Філософія

УДК 1 (091): 004.8: 165: 316.77

DOI https://doi.org/10.5281/zenodo.20599450

Штучний інтелект як тririvневий Lethe: феноменологічна типологія AI-загроз у контексті досягнення істини в епоху постправди

Шевчук Андрій Вікторович,
аспірант кафедри філософії та політології,
Житомирський державний університет імені Івана Франка,
м. Житомир, Україна, https://orcid.org/0009-0007-1992-2741

Прийнято: 19.05.2026 | Опубліковано: 30.05.2026

Анотація. *Метою статті є обґрунтування феноменологічної типології AI-загроз у контексті досягнення істини на основі гайдеґґерівських концептів Lethe та Gestell. Для реалізації поставленої мети використано метод феноменологічного аналізу у тлумаченні відношень між технологією та істиною, метод типологізації для диференціації якісно різних рівнів AI-загроз в сучасному інформаційному просторі та метод герменевтичного аналізу у площині дослідження наявних філософських підходів до розуміння проблематики співвідношення істини, правди та постправди.*

У статті доведено, що штучний інтелект (AI) трансформує онтологічну структуру постправди, формуючи три якісно різні рівні загроз для істини: інструментальний (великі мовні моделі як засіб масштабування людської дезінформації), алгоритмічний (платформні рекомендаційні системи як структурно байдужі до правдивості контенту) та квазі-автономний /

генеративний (deepfakes і синтетичні медіа як сфера продукування реалістичного контенту без оригіналу).

Показано, що кожен із цих рівнів реалізує свою модальність гайдеґґерівського концепту Lethe: інструментальний Lethe (масштабування приховування через LLM), системний Lethe (структурна індиферентність алгоритмічних платформ до питання про відкритість) та генеративний Lethe (виробництво симулякрів deepfake-технологіями). Обґрунтовано, що метафора «синтетичної квазі-інтенціональності» є концептуально недостатньою, адже в її межах відбувається антропологізація AI і безпідставне приписування йому людського типу продукування істини на основі соціокультурного досвіду. На противагу цьому запропоновано концептуальні межі «трирівневого Lethe» як інструменту осмислення AI як нової фази постправди та сфери розширення впливу логіки Gestell на мову, увагу та публічну реальність. Наукова новизна дослідження полягає в обґрунтуванні трирівневої феноменологічної типології AI-загроз, що дозволяє диференціювати їх за онтологічним статусом і механізмами впливу. Запропонований підхід розширює застосування гайдеґґерівських концептів Lethe та Gestell до аналізу сучасних інформаційних технологій. Отримані результати можуть бути використані для подальших досліджень постправди та розробки концептуальних моделей регуляції AI.

Ключові слова: *штучний інтелект (AI), постправда, дезінформація, deepfakes, великі мовні моделі, типологія, Lethe, Gestell, М. Гайдеґґер, феноменологія.*

Artificial Intelligence as a Three-Level Lethe: A Phenomenological Typology of AI Threats to Truth in the Post-Truth Era

Andrii Shevchuk,
PhD Student, Department of Philosophy and Political Science,
Zhytomyr Ivan Franko State University, Zhytomyr, Ukraine,
https://orcid.org/0009-0007-1992-2741

Abstract. *The article aims to construct a phenomenological typology of AI threats to truth based on Heidegger's concepts of Lethe and Gestell. Phenomenological analysis, typological method, and hermeneutical analysis of existing philosophical conceptualizations are employed.*

The article demonstrates that AI transforms the ontological structure of post-truth by generating three qualitatively different levels of threats to truth: instrumental (LLMs as tools for scaling human disinformation), algorithmic (platform recommendation systems as structurally indifferent to truthfulness), and quasi-autonomous / generative (deepfakes and synthetic media as the production of realistic content without an original).

Each level is shown to realize its own modality of Heideggerian Lethe: instrumental Lethe (scaling of concealment via LLMs), systemic Lethe (structural indifference of algorithmic platforms to the question of disclosure), and generative Lethe (production of simulacra by deepfake technologies). The metaphor of «synthetic quasi-intentionality» is shown to be conceptually insufficient, since it anthropomorphizes AI and ascribes to it a human type of truth-production rooted in sociocultural experience. The conceptual framework of «three-level Lethe» is proposed as an instrument for understanding AI as a new phase of post-truth and as an extension of the logic of Gestell to language, attention, and public reality. The scientific novelty of the article consists in constructing a three-level

phenomenological typology of AI threats that differentiates them by their ontological status and mechanisms of influence. The proposed approach extends the application of Heideggerian concepts of Lethe and Gestell to the analysis of contemporary information technologies. The findings can be used for further research on post-truth and for the development of conceptual models of AI regulation.

Keywords: *artificial intelligence, post-truth, Lethe, Gestell, deepfakes, large language models, phenomenology, Heidegger, typology, disinformation.*

Постановка проблеми. У сучасній філософській парадигмі проблематика дослідження співвідношення між штучним інтелектом (AI) та процесами досягнення істини в умовах домінування постправди є однією з найбільш затребуваних і водночас найменш концептуально розроблених наукових проблем. Стрімкий розвиток генеративного AI – великих мовних моделей (LLM), технологій синтетичних медіа (deepfakes), алгоритмічних систем рекомендацій – переводить загрозу постправди з рівня розповсюдження інформації на рівень її виробництва, трансформуючи саму онтологічну структуру відносин між суб'єктом, інструментом і актом дезінформації. Якщо в період домінування «класичної» постправди центральним актором залишалася людина-пропагандист, яка свідомо маніпулювала фактами, то AI створює принципово нову ситуацію: технологія сама стає джерелом контенту, що не відповідає реальності, причому без будь-якої інтенціональності у традиційному феноменологічному сенсі.

Для адекватного філософського осмислення цієї трансформації ми звертаємося до гайдеґґерівської концепції істини як aletheia (ἀ-λήθεια). Давньогрецьке слово aletheia етимологічно розкривається як «не-прихованість», «від-криття»: префікс ἀ- є запереченням, а слово λήθη (Lethe) позначає «прихованість», «закритість», «забуття». Отже, для давніх греків і для Мартіна Гайдеґґера, який реактуалізував цю етимологію, істина є не

«відповідністю судження стану речей» (як у класичній адекваціоністській теорії), а онтологічною подією відкриття, в якій буття розкриває себе лише тому, хто ставить запитання. У такому розумінні Lethe є не просто «помилкою» або «брехнею», а протилежною онтологічною подією - закриттям горизонту відкриття, станом, в якому буття приховує себе від того, хто мав би його відкрити [19, S. 12–18]. Ця відмінність є принциповою: брехня передбачає суб'єкта, який знає правду і свідомо її приховує; Lethe описує ситуацію, де саме питання про правду стає неможливим, адже закритість передує будь-якому індивідуальному акту приховування.

Саме тому Lethe, а не «дезінформація» чи «маніпуляція», є адекватним концептом для осмислення AI-загроз для досягнення істини: AI створює форми закритості, де немає жодного суб'єкта-брехуна, але горизонт відкриття все одно закривається системно, масштабно й, і часто незворотно.

Продукована AI дезінформація дедалі більше стає глобальним явищем, яке зачіпає демократичні процеси в усьому світі. Це проявляється в процесах алгоритмічного підсилення контраверсійного контенту під час виборчих кампаній, у deepfake-відео публічних осіб та діяльності LLM-генерованих ботоферм. Різні форми AI-опосередкованої дезінформації вже фіксуються у сучасних демократичних процесах, а їх поєднання дедалі частіше створює багаторівневі загрози для досягнення істини. Водночас наявні філософські концептуалізації AI як загрози для істини залишаються або надмірно метафоричними (тенденція описувати AI через єдину метафору «квазі-інтенціональності», яку можна реконструювати, зокрема, з політико-філософських праць бельгійського дослідника М. Коекельберґа [11]), або емпірично обмеженими через зосередження на окремих технологіях без обґрунтування типології. У підсумку це засвідчує науково-теоретичну лакуну, заповнення якої і є завданням репрезентованої статті.

 ISSN: 3083-5712

Аналіз останніх досліджень і публікацій. Дослідження проблематики AI як загрози для досягнення істини стало предметом зацікавленості для значної кількості як зарубіжних, так і українських науковців. Зокрема британський дослідник Л. Вайдінгер та його співавтори (2022) [1] запропонували одну з найґрунтовніших таксономій ризиків великих мовних моделей, у якій виокремили шість класів шкоди: дискримінація та мова ворожнечі, інформаційні ризики, шкода від дезінформації, зловживання, ризики взаємодії «людина-комп'ютер» і екологічно-соціоекономічні ризики; саме ця робота окреслила ширину проблемного поля у вирішенні протистояння технологій AI і можливості досягнення істини. Американські дослідники Дж. Ґолдштейн та ін. (2023) [2] розробили концепцію AI-powered influence operations і довели, що генеративні моделі різко здешевлюють масштабування дезінформаційних кампаній, переводячи їх у напівавтоматичний режим. Американський дослідник М. Мюссер (2023) [3] кількісно оцінив, як LLM змінюють економіку дезінформаційних кампаній, а група британських дослідників на чолі з А. Р. Вільямс та ін. (2024) [4] експериментально перевірили, наскільки переконливим для користувачів є AI-генерований дезінформаційних контент.

Дж. Ґолдштейн, Дж. Сестрі, М. Мюссер та Р. ДіРеста (2023) [2] в одному з перших досліджень зв'язку між LLM і операціями впливу виокремили чотири стадії «конвеєра» дезінформації: побудова/адаптація моделі, отримання доступу, дисемінація AI-генерованих текстів і формування переконань. Автори розрізняють підсилення впливу людей-пропагандистів та автоматизацію операцій як дві функції LLM у контексті упровадження практик дезінформації [2, p. 7–12]. Натомість економічна методологія аналізу М. Мюссера (2023) [3] деконструює міф про AI як «магічну зброю» дезінформації. На думку дослідника, зниження витрат на AI у сфері дезінформації може бути суттєвим, а в граничному випадку використання високонадійних моделей – воно

становить 70%, що робить AI не «абсолютною зброєю», а економічно раціональним інструментом, який зміщує cost-benefit-баланс на користь пропагандистів [3, p. 3–5].

Група дослідників на чолі з А. Вільямс (2024) [4] у серії з 2340 експериментів показала, що практично всі великі мовні моделі після 2022 року здатні генерувати дезінформаційний контент, який для понад 50% оцінювачів виявився нерозрізненним від людського – за винятком тих випадків, коли модель відмовлялася відповідати на шкідливий запит [4, p. 8–10]. Це означає, що навіть без спеціального налаштування «стандартні» LLM уже є достатньо потужними для промислового виробництва переконливої дезінформації. Поряд з цим американський дослідник Е. Феррара (2024) [5] на основі реальних кейсів виборчої інтерференції 2023-2024 рр. виокремив три вектори інструментального зловживання AI та довів структурну асиметрію: вартість подібної атаки різко знижується, тоді як вартість захисту зростає [5, p. 4–6].

Британські дослідники К. Ваккарі та Е. Чедвік (2020) [6] у репрезентативному експерименті на основі британської вибірки виявили показовий механізм: найбільшою загрозою deepfakes є не прямий обман, а продукована ними «узагальнена індетермінантність» - загальна втрата довіри до будь-якого відеоконтенту. У контексті цього знахідка Ваккарі та Чедвіка перегукується із концептом американських юристів Р. Чезні та Д. Цитрон (2019) «liar’s dividend» («дивіденд брехуна»): коли будь-яке відео може бути deepfake-ом, будь-який реальний компрометуючий відеодоказ може бути відкинутий як «фейк» [7, p. 1758]. Німецька дослідниця М. Павелець (2022) [8] системно проаналізувала загрози deepfakes для демократії через призму нормативних теорій і виявила три структурні виміри цієї загрози: порушення інклюзивного представництва, підваження деліберативного обміну аргументами та делегітимізація колективних рішень [8, p. 5–10].

Канадська група дослідників на чолі з Дж. Кітцманном (2020) [9] запропонувала ключове для правової класифікації розрізнення маніпуляції інформацією (зміна існуючого матеріалу) та генерації (створення повністю синтетичного контенту). Маніпуляція при цьому є фальсифікацією доказу, а генерація постає створенням повністю синтетичного контенту, що не має відповідника у реальності [9, p. 138–140]. Поряд з цим бельгійський філософ М. Коекельберг (2023) обґрунтовує: AI-генерований контент становить загрозу для епістемологічної агентності демократичних суспільств, оскільки підриває здатність громадян формувати обґрунтовані судження на підставі двох якісних відмінностей: безпрецедентного масштабу і системного характеру AI-виходів [10, p. 1343–1347].

Серед філософських осмислень AI у контексті постправди особливе місце посідає гайдеґґерівський підхід. Британські дослідники М. Хікс, Дж. Хамфріс та Дж. Слейтер (2024) [23] концептуалізували LLM як «bullshit-машини» в сенсі американського філософа Г. Франкфурта (2005) [20], а М. Коекельберг (2022) [11] обґрунтував структурну байдужість AI до розрізнення правди і брехні як принцип; у подальшій роботі (2023) [10] він розкрив наслідки цього для епістемологічної агентності демократичних суб'єктів. Британський дослідник І. Томсон (2025) своєю чергою показує, що гайдеґґерівське питання про техніку не втрачає аналітичної сили в добу AI, якщо розглядати технологію як історичний спосіб онтологічного розкриття, що перебудовує наші стосунки зі світом, іншими та самими собою [12, p. 1–2, 52]. Саме ця позиція є безпосередньо важливою для нашої статті, адже вона дозволяє перейти від опису окремих ризиків до осмислення AI як новітнього способу творення різного типу інформації та знань.

Британський філософ І. Ґабріел (2020) у контексті дослідження системи цінностей та узгодження AI з ними обґрунтував, що AI-системи потребують узгодження не з інструкціями чи преференціями користувачів, а з принципами

справедливості, які отримують рефлексивну підтримку попри різноманітність моральних переконань у суспільстві [13, p. 418–425]. Це має пряме значення для вирішення проблеми постправди: якщо AI-системи узгоджуються лише з метриками залученості аудиторії, вони неминуче підсилюють дезінформацію. Польські дослідники А. Лабуз і К. Негрінг (2024) [14] у критичній рецензії щодо розвитку технологій AI аргументували, що тривоги щодо AI-дезінформації є значною мірою перебільшеними: значна частина дискурсу оперує антиципованими загрозами як вже реалізованими, натомість емпіричні дані про ефективність AI-дезінформації досі залишаються обмеженими і суперечливими [14, p. 3–6].

Серед українських дослідників проблематику постправди в контексті трансформації публічної сфери аналізували В. Волковський (2019) [15], Н. Ковтун (2025) [16] та Я. Коркос (2023) [17]. Зокрема В. Волковський поставив питання про долю філософії в епоху постправди, наголошуючи на загрозі системної ерозії поняття істини [15, с. 85–92]. Натомість Н. Ковтун розглядає феномен постправди через його співвідношення з критичним мисленням, акцентуючи увагу на тому, що протистояння механізмам постправди можливе через «оцінку достовірності джерел інформації, розрахунок витрат і прибутків, розпізнавання регресії до середнього значення, розуміння меж екстраполяції, нівеляцію реактивних реакцій, використання аналогій у міркуваннях, оцінювання сили аргументації, що підтверджують чи заперечують висновки, розпізнавання упередженості у міркуваннях» [16, с. 2]. Я. Коркос проаналізувала витоки та причини епохи постправди у сучасному світі, зосередившись на її соціально-політичних передумовах [17, с. 540–548].

Виділення невирішених раніше частин загальної проблеми. Незважаючи на значний масив досліджень, присвячених окремим аспектам AI-загроз для досягнення істини, залишається невирішеною проблема побудови цілісної феноменологічної типології, яка б: (а) чітко диференціювала якісно

різні рівні AI-загроз за критеріями суб'єкта дії, механізмів впливу та адекватної регулятивної відповіді; (б) здійснила послідовне феноменологічне осмислення кожного рівня через гайдеґґерівські концепти Lethe та Gestell; (в) інтегрувала критику наявних підходів (обмеженість концепту «квазі-інтенціональності», межі аплікативності Gestell до AI, перебільшення загроз AI-дезінформації) у конструктивний теоретичний концепт. Крім того, залишається недостатньо дослідженою проблематика приписування AI редукованої форми людської інтенціональності – від політико-філософських реконструкцій М. Коекельберґа [11] до ширших дискусій про epistemic agency [10]. Саме реалізація цієї проблематики визначає наукову новизну запропонованого дослідження.

Мета та завдання дослідження. Метою статті є розробка феноменологічної типології AI-загроз для досягнення істини, що інтегрує емпіричні дані 2022-2026 рр. у гайдеґґерівські онтологічні концепти Lethe/Gestell. Для досягнення цієї мети поставлено такі завдання: 1) обґрунтувати аналітичну типологію трьох рівнів AI-загроз для досягнення істини (інструментальний, алгоритмічний, квазі-автономний / генеративний як граничний аналітичний тип), чітко диференціювавши їх за суб'єктом дії, механізмами впливу та регулятивною відповіддю; 2) здійснити феноменологічне осмислення кожного з означених трьох рівнів AI-загроз через модальності Lethe; 3) обґрунтувати інтерпретацію AI як розширення впливу логіки Gestell на мову, увагу та публічну видимість; 4) критично оцінити межі запропонованої типології та альтернативні підходи до неї.

Дискусії та результати дослідження.

1. Від метафори до типології: три рівні AI-загроз для досягнення істини. Першим методологічним кроком до розробки типології є відмова від нерозчленованого уявлення про AI як «єдину» загрозу для правди і побудова диференційованої типології. Дослідження 2022-2026 рр. - від таксономії

ризиків Л. Вайдінгера та ін. (2022) [1] до аналізу операцій впливу Дж. Ґолдштейна та ін. (2023) [2] – дозволяють реконструювати принаймні три якісно різні рівні, на яких AI може взаємодіяти з дезінформацією. Ці три рівні – інструментальний, алгоритмічний і квазі-автономний / генеративний як граничний – розрізняються за чотирма вимірами: а) суб'єкт дії, б) механізми впливу, в) типові приклади, г) адекватна регулятивна відповідь.

На інструментальному рівні суб'єктом дезінформації залишається людина-пропагандист, натомість AI виконує роль «промислового обладнання» для масштабування операцій впливу. Типовим прикладом цього є використання LLM для масового виробництва текстів для ботоферм. За твердженням Дж. Ґолдштейна, регулятивні відповіді тут спрямовані на моніторинг, атрибуцію та контроль доступу до API великих моделей [2, p. 25-30]. На алгоритмічному рівні суб'єктом постає платформа як системний актор: не конкретна людина з конкретним наміром, а технологічна архітектура, спроєктована для максимізації залученості користувачів. Алгоритм рекомендацій YouTube, TikTok чи Facebook не «обирає» дезінформацію свідомо, а системно надає перевагу контенту з вищою залученістю аудиторії, що непропорційно часто має дезінформаційний характер. Американські дослідники на чолі з Р. Патаком (2023) [22] на основі емпіричного моделювання довели, як рекомендаційні алгоритми сприяють поширенню дезінформації в соціальних мережах. При цьому регулятивна відповідь потребує аудиту і прозорості алгоритмічних систем, що, зокрема, закладено в європейському Digital Services Act. На квазі-автономному / генеративному рівні – як граничному аналітичному типі, що описує тенденцію, а не повністю реалізовану емпіричну дійсність – AI-система сама стає генератором контенту: deepfake-відео, синтетичних зображень, штучних голосів, які, на думку К. Ваккарі та Е. Чедвік (2020) і Дж. Кітцманна та ін. (2020), неможливо онтологічно відрізнити від автентичних [6; 9]. При цьому регулятивна

відповідь потребує принципово нових підходів: цифрового маркування контенту (watermarking), автоматизованої детекції та нових правових норм.

Ця типологія є не лише дескриптивною, а й аналітично конститутивною. У парадигмі феноменологічного аналізу три типи розташовані на шкалі зростаючого відчуження від людської інтенціональності: від інструменту через систему до граничного випадку квазі-автономної генерації. Її перевага полягає у точнішому розрізненні режимів відповідальності: людина-пропагандист на інструментальному рівні, платформа – на алгоритмічному, генеративна AI-система – на квазі-автономному / генеративному як граничному аналітичному типі. Водночас ці рівні не є взаємовиключними: у реальних дезінформаційних кампаніях вони часто поєднуються, і саме ця синергія робить згенеровану AI постправду небезпечнішою, ніж будь-який із рівнів окремо.

2. Інструментальний рівень: LLM як зброя масштабування. Інструментальний рівень AI-загроз є найкраще дослідженим і задокументованим: людина-пропагандист свідомо використовує LLM як інструмент для масштабування операцій впливу. При цьому суб'єктом дезінформації залишається людина; натомість AI виконує роль «промислового обладнання».

Американські дослідники Дж. Ґолдштейн та ін. (2023) [2] виокремили чотири стадії «конвеєра» дезінформації: побудова/адаптація моделі; отримання доступу через API або відкритий код; дисемінація AI-генерованих текстів через ботоферми і фейкові медіа; формування переконань через повторення, фреймінг та емоційну маніпуляцію. На інструментальному рівні переважає підсилення: людина залишається «архітектором» кампанії, але LLM дозволяє їй працювати у масштабах, що раніше вимагали десятків операторів [2, p. 9–12]. М. Мюссер (2023) переконливо деконструював міф про AI як «магічну» технологію дезінформації: навіть за відносно скромної частки доданої вартості економія на виробництві контенту AI може бути досить

суттєвою, а в граничному випадку високоякісних моделей - сягати приблизно 70% [3, p. 3–5]. Відтак LLM є не універсальною «абсолютною зброєю», а економічно виправданим раціональним інструментом. Водночас контроль через відстеження запитів до API комерційних моделей – з метою виявлення масових автоматизованих кампаній – має обмежений ефект, оскільки дедалі доступнішими для локального запуску без централізованих обмежень стають відкриті моделі на кшталт Llama [3, p. 5–8].

Група дослідників на чолі з А. Вільямс у 2024 році у серії з 2340 експериментів встановила, що практично всі великі мовні моделі після 2022 року здатні генерувати дезінформаційний контент, який у понад 50% випадків для споживачів неможливо відрізнити від створеного людиною – за винятком тих випадків, коли модель відмовляється від потенційно шкідливого запиту [4, p. 8–10]. Відтак навіть без спеціальних налаштувань «стандартні» LLM є достатньо потужними для виробництва у значних масштабах переконливого дезінформаційного продукту. Водночас під час експериментів було виявлено парадокс: моделі з захисними обмеженнями (guardrails) відмовляються і від нешкідливих запитів подібної структури, що ставить під сумнів можливість ефективного захисту від дезінформації без шкоди для функціональності. У контексті цього Е. Феррара на основі реальних кейсів виборчої інтерференції виокремив три вектори зловживання у сфері поширення дезінформації: масштабоване генерування контенту ботофермами, створення синтетичних ідентичностей, виробництво персоналізованих дезінформаційних повідомлень. Він також наголошує на структурній асиметрії між вартістю атаки і вартістю захисту [5, p. 5–7].

Доречним є звернення і до критики перебільшення загрози використання ШІ в різних сферах інформаційної комунікації. Польські дослідники А. Лабуз та К. Негрінг у публікації 2024 року зауважили, що значна частина тривог щодо інструментального зловживання AI може виявитися перебільшеною [14].

Однією з причин такого перебільшення є те, що емпіричні дані про реальну ефективність AI-генерованої дезінформації (на відміну від лабораторних умов) залишаються обмеженими. У зв'язку з цим М. Мюссер зауважує, що цифра 70% ефективності стосується не «середньостатистичного» сценарію, а граничного випадку функціонування високонадійної моделі ШІ [3]. Відтак зниження вартості виробництва інформаційного продукту не можна автоматично прирівнювати до зростання його переконливості [14, p. 4–5]. Ця поправка не скасовує загрозу, але дає змогу точніше оцінювати її масштаб і не скочуватися до масованої моральної паніки щодо використання технологій AI.

3. Алгоритмічний рівень: платформне підсилення як системна загроза для поширення дезінформації. Алгоритмічний тип AI-загроз принципово відрізняється від інструментального тим, що в ньому відсутня людина-пропагандист як свідомий суб'єкт діяльності. На алгоритмічному рівні загрозу для істини створює сам алгоритм рекомендацій – не як інструмент у руках людини, а як безсуб'єктна система оптимізації. Суб'єктом активності постає платформа як системний актор: не конкретна людина з її переконаннями і намірами, а технологічна архітектура, спроектована для максимізації залученості. Американські дослідники на чолі з Р. Патаком на основі емпіричного моделювання продемонстрували, як рекомендаційні алгоритми структурно сприяють поширенню дезінформації в соціальних мережах [22]. Інша група дослідників на чолі з Л. Вайдінгером окреслила ширший діапазон інформаційних ризиків, в межах якого алгоритмічне підсилення дезінформації постає як одна з його важливих форм [1].

LLM, інтегровані в пошукові системи і рекомендаційні алгоритми, формують «інформаційний ландшафт» не як нейтральні посередники, а як активні конструктори того, яка інформація для споживача є «видимою».

На відміну від інструментального рівня, на алгоритмічному рівні немає конкретного «зловмисника», який маніпулює інформацією. Дезінформація

виникає як побічний продукт процесів оптимізації залученості споживачів інформації. Саме тому правові межі регуляції, що базуються на презумпції суб'єктного наміру, не пристосовані для оцінки функціонування знеособлених алгоритмів. Новітнє законодавство ЄС (Digital Services Act та AI Act) зміщує акценти від намірів суб'єкта до системних ризиків функціонування архітектоніки платформ ШІ, що вказує на перехід від індивідуальної до системної відповідальності.

Це означає, що слід звернутися до досліджень британських дослідників М. Хікса, Дж. Хамфріса та Дж. Слейтера (2024) [23] та американського філософа Г. Франкфурта (2005) [20], які говорять про нейтральний характер AI. Такі системи у широкому розумінні «не брешуть і не говорять правду», а є структурно байдужими до самого розрізнення правди/брехні. Їх підходи узгоджуються з концептуалізацією М. Коекельберга (2022) про обмеженість функції AI щодо розрізнення правди і брехні [11]. Вони добре описують алгоритмічний рівень AI-загроз, проте не охоплюють інструментального використання LLM людьми-пропагандистами і не вичерпують проблеми синтетичного продукування реалістичного контенту [23]. Саме тому спроби розв'язання означеної проблеми потребують доповнення типологією різних модальностей Lethe.

4. Квазі-автономний / генеративний рівень як граничний аналітичний тип: deepfakes, «epistemic corrosion» та liar's dividend. Третій рівень AI-загроз ми розуміємо як граничний аналітичний тип, що окреслює теоретичні напрямки розв'язання поставленої проблеми, але ще не повністю базується на емпірично дослідженій дійсності. Важливо підкреслити: ми свідомо уникаємо визнання онтологічного статусу «повністю автономної» AI-дезінформації. Квазі-автономний / генеративний рівень є граничним аналітичним типом у веберівському сенсі - концептом, що не реалізується у чистій формі, а лише окреслює напрямки розвитку відповідних алгоритмів AI.

Його включення до типології є методологічно необхідним: по-перше, він логічно завершує шкалу зростаючого відчуження AI від людської інтенціональності; по-друге, без нього лишаються недоописаними вже задокументовані ефекти «liar's dividend» («дивідендів брехуна») [7] і структурні демократичні загрози поширення deepfakes [8]. На цьому рівні генеративна AI-система продукує інформаційний контент, який може функціонувати як реалістична заміна автентичного. На цьому рівні немає ні людини-пропагандиста, ні платформної системи, що лише побічно підсилює загрозу дезінформації. В онтологічній реальності існує генеративна AI-система, здатна продукувати синтетичний контент із мінімальним людським втручанням (deepfake-відео, згенеровані зображення, штучні голоси), який дуже часто неможливо відрізнити від автентичного. При цьому слід визнавати і те, що поширення і політичне використання таких продуктів все ще потребує людського рішення.

Британські дослідники К. Ваккарі та Е. Чедвік зафіксували показовий механізм: найбільшою загрозою deepfakes є не прямий обман, а «узагальнена індетермінантність» – загальне зниження довіри до будь-якого відеоконтенту. Учасники експерименту, зіткнувшись із deepfake-відео, переважно не вірили у його справжність, але їхня загальна довіра до відеоконтенту значно знижувалася [6, p. 5–6]. Це спостереження є онтологічно, а не лише епістемологічно значущим: руйнується не конкретне переконання, а сама здатність людини розрізняти реальне і нереальне. Сам механізм «liar's dividend», як показують Р. Чезні та Д. Цитрон, інвертує нормальну презумпцію автентичності: досить лише промовити «це – deepfake», і тягар доведення зміщується на бік обвинувача [7, p. 1758].

Німецька дослідниця М. Павелець виявила три структурних виміри загрози deepfakes для демократії: порушення інклюзивного представництва, коли deepfakes ускладнюють підзвітність і непропорційно вражають вразливі

групи; підрив деліберації, коли у процесі комунікації руйнується доказова база обміну аргументами; а також делегітимація рішень, коли виборці формують свої переконання на основі контенту з невизначеною автентичністю [8, p. 6–10].

Канадська група дослідників на чолі з Дж. Кітцманном запропонувала ключове розрізнення маніпуляції як зміни наявного матеріалу та генерації як створення повністю синтетичного контенту. Маніпуляція є фальсифікацією доказів, а генерація - створенням «симулякрів» як об'єктів без оригіналу. Правові системи, які використовують поняття «фальсифікації», не пристосовані для регулювання «симулякрів», створених AI, адже неможливо «підробити» те, чого не існувало [9, p. 140–142]. Австрійсько-бельгійський філософ М. Коекельберг зауважує, що згенерований AI контент підриває епістемологічну агентність демократичних суб'єктів - їх здатність формувати і переглядати політичні переконання - завдяки безпрецедентному масштабу і системному характеру AI-виходів [10, p. 1345–1348]. Більш того, він відкриває перспективу квазі-автономних AI-систем, що потенційно можуть маніпулювати людьми задля цілей, визначених не людськими намірами, а внутрішньою логікою оптимізації самої системи AI. Ця тенденція на сьогодні постає радше аналітичним горизонтом, а не онтологічною реальністю.

Водночас слід враховувати, що саме усвідомлення існування deepfakes може підривати довіру навіть до легітимного відеоконтенту. Навіть «проінформований» споживач зазнає шкоди: знання про deepfakes не захищає, а підриває базову довіру до візуальних свідчень – будь-який контент може розглядатися як потенційно фальшивий [6, p. 5–6]. Цей ефект «узагальненої індетермінантності» знаходить теоретичне відображення у концепті «liar's dividend» («дивідендів брехуна») Р. Чезні та Д. Цитрон: коли будь-яке відео може виявитися deepfake-ом, тягар доведення автентичності зміщується на бік обвинувача [7, p. 1758–1760]. Дослідник з Королівства Бахрейн Р. Мубарак та

ін. у комплексному огляді методів виявлення deepfakes підкреслюють, що технічні рішення, публічна обізнаність та законодавчі заходи мають розглядатися як взаємодоповнювальні, а не альтернативні відповіді на цю загрозу [18, p. 144497–144499].

Це ускладнює стратегії критичного мислення як ключові механізми подолання дезінформації: зростання знання про deepfakes продукує зростання недовіри до всього, що унеможливлює нормальне сприйняття фактів.Перш ніж безпосередньо застосовувати гайдеґґерівські концепти Lethe та Gestell до пояснення функціонування AI, слід окреслити межі такого перенесення. Для осмислення AI-загроз для досягнення істини можна залучати й інші теоретичні концепти. Зокрема, французький філософ Ж. Бодрійяр запропонував концепт симулякрів для дескрипції медіа-реальності [21]. Утім гайдеґґерівські підходи мають методологічну перевагу: концепт Lethe дозволяє розкрити якісну різницю між промисловим масштабуванням, системною байдужістю та генеративним виробництвом симулякрів. Зауважимо, що М. Гайдеґґер формулював свою філософію техніки у 1950-х роках, маючи на увазі передусім індустріальну технологію. Однак за допомогою концепту Gestell він описує не окремий пристрій, а спосіб, у який дійсне розкривається як Bestand – «наявний запас», тобто зарезервований ресурс, мобілізований для подальшого використання [19, S. 24–26, 30–31]. Цей концепт зберігає аналітичну силу і для епохи цифрових технологій.

Британський дослідник І. Томсон підкреслює, що сучасні технології варто мислити не лише на рівні пристроїв, а як частину ширшої трансформації смислових світів сучасної людини [12, p. 1–2, 52]. Проблема полягає не в «застарілості» концепту, а в тому, що в цифрову добу об’єктом конституювання дедалі частіше стають не природні ресурси, а мова, увага і публічна видимість людини.

 ISSN: 3083-5712

У такому розумінні Lethe (λήθη), як антипод aletheia (ἀ-λήθεια, не-прихованість), є не просто «забуванням» або «помилкою», а онтологічним закриттям – станом, в якому буття приховує себе від того, хто мав би його відкрити. Поширення AI додає до цієї структури якісно новий вимір. На кожному з трьох рівнів нашої типології AI реалізує різну модальність Lethe.

На інструментальному рівні AI постає «механізмом масштабування Lethe»: людина-пропагандист масштабує закриття горизонту відкриття. При цьому Lethe є людським творінням, але технологічно воно майже безмежне за масштабом. LLM перетворює індивідуальний акт брехні на промислове виробництво прихованості. У розумінні М. Гайдеґґера мова йде про перетворення людської здатності до брехні у режим масового виробництва. AI не скасовує логіку Gestell, а лише поширює її на сферу мовного виготовлення контенту.

На алгоритмічному рівні AI є «системним Lethe»: закриття горизонту відкриття відбувається без жодної людської інтенції до приховування. Алгоритм рекомендацій не «приховує» правду – він взагалі не оперує категорією правди. Для алгоритму відмінність між правдивим і хибним контентом є такою ж ірелевантною, як відмінність між контентом різного кольору: обидва оцінюються винятково за критеріями залученості. Ми пропонуємо назвати це структурним Lethe: формою прихованості буття, що виникає не з акту приховування, а зі структурної байдужості системи AI до самого питання про відкритість. У термінах М. Гайдеґґера йдеться про Gestell у його цифровій модифікації: технологічну форму, яка перетворює суще – тобто людські думки, бажання, реакції – на Bestand [19, S. 24–26, 30–31]. І. Томсон пропонує бачити в цьому не окремий збій, а ширшу пізньомодерну логіку оптимізації реальності як інформаційного ресурсу [12, p. 52].

На квазі-автономному / генеративному рівні AI постає «генеративним Lethe»: він не масштабує людську брехню і не ігнорує розрізнення

правди/брехні – він просто генерує нову реальність, що не має жодного відношення до будь-якої «первинної» людської реальності. Deepfake-відео не є ні «правдою», ні «брехнею» у традиційному сенсі: воно є симулякром – об'єктом без оригіналу, знаком без референта. Lethe у такому розумінні – це не просто «відсутність правди», а структурна закритість, що робить саме питання про правду неможливим. В середовищі, насиченому синтетичним контентом, не лише «правду» неможливо відрізнити від «брехні» – сама потреба у такому розрізненні починає здаватися зайвою [9; 10].

Водночас критика концепту «синтетичної квазі-інтенціональності» залишається вкрай важливою. Тенденція до приписування AI редукованої форми людської інтенціональності – зокрема у політико-філософських реконструкціях М. Коекельберґа, де йдеться про структурну байдужість AI до істини та про епістемологічного суб'єкта у демократичному контексті [11; 10] на сучасному етапі розвитку є концептуально проблематичною. Вона розмиває відмінності між різними типами AI-загроз і водночас надмірно антропоморфізує систему, безпідставно переносячи на неї людську структуру пізнавальної інтенціональності.

Запропонована нами типологія показує іншу картину: на різних рівнях AI вступає у різні відношення з істиною, а його принципова новизна полягає не в появі «штучного суб'єкта», а в технологічному виробництві режимів закритості інформації. Саме тому ми пропонуємо типологію «трирівневого Lethe»: AI як механізм, що реалізує закриття горизонту відкриття на інструментальному, системному та генеративному рівнях.

Gestell у розумінні М. Гайдеґґера є не апаратом чи пристроєм, а способом розкриття дійсного – тим рамковим конституюванням, в якому суще відкривається виключно як Bestand [19, S. 24–26, 30–31, 37–38]. Якщо класичний Gestell конституював природу як Bestand (вітер - як енергетичний ресурс, природу – як сировину), то в цифрову добу та сама логіка

конституювання поширюється на нові об'єкти. Сьогодні в цю логіку залучається людська мова – через LLM, що перетворює висловлювання на статистичний ресурс; людська увага – через алгоритми, що оптимізують її як рекламний інвентар; та сама публічна присутність особи – через deepfake-технології, що перетворюють людське обличчя і голос на синтетично відтворюваний матеріал. У розумінні І. Томсона генеративний AI позначає ту точку, в якій логіка конституювання охоплює вже не лише природні, а й семантичні та перцептивні ресурси людини [12, p. 52]. Саме тому AI доцільно тлумачити не як «новий Gestell», а як нову історичну фазу розгортання тієї самої логіки конституювання - на терені мови, уваги і видимості. Перехід від індустріальної до цифрової фази Gestell і є одним із теоретичних результатів цієї статті.

6. Критика і межі використання запропонованої типології. Обґрунтування будь-якої типології неминуче ставить питання про її межі використання. Нижче ми окреслимо ті застереження, які роблять запропоновану рамку чесною і науково відповідальною.

Ми частково визнаємо критику А. Лабуз та К. Негрінг про «перебільшення загроз» AI-дезінформації і те, що ці загрози стосуються передусім інструментального рівня [14]. Також А. Вільямс зауважує, що зниження вартості виробництва дезінформації не є тотожним зростанню її ефективності [4], а відтак не переноситься автоматично на реальну ефективність кампаній впливу. Запропонована трирівнева типологія не виходить з позиції, що всі три рівні загроз є однаково небезпечними: інструментальний рівень є найкраще дослідженим і доведеним – і він же найкраще контрольований [14, p. 5–6].

Не менш важливо, що наша інтерпретація загроз AI як розширення логіки Gestell в епоху цифровізації спирається на оригінальні тексти М. Гайдеґґера [19, S. 24–31, 37–38] і сучасні інтерпретації праць німецького

мислителя британського дослідника І. Томсона [12, p. 1–2, 52]. При цьому вона є власним теоретичним науковим здобутком, хоча й потребує подальшого обґрунтування - зокрема, в контексті співвідношення з концептуальними підходами, що по-іншому описують симуляцію, пам'ять і технічне посередництво в процесі поширення інформації.

Квазі-автономний / генеративний рівень нашої типології свідомо сконструйований як граничний аналітичний тип: «повністю автономна» AI-генерація дезінформації без людської ініціативи залишається швидше антиципованою загрозою, ніж масово документованою практикою. Р. Мубарак та його співавтори справедливо зауважують, що навіть deepfakes потребують людського рішення про їх поширення; а повністю автономна AI-дезінформація є поки що теоретичним граничним випадком [18, p. 144510–144515]. Відтак ми зберігаємо цей рівень у типології саме як граничний: він логічно завершує шкалу зростаючого відчуження від інтенціональності і водночас достатньо підкріплений емпірично засвідченими deepfake-кейсами.

Окремо слід враховувати концептуалізацію М. Хікса, Дж. Хамфріса та Дж. Слейтера, які обґрунтували, що LLM є «bullshit-машинами», а не «агентами» у феноменологічному сенсі - системами, чия робота є структурно байдужою до розрізнення істини й хиби [23]. Ці підходи узгоджуються з ширшою критикою М. Коекельберга щодо структурної байдужості AI до істини [11]. Саме тому ми відмовляємося від метафори «квазі-інтенціональності» на користь типології модальностей Lethe, яка не потребує приписування AI будь-якої форми інтенціональності.

Висновки. Запропонована типологія дозволяє точніше описати спосіб, в який технології AI змінюють умови досягнення істини в епоху постправди. Встановлено, що AI-загрози для досягнення істини не утворюють єдиного феномена. Інструментальний, алгоритмічний та квазі-автономний /

генеративний рівні розрізняються за суб'єктом дії, механізмом впливу та типом регулятивної відповіді.

Запропонована в статті типологія демонструє, що AI-загрози для досягнення істини не є однорідним явищем, а формують три якісно різні режими впливу. Інструментальний рівень пов'язаний із економічним здешевленням виробництва дезінформації, алгоритмічний – із системним перекроюванням інформаційного ландшафту рекомендаційними системами, а квазі-автономний / генеративний – із продукуванням симулякрів, що підривають саму можливість розрізнення реального й штучного. У цьому сенсі штучний інтелект не просто підсилює наявні механізми маніпуляції, а змінює онтологічні умови відношення до істини. Феноменологічне прочитання цих трьох рівнів через модальності Lethe дозволяє описати, як саме AI розгортає різні форми прихованості. Інструментальний Lethe постає як масштабування людської практики брехні до промислових масштабів, системний Lethe – як байдужість алгоритмів до питання істинності, генеративний Lethe – як виробництво нової, синтетичної «реальності» без оригіналу. Такий підхід дає змогу сфокусуватися не лише на окремих ризиках, а на глибинних змінах у структурі відкритості й закритості буття.

Інтерпретація AI як нової фази розгортання логіки Gestell показує, що у цифрову добу в якості ресурсів починають розглядатися не тільки природні об'єкти, а й мова, увага та публічна присутність людини. Мовні висловлювання, клік-динаміка і навіть зображення обличчя стають тим, що можна стандартизувати, комбінувати й відтворювати у вигляді синтетичного матеріалу. Це змушує по-новому поставити питання про межі технічного конструювання публічної реальності та про відповідальність за її архітектуру.

Практичний сенс запропонованої типології полягає в можливості диференційованого реагування на виклики постправди. Інструментальний рівень потребує передусім інструментів моніторингу, атрибуції та

стримування масштабування дезінформаційних кампаній; алгоритмічний – прозорості й аудиту платформи як цілого; генеративний – нових стандартів доказовості, етики використання синтетичних медіа та розвитку культурних практик розрізнення. Розмежування цих рівнів дозволяє уникнути як недооцінки, так і панічного перебільшення AI-загроз.

Список використаних джерел

1. Weidinger L., Mellor J., Rauh M. et al. Taxonomy of risks posed by language models. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022. P. 214–229. URL: https://doi.org/10.1145/3531146.3533088 (дата звернення: 18.03.2026).

2. Goldstein J., Sastry G., Musser M., DiResta R., Gentzel M., Sedova K. Generative language models and automated influence operations: Emerging threats and potential mitigations. 2023. URL: https://doi.org/10.48550/arXiv.2301.04246 (дата звернення: 18.03.2026).

3. Musser M. A cost analysis of generative language models and influence operations. 2023. URL: https://doi.org/10.48550/arXiv.2308.03740 (дата звернення: 19.03.2026).

4. Williams A. R., Burke-Moore L., Chan R. S.-Y. et al. Large language models can consistently generate high-quality content for election disinformation operations. 2024. URL: https://doi.org/10.48550/arXiv.2408.06731 (дата звернення: 20.03.2026).

5. Ferrara E. GenAI against humanity: nefarious applications of generative artificial intelligence and large language models. Journal of Computational Social Science. 2024. Vol. 7. P. 549–569. URL: https://doi.org/10.1007/s42001-024-00250-1 (дата звернення: 20.03.2026).

6. Vaccari C., Chadwick A. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news.

Social Media + Society. 2020. Vol. 6(1). P. 1–13. URL: https://doi.org/10.1177/2056305120903408 (дата звернення: 17.03.2026).

7. Chesney R., Citron D. K. Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*. 2019. Vol. 107(6). P. 1753–1820. URL: https://doi.org/10.15779/Z38RV0D15J (дата звернення: 24.04.2026).

8. Pawelec M. Deepfakes and democracy (theory): How synthetic audio-visual media for disinformation and hate speech threaten core democratic functions. *Digital Society*. 2022. Vol. 1. Art. 19. URL: https://doi.org/10.1007/s44206-022-00010-6 (дата звернення: 17.03.2026).

9. Kietzmann J., Lee L. W., McCarthy I. P., Kietzmann T. C. Deepfakes: trick or treat? *Business Horizons*. 2020. Vol. 63(2). P. 135–146. URL: https://doi.org/10.1016/j.bushor.2019.11.006 (дата звернення: 18.03.2026).

10. Coeckelbergh M. Democracy, epistemic agency, and AI: political epistemology in times of artificial intelligence. *AI and Ethics*. 2023. Vol. 3. P. 1341–1350. URL: https://doi.org/10.1007/s43681-022-00239-4 (дата звернення: 19.03.2026).

11. Coeckelbergh M. The political philosophy of AI: An introduction. Cambridge: Polity Press, 2022. 176 p.

12. Thomson I. D. Heidegger on Technology's Danger and Promise in the Age of AI. Cambridge: Cambridge University Press, 2025. 74 p. URL: https://doi.org/10.1017/9781009629423 (дата звернення: 25.03.2026).

13. Gabriel I. Artificial Intelligence, values and alignment. *Minds and Machines.* 2020. Vol. 30. P. 411–437. URL: https://doi.org/10.1007/s11023-020-09539-2 (дата звернення: 19.03.2026).

14. Łabuz A., Nehring E. Information apocalypse or overblown fears - what AI mis- and disinformation is all about? Shifting away from technology toward

human reactions. *Politics & Policy*. 2024. Vol. 52(4). P. 874–891. URL: https://doi.org/10.1111/polp.12617 (дата звернення: 20.03.2026).

15. Волковський В. Доля філософії в епоху постправди: слуга короля чи клоун на агорі. *Україна модерна*. 2019. № 26. С. 80–98.

16. Ковтун Н. М. Критичне мислення у період постправди в розумінні Даяни Халперн: епістемологічні та методологічні аспекти. *Вісник гуманітарних наук.* 2025. № 9. С. 1–21. ISSN 3083-5712. DOI: https://doi.org/10.5281/zenodo.16520516.

17. Коркос Я. О. Епоха постправди у сучасному світі: витоки та причини. *Наукові перспективи*. 2023. № 4(34). С. 537–554. URL: https://doi.org/10.52058/2708-7530-2023-4(34)-537-554 (дата звернення: 17.03.2026).

18. Mubarak R., Alsboui T. A. A., Alshaikh O. et al. A survey on the detection and impacts of deepfakes in visual, audio, and textual formats. *IEEE Access*. 2023. Vol. 11. P. 144497–144529. URL: https://doi.org/10.1109/ACCESS.2023.3344653 (дата звернення: 19.03.2026).

19. Heidegger M. Die Frage nach der Technik. Vorträge und Aufsätze. Pfullingen: Günther Neske, 1954. S. 9–40.

20. Frankfurt H. On Bullshit. Princeton: Princeton University Press, 2005. 80 p.

21. Baudrillard J. Simulacra and Simulation / Trans. by Sh. F. Glaser. Ann Arbor: University of Michigan Press, 1994. 164 p.

22. Pathak R., Spezzano F., Pera M. S. Understanding the contribution of recommendation algorithms on misinformation recommendation and misinformation dissemination on social networks. *ACM Transactions on the Web*. 2023. Vol. 17(4). Art. 35. URL: https://doi.org/10.1145/3616088 (дата звернення: 20.03.2026).

23. Hicks M. T., Humphries J., Slater J. ChatGPT is bullshit. *Ethics and Information Technology.* 2024. Vol. 26. Art. 38. URL: https://doi.org/10.1007/s10676-024-09775-5 (дата звернення: 20.03.2026).

 ISSN: 3083-5712