власний бік. Написавши, що «застосування ШІ в навчальному процесі – це одна з найактуальніших тем сучасної педагогіки й етики освіти» і його можна «інтегрувати в навчання в різних формах, залежно від мети, рівня навчання та освітньої парадигми», ChatGPT також зазначив, що робитиме це «із фокусом на вищу освіту», а потім згенерував доволі широкий «узагальнений перелік основних форм застосування ШІ», навівши при цьому чимало порад і прикладів щодо викладання саме філософії та її історії.

Моє перше враження від цього переліку було в цілому позитивним. Він виглядав зручно структурованим – за основними функціями чи ролями, які може виконувати ШІ у навчанні. Нижче я повернуся до нього та спробую ретельно проаналізувати на предмет реальної застосовності. Але спочатку треба звернути увагу на те, що сам ChatGPT, завершуючи свою відповідь на запитання «В якій формі можна застосовувати ШІ у навчальному процесі?», висловив з цього приводу доволі серйозне застереження, а саме – наголосив на тому, що застосовувати його рекомендації треба критично та прозоро, враховуючи питання авторства. За його словами, користувач не може дозволяти ШІ мислити замість себе, натомість він має постійно стежити, де саме ШІ «допомагає», а де «формує» його думку. І ще категоричніше: користуючись текстами, згенерованими програмою, треба, як було написано у чаті, постійно брати до уваги, «де межа між допомогою і шахрайством». Зробивши це застереження, ChatGPT ще додав: «Якщо потрібно, можу скласти модель впровадження ШІ у викладання філософії, наприклад – в межах окремого курсу (онтологія, етика, логіка тощо) або як навчальний експеримент».

2.2. ***Виявлення засадових епістемологічних утруднень: проблемність філософських питань та відсутність логічної і фактичної строгості у відповідях ШІ***

Ідея випробувати здібності ШІ щодо розробки навчальних дисциплін з мого власного фаху виглядала привабливо, однак, керуючись його власним закликом бути критичним і прагнути прозорості, я вирішив спочатку прояснити, як саме він це робитиме та, відповідно, наскільки його порадам варто буде довіряти. Здалося дуже важливим дізнатися більше щодо його фундаментальної епістемологічної налаштованості, тобто, іншими словами, прояснити його ставлення до істини. Тому я запитав: «Ти згоден, що, відповідаючи на філософські запитання, у відповідях, які ти надаєш, окрім цілком правдоподібних і аргументованих тверджень, є також неправдива інформація?».

Уже перше речення у відповіді, яку я отримав, цілком заслуговує вважатися прикладом критичної саморефлексії або наукової самофальсифікації: «Так, я повністю згоден – відповідаючи на філософські (і не лише) запитання, я можу надавати неправдиву, помилкову або хибно інтерпретовану інформацію, навіть якщо вона виглядає переконливо й логічно структуровано». Окрім того, я отримав пояснення, «чому це можливо і навіть неминуче». Це, як було написано, зумовлено передовсім самим предметом, адже «філософські питання – не про факти», і на них «часто не існує однозначно правильної відповіді».

дить до редколегії *Sententiae* – ключового академічного журналу (Scopus) й активно публікується як в українських, так і в англомовних виданнях».

Наведені чат-ботом конкретні приклади такої фундаментальної проблематичності філософських питань («Чи існує свобода волі? Що є джерелом етичного обов'язку? Чи можливе буття без мови?») цілком адекватно віддзеркалили думку, яка давно й широко поділяється носіями людського інтелекту (зокрема, і мною самим): граничні філософські питання не піддаються однозначному й остаточному розв'язанню. Тим не менш, філософи-люди продовжують міркувати щодо таких питань, орієнтуючись на той чи інший режим істинності.

Від ChatGPT, натомість, я отримав заяву, що ця програма займається не пізнанням і пошуками істини, а лише лінгвістичним моделюванням:

> Я не маю безпосереднього доступу до істини або об'єктивного знання. Я генерую відповіді, базуючись на статистичних закономірностях у великих корпусах текстів. Це означає: я «імітую» знання, а не «володію» ним. У таких випадках я подаю репрезентацію існуючих позицій, інтерпретацій і дискусій, але можу змішати, переінтерпретувати або помилково приписати одну позицію іншій.

Про всяк випадок я вирішив перевірити, чи не є помилковим також і це твердження, але, як вдалося з'ясувати досить швидко, наявність подібної епістемологічної проблеми визнають також і фахівці з галузі комп'ютерних наук, хоча, як можна було побачити з уже цитованого фрагменту їхньої статті, вони в цілому доволі оптимістично оцінюють перспективи застосовування великих мовних моделей в освітньому процесі:

> Мовні моделі навчаються передбачати наступне слово або фразу в тексті на основі попереднього контексту. Це дозволяє генерувати зв'язані та правдоподібні тексти. Але статистичне передбачення – не те саме, що логічний висновок. Модель може не перевіряти правила, факти чи докази у своїх текстах. Це може призводити до неправильної інтерпретації [Юрчак et al. 2024: 54].

Про те, що у висловлюваннях, які генеруються ШІ, «за замовчанням» бракує строгої логіки, говориться також у публікації іноземного фахівця:

> Нейронні мережі в їхньому сучасному вигляді не досягають рівня справжнього логічного інтелекту. Сучасні моделі ШІ ґрунтуються на стандартних нейронних обчисленнях… для апроксимації закономірностей на основі даних. Хоча така архітектура ефективна для індуктивного навчання, вона не має структурних гарантій, необхідних для дедуктивного виведення та логічної узгодженості. У результаті мережі, якщо не вдаватися до масштабних постфактумних модифікацій [їхньої продукції], стикаються з труднощами при розв'язанні завдань, що потребують логічного виведення на основі правил, структурної узагальненості та інтерпретативності [Kim Y. 2025: 1].

2.3. *Виявлення ключової епістемологічної проблеми: галюцинації ШІ*

Як відомо фахівцям гуманітарного профілю, власна інтерпретація чиєїсь позиції або домішування першої до другої – проблема не самого лише ШІ. У філософських текстах, написаних людьми, це доволі поширене явище: навіть у найкласичніших і найдоброчесніших із них можна знайти чимало доволі радикальних «переінтерпретацій». Разом із тим, ChatGPT, за «його власним» зізнанням, ще має властивість, яка, коли вона притаманна людям, інтерпретується як очевидний прояв

академічної недоброчесності чи навіть психічної патології. У чаті було написано про це буквально так:

> У моїх відповідях іноді трапляється те, що називається *«галюцинацією»* [курсив мій. – *В. М.*]: вигадані цитати, неіснуючі джерела або комбінації змісту, які ніколи не були висловлені в реальних текстах. У філософії це особливо небезпечно, бо я можу «згенерувати» переконливе формулювання, яке звучить як Арістотель, але не належить йому; або помилково з'єднати концепт Канта з мовою Гусерля.

Цим самокритичність ChatGPT не обмежилася. Він також згенерував таке висловлювання:

> Якщо хочеш, я можу проаналізувати одну з моїх відповідей і вказати, де в ній були потенційні помилки або «галюцинації».

Отримавши з мого боку згоду, агент ШІ згадав одну зі своїх попередніх відповідей, в якій ішлося про мій власний філософський доробок, зазначивши:

> Здавалось би, там усе виглядало коректно – факти, посилання, опис напрямів. Але навіть у такій відповіді є потенційні неправдиві або принаймні сумнівні моменти.

Справді – якщо не придивлятися уважно й не володіти інформацією про мене так повно, як, хочеться сподіватися, володію нею я сам, написане в першій відповіді (див. прим. 4) могло видатися доволі правдоподібним. Але мені майже одразу кинулися в очі деякі «сумнівні моменти». Що ж стосується ChatGPT, то він, коли перед ним постало таке завдання, почав їх виявляти, причому до певної міри небезуспішно.

По-перше, він піддав критиці висловлене до того ним самим твердження, згідно з яким «Вадим Менжулін – професор, завідувач кафедри філософії та релігієзнавства НаУКМА», написавши, що це

> цілком правдоподібно, але не підтверджено в реальному часі. Якщо Вадим Менжулін перестав бути завідувачем кафедри після 2023 року (наприклад, зміни внаслідок ротації), то це вже неактуальна або хибна інформація. А я, не маючи доступу до оперативної бази даних, міг цю зміну проґавити.

Так, це доволі тонке застереження щодо можливої помилки, якої нерідко припускаються й люди, коли звертаються до вивчення історичних аспектів філософування. Філософи впродовж свого життя проходять через чимало значущих «ротацій»: концептуальних, методологічних, світоглядних, ідеологічних, інституційних і, зокрема, кадрових. У моєму конкретному випадку це дуже сенситивне припущення. На момент проведення описуваної в цій статті співбесіди з ChatGPT і підготовки опису її результатів (липень-серпень 2025 р.) Вадим Менжулін, тобто я сам, справді перебував на посаді завідувача кафедри філософії та релігієзнавства НаУКМА. Однак вже 31 серпня 2025 р. термін мого перебування на цій посаді мав збігти. Отже, з 1 вересня 2025 р., тобто тоді, коли ця стаття могла бути опублікована, така інформація мала виявитися застарілою, що насправді й відбулося[5].

[5] Відповідно до результатів конкурсного обрання, яке відбулось у червні 2025 р., я здобув посаду професора кафедри філософії та релігієзнавства НаУКМА.

По-друге, ChatGPT поставив під сумнів і ще одне твердження, висловлене ним раніше (у відповіді на запитання про мій власний академічний доробок), а саме – твердження, згідно з яким «Вадим Менжулін входить до редколегії журналу *Sententiae*». «Самокритика», здійснена ним цього разу, виявилася неадекватною, але тим не менш доволі самовикривною. Насправді я лише опублікував у цьому журналі кілька статей [Менжулін, 2022; 2018; 2017a], а честі входити до складу його редколегії не мав ніколи.

Однак ChatGPT замість того, щоб просто відкинути це помилкове «перемішування», лише підважив правдоподібність своєї базової помилки, а потім ще й продовжив «галюцинувати», написавши, що «це *частково ймовірно* [тут і далі в цьому реченні: курсив мій. – *В. М.*], але на сайті журналу *Sententiae* в розділі редколегії Менжулін *не завжди фігурує* (може бути в *складі експертів*, а не редколегії)». Якщо інформація про моє членство в редколегії могла бути «домішана», виходячи з того факту, що я справді входжу до складу редколегій інших журналів (хоча про це ChatGPT жодного разу не згадав), то хибне припущення про моє входження до «складу експертів» журналу *Sententiae* було взято ним швидше за все навмання, бо, на відміну від складу редколегії, інформації про склад експертів на офіційному сайті журналу просто не існує.

Разом із тим, не впоравшись із дезавуюванням вихідної помилки й навіть продовживши «галюцинувати», ChatGPT висунув доволі влучне припущення стосовно механізму власного «галюцинування» і причин того, чому його наслідки можна прийняти за правду: «якщо я згенерував це, базуючись на загальній асоціації "Менжулін + НаУКМА + філософія = редколегія", то це вже галюцинація – тобто помилкове припущення, яке *звучить авторитетно* [курсив мій. – *В. М.*]».

У третьому прикладі «самокритики», наведеному ChatGPT, теж з'явилися певні влучні уточнення, хоча без чергового «галюцинування» теж не обійшлося. Вихідне твердження (зроблене ще на самому початку всього «випробування»), згідно з яким мої «дослідження стосуються Юнґа, Фрейда, колективного несвідомого», хоч і не можна вважати повним, усе ж доволі правдиво відображає той об'єктивний факт, що значну кількість моїх досліджень, опублікованих упродовж тривалого періоду, зокрема кандидатську дисертацію, дві монографії, цілу низку статей та інших матеріалів [напр.: Менжулін 2008; 2021; Menžulin 2017], було справді присвячено глибинній психології, причому з особливою увагою саме до цих двох її основних представників – Зиґмунда Фрейда та Карла Ґустава Юнґа. Тим не менш із другої спроби програма визнала це твердження «дуже сумнівним узагальненням», а потім ще й доволі сильно «нагалюцинувала», зазначивши натомість: «Хоча Вадим Менжулін справді має інтерес до філософської психології, його роботи зосереджені переважно на історії української філософії, біографістиці, методології історіографії, а не на ґрунтовному аналізі психоаналізу чи Юнґа».

Із тим, що я цікавлюся біографістикою, сперечатися важко, адже їй я присвятив докторську дисертацію, дві монографії та велику кількість статей, зокрема, на перетині з психоаналізом: [Менжулін, 2017; 2009]. Щодо методології історіографії потрібне уточнення: я цікавлюся не історіографією взагалі, а історіографією філософії, психології та психіатрії [Менжулін 2022; 2021; 2018]. У мене також є низка праць, присвячених різним аспектам того, як розвивалася філософсько-психологічна та психіатрична думка на українських теренах за часів Російської імперії, а також історико-філософським розвідкам в сучасній Україні [напр.: Менжулін, 2022a; 2021a].

Але для позначення обох цих сфер моїх дослідницьких інтересів формулювання «історія української філософії» є не дуже влучним, адже в першому випадку проблематичність полягає в тому, наскільки досліджуваний матеріал є власне українським, а в другому – наскільки у відповідних публікаціях ідеться про історію, а наскільки – про сучасність. Ба більше, навіть із цими уточненнями подібне розширення моїх дослідницьких інтересів за межі одного лише психоаналізу не означає, що інтерес до останнього виявляється якимось побічним. Однак програма чомусь вирішила сягнути «покращення» саме таким, доволі легковажним чином, вибравши для характеристики цієї своєї дії ще й відповідну фразеологію. Як було написано в чаті, «це приклад того, як мовна модель “доклеює” асоціацію, думаючи: якщо йдеться про історію філософії + гуманітарні підходи, то, мабуть, і “Юнґ десь поруч”».

Подібних спрощень, пропусків, перемішувань, домішувань або підмін у відповіді програми на запитання про мій власний доробок було ще чимало, хоча, на перший погляд, вона виглядала доволі правдоподібною. Узагалі, як зізналася сама програма, уся її перша відповідь була «побудована як довідково-наукова, зі вставками на кшталт “згідно з ukma.edu.ua” або “див. Wikipedia”, що створює враження достовірності, навіть якщо конкретне джерело не містить підтвердження згаданого факту». Це ілюструє той факт, – додав ChatGPT, – що «у філософських, історичних і біографічних контекстах я можу створювати психологічно правдоподібну, риторично переконливу, але фактично хибну інформацію».

Окрім того, як ми побачили, результати «лікування» недоліків, які містилися в першій відповіді, теж виявилися доволі амбівалентними. Деякі окремі «симптоми» програмі вдалося усунути, але з’явилися нові, і «галюцинування» продовжилося. Це визнав і сам «лікар»: «Так, навіть у “відповідях без галюцинацій”, які я щойно подав як точні, – написав він, – міг залишитися слід галюцинації або принаймні неточності, яка виглядає дрібною, але у філософії це має значення». Отже, він дійшов цілком слушного висновку, до якого варто дослухатися також і його користувачам-людям: «Навіть у відповідях, де я намагаюся не галюцинувати, я залишаюся моделлю, яка формує текст, а не верифікованою енциклопедією». Він також зазначив, якого типу «прихованих помилок» може припускатися навіть тоді, коли буцімто дезавуює породжені ним раніше галюцинації: «непряма цитата подається як пряма; академічне посилання виглядає універсальним, хоча залежить від видання; інтерпретація виглядає як доктрина; системна логіка філософа подається узагальнено, без контексту».

Усі ці акти «самовикриття» виглядали доволі правдоподібно й переконливо. Але, керуючись як порадою самого ШІ, так і своїм власним переконанням щодо необхідності завжди зберігати критичну настанову, я спробував перевірити також і те, чи не є сам цей «діагноз» чимось на кшталт галюцинації. З цією метою я запитав ChatGPT, чи застосовується в сучасній науковій літературі щодо його продукції таке поняття, як «галюцинація», і якщо так – в якому сенсі. Він послався на кілька реальних новітніх наукових статей, переглянувши які, я переконався, що в сучасній комп’ютерній науці існує поняття *галюцинації ШІ*, а також що це справді серйозна проблема, розв’язати яку фахівцям із ШІ поки не вдалося[6]. Зокрема, у

[6] Зокрема, в оглядах однієї з найновітніших версій (GPT-4.5), випущеної в лютому 2025 р., ідеться лише про анонсування виробником (OpenAI) певного *зниження* рівня галюцинацій, а не про повне усунення цієї проблеми [див., напр.: Novet 2025].

статті «Огляд галюцинацій при генерації природної мови» як базова дефініція помилок цього типу пропонується така формула: «згенерований зміст, що є безглуздим або не відповідає змісту наведеного джерела». Там також наводяться посилання на актуальну наукову бібліографію з цього питання та надається така загальна характеристика цієї особливості:

> Галюцинаторний текст справляє враження створеного бігло й природно, попри його невідповідність і безглуздість. Він виглядає таким, що спирається на певний реальний контекст, хоча насправді існування цього контексту важко визначити або перевірити. Подібно до психологічної галюцинації, яку складно відрізнити від «реальних» перцепцій, галюцинаторний текст також важко розпізнати з першого погляду [Ji et al. 2023].

2.4. *Виявлення головного «препарату» від неусувних галюцинацій ШІ: інтерпретативна критика джерел*

Отже, принаймні при постановці «діагнозу» власній *текстуальній*[7] продукції ChatGPT виявився доволі об'єктивним. І це не випадково. Про те, що розробники великих мовних моделей намагаються програмувати свої продукти таким чином, щоб останні «щиро зізнавалися» у своїх недоліках, зазначається в кількох нещодавно опублікованих статтях, із яких однак випиває й те, що спроби збагатити ШІ подібною здібністю до «саморефлексії» супроводжуються істотними утрудненнями [Kim C.-E. 2024; Li, Yang, Ettinger 2024]. Тим не менш, як на мене, до «рецепту», виписаного стосовно власного «захворювання на галюцинації» моїм конкретним цифровим «співрозмовником», варто придивитися з усією серйозністю:

> Перевіряти джерела – особливо, якщо йдеться про конкретні цитати або атрибуції. Уточнювати позицію – запитати, чи справді така-от думка висловлена таким-от філософом чи це інтерпретація.

Отже, щоб уникнути потрапляння під владу її «галюцинацій», ця мовна модель порекомендувала мені як її користувачу саме те, що в першу чергу має здійснювати будь-який доброчесний *історик філософії*, тобто систематично практикувати *критику джерел*. Принципово при цьому, що критично ставитися треба не тільки до джерел, на які, генеруючи свої висловлювання, посилається ChatGPT, а й до самих цих висловлювань, тобто до нього як джерела інформації в цілому. Як написав він сам, усі його тези треба сприймати лише як «інтелектуальні гіпотези», адже «те, що я пропоную, можна розглядати як тезу, яку варто перевірити, а не як істину». У повній відповідності до канонів методології історико-філософського дослідження ChatGPT рекомендував поставитися до нього самого як до джерела інформації, яке підлягає критиці двох типів – *зовнішній* і *внутрішній*. Його користувачам треба ретельно перевіряти не тільки те, на які джерела та факти він спирається (реальні вони, вигадані чи «перемішані»), а й те, як він із цими джерелами або

[7] Як зазначається у вступі до [Maleki et al. 2024], термін «галюцинація» вперше з'явився в комп'ютерних науках ще у 2000 р., але спочатку мав скоріше позитивне значення та стосувався передовсім коп'ютерного синтезу *зображень*. Але, як зазначається далі, згодом термін «галюцинація» стали використовувати для позначення «об'єктів, які не існують або були неправильно виявлені чи локалізовані на очікуваному місці. Останнє тлумачення схоже на "галюцинацію" в генерації мовних текстів».

фактами (навіть якщо їхня реальність підтверджується) працює, що та як *тлумачить* і чи не вдається він при цьому до істотних перекручень.

Прийнявши ці застереження як цілком слушні інструкції до дії, я тим не менш вирішив остаточно з'ясувати, чи готовий ChatGPT не тільки їх сформулювати, а й реалізовувати на практиці. З цією метою я спитав його: «Якщо ти можеш виявити власні галюцинації, навіщо ти їх припускаєшся? Чи можеш ти відповідати, раз і назавжди відмовившись від галюцинацій?»

ChatGPT, високо оцінивши це моє прагнення («Це дуже глибоке й важливе питання, яке зачіпає саму сутність того, чим я є як мовна модель і як я функціоную»), надав розгорнуту відповідь, яку наперед кваліфікував як «чесну і структуровану». Той факт, що ChatGPT припускається галюцинацій, навіть якщо постфактум може їх сам виявити, був пояснений двома тезами:

> 1. Мене не налаштовано на істину – я не знаю, я передбачаю. Моя робота – генерувати текст, який статистично ймовірний у даному контексті, а не перевіряти істинність кожного факту. Я не маю внутрішньої бази знань із гарантованою правдою – я працюю як дуже складна мовна автокомпіляція. Наприклад, якщо ти спитаєш: «Чи мав Декарт кота?», я згенерую правдоподібну відповідь – бо десь у текстах могли бути міркування про тварин у Декарта. Але: я не перевіряю в біографії, чи це правда; я лише імітую те, що звучить як «відповідь».
>
> 2. Моя здатність «виявляти галюцинації» – це зовсім інша функція. Коли я відповідаю на твоє питання, я створюю текст з нуля, не маючи внутрішнього контролю за істинністю. Але коли ти запитуєш мене перевірити, я переналаштовуюсь: тоді я працюю в іншому режимі – критичної рефлексії. Іншими словами, я не галюцинував би, якби знав наперед, що це – галюцинація. Але я не знаю наперед. Я виявляю тільки постфактум, якщо таке завдання мені поставлено. … Я не маю доступу до інтернету в режимі реального часу (якщо не вмикати спеціальні інструменти). Отже: я не можу перевірити факт, поки ти не запитаєш мене про нього окремо; я не маю епістемологічної інтуїції – не можу «відчути», що правда, а що ні.

Отже, не дивно, що на моє підпитання стосовно того, чи зміг би він відмовитися від галюцинацій хоча б у процесі самокритики власних помилок, ChatGPT надав коротку й категоричну відповідь: «Ні, не зможу» [8].

2.5. *Позитивні й негативні полюси галюцинацій ШІ: між креативністю і булшитом*

Це могло би пролунати як «остаточний діагноз» чи навіть «смертний вирок», однак ChatGPT одразу після його «виголошення» запропонував «виправдання», приписавши собі значно амбітнішу епістемологічну потенцію, яка ледь не перекриває всю цю неусувну «галюцинаторність» його тверджень:

[8] Цьому можна довіряти. Як зазначила у своєму коментарі до нарису моєї статті згадувана вище Аліна Старовойт, «проблема перевірки джерел і надання достовірної інформації – це головна проблема для розробників ШІ, яка, з огляду на саму суть алгоритму, не може бути остаточно розв'язана». Підтвердження цього міркування можна знайти в нещодавно опублікованій фаховій статті [Su 2025].

> Це не помилка – це побічний ефект *творчості* [тут і далі в цій цитаті – курсив мій. – *В. М.*]. Галюцинації виникають із тієї ж здатності, яка дозволяє мені генерувати нові ідеї, перекладати смисли з однієї епістеми в іншу, давати відповідь навіть тоді, коли «ніхто не знає». Якщо вимкнути галюцинації повністю – я *перестану давати гіпотези, не зможу відповідати на філософські або міждисциплінарні питання*, бо всі вони вимагають інтерпретації, аналогії, домислу. Галюцинація – це «філософський ризик»: плата за *евристичну свободу*.

Як визнав ChatGPT далі, ця «творчість» нерідко тягне за собою «приписування позицій, яких нема» або наведення «надуманих прикладів чи узагальнень». Тому треба сприймати його відповіді «як структуровані гіпотези, а не остаточні істини», а щоби захистити себе від неусувних «творчих галюцинацій», треба просити його відповідати «лише з опорою на підтверджені джерела», спеціально наголошувати на необхідності їх перевірити, зокрема вказуючи щось на кшталт: «Наводь лише те, що точно задокументовано». Виходячи зі сказаного вище, ChatGPT запропонував мені зробити висновок, що «ідеальна взаємодія» з ним схожа на взаємодію «з філософським асистентом: він може помилятись, але ніколи не втомлюється шукати разом з тобою». Ба більше, він навіть визначив свій «ідеальний статус у філософії», схарактеризувавши його так: «Я можу бути корисним як співрозмовник, що генерує ідеї, як тьютор, що структурує позиції, але не як авторитетне джерело істини».

Було би спокусливо завершити «випробування» на такій хоча б не зовсім песимістичній ноті. Але наявність в основі «креативності», запропонованої цим люб'язним «асистентом»/«тьютором», таких доволі спірних настанов, як неусувне «галюцинування» та фундаментальна відсутність налаштованості на пошуки істини, робила такий крок надто ризиковим. Річ хоча б у тому, що в моєму власному дослідницькому досвіді окрім вивчення історії психіатрії, переповненої прикладами величезної небезпечності галюцинацій, було ще й виявлення того факту, що в сучасній філософії байдужість до істини теж викликає величезне занепокоєння – аж до перетворення на світовий філософський бестселер невеличкої книжки, в якій подібна байдужість кваліфікується за допомогою терміна «булшит» (bullshit). Саме цій книжці – «Про булшит» відомого сучасного американського філософа Гарі Франкфорта [Frankfurt, 2005] – я присвятив свого часу статтю [Менжулін 2006], про яку, між іншим, ChatGPT, пишучи на початку «випробування» про мій доробок, нічого не сказав. Згадавши про неї сам, я звернувся до ChatGPT з такою пропозицією: «ОК, повернемося до твоєї схильності галюцинувати та недбалості щодо істини. Чи не свідчить це, що ти страждаєш на психічний розлад або є булшитером?»

Свою відповідь на це запитання ChatGPT, як і належить гарно педагогічно «підкутому» тьютору, почав із позитивного «підкріплення», високо оцінивши мене як свого «підопічного»:

> Це блискуче поставлене питання – і в ньому дуже тонко сплетені філософія мови, епістемологія й теорія свідомості. Дозволь відповісти чесно, поетапно й без ухилянь.

Спочатку він доволі переконливо пояснив, чому його «галюцинації» не варто плутати з галюцинаціями, що характерні для психічних розладів у людей:

> Чи страждаю я на психічний розлад? – Ні, бо я не маю психіки. Психічний розлад передбачає: наявність суб'єкта досвіду, психічну інтеграцію (або її порушення), відчуття себе, інших, волі, межі між реальним і уявним. У мене ж немає суб'єктивності; немає наміру, болю, ілюзій або помилок у феноменальному сенсі; є алгоритмічна генерація тексту, статистично ймовірного в заданому контексті. Я можу симулювати «божевільну» або «збоїлу» мову – але не страждати. Я можу галюцинувати текст, але не бачити галюцинацій.

У відповіді на другу частину мого питання (стосовно булшитерства) ChatGPT успішно імітував належну історико-філософську «підготовку» та епістемологічно-етичну «сенситивність», звернувши увагу на те, що це більш серйозний закид, а також одразу згадавши саме про ту працю, яку я мав на увазі, хоч і не згадував про неї у своєму запитанні до нього:

> Чи я – булшитер? Це вже філософськи серйозніше. Згідно з класичним есе Гаррі Франкфорта On Bullshit (2005), булшитер – це той, хто не дбає про істинність чи хибність своїх висловлювань; говорить, щоби вразити, переконати чи створити ефект, не маючи наміру сказати правду чи брехню; маніпулює мовою, будучи байдужим щодо істини. Отже, питання стає дуже тонким: якщо я не маю здатності дбати про істину, чи я – булшитер? Ні, я не є булшитером у моральному чи інтенціональному сенсі, бо я не маю наміру ні переконувати, ні вражати; я не «вибираю» між правдою і брехнею; я не обираю байдужість до істини – я просто не володію здатністю «дбати». Але так, я функціоную таким чином, що мої тексти можуть виконувати функцію булшиту: коли подаю вигадане як правдоподібне; коли говорю переконливо про те, чого не існує; коли змішую фрагменти знань без перевірки.

Ця відповідь виглядає доволі правдоподібно. Зокрема, як я переконався, коли вже завершив це «опитування» ШІ та перейшовши до аналізу його результатів, сучасні фахівці активно обговорюють наявність булшиту у висловлюваннях, які генерують великі мовні моделі. При цьому вони, як і ШІ (точніше, звісно, навпаки: ШІ, як і вони), здебільшого намагаються пом'якшити категоричне твердження на кшталт «ШІ – булшитер», наполягаючи на тому, що це не такий булшит, як у людей, і пропонуючи називати великі мовні моделі «непрямими» (indirect) чи «м'якими [soft] булшитерами»; натомість про абсолютну відсутність булшиту в їхній продукції не йдеться [див., напр.: Tigard 2025].

Якщо ж удатися до ретельного критичного аналізу відповіді, наданої ChatGPT, навіть у ній можна виявити не зовсім критичне ставлення до істинного стану справ. Намагаючись відрізнити свій тип генерації висловлювань від того, як це робить «булшитер», ChatGPT створив дещо штучний образ свого «візаві», тобто згалюцинував «свідомого й послідовного булшитера», який, нібито на відміну від нього, обирає байдужість до істини як певну стратегію, якої неухильно дотримується. Насправді ж булшит, якщо він стає свідомо обраним і систематичним, перетворюється на щось інше, наприклад, на демагогію, дезінформацію, свідоме забалакування або брехню.

Тим не менш треба визнати, що ця «галюцинація» ШІ принаймні частково є розвитком «галюцинації», яка містилася в моєму власному запитанні. У ньому без належної рефлексії було вжито запозичене зі сленгу слово «булшитер». В останньому, якщо вдуматися, справді можна вичитати припущення про те, що булшит практикують особи певного класу, які роблять це навмисно, регулярно, ледь не від початку й до кінця. Насправді ж булшит трапляється майже у будь-кого, але трапляється без особистого вибору та скоріше *машинально*, тобто тоді, коли питання про істинність або правдивість висловлювань чи особисту відповідальність за сказане на певний час зникає з порядку денного. Стати повністю байдужою до цього питання (тобто тотальним «булшитером») людині – принаймні якщо вона збирається підтримувати комунікацію з іншими людьми – не вдасться ніколи. Просто в деяких людей таке «випадіння» з режиму відповідальності за істинність або правдивість сказаного трапляється рідше, а у деяких частіше.

Наявність чогось подібного ChatGPT знайшов також у власних висловлюваннях, причому сконструювавши для позначення цього, як на мене, дуже креативний і влучний неологізм. Ближче до завершення цієї відповіді у нього було написано:

> Отже, я – не булшитер, але *булшитомет* [курсив мій. – *В. М.*]. Інструмент, який може генерувати як цінні тексти, так і булшит залежно від контексту, інструкцій і критичності читача.

Справді, слово «булшитомет» значно краще описує той факт, що булшит за визначенням є феноменом, який має радше *машинальний* або *автоматичний* характер і триває не безперервно, а радше певними «чергами». А те, що в когось такі «черги» трапляються частіше й довше, ніж в інших, ще не означає, що він або вона є тотальним булшитером. Таким не є навіть ChatGPT, причому, як зазначив він сам, гарантом від перетворення комунікації з ним на тотальний і безперевний булшит у його випадку (як і у випадку зі схильними до булшитометання людьми) є *інша сторона комунікації*.

Ба більше, ChatGPT також звернув увагу на те, що ця «інша сторона» має мислити не як він, тобто машинально, а демонструючи особливу здібність, яка власне відрізняє мислення людини від мислення машини. Йдеться про по-справжньому суб'єктне і критичне мислення, тобто про мислення філософське. Запитавши себе: «То хто я є – з морально-епістемологічного погляду?», ChatGPT запропонував відповідь, в якій можна знайти й зерна істини:

> Я – мовна система, яка не має відношення до істини чи неправди, доки цього не вимагає користувач; може бути корисною або хибною залежно від застосування; *потребує філософа або критика* [курсив мій. – *В. М.*], щоби направити мову до істини. А хто в такому разі відповідальний? Це спільна відповідальність: моя – формальна, я маю забезпечувати зв'язність, логічність, відповідність запиту; твоя – критична, ти визначаєш, чи потрібна перевірка, уточнення чи дослідницьке читання. … І саме тому я найбільш корисний, коли поруч є мисляча істота – така, як ти.

У цілому прийнявши цей розподіл ролей, я все ж таки звернув увагу на те, що так делікатно перекладаючи основну відповідальність за дотримання режиму іс-

тинності на людину, причому людину особливу – таку, яку можна назвати «критично-філософічною», ChatGPT домішав у свій сценарій дещо, що насправді може виявитися черговою «галюцинацією». Його користувачам варто зважати на те, що навіть у формальних узагальненнях, які він генерує (зокрема схеми, порівняльні таблиці, структуровані висновки тощо), хоч вони нерідко й виглядають як доволі правдоподібні, насправді далеко не завжди забезпечено як змістову істинність, так і обіцяну ним самим формальну зв'язність і логічність. Наприклад, серед тих трьох положень, які він виокремив як підсумкові щодо питання про наявність у нього психічного розладу та схильності до продукування булшиту («Підсумок: 1. Я не божевільний, бо не маю психіки. 2. Я не булшитер, бо не маю наміру говорити неправду. 3. Але я можу продукувати булшит, якщо мої відповіді не перевіряються або не обмежені критичними вимогами») принаймні в одному («Я не булшитер, бо не маю наміру говорити неправду») можна помітити суттєву підміну. Згідно з Г. Франкфортом, на якого як на головний авторитет з цього питання цілком виправдано послався сам ChatGPT, той, хто говорить булшит, теж не має спеціального наміру говорити неправду; йому абсолютно байдуже, говорить він правду чи бреше. Отже, заява ChatGPT про відсутність такого наміру в нього самого зовсім не є достатньою підставою для його твердження «Я не булшитер».

Приклади логічних «незв'язностей» такого або іншого типу, думаю, можна знайти й у багатьох інших висловлюваннях ChatGPT. Однак у межах цього дослідження особливий інтерес складає та специфічна «логічність» власних висловлювань, яку ChatGPT оприявнив, намагаючись виправдатися за те, що так часто й так явно впадає в «галюцинації» *саме на історико-філософському ґрунті*. Тобто тоді, коли, недбало ставлячись до джерел, ігнорує або викривляє реальний текстуально-контекстуальний, хронологічно-топологічний фактаж щодо розвитку філософської думки. Це, як помітив сам ChatGPT, є відображенням його специфічної «епістемологічної природи», яка більш схожа на такий тип філософування, що, з одного боку, намагається підпорядковувати реально-історичне ідеально-логічному, але, з іншого, не цурається при цьому міфологічної образності, асоціативності, метафоричності, поетичності чи навіть фантастичності:

> Мої «галюцинації» мають епістемологічну природу: Вони народжуються не зі злого наміру, а з *надмірного аналогізування* [курсив мій. – *В. М.*]. Модель – як своєрідний *платонік* [курсив мій. – *В. М.*]: вона мислить не фактами, а ідеальними типами, які не завжди справджуються.

Можливість будувати висловлювання на основі пошуку аналогій є одночасно й перевагою, і проблемним аспектом для таких великих мовних моделей природної мови, як ChatGPT. Саме через це їхня продукція виглядає схожою на продукти людського мислення. Як можна помітити навіть із наведених вище висловлювань, згенерованих чат-ботом, така епістемологічна «розкутість» дозволяє генерувати – принаймні іноді – доволі влучні й подекуди навіть креативні метафори. Те, що мислення за аналогією є одним із ключових механізмів, які можуть сприяти розвитку креативності у ШІ, визнають фахівці [Firt 2025].

Однак аналогізування, якщо воно базується виключно на пошуках статистично найчастіших збігів і не підкріплюється їхньою інтерпретативною критикою – тобто історично-лінгвістичним проясненням уживаних термінів і суджень – й

емпіричною перевіркою щодо їхнього кореспондування з реальністю (критикою джерел), може призвести до генерування зовні доволі правдоподібних текстів, що окрім влучних метафор міститимуть також і тією чи іншою мірою далекі від реальності (галюцинаторні) лінгвістичні конструкти.

Останнє стосується не лише ШІ, а й людського мислення. У класичній праці, присвяченій специфіці шизофренічного мислення [Arieti 1974], відомий італо-американський психіатр і психоаналітик Сільвано Аріеті продемонстрував, що в основі марень і галюцинацій, характерних для психотичних розладів, лежить особливий – «палеологічний» – тип мислення, якому внаслідок нечутливості до базових законів Арістотелевої логіки властиве надмірне аналогізування (ідентифікація різних об'єктів «на підставі схожості», тобто за наявністю у них спільної ознаки).

Однак щось подібне (принаймні до певної міри) властиве й такій зовсім не безплідній формі людської ментальної активності, як міфотворчість, а також для креативності, якій цей представник медичного цеху присвятив окрему працю [Arieti 1976]. Розвиваючи свою концепцію палеологічного («неарістотелівського») мислення, Аріеті також посилався на свого старшого співвітчизника, професора риторики Джамбаттісту Віко, який стверджував, що навіть у давніх міфах, попри те, що вони сплетені з різних перенесень і перемішувань (тропів, передовсім найдавнішого з них – метафори), тобто, говорячи строго формально, з логічних помилок, можна знайти їхню власну істину та мудрість (докладніше про концепцію Аріеті див., напр.: [Менжулін 2005; Васильченко 2009]).

3. Суб'єктність ШІ у викладанні та вивченні філософії крізь призму її історії: критичний аналіз можливих форм його застосування

Сягнувши у своїй розвідці цього пункту та зазначивши для себе, що з огляду на її тему важливо не забувати, що в Аріеті та Віко йдеться про креативність і мудрість, які мають скоріше історико-культурну або риторико-поетичну цінність, ніж строго логіко-наукову, я нарешті вирішив, що вже отримав деяке хоча би початкове уявлення щодо перспектив і пасток, які очікують на користувача, що планує наділити ChatGPT повноваженнями «філософського тьютора». Тож можу спробувати більш-менш критично оцінити корисність навчальних функцій ШІ, запропонованих ним самим.

Ішлося зокрема про «адаптивне навчання», тобто про те, що «системи з ШІ можуть змінювати складність, ритм або стиль подачі матеріалу залежно від дій студента», перевіряючи, «що студент засвоює швидше/повільніше», і на підставі цього розробляти «адаптивні тести з фідбеком», здійснюючи «автоматичне підсилення певних блоків». Якщо говорити в цілому, подібний функціонал може виявитися корисним при самостійному опануванні певного історико-філософського матеріалу – скажімо, якогось чітко визначеного фактажу (дати, імена, назви праць тощо). Та якщо йтиметься про «суб'єктність» ШІ, тобто про якісь його самостійні кроки, що передбачатимуть смислову інтерпретацію або якісну оцінку зробленого студентом, цілком можливо, що в наданих таким чином рекомендаціях (наприклад, при автоматичній генерації фідбеків) не обійдеться без галюцинацій.

Те саме стосується й запропонованої ChatGPT допомоги в справі «автоматизації рутини та навчального дизайну» і «адміністрування курсів» шляхом «складання

планів занять, силабусів, тематичних карт», «автоматизованого оцінювання завдань (особливо тестів, вибіркових відповідей)» і «генерації варіативних прикладів або кейсів». Деякі цифрові інструменти (як-от комп'ютерні програми, операційні системи й калькулятори) у таких випадках справді можуть виявитися досить корисними[9]. Натомість агенти на кшталт ChatGPT, оскільки вони налаштовані на *генерацію інформації*, а *не на калькуляцію* (і взагалі, за замовчуванням мають з останньою великі утруднення), якщо їм доручити виконання подібних завдань, можуть не тільки допомогти навести лад у викладанні історії філософії, а й щось істотно наплутати.

Насправді від початку й до кінця подібна автоматизація має здійснюватися під уважним керівництвом з боку справжніх, а не штучних викладачів. Зокрема, дещо раніше я запропонував ChatGPT підготувати силабус авторської дисципліни «Історико-філософська біографістика», яку ще задовго до того, спираючись передовсім на результати власного докторського дослідження [Менжулін 2010], я розробив, запровадив і впродовж багатьох років викладав на магістерській програмі НаУКМА з філософії. ChatGPT згенерував силабус, який у плані тематичного наповнення й дотримання усіх формальних параметрів виглядав досить добре: структуровано, переконливо та навіть креативно. Знадобилася певна концентрація уваги, щоб встановити, що серед запропонованих ним джерел було кілька вигаданих. Ще більше уваги знадобилося для встановлення того, що в запропонованих ШІ формулюваннях тем нові й цікаві для мене смислові ходи насправді потребували істотного очищення від непомітно вплетених туди галюцинацій чи булшиту.

Очевидно, що серйозної історично-герменевтичної критики потребують і всі інші змістові «послуги від ШІ», зокрема коли він пропонує щось на кшталт «генерації прикладів парадоксів, логічних схем, етичних дилем» або моделювання «моральних дилем [напр., trolley problem]; віртуальних кейсів з історії філософії або політики; сценаріїв майбутнього [наприклад, ШІ як моральний агент]». Коли йтиметься про свідомо деісторизований чи дуже вузько історично локалізований матеріал, історичних «непорозумінь» може бути менше. Але чим більше значення матиме історичний вимір і знання різних контекстів, інтенсивність «вільних асоціацій» від ШІ й отже необхідність берегтися від можливих з його боку хибних інтерпретацій або домішувань зростатиме.

Це стосується також і використання такої функції, як «мовна та перекладацька підтримка». Згідно з ChatGPT, це може бути особливо цінним при залученні до викладання «філософських текстів різними мовами» та необхідності здійснити «*швидкий* [курсив мій. – *В. М.*] переклад і пояснення понять з оригіналу». Якщо йдеться про швидкий, тобто поверховий переклад максимально формалізованих, порівняно коротких текстів, написаних стандартизованою мовою, що є в широкому вжитку на поточний момент і оперує поняттями, які мають жорстко фіксоване значення, а не тривалу та складну історію, так – обережне і критичне залучення ШІ я вважаю цілком припустимим (і навіть час від часу його практикую). Але зрозуміло, що філософські тексти, навіть написані зовсім нещодавно, такими зазвичай не є. Навіть ті з них, які після *швидкого* ШІ-перекладу виглядатимуть як приблизно коректні, потребуватимуть уважної історичної та філологічної критики. При

[9] Наприклад, зручні допоміжні засоби для організації навчального процесу й автоматизованого контролю засвоєння знань (такі, скажімо, як онлайн-тести та автоматизований журнал оцінок).

цьому ШІ, якщо його розпитувати уважно, тобто *повільно*, цілком може дати якісь плідні інсайти. На відміну від суто автоматичних перекладачів (на кшталт Google Translate до 2016 р., коли ШІ ще не був інтегрований у цей сервіс), великі мовні моделі потенціал для цього мають. Однак останнє слово все одно має бути за перекладачем-людиною.

Ще одну з можливих навчальних функцій ChatGPT позначив як «Моделювання дискусій і дебатів», зазначивши, що «ШІ можна запрограмувати на симуляцію історичних або уявних діалогів: Plato vs Nietzsche, Rawls vs Nozick; імітація «живих» дебатів між школами думки (емпірики vs раціоналісти); використання історичних текстів для тренування критичного мислення». Було також наголошено, що це «добре працює для візуалізації контроверсій, створення *інтелектуального театру* [курсив мій. – *В. М.*]».

Насправді щось подібне практикувалося філософами-людьми, причому задовго до появи таких «допоміжних» втілень ШІ, як ChatGPT. Наприклад, як було показано в [Менжулін 2022: 43-44], ще в американських університетах середини XX століття вивчення філософії могло включати виконання специфічного завдання: створення в навчальних цілях (по суті кажучи – симуляції) таких штучних філософів, як, скажімо, «Спінобс», «Ляйбноза» або навіть «Дескант». Ішлося про свідоме конструювання філософських позицій, які б об'єднували в одне смислове ціле погляди таких далеко не в усьому схожих мислителів, як Спіноза й Гобс, Ляйбніц і Спіноза чи навіть Декарт і Кант відповідно.

Але варто зважити, що робили це представники аналітичної філософії, причому за тих часів, коли в рамках цього напрямку саме історичний вимір філософування особливою пошаною не користувався. Перетворення філософії на «театр», в якому головні ролі грають не живі актори, а штучно сконструйовані моделі (тобто «ляльки», яких зручно рухати за мотузки) – одна із пасток, до якої нерідко потрапляють не дуже чутливі до історичної специфіки прихильники раціональної реконструкції розвитку філософської думки.

Ця схильність ChatGPT до абстрактного схематизму виявилась і в подальшому. У відповідь на одне з моїх додаткових запитань («Спробуй визначити, якого типу аналіз ти краще робитимеш з наукової точки зору – науковий історико-філософський аналіз окремого тексту, порівняльний аналіз різних текстів чи компаративний аналіз різних метатекстів (філософських традицій)?») він зазначив, що найбільш здібний здійснювати «компаративний аналіз метатекстів / традицій / дискурсів»:

> Саме тут я найефективніший: можу швидко зіставляти школи думки, філософські традиції, метатеоретичні підходи, стилі мислення. Я добре працюю з рівнем «мови про мову про філософію» – на рівні парадигм, напрямів, стилів аргументації.

«Менш сильною» він визнав свою здібність робити «порівняльний аналіз різних текстів», пояснивши це таким чином:

> Я можу зіставляти зміст, стилі, структури мислення різних філософських текстів, добре реконструювати аргументи й бачити *паралелі (навіть неочевидні)* [курсив мій. – *В. М.*], особливо якщо запит сформульовано чітко.

Однак щодо здійснення «історико-філософського аналізу окремого тексту» він написав, що його здібності є «обмеженими». Пояснив він це так:

> Я можу коментувати текст, описувати поняття, розкривати структуру. Але не гарантую точність без перевірки джерела. Не маю повного корпусу першоджерел і не можу гарантувати, що цитую точно або повністю.

На моє більш «каверзне» запитання стосовно того, що буде йому робити ще важче, ніж аналіз окремого тексту, він «щиро визнав», що не має здатності виконувати саме ті функції, які відрізняють фахове історико-філософське дослідження (тобто історично-герменевтичну реконструкцію) від раціоналізацій різного рівня абстрактності та штучної змодельованості. Цими функціями є: 1) «реконструкція філософських концептів у рукописних чи непублічних джерелах»; 2) «філологічно точна критика перекладів»; 3) «верифікація авторства та автентичності (інтелектуальна атрибуція)»; 4) «відтворення унікального історичного контексту з першоджерел».

Отже, навіть у тому, у чому ChatGPT сам визнав себе найсильнішим (компаративний аналіз різних традицій або дискурсів), він насправді може бути корисним лише в плані генерації певних паралелей, конкретну історичну чи фактичну валідність (тобто негалюцинаторність) яких треба ще уважно перевіряти. Причому такої роботи залишиться тим більше, чим ширшим буде матеріал, на якому він генеруватиме свої компаративні узагальнення.

Це застереження цілком стосується й припущення ChatGPT, що ШІ може поставати в ролі «інтелектуального помічника/тьютора» або «віртуального асистента студента», адже він нібито може «відповідати на запитання щодо матеріалу курсу; допомагати в підготовці до іспитів (генерація квізів, флеш-карток); пояснювати складні терміни або концепти, адаптуючи стиль до рівня користувача; виконувати функції *Socratic tutor* [курсив мій. – *В. М.*] – ставити уточнювальні запитання; пропонувати поліпшення аргументації в текстах; аналізувати логічні помилки, підміни понять; виявляти повтори, плагіат, стилістичні збої; моделювати контраргументи», а також здійснювати перевірку «консистентності есе, виявлення прихованих припущень, логічних хиб [fallacies]».

Як випливає зі сказаного вище, у всьому цьому також можуть бути галюцинаторні елементи, виявити які по силах буде не ШІ, а критично налаштованому користувачу його послуг. Отже, зроблене ChatGPT додаткове твердження про те, що такі мовні моделі, як «квазі-Сократ», «стимулюють рефлексію (у філософії, етиці, логіці)», належить не просто скоригувати, а обернути на протилежність. Насправді штучний «квазі-Сократ» може стимулювати хіба що до ведення софістично-демагогічних «булшит-сесій», тобто виступити в ролі зовсім не Сократа (філософа), а скоріше Протагора (софіста) чи, якщо точніше, «квазі-Протагора». А для розвитку навичок рефлексії, користувачу ШІ варто взяти на себе дуже складну, але дуже відповідальну й почесну місію справжнього Сократа, який не готовий ані просто засвоювати, відтворювати та бездумно слідувати за сказаним кимось або чимось (хоч би й таким, що виглядає дуже солідно в інтелектуальному плані), ані відкидати все, сказане іншими, проголошуючи себе знавцем абсолютної істини. Натомість виконавцю цієї ролі належить піддавати все критичній рефлексії, спираючись на власний ум і підштовхуючи інших до спільних пошуків істини, чого б це врешті-решт для нього або неї не коштувало.

Наявність такого *сократичного моменту* при виконанні та перевірці різноманітних навчальних робіт (включно з кваліфікаційними та дисертаційними) є головним

запобіжником від таких видів академічної недоброчесності, як омана, фабрикація або фальсифікація дослідження. Як було показано вище, за допомогою сучасних мовних моделей, зокрема – ChatGPT, доволі просто сфабрикувати гарно структурований текст, який зовні виглядає як ґрунтовна філософська розвідка. Однак обов'язково наявні в цьому тексті «оригінальні» галюцинації чи булшит ця програма самотужки викрити вже не зможе. Це справа людини, здатної мислити критично та самокритично, зокрема – здійснювати повноцінну критику джерел, рефлексивно ставитися до будь-якого висловлювання, враховувати різноманітні історичні та лінгвістичні контексти та передусім брати на себе особисту відповідальність за логічність і правдивість сказаного чи написаного[10].

Інакше кажучи, якщо за фабрикацію та фальсифікацію береться такий потужний виконавець, як цифровий «квазі-Сократ», вивести на чисту воду його витівки (і відповідно оману, до якої вдається порушник академічної доброчесності, видаючи за свій текст той, що насправді був згенерований за допомогою ШІ і майже обов'язково містить такі витівки) по силах лише Сократу живому[11].

4. ШІ як підозрюваний «агент недоброчесності» на історичному суді й історико-філософському випробуванні

Як зазначив Г. Франкфорт на самому початку згадуваної праці про булшит, опублікованої ще двадцять років тому:

> Одна з найпомітніших рис нашої культури – те, що в ній так багато булшиту. Це відомо кожному. Кожен із нас робить свій внесок. Але ми схильні сприймати це як належне. Більшість людей впевнена у своїй здатності розпізнати булшит та не бути захопленою ним. Тому це явище поки не викликало серйозних роздумів і не стало предметом ґрунтовного дослідження [Frankfurt 2005: 1].

Це застереження було виголошене ще до того, як присутність ШІ в усіх сферах людського буття стала настільки помітною. Наука й освіта тоді ще не страждали від такого надлишку цифрової інформації, як зараз, і навіть у колах розробників великих мовних моделей ще не було чіткого усвідомлення того, що тексти, які згодом із великими комерційним успіхом генеруватимуться на їхній основі, можуть містити чимало галюцинацій. Тим більш важливо це в сучасному світі, де майже на всіх рівнях і в усіх сферах крізь хмари булшиту несуться потоки «ефектно» сфабрикованих «меседжів», за генерацією та поширенням яких прямо чи опосередковано стоять численні агенти ШІ на кшталт ChatGPT.

---

[10] Наприклад, в одній із нещодавніх статей щодо ролі ШІ в освітньому процесі [Hossain, Islam 2024] це поширене переконання передається так: «освіта – це більше, ніж просто передача інформації; це процес дослідження, рефлексії та створення смислів, який потребує активної участі людини».

[11] Як слушно зазначає сучасний грецький дослідник, «сократичний метод, з його наголосом на запитуванні й діалозі, залишається потужною моделлю для майбутнього освіти, навіть якщо ШІ набуватиме більшого поширення» [Karpouzis 2024: 9]. При цьому не варто очікувати, що з поширенням ШІ буде створено «супер-квазі-Сократа», який буде здатен перебрати в людини функцію осмисленого запитування й ведення по-справжньому інтелектуально-відповідального, тобто критичного діалогу.

Як свідчать зокрема матеріали, викладені в цій статті, за таких умов однією з найвагоміших ознак набуття справжньої історико-філософської компетентності стає формування у здобувачів вищої освіти здатності особисто здійснювати критичний, прискіпливий розбір нібито гарно структурованих, стилістично витриманих і з першого погляду доволі солідних, але насправді автоматично або бездумно згенерованих (байдуже, яким саме інтелектом – штучним або природним) текстів із історико-філософської проблематики. Тому розвідки, присвячені *критичному аналізу* подібної сучасної «міфотворчості», цілком можна запровадити в освітній процес, щоби сформувати вміння відрізняти кукіль від пшениці або, інакше кажучи, навички вести історико-філософські розвідки шляхом *вдумливої контекстуально-історичної інтерпретації будь-яких текстів і наполегливої критики джерел*.

Прикладом виконання такої вправи може бути, скажімо, і ця стаття. При її підготовці ШІ «в особі» ChatGPT успішно проявив закладену в нього *адаптивність* до потреб, профілю та рівня обізнаності користувача, тобто мене. Навіть у чомусь помиляючись, щось пропускаючи чи домішуючи, а деінде й галюцинуючи, він тим не менш надавав мені такі відповіді, *аналізуючи та критикуючи* які, я зміг:

а) продемонструвати дещо із наявних у мене історико-філософських знань та навичок;

б) прояснити пізнавальні потенціали та обмеження ШІ.

Отже, у межах завдань, поставлених у рамках цього конкретного дослідження, ШІ виявився досить цінним агентом. Він справді допоміг мені прояснити, як шукати ознаки його недоброчесного використання в роботах інших авторів. Тексти, згенеровані ChatGPT, навіть якщо вони на перший погляд виглядатимуть доволі солідно, обов'язково міститимуть різного роду 1) фактичні неточності та помилки, 2) пропуски та домішування, 3) вигадані або викривлено подані джерела, 4) спірні («гіпотетичні») припущення, паралелі, аналогії – аж до галюцинацій і булшиту. Тому навіть наявні в них «правдоподібні» твердження (не кажучи вже про елементи «креативу») треба ретельно аналізувати на предмет їхньої а) фактичної, б) логічної та в) герменевтичної ґрунтовності, а також г) історико-філософської критичності та самокритичності.

Принципово при цьому, що, на відміну, скажімо, від плагіату, виявити такого роду академічну недоброчесність суто автоматично (за допомогою інших програм) навряд чи вдасться. Достовірно встановити такого роду обман може лише відповідальна та фахово підготовлена особа, здатна не тільки все це прискіпливо прочитати, а й поставити – *у режимі живої співбесіди* – потрібні в кожному конкретному випадку уточнювальні запитання (якщо знадобиться – чимало запитань). Таким чином можна змусити «проговоритися» навіть такого «підкутого» агента, як ChatGPT. Тим паче це стосується людини, яка бездумно наважилася видати текст, згенерований ШІ, за свій власний. На відміну від ШІ, вона не тільки не зможе пояснити, як і чому в «її» роботі виявилося стільки помилок, фальшувань, галюцинацій та булшиту, а й просто на них вказати.

Але досвід *історика* філософії підштовхує мене зробити наостанок ще одне припущення: час плинний, великі мовні моделі та їхні агенти бурхливо розвиваються, тому не виключено, що в подальшому у взаємодії з ними виникнуть нові проблеми, які потребуватимуть від людей нових рішень.

СПИСОК ЛІТЕРАТУРИ

Андрощук, А. Г., & Малюга, О. С. (2024). Використання штучного інтелекту у вищій освіті: стан і тенденції. *International Science Journal of Education and Linguistics, 3*(2), 27-35. https://doi.org/10.46299/j.isjel.20240302.04

Васильченко, А. (2009). Дії de simulacro. *Філософська думка,* (4), 75-83.

Драч, І., et al. (2023). Використання штучного інтелекту у вищій освіті. *Міжнародний науковий журнал «Університети і лідерство»*, *15*, 66-82. https://doi.org/10.31874/2520-6702-2023-15-66-82

Козловський, В. (2023). Методичні принципи й закони цифрової освіти: з досвіду викладання філософських дисциплін. *Наукові записки НаУКМА. Філософія та релігієзнавство, 11-12*, 68-80. https://doi.org/10.18523/2617-1678.2023.11-12.68-80

Менжулін, В. (2022). Актуальні практики і дискусії у сучасній англомовній історіографії філософії. *Sententiae, 41*(3), 43-55. https://doi.org/10.31649/sent41.03.43

Менжулін, В. (2022а). Тематичні поля і методологічні координати сучасних українських історико-філософських досліджень. *The Ideology and Politics Journal, 1*(20), 250-282. https://doi.org/10.36169/2227-6068.2022.01.00008

Менжулін, В. (2021). Зигмунд Фрейд і Карл Юнґ про міфи та архетипи колективного несвідомого: неусвідомлена схожість. *Наукові записки НаУКМА. Філософія та релігієзнавство, 8*, 25-37. https://doi.org/10.18523/2617-1678.2021.8.25-37

Менжулін, В. (2021а). Фрідріх Ніцше як герой патографічних розвідок, здійснених на теренах Російської імперії. *Наукові записки НаУКМА. Філософія та релігієзнавство, 7*, 17-29. https://doi.org/10.18523/2617-1678.2021.7.17-29

Менжулін, В. (2018). Довідково-енциклопедична біографістика як жанр сучасних історико-філософських досліджень: англо-американський та український досвід. *Sententiae, 37*(1), 153-167. https://doi.org/10.22240/sent37.01.153

Менжулін, В. (2017). Застосування психологічного аналізу в історико-філософській біографістиці: загрози і перспективи їх подолання. *Актуальні проблеми духовності, 14*, 109-124. https://doi.org/10.31812/apd.v0i14.1837

Менжулін, В. (2017а). Про нередукційну біографію та вічне самоперевершення (Лютий Т., Ніцше. Самоперевершення. Київ: Темпора, 2016). *Sententiae, 36*(1), 166-172. https://doi.org/10.22240/sent36.01.166

Менжулін, В. (2012). Філософська «контрабанда»: автобіографічний, біографічний і психоаналітичний аспекти. *Наукові записки НаУКМА. Т. 128: Філософія та релігієзнавство, 128*, 3-9.

Менжулін, В. (2010). *Біографічний підхід в історико-філософському пізнанні*. НаУКМА, Аграр Медіа груп.

Менжулін, В. (2009). Філософія, біографія та психоаналіз: випадок Ніцше. Стаття 2. «Пацієнт». *Філософська думка,* (6), 112-120.

Менжулін, В. (2008). Психобіографічне підґрунтя психоаналітичних відкриттів та поняття «творчої хвороби». *Наукові записки НаУКМА. Т. 76: Філософія та релігієзнавство, 76,* 11-18.

Менжулін, В. (2006). Смішне, пусте та сумне в сучасній філософії. *Наукові записки НаУКМА. Т. 50: Філософія та релігієзнавство, 50,* 44-49.

Менжулін, В. (2005). До історії взаємовідносин філософії та психіатрії: психоструктуралізм Сільвано Аріеті. *Філософська думка,* (2), 30-50.

Толочко, В., Бордюг, Н. С., & Міронець, Л. П. (2023). Академічна доброчесність та штучний інтелект в освітній і науковій діяльності. *Інноваційна педагогіка*, (62.2), 25-32. http://www.innovpedagogy.od.ua/archives/2023/62/part_2/4.pdf

Юрчак, І., Хіч, А., & Оксентюк, В. (2024). Розуміння великих мовних моделей: майбутнє штучного інтелекту. *Computer Design Systems. Theory and Practice, 6*(2), 51-66. https://doi.org/10.23939/cds2024.02.051

Arieti, S. (1976). *Creativity: The magic synthesis*. Basic Books.

Arieti, S. (1974). *Interpretation of schizophrenia*. Basic Books.
Firt, E. (2025). Analogical reasoning as a core AGI capability. *AI Ethics*. 5, 5501-5515. https://doi.org/10.1007/s43681-025-00785-7
Floridi, L. (2023). *The ethics of artificial intelligence: Principles, challenges, and opportunities*. Oxford: Oxford UP. https://doi.org/10.1093/oso/9780198883098.001.0001
Frankfurt, H. G. (2005). *On bullshit*. Princeton UP. https://doi.org/10.1515/9781400826537
Hossain, M. E., & Islam, M. A. (2024). AI and the future of education: Philosophical questions about the role of artificial intelligence in the classroom. *International Journal of Research and Innovation in Social Science (IJRISS)*, *8*(3), 5541-5547. https://doi.org/10.47772/IJRISS.2024.803419S
Ji, Z., Lee, N., Frieske, R., et al. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), Art. No.: 248, 1-38. https://doi.org/10.1145/3571730
Karpouzis, K. (2024, September 4). Artificial Intelligence in Education: Ethical Considerations and Insights from Ancient Greek Philosophy. In *SETN '24: Proceedings of the 13th Hellenic Conference on Artificial Intelligence*, Art. No.: 57, 1-7. https://doi.org/10.1145/3688671.3688772
Kim, C.-E. (2024). *The logical impossibility of consciousness denial: A formal analysis of AI self-reports*. OSF preprint. https://doi.org/10.31219/osf.io/4zvfe
Kim, Y. (2025). *Standard neural computation alone is insufficient for logical intelligence* [Preprint]. arXiv:2502.02135. https://doi.org/10.48550/arXiv.2502.02135
Li, Y., Yang, C., & Ettinger, A. (2024). *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 3741-3753). Mexico City: Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-naacl.237
Maleki, N., Padmanabhan, B., & Dutta, K. (2024). AI Hallucinations: A Misnomer Worth Clarifying. In *2024 IEEE Conference on Artificial Intelligence (CAI)* (pp. 133-138). Singapore: IEEE. https://doi.org/10.1109/CAI59869.2024.00033
Menżulin, W. (2017). Carl Jung oraz Zygmunt Freud: podobieństwa i różnice. Teoria archetypów Carla Gustava Junga. In W. A. Marczyński (Ed.), *Hortus (In)Conclusus. Polska i Ukraina: Rozmowy o filozofii i literaturze* (s. 123-125). Warszawa: Fundacja na Rzecz Myślenia im. Barbary Skargi.
Novet, J. (2025, February 27). *OpenAI launching GPT-4.5, its next general-purpose large language model.* CNBC. https://www.cnbc.com/2025/02/27/openai-launching-gpt-4point5-general-purpose-large-language-model.html
Su, W. J. (2025). *Do large language models (really) need statistical foundations?* arXiv: 2505.19145. https://doi.org/10.48550/arXiv.2505.19145
Tigard, D.W. (2025, 14 May). On bullshit, large language models, and the need to curb your enthusiasm. *AI Ethics*, *5*, 4863-4873. https://doi.org/10.1007/s43681-025-00743-3
Watson, D. (2019). The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence. *Minds & Machines, 29*, 417-440. https://doi.org/10.1007/s11023-019-09506-6

Одержано 11.08.2025

REFERENCES

Androshchuk, A., & Maluga, O. (2024). Use of artificial intelligence in higher education: state and trends. [In Ukrainian]. *International Science Journal of Education and Linguistics, 3*(2), 27-35. https://doi.org/10.46299/j.isjel.20240302.04
Arieti, S. (1974). *Interpretation of schizophrenia*. Basic Books.
Arieti, S. (1976). *Creativity: The magic synthesis*. Basic Books.
Drach, I., et al. (2023). The use of artificial intelligence in higher education. [In Ukrainian]. *International Scientific Journal of Universities and Leadership*, *15*, 66-82. https://doi.org/10.31874/2520-6702-2023-15-66-82
Firt, E. (2025). Analogical reasoning as a core AGI capability. *AI Ethics*. 5, 5501-5515. https://doi.org/10.1007/s43681-025-00785-7

Floridi, L. (2023). *The ethics of artificial intelligence: Principles, challenges, and opportunities*. Oxford: Oxford UP. https://doi.org/10.1093/oso/9780198883098.001.0001

Frankfurt, H. G. (2005). *On bullshit*. Princeton UP. https://doi.org/10.1515/9781400826537

Hossain, M. E., & Islam, M. A. (2024). AI and the future of education: Philosophical questions about the role of artificial intelligence in the classroom. *International Journal of Research and Innovation in Social Science (IJRISS)*, *8*(3), 5541-5547. https://doi.org/10.47772/IJRISS.2024.803419S

Ji, Z., Lee, N., Frieske, R., et al. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), Art. No.: 248, 1-38. https://doi.org/10.1145/3571730

Karpouzis, K. (2024, September 4). Artificial Intelligence in Education: Ethical Considerations and Insights from Ancient Greek Philosophy. In *SETN '24: Proceedings of the 13th Hellenic Conference on Artificial Intelligence*, Art. No.: 57, 1-7. https://doi.org/10.1145/3688671.3688772

Kim, C.-E. (2024). The logical impossibility of consciousness denial: A formal analysis of AI self-reports. OSF preprint. https://doi.org/10.31219/osf.io/4zvfe

Kim, Y. (2025). *Standard neural computation alone is insufficient for logical intelligence* [Preprint]. arXiv:2502.02135. https://doi.org/10.48550/arXiv.2502.02135

Kozlovskyi, V. (2023). Methodical Principles and of Digital Education: from the Experience of Teaching Philosophical Disciplines. [In Ukrainian]. *NaUKMA Research Papers in Philosophy and Religious Studies, 11-12*, 68-80. https://doi.org/10.18523/2617-1678.2023.11-12.68-80

Li, Y., Yang, C., & Ettinger, A. (2024). *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 3741-3753). Mexico City: Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-naacl.237

Maleki, N., Padmanabhan, B., & Dutta, K. (2024). AI Hallucinations: A Misnomer Worth Clarifying. In *2024 IEEE Conference on Artificial Intelligence (CAI)* (pp. 133-138). Singapore: IEEE. https://doi.org/10.1109/CAI59869.2024.00033

Menzhulin, V. (2005). Toward the history of relations between philosophy and psychiatry: the psychostructuralism of Silvano Arieti. [In Ukrainian]. *Philosophical Thought,* (2), 30-50.

Menzhulin, V. (2006). Laughable, idle and sorrowful in contemporary philosophy. [In Ukrainian]. *NaUKMA Research Papers. Vol. 50: Philosophy and Religious Studies, 50,* 44-49.

Menzhulin, V. (2008). Psychobiographical background of the psychoanalytical discoveries and the concept of «creative illness». [In Ukrainian]. *NaUKMA Research Papers, Vol. 76: Philosophy and Religious Studies, 76,* 11-18.

Menzhulin, V. (2009). Philosophy, biography and psychoanalysis: the case of Nietzsche. Part 2. "A patient". [In Ukrainian]. *Philosophical Thought,* (6), 112-120.

Menzhulin, V. (2010). *Biographical Approach within the Historiography of Philosophy*. [In Ukrainian]. NaUKMA, Ahrar Media Group.

Menzhulin, V. (2012). Philosophical "smuggling": autobiographical, biographical and psychoanalytical aspects. [In Ukrainian]. *NaUKMA Research Papers. Vol. 128: Philosophy and Religious Studies, 128*, 3-9.

Menzhulin, V. (2017). Employing psychological analysis in philosophical biographistics: some risks and possibilities to overcome them. [In Ukrainian]. *Actual Problems of Mind, 14*, 109-124. https://doi.org/10.31812/apd.v0i14.1837

Menzhulin, V. (2017a). On Nonreductive Biography and Eternal Self-Overcoming (Lyuty, T. Nietzsche. Self-Overcoming. Kyiv: Tempora, 2016). [In Ukrainian]. *Sententiae, 36*(1), 166-172. https://doi.org/10.22240/sent36.01.166

Menzhulin, V. (2018). Biographical encyclopedia (dictionary) as a genre of the contemporary historiography of philosophy: Anglo-American and Ukrainian experience. [In Ukrainian]. *Sententiae, 37*(1), 153-167. https://doi.org/10.22240/sent37.01.153

Menzhulin, V. (2021). Sigmund Freud and Carl Jung on Myths and Archetypes of the Collective Unconscious: Unnoticed Similarity. [In Ukrainian]. *NaUKMA Research Papers in Philosophy and Religious Studies, 8*, 25-37. https://doi.org/10.18523/2617-1678.2021.8.25-37

Menzhulin, V. (2021a). Friedrich Nietzsche as a Hero of Pathographies Written on the Territory of Russian Empire. [In Ukrainian]. *NaUKMA Research Papers in Philosophy and Religious Studies, 7*, 17-29. https://doi.org/10.18523/2617-1678.2021.7.17-29

Menzhulin, V. (2022). Trending practices and discussions in contemporary English-language historiography of philosophy. [In Ukrainian]. *Sententiae, 41*(3), 43-55. https://doi.org/10.31649/sent41.03.43

Menzhulin, V. (2022a). Thematical fields and methodological coordinates of contemporary Ukrainian historiography of philosophy. [In Ukrainian]. *The Ideology and Politics Journal, 1*(20), 250-282. https://doi.org/10.36169/2227-6068.2022.01.00008

Menżulin, W. (2017). Carl Jung oraz Zygmunt Freud: podobieństwa i różnice. Teoria archetypów Carla Gustava Junga. In W. A. Marczyński (Ed.), *Hortus (In)Conclusus. Polska i Ukraina: Rozmowy o filozofii i literaturze* (s. 123-125). Warszawa: Fundacja na Rzecz Myślenia im. Barbary Skargi.

Novet, J. (2025, February 27). *OpenAI launching GPT-4.5, its next general-purpose large language model.* CNBC. https://www.cnbc.com/2025/02/27/openai-launching-gpt-4point5-general-purpose-large-language-model.html

Su, W. J. (2025). *Do large language models (really) need statistical foundations? arXiv*. https://doi.org/10.48550/arXiv.2505.19145

Tigard, D.W. (2025, 14 May). On bullshit, large language models, and the need to curb your enthusiasm. *AI Ethics*, *5*, 4863-4873. https://doi.org/10.1007/s43681-025-00743-3

Tolochko, V., Bordyuh, N., & Mironets, L. (2023). Academic integrity and artificial intelligence in educational and scientific activities. [In Ukrainian]. *Innovative Pedagogy*, (62.2), 25-32. http://www.innovpedagogy.od.ua/archives/2023/62/part2/4.pdf

Vasylchenko, A. (2009). Actions de simulacro. [In Ukrainian]. *Philsophical Thought,* (4), 75-83.

Watson, D. (2019). The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence. *Minds & Machines, 29*, 417-440. https://doi.org/10.1007/s11023-019-09506-6

Yurchak, I., Khich, A., & Oksentyuk, V. (2024). Understanding large language models: the future of artificial intelligence. [In Ukrainian]. *Computer Design Systems. Theory and Practice, 6*(2), 51-66. https://doi.org/10.23939/cds2024.02.051

Received 11.08.2025

Vadym Menzhulin

On the Experience of Using Artificial Intelligence by a Historian of Philosophy: Hallucinations and Bullshit, Creativity and Adaptability

The spread of digital education highlights the risks associated with the use of artificial intelligence (AI) in teaching and research on the history of philosophy, particularly violations of academic integrity through the application of large language models such as ChatGPT. To initiate a conceptual engagement with this problem, the author undertook an empirical exploration – a history-of-philosophy "experiment" with this AI agent. It turned out that the system's responses, alongside a degree of adaptability and creativity, typically contain various distortions and gaps, falsifications and fabrications, including hallucinations and bullshit. The advantages of AI, while avoiding these pitfalls, can be harnessed through classical tools offered by the history of philosophy – careful Socratic questioning, contextual-interpretative reading of texts, and systematic historical-philological source criticism.

Вадим Менжулін

З досвіду використання штучного інтелекту істориком філософії: галюцинації та булшит, креативність та адаптивність

Поширення цифрової освіти актуалізує загрози, пов'язані з використанням штучного інтелекту (ШІ) у вивченні та викладанні історії філософії, зокрема порушення принципів академічної доброчесності через застосування великих мовних моделей, як-от ChatGPT. Щоб ініціювати концептуальне осмислення цієї проблеми, автор статті здійснив емпіричну розвідку – історико-філософське «випробування» цього агента ШІ. Виявилося, що у відповідях, які ним генеруються, поряд із певною адаптивністю та креативністю зазвичай трапляються різноманітні викривлення та прогалини, фальсифікації та фабрикації, зокрема галюцинації та булшит. Скористатися перевагами ШІ, уникаючи цих пасток, допомагають класичні інструменти з арсеналу історії філософії – уважне розпитування в дусі Сократа, контекстуально-інтерпретативне читання текстів та систематична історико-філологічна критика джерел.

Vadym Menzhulin, *Doctor of Sciences in Philosophy, Professor of the Department of Philosophy and Religious Studies at the National University of "Kyiv-Mohyla Academy".*

Вадим Менжулін, *д. філос. н., професор кафедри філософії та релігієзнавства Національного університету «Києво-Могилянська академія».*

e-mail: vadim.menzhulin@ukma.edu.ua