

ІМПЛІЦИТНІ ТЕОРІЇ ШТУЧНОГО ІНТЕЛЕКТУ ЯК ПРОБЛЕМНЕ ПОЛЕ ПСИХОЛОГІЧНИХ ДОСЛІДЖЕНЬ

IMPLICIT THEORIES OF ARTIFICIAL INTELLIGENCE AS A PROBLEM FIELD OF PSYCHOLOGICAL RESEARCH

У статті розглянуто феномен імпліцитних теорій штучного інтелекту (ІТШІ) як нового міждисциплінарного напрямку психологічних досліджень, що формується на перетині когнітивної, соціальної та медіапсихології. Обґрунтовано, що стрімке поширення технологій штучного інтелекту спричинило появу не лише нових експліцитних (наукових) моделей інтелекту, а й значного пласта неусвідомлених, інтуїтивних, емоційно забарвлених уявлень про «розумні машини». Імпліцитні теорії визначають довіру, ставлення та поведінку користувачів у цифровому середовищі, впливають на сприйняття автономності, морального статусу й надійності технологій. У статті систематизовано підходи до аналізу ІТШІ, розглянуто їхню структуру (когнітивний, емоційний, ціннісно-смысловий та поведінковий компоненти), джерела формування (масова культура, освіта, медіа, особистий досвід) та основні функції – регуляторну, мотиваційну, соціально-ідентифікаційну й антиципаційну. Описано типові прояви імпліцитних теорій – лінгвістичну, візуальну та моральну антропоморфізацію ШІ, що спрощує розуміння технологій, але водночас спотворює їх оцінювання. Показано зв'язок ІТШІ з феноменами надмірної довіри, маніпулятивної вразливості та прийняття алгоритмічних рішень. Запропоновано розглядати ІТШІ як когнітивно-афективні конструкції, що формують «ментальні моделі» взаємодії людини з інтелектуальними агентами та визначають психологічну готовність до використання ШІ. Окремлено напрями подальших досліджень – розроблення психодіагностичних інструментів вимірювання ІТШІ, міжкультурні порівняння, а також інтеграцію здобутих результатів у психологію цифрової компетентності, освіти й політику відповідального впровадження ШІ. Підкреслено, що розуміння імпліцитних уявлень про ШІ є необхідною умовою формування реалістичної та етично виваженої взаємодії між людиною і технологіями в цифрову епоху.

Ключові слова: імпліцитні теорії, імпліцитні теорії штучного інтелекту, експліцитні теорії інтелекту, інтелект, особистість, когнітивна психологія, довіра до ШІ, антро-

поморфізація, когнітивно-афективні моделі, психологія технологій, цифрова компетентність.

The article examines the phenomenon of implicit theories of artificial intelligence (ITAI) as a new interdisciplinary direction of psychological research emerging at the intersection of cognitive, social, and media psychology. It is argued that the rapid expansion of artificial intelligence technologies has led not only to the creation of new explicit (scientific) models of intelligence but also to the formation of a vast layer of unconscious, intuitive, and emotionally charged representations of “intelligent machines”. Implicit theories determine users’ trust, attitudes, and behavior in the digital environment, influencing their perception of autonomy, moral status, and reliability of AI systems. The paper systematizes theoretical approaches to ITAI analysis and describes their structure (cognitive, emotional, value-semantic, and behavioral components), sources of formation (mass culture, education, media, and personal experience), and main functions – regulatory, motivational, social-identification, and anticipatory. Typical manifestations of ITAI are outlined, including linguistic, visual, and moral anthropomorphism, which simplifies understanding of AI but also distorts its evaluation. The relationship between ITAI and phenomena such as overtrust, manipulative vulnerability, and algorithmic decision-making is revealed. ITAI are proposed to be understood as cognitive-affective constructs that form “mental models” of human interaction with intelligent agents and determine psychological readiness for AI use. The paper outlines directions for future research – developing psychodiagnostic tools for measuring ITAI, conducting cross-cultural comparisons, and integrating findings into the psychology of digital competence, education, and responsible AI implementation policy. It is emphasized that understanding implicit beliefs about AI is a prerequisite for developing realistic and ethically sound human-technology interaction in the digital age.

Key words: implicit theories, implicit theories of artificial intelligence, explicit theories of intelligence, intelligence, personality, cognitive psychology, trust in AI, anthropomorphism, cognitive-affective models, psychology of technology, digital competence.

УДК 159.955.4:004.8
DOI <https://doi.org/10.32782/hbts.78.1.7>
Стаття поширюється на умовах
ліцензії CC BY 4.0

Батраченко І.Г.

д.психол.н., професор кафедри
загальної та соціальної психології
Дніпровський національний університет
імені Олеся Гончара
ORCID ID: 0000-0003-1699-7047

Кваша А.О.

аспірант кафедри загальної
та соціальної психології
Дніпровський національний університет
імені Олеся Гончара
ORCID ID: 0009-0002-5742-5045

Постановка проблеми. Бурхливий розвиток штучного інтелекту (ШІ) у XXI столітті спричинив не лише технологічну, а й когнітивно-культурну революцію. Якщо раніше людський інтелект був єдиною моделлю мислення, то нині поряд із ним з'являються алгоритмічні форми оброблення інформації, здатні до навчання, адаптації та творчих активів. Проте поряд з експліцитними (науковими) моделями штучного інтелекту формується й потужний пласт імпліцитних уявлень, що вкорінені в масовій свідомості, культурних міфах, художніх образах, інтуїтивних судженнях про «розумні машини». Саме ці імпліцитні теорії визначають емоційне ставлен-

ня людей до штучного інтелекту, рівень довіри, страху або прийняття технологій, і тим самим суттєво впливають на соціально-психологічні процеси. Отже, актуальним стає питання: яким чином психологія може досліджувати, концептуалізувати та змінювати імпліцитні уявлення про штучний інтелект, що формуються поза межами академічного дискурсу, але мають потужний регулятивний потенціал для суспільної поведінки.

Аналіз останніх досліджень і публікацій. У сучасній психологічній літературі поняття імпліцитних теорій активно використовується для пояснення неусвідомлених когнітивних струк-

тур, що впливають на сприймання, оцінювання і прогнозування поведінки [1; 16]. Дослідження імпліцитних теорій інтелекту людини показали, що уявлення про «фіксованість» чи «змінність» здібностей визначає мотивацію, самооцінку і навчальну успішність. Водночас у сфері штучного інтелекту наукові студії зосереджувалися переважно на експліцитних моделях – архітектурах нейромереж, алгоритмах машинного навчання, когнітивних симуляціях. Лише останніми роками почали з'являтися праці, що вивчають людське сприйняття штучного інтелекту, його антропоморфізацію, формування довіри до машин, страхи перед автономністю технологій.

Проблематика імпліцитних теорій штучного інтелекту (ШІ) поступово виокремлюється в самостійне міждисциплінарне поле на перетині когнітивної, соціальної та медіапсихології, етики технологій і правових студій. Її методологічним підґрунтям слугують класичні підходи до імпліцитних теорій інтелекту та особистості, що розглядають наївні уявлення як відносно стійкі когнітивні схеми, котрі скеровують інтерпретацію досвіду, емоційні реакції та поведінку [1; 6; 16; 17]. У межах цих підходів доведено зв'язок «теорії зростання/сталості» з навчальними досягненнями, мотивацією та саморегуляцією [14; 15]. Перенесення цієї рамки на сферу ШІ дозволяє трактувати «імпліцитні теорії ШІ» як сукупність буденних переконань про природу, можливості, межі та моральний статус інтелектуальних систем, що впливають на довіру, прийняття рішень і готовність до взаємодії з алгоритмічними агентами.

Окреслюючи емпіричний стан питання, слід виокремити принаймні три лінії досліджень. Перша – це роботи про довіру та відновлення довіри до інтелектуальних агентів. Експерименти Т. Кіма та Х. Сонга показали, що вірування користувачів щодо «здатності ШІ вчитися на помилках» модулюють процес «trust repair» після збоїв: прихильники «теорії зростання» щодо ШІ швидше відновлюють довіру, якщо агент демонструє ознаки навчання та виправлення [2; 3]. Друга лінія – наївні/народні теорії («folk theories») та медіанаративи: дослідження в різних культурних контекстах (США, Бразилія) фіксують, що масові уявлення про ШІ формуються з міксу медійних метафор, персональних історій і культурних тропів, часто далекого від технічної реальності [5; 7; 18]. Третя лінія стосується прийняття алгоритмічних рішень: вірування щодо компетентності/об'єктивності ШІ пов'язані з готовністю покладатися на алгоритмічні поради в прикладних доменах (маркетинг, управління, публічна політика) [4; 11].

Концептуально важливий внесок роблять праці про антропоморфізацію ШІ як механізм формування імпліцитних уявлень. Н. Барроу показує, що мова («ШІ бачить/чує/говорить», «галюцинує»), ілюстративні штампи (людиноподібні роботи, «термінатори») та моральні атрибуції (приписування відчуттів, страждання чи прав) спричиняють систематичне перебільшення можливостей ШІ і під-

живлюють «AI-hype» [12]. Це безпосередньо змикається з психологічними спостереженнями Насс і Мун про «соціальні відповіді на комп'ютери», коли люди інтуїтивно застосовують до машин соціальні скрипти, призначені для взаємодії з людьми [11]. Таким чином, антропоморфізація виступає когнітивною «швидкою евристикою», яка полегшує розуміння нового явища, але водночас спотворює оцінку ризиків/користі й ускладнює етичну та правову розвідку.

У медіа- та соціально-психологічній перспективі вказані імпліцитні моделі підсилюються популярними культурними наративами. Огляди, присвячені суспільній рецепції ШІ, фіксують швидке поширення «великих історій» – від романтизованої автоматизації до сценаріїв штучного загального інтелекту (AGI) і трансгуманізму [7; 8]. Водночас критичні роботи підкреслюють, що теперішні наслідки інтеграції ШІ (упередження даних, непрозорість, процедурна справедливість) мають більше практичне значення, ніж віддалені спекуляції [12]. На перетині філософії та гуманітаристики осмислюється антропотехнологічний вимір ШІ як чинник трансформації людських практик і уявлень про суб'єктність [23].

Український сегмент досліджень додає важливий контекст – зв'язок імпліцитних теорій з мотиваційно-регуляторними та копінг-механізмами. Показано асоціації між імпліцитними теоріями інтелекту/особистості та губристичною мотивацією, саморегуляцією, копінг-стратегіями, а також окреслено перспективність виділення «імпліцитної психології» як напрямку [13; 20; 21; 24]. Низка праць актуалізує проблематику імпліцитних теорій у суміжних доменах – від правосвідомості («імпліцитні концепції права») до ментального здоров'я студентів та командної взаємодії в умовах воєнного стану [25–27]. Це створює підґрунтя для екстраполяції методик і аналітичних рамок на сферу ШІ – з урахуванням локальних культурно-історичних чинників.

Синтезуючи здобуте, можна виділити кілька найбільш емпіричних закономірностей. По-перше, імпліцитні теорії ШІ впливають на динаміку довіри: очікування щодо «навчованості» і «покращуваності» агента підвищують імовірність прийняття вибачень/виправлень системи після помилок [2; 3]. По-друге, медійні та культурні наративи й образи (у т.ч. антропоморфні) конституують «наївні моделі», які задають рамку сприйняття функцій і ризиків ШІ, часто відсуваючи «приземлені» питання упереджень і підзвітності [5; 7; 12; 18]. По-третє, лей-переконання щодо об'єктивності/компетентності алгоритмів корелюють із готовністю використовувати алгоритмічні поради, причому цей ефект модулюється ситуаційними та індивідуально-психологічними факторами [4; 11].

Водночас існують помітні прогалини. Зокрема, методологічно бракує валідованих інструментів саме для вимірювання імпліцитних теорій ШІ (на відміну від класичних шкал щодо людського інтелекту). Поточні дослідження часто спираються на

адаптовані опитувальники або авторські шкали з обмеженою перевіркою інваріантності вимірювання в різних культурах [2–5; 18]. У теоретичному плані потрібна більш чітка диференціація конструктів – відрізнення вірувань про «можливості системи» (competence), «моральний статус» (moral patiency/agency) і «соціальність» (intention-ascription), які сьогодні часто змішуються під парасолькою «антропоморфізму» [11; 12]. З прикладної точки зору недостатньо робіт про зв'язок імпліцитних теорій ШІ з адаптивністю в цифрових організаціях, зокрема в умовах високої невизначеності, стресу й інформаційного перевантаження (BANI-контекст), кроскультурно: обмежена репрезентація регіонів Центрально-Східної Європи. Перспективним є порівняльний аналіз українських, польських та інших європейських вибірок [19–27].

Отже, сучасний стан поля свідчить: імпліцитні теорії ШІ є реально дієвими когнітивними детермінантами довіри, прийняття рішень і поведінки користувачів; вони живляться антропоморфними метафорами та медіаобразами, а також віддзеркалюють ширші культурні ідеології щодо технологій [4; 7; 11; 12; 18; 23]. Подальший поступ вимагає: а) розроблення та валідації інструментів вимірювання імпліцитних теорій ШІ (з перевіркою кроскультурної інваріантності); б) розрізнення рівнів – технічних вірувань, моральних атрибутів і соціальної персоніфікації; в) дослідження впливу імпліцитних теорій ШІ на адаптацію фахівців і організацій у цифровій екосистемі; г) інтеграції локальних (українських/регіональних) даних у глобальний науковий дискурс [1–6; 11–12; 18–27]. Така програма дозволить перевести обговорення від спекулятивних тропів до операціоналізованих конструктів і практик інтервенції – в освіті, організаційному розвитку та політиці відповідального впровадження ШІ.

Метою статті є теоретичне обґрунтування та системний аналіз імпліцитних теорій штучного інтелекту (ІТШІ) як нового об'єкта психологічного дослідження; визначення їх структури, змісту, джерел формування та функцій у когнітивній, емоційній і соціально-регулятивній сферах; а також розроблення концептуальних підходів до їх експлікації, операціоналізації та психодіагностичного вимірювання в контексті сучасної психології технологічної взаємодії.

Основний виклад матеріалу. Імпліцитні теорії ШІ можуть бути розглянуті як когнітивно-афективні конструкції, що поєднують елементи повсякденних знань про технології, культурні наративи (наприклад, міфи про «бунт машин» або «цифрове спасіння»), а також індивідуальний досвід взаємодії з інтерфейсами, ботами, голосовими асистентами; соціальні дискурси щодо етики, автономії та майбутнього людини. Такого роду ментальні моделі критично важливі, оскільки вони визначають очікування користувачів щодо поведінки системи, впливають на їхні рішення щодо довіри та прийняття технологій, а також модулюють реакцію на

системні помилки. При цьому дослідницький інтерес зосереджений на тому, чи впливають імпліцитні теорії штучного інтелекту на сприйняття інтелектуального агента так само, як імпліцитні теорії про людину впливають на міжлюдську взаємодію, хоча вже встановлено, що прояви та наслідки можуть відрізнитися.

Імпліцитні теорії штучного інтелекту мають багатовимірну структуру. Зокрема, вони включають до свого складу когнітивний компонент (знання та уявлення про можливості, обмеження, принципи дії ШІ), емоційний компонент (страх, цікавість, довіра, відраза, захоплення), ціннісно-смысловий компонент (етичне оцінювання ролі ШІ в суспільстві, питання гідності, контролю, відповідальності) та поведінковий компонент як готовність використовувати або відмовлятися від ШІ, делегувати йому частину своїх функцій. Джерелами формування імпліцитних теорій зазвичай є масова культура (кіно, література, реклама), яка формує архетипи «друга-робота» або «механічного ворога». Освітнє середовище також створює уявлення про технологічну компетентність і межі людського контролю, а медіа і соціальні мережі продукують гібридні образи, де ШІ сприймається одночасно як інструмент і як суб'єкт. Істотне значення також має й особистий досвід особистості щодо взаємодії з чат-ботами, голосовими асистентами, системами рекомендацій.

У функціональному плані імпліцитні теорії ШІ визначають поведінкові патерни в контакт з технологіями, впливають на бажання чи опір щодо використання ШІ, дозволяють людині окреслити своє місце в «тандемі людина – машина», а також формувати очікування щодо майбутнього ШІ (оптимістичні або катастрофічні сценарії). Тобто життєве, повсякденне значення імпліцитних теорій ШІ не менше, а можливо, у певному сенсі й більше, ніж строгих наукових експліцитних теорій, які в повноцінному форматі засвоює значно менша кількість суб'єктів із високим рівнем відповідного напрямку освіти.

Уперше розмежування експліцитних (ETI) та імпліцитних (ITI) теорій інтелекту було проведено Робертом Стернбергом [16]. Експліцитні теорії інтелекту (ETI) постають як структуровані, формалізовані та валідизовані академічні концепції, підтверджені дослідницькими та психометричними інструментами. Їхні основні функції – опис, вимірювання та прогнозування інтелектуальних здібностей у заданих рамках. Натомість імпліцитні теорії інтелекту (ITI) репрезентують персональні, часто неусвідомлені переконання індивіда щодо сутності інтелекту (наприклад, дихотомія Двек: нарощуваний чи заданий інтелект). Їхня ключова функція – інтерпретація та регуляція поведінки та мотивації в досягненні цілей. Імпліцитні смисли виникають на базі експліцитних значень, які взаємодіють із фоновими знаннями, контекстом і ситуацією. Це свідчить про їх тісну взаємозалежність, що дає змогу глибше проникати у формування змісту мислення.

Найбільш помітним прикладом впливу ETI на ITI є робота, заснована на дихотомії Керол Двек. Хоча сама дихотомія (нарощувана vs. задана установка) є моделлю ITI, психологічні інтервенції, розроблені на її основі, використовують експліцитну наукову інформацію для зміни внутрішніх переконань. Тобто навчання індивідів тому, що здібності можна розвивати через зусилля (експліцитна інформація), змінює їхню наївну теорію інтелекту, що, своєю чергою, позитивно впливає на мотивацію, самоефективність та успішність. Як наслідок, нарощувана установка позитивно корелює з показниками інноваційного потенціалу особистості. Такого роду інтервенції, часто реалізовані в освітньому контексті, є прямим доказом того, що науково обґрунтовані (експліцитні) ідеї можуть бути використані для модифікації внутрішніх (імпліцитних) переконань.

З іншого боку, експліцитні концепції, як-от емоційний інтелект або креативність, котрі були формалізовані в академічній психології, часто проникають у масову культуру та освіту. Коли ці концепції популяризуються, вони можуть змінити загальний імпліцитний погляд на те, що означає бути «розумним». Популяризація емоційного інтелекту спонукає людей звертатися до таких аспектів, як емпатія та соціальні навички, що формує більш емпатичний і адаптивний імпліцитний світогляд. Тобто імпліцитні значення виникають із взаємодії експліцитних значень і фонових знань, що підтверджує, що оприлюднена академічна інформація (ETI) стає частиною фонового знання, що формує ITI.

Однак, незважаючи на те, що ETI описують та формалізують інтелект, саме ITI мають найбільш сильний прямий вплив на поведінку та навчання індивіда, бо саме вони створюють систему значень, яка впливає на індивідуальну інтерпретацію навчальних та життєвих ситуацій. Таким чином, ETI надають наукові інструменти для оцінювання інтелекту, але ITI є когнітивними предикторами реалізації потенціалу людини, оскільки вони визначають, чи буде індивід застосовувати наявні інтелектуальні компоненти (метакомпоненти, компоненти виконання) для розвитку та розв'язання проблем.

Імпліцитні теорії ШІ значною мірою впливають на те, як публіка сприймає різні класи інтелекту. Існує поширена когнітивна плутанина, де поточні системи, які є вузьким ШІ (Narrow AI, або Weak AI), як-от ChatGPT, Siri чи рекомендаційні системи, часто помилково приймаються за ШІ загального призначення AGI (Artificial General Intelligence) – аббревіатура від «штучний загальний інтелект» (українською також перекладають як загальний, універсальний або повноцінний штучний інтелект). Це поняття протиставляється нинішнім системам вузького (спеціалізованого) ШІ, які виконують лише конкретні завдання, наприклад, розпізнавання зображень, переклад, прогнозування текстів тощо. Технічно Narrow AI здатний виконувати вузький ді-

апазон завдань, проте йому бракує універсальної гнучкості, самосвідомості, загального мислення чи здатності застосовувати знання в непов'язаних доменах, які є ключовими характеристиками теоретичного AGI.

Таким чином, AGI – це гіпотетичний рівень розвитку штучного інтелекту, на якому система здатна розуміти, навчатися та розв'язувати завдання будь-якого типу так само гнучко й ефективно, як людина. Іншими словами, AGI мав би володіти загальною когнітивною здатністю – переносити знання з однієї галузі в іншу, демонструвати самосвідомість, мотивацію та розуміння контексту, здатність до метапізнання (усвідомлення власних процесів мислення) та мати адаптивну поведінку в нових і невизначених ситуаціях.

З огляду на сказане вище в сучасному науковому дискурсі важливим контекстом аналізу імпліцитних теорій штучного інтелекту є розмежування між вузьким (спеціалізованим) та загальним (універсальним) інтелектом. Під AGI (Artificial General Intelligence) розуміють гіпотетичний рівень розвитку штучних систем, які володіють загальними когнітивними здібностями, порівняними з людськими, здатністю вчитися, узагальнювати досвід і розв'язувати нові типи завдань у різних доменах. На відміну від сучасних систем вузького штучного інтелекту (ANI – Artificial Narrow Intelligence), що виконують окремі функції (розпізнавання образів, переклад, аналіз даних тощо), AGI передбачає гнучкість, контекстну обізнаність і метакогнітивну саморегуляцію.

Отже, поняття AGI має не лише технічний, а й психологічний вимір: воно виступає каталізатором формування імпліцитних уявлень про «машинну свідомість», «розумність» і «моральну суб'єктність» технологій. У масовій свідомості та медійних наративах AGI часто постає у вигляді символу майбутнього злиття людини й машини або як джерело екзистенційних побоювань – втрати контролю, самостійності машин, «сингулярності». Саме тому аналіз уявлень про AGI дає змогу виявити глибинні когнітивно-культурні схеми, що лежать в основі імпліцитних теорій штучного інтелекту та визначають психологічні механізми довіри, страху й очікувань щодо цифрового майбутнього.

Імпліцитний антропоморфізм, вбудований у народні теорії, часто ґрунтується на припущенні, що людський (загальний) інтелект є «реальною» та кінцевою формою інтелекту, яку ШІ має імітувати. Це сприйняття, де термін «штучний інтелект» уже неявно припускає, що він «не цілком реальний», одночасно підживлює високі очікування щодо AGI та скептицизм щодо його здатності досягти істинної людської свідомості (таблиця 1).

Як бачимо, N. Barrow виходить із тези, що антропоморфізація – вроджена когнітивна тенденція людини, через яку люди зазвичай схильні приписувати нелюдським об'єктам людські риси, емоції та інтенції. Це, по суті, і є імпліцитна ментальна мо-

дель ШІ, що виникає автоматично, без усвідомленого аналізу – аналогічно до того, як люди інтуїтивно «розуміють» інших людей. Це збігається з ідеєю імпліцитних теорій інтелекту (natural & artificial) у тому сенсі, що вони є неусвідомленими схемами пояснення, що формують ставлення, довіру, страхи та поведінку щодо ШІ. При цьому він зазначає, що антропоморфізація невідворотна, оскільки допомагає людям інтерпретувати незрозуміле через знайоме, що цілком узгоджується з психологічним підходом до імпліцитних теорій. Останні слугують адаптивними когнітивними схемами, які забезпечують розуміння складних і нових явищ (тут – ШІ) у термінах людського досвіду [12].

Імпліцитні, народні теорії ШІ нечасто ґрунтуються на технічних документах. Більше того, вони значною мірою формуються під впливом зовнішніх культурних наративів, ключовим джерелом яких є наукова фантастика (sci-fi media). У цьому контексті саме медіа та наукова фантастика є потужними джерелами, які можуть посилювати суспільні хибні уявлення та гіперболізовані очікування або страхи щодо ШІ, оскільки наукова фантастика часто зображує ШІ як високорозвинену, свідому (AGI) або навіть суперагентивну сутність і створює розрив між вигаданими, гіперболічними можливостями та фактичними обмеженнями сучасного Narrow AI. Саме про це свідчать дослідження, які вивчають тематичні паралелі між сюжетами наукової фантастики та дискусіями в реальних медіа, підкреслюючи, що ці наративи значною мірою формують суспільні погляди та емоційне сприйняття технології.

Цей культурний вплив створює петлю зворотного зв'язку, де сформовані народні теорії (висока ITAI) призводять до плутанини AGI/Narrow AI. Ця плутанина, у свою чергу, підвищує суспільний інтерес до абстрактних етичних питань, як-от свідомість і моральний статус. Хоча ці філософські питання є важливими, вони можуть відволікати увагу від більш нагальних проблем, які стосуються функціонування сучасного вузького ШІ, – вбудовані упередження, надмірне покладання та ризик маніпуляцій. При цьому ITAI є критичною змінною, що модулює прийняття рішень людиною-оператором у системах ШІ. Так, зокрема, якщо користувач дотримується ентитетної теорії, сприймаючи ШІ як ідеальний, фіксований алгоритм, це може призвести до надмірного покладання (over-reliance) та некритичного прийняття його рекомендацій.

Таким чином, ITAI діє як канал, через який алгоритмічне упередження взаємодіє з когнітивним упередженням людини-оператора. Надмірна довіра, спричинена певною ITAI, зменшує критичну оцінку, дозволяючи упередженості ШІ безперешкодно підсилювати вже наявні упередження людини. Імпліцитні теорії можуть створювати критичні вразливості у взаємодії людини та ШІ. Користувачі часто схильні до надмірної довіри до сприйнятої об'єктивності ШІ, особливо в доменах, які вважаються раціональними, як-от фінанси. ITAIs, які припускають фіксовану, ідеальну або «бездоганну» природу ШІ, створюють ідеальну психологічну мішень для маніпуляцій. Оскільки маніпуляція експлуатує когнітивні упередження та емоційні вразливості ко-

Таблиця 1

Антропоморфізація як прояв імпліцитних теорій штучного інтелекту (за Barrow, 2024)

Ідея з Barrow (2024)	Прояв імпліцитної теорії ШІ	Психологічна інтерпретація
Антропоморфізм як вроджена когнітивна тенденція – люди приписують нелюдським об'єктам людські властивості (емоції, наміри, свідомість)	Імпліцитна теорія «AI як людина» – несвідома схема, що порівнює інтелект ШІ до людського	Когнітивна евристика: спрощене осмислення складного явища через знайомі категорії
Лінгвістична антропоморфізація – використання слів «бачить», «чує», «говорить», «галюцинує» щодо ШІ	Формування імпліцитних уявлень про наявність у ШІ сенсорного досвіду та емоційної компетентності	Мовні патерни відображають глибинні ментальні моделі; повторювані метафори підсилюють віру в «людяність» ШІ
Візуальна антропоморфізація – зображення ШІ у вигляді людоподібних роботів, «розумних» істот, термінаторів	Імпліцитна теорія ШІ як «іншого розуму», здатного до емоцій і волі	Візуальні символи активують соціальні сценарії «зустрічі з Іншим», провокуючи емпатію або страх
Моральна антропоморфізація – приписування ШІ моральних якостей або статусу (як у випадку LaMDA)	Імпліцитна моральна теорія ШІ: несвідоме віднесення до ШІ морального пацієнта або агента	Механізм моральної проєкції: люди інтерпретують поведінку систем через категорії добра, болю, справедливості
Антропоморфізація як джерело «AI гуре» – переоцінювання можливостей ШІ через людські порівняння	Імпліцитна віра в «надлюдський інтелект» або «загрозу заміщення»	Емоційно-когнітивний феномен: страх і захоплення як дві сторони тієї ж проєкції – уявлення про себе в машині
Неминучість антропоморфізації – навіть якщо вона спотворює реальність, вона допомагає зрозуміти нові явища	Імпліцитна адаптаційна функція: схеми антропоморфізації як засіб когнітивного наближення до невідомого	Антропоморфні уявлення виконують інтегративну функцію – підтримують психологічну рівновагу в умовах технологічної невизначеності

Таблиця 2

Взаємозв'язок експліцитних та імпліцитних теорій природного і штучного інтелекту

ЕКСПЛІЦИТНІ ТЕОРІЇ ПРИРОДНОГО ІНТЕЛЕКТУ	ЕКСПЛІЦИТНІ ТЕОРІЇ ШТУЧНОГО ІНТЕЛЕКТУ
Психологічні моделі, що науково пояснюють людський інтелект (Піаже, Кеттелл, Гарднер, Стернберг). Формальні тести IQ, EQ, когнітивних здібностей	Алгоритмічні, нейромережеві та когнітивні моделі машинного інтелекту (машинне навчання, AGI, оброблення природної мови, архітектури ШІ)
ІМПЛІЦИТНІ ТЕОРІЇ ПРИРОДНОГО ІНТЕЛЕКТУ	ІМПЛІЦИТНІ ТЕОРІЇ ШТУЧНОГО ІНТЕЛЕКТУ
Інтуїтивні уявлення про «розумність», «мудрість», «кмітливність»; культурні образи геніальності, таланту, «здорового глузду»	Масові уявлення, міфи, страхи та очікування щодо «розумних машин»: від ідеї цифрового помічника до образу небезпечного ШІ

ристувача для досягнення прихованих цілей, вона може призвести до серйозних наслідків, включаючи фінансові втрати та експлуатацію даних.

Виявлення того, що вразливості, створені ITAIs, не завжди вимагають складних технічних експлойтів, а скоріше експлуатують психологічні механізми, має глибокі наслідки. Це переводить фокус відповідального розроблення ШІ (Responsible AI) з чисто технічних викликів (наприклад, підвищення точності) на виклики поведінкового та комунікаційного дизайну.

Для повноти нашого аналізу проблеми імпліцитних концепцій звернемося також до метатеоретичного рівня і розглянемо цей концепт у контексті двох вимірів: тип теорії (експліцитна ↔ імпліцитна) та об'єкт пізнання (природний ↔ штучний інтелект). В узагальненому форматі координати концептуального контексту поняття «імпліцитні теорії штучного інтелекту» можна представити наочно (таблиця 2).

У таблиці продемонстровано взаємозв'язки між чотирма типами теорій, які репрезентують те, що імпліцитні уявлення про людський інтелект формують експліцитні психологічні теорії, а експліцитні теорії природного інтелекту стають основою для експліцитних теорій ШІ. Окрім того, експліцитні теорії ШІ породжують нові імпліцитні образи «розумних машин», а імпліцитні теорії ШІ змінюють людські імпліцитні уявлення про людський інтелект. Цей цикл утворює спіраль когнітивно-культурної еволюції, в якій людина та ШІ взаємно моделюють одне одного. Тобто взаємозв'язок між експліцитними (науковими, формалізованими) та імпліцитними (наївними, побутовими) теоріями інтелекту є динамічним циклом, де кожен тип теорії виконує свою унікальну функцію і впливає на інший. Ця взаємодія є ключовою для розуміння того, як академічна думка інтегрується в повсякденне життя та як повсякденні уявлення формують науковий підхід.

Висновки. Імпліцитні теорії штучного інтелекту становлять нове міждисциплінарне проблемне поле, в якому поєднуються когнітивна, соціальна, культурна та екзистенційна психологія. Вони є дзеркалом суспільної свідомості, де відбиваються страхи, очікування та надії людства щодо майбутнього розуму. ITAI є критичною когнітивною лінзою, через яку суспільство сприймає та взаємодіє зі штучним інтелектом. Розуміння цих неявних переконань є фундаментальною вимогою для за-

безпечення прозорості, надійності та етичної відповідності в розробленні та розгортанні інтелектуальних систем.

Подальші дослідження мають бути спрямовані на створення психодіагностичних інструментів для вимірювання імпліцитних уявлень про ШІ; розроблення освітніх і психопрофілактичних програм, що сприяють формуванню реалістичного й етично відповідального ставлення до технологій; інтеграцію цих досліджень у психологію цифрової компетентності та когнітивну екологію людини цифрової епохи.

ЛІТЕРАТУРА:

1. Dweck C. Implicit Theories of Intelligence. *Personality and Social Psychology Review*. 1999. Vol. 6, № 2. P. 201–215.
2. Kim T., Song H. «I Believe AI Can Learn from the Error! Or Can It Not?»: The Effects of Implicit Theories of Artificial Intelligence on Trust Repair of the Intelligent Agent. *PsyArXiv*. 2021. August 16.
3. Kim T., Song H. «I believe AI can learn from the error. Or can it not?»: The effects of implicit theories on trust repair of the intelligent agent. *International Journal of Social Robotics*. 2023. Vol. 15, № 1. P. 115–128.
4. von Walter B., Kremmel D., Jäger B. The impact of lay beliefs about AI on adoption of algorithmic advice. *Marketing Letters*. 2022. Vol. 33. P. 143–155. DOI: <https://doi.org/10.1007/s11002-021-09589-1>.
5. Eder M., Palhano Lhamby A. L. Navigating AI Narratives: Exploring Folk Theories about AI in Brazil. *Artificial Intelligence and Human Perception: Media Discourse and Public Opinion*. Milan : Franco Angeli, 2025. P. 145–160.
6. Sternberg R. J. Implicit Theories of Intelligence, Creativity, and Wisdom. *Journal of Personality and Social Psychology*. 1985. Vol. 49, № 3. P. 607–627.
7. Pohl B., Goldkind L. AI folk tales: How nontechnical publics make sense of artificial intelligence. *Research Handbook on Artificial Intelligence and Communication*. Cheltenham : Edward Elgar Publishing, 2023. P. 246–266.
8. Yenduri G. et al. Artificial General Intelligence: Advancements, Challenges, and Future Directions in AGI Research. *IEEE Access*. 2025.
9. Chaulet J. Sherry Turkle, Alone Together: Why We Expect More from Technology and Less from Each Other. New York : Basic Books, 2024. 384 p.
10. Arnd-Caddigan M. Sherry Turkle: Alone Together: Why We Expect More from Technology and Less from

Each Other. *Clinical Social Work Journal*. 2015. Vol. 43, № 3. P. 247–248.

11. Nass C., Moon Y. Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*. 2000. Vol. 56, № 1. P. 81–103.

12. Barrow N. Anthropomorphism and AI Hype. *AI and Ethics*. 2024. Vol. 4. P. 707–711. DOI: <https://doi.org/10.1007/s43681-024-00454-1>.

13. Фоменко К.І. Імплицитні теорії інтелекту й особистості у зв'язку з губристичною мотивацією в студентів. *Науковий вісник Херсонського державного університету. Серія: Психологічні науки*. 2016. Вип. 5, № 2. С. 71–75.

14. Costa A., Faria L. Implicit theories of intelligence and academic achievement: A meta-analytic review. *Frontiers in Psychology*. 2018. Vol. 9. P. 829.

15. Blackwell L.S., Trzesniewski K.H., Dweck C.S. Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Development*. 2007. Vol. 78, № 1. P. 246–263.

16. Sternberg R. J. Implicit theories of intelligence, creativity, and wisdom. *Journal of Personality and Social Psychology*. 1985. Vol. 49, № 3. P. 607–627.

17. Hong Y.-Y., Chiu C.-Y., Dweck C. S. Implicit theories of intelligence: Reconsidering the role of confidence in achievement motivation. *Efficacy, Agency, and Self-Esteem*. Boston : Springer US, 1995. P. 197–216.

18. Eder M., Palhano Lhamby A. L. Navigating AI narratives: Exploring folk theories about AI in Brazil. *Artificial Intelligence and Human Perception: Media Discourse and Public Opinion* / eds. E. Lupano, P. Orrù. Milan : Franco Angeli, 2025. P. 145–160.

19. Гальцева Т.О., Величко Р.В. Імплицитні теорії інтелекту й особистості як проблемне поле психологічних досліджень. *Психологія та соціальна робота*. 2024. № 1. С. 63–70.

20. Салюк М., Тарасюк Х. Роль імплицитних теорій у забезпеченні психічної саморегуляції діяльності та адаптаційного потенціалу особистості. *Молодий вчений*. 2020. № 12(88). С. 288–292.

21. Аршава І.Ф., Салюк М.А. Особливості взаємозв'язку імплицитних теорій інтелекту та особистості з копінг-стратегіями. *Дніпровський науковий часопис публічного управління, психології, права*. 2022. № 1. С. 34–41.

22. Горбунова В.В. Проблеми дослідження імплицитних теорій. *Соціальна психологія*. 2008. № 4(30). С. 17–28.

23. Богачов М. Штучний інтелект як антропотехнологія. *Філософська думка*. 2021. № 3. С. 180–200.

24. Величко Р.В., Салюк М.А. Імплицитна психологія як науковий напрям психологічних досліджень. *Габітус*. 2023. Вип. 46. С. 49–54. URL: <http://habitus.od.ua/journals/2023/48-2023/8.pdf> (дата звернення: 27.10.2025).

25. Кириченко В.В. Імплицитна теорія психічного здоров'я у студентської молоді. *Науковий вісник Ужгородського національного університету. Серія: Психологія*. 2023. Вип. 4. С. 14–18.

26. Ольмезов В. Імплицитні концепції права у правосвідомості громадян та їх врахування в державному управлінні. *Державне управління та місцеве самоврядування*. 2014. Вип. 3. С. 52–58.

27. Ющишина Л.О. Імплицитні теорії командної інтеракції через призму воєнного стану. *Соціально-економічний стан в умовах воєнного часу*. Research Europe, 2024.

28. Дерід К., Додонова В. Деніел Деннет – філософ у цифрову епоху: від еволюції свідомості до штучного інтелекту. *Грані*. 2024. Т. 27, № 3. С. 7–13.

Стаття надійшла у редакцію: 14.10.2025

Стаття прийнята: 29.10.2025

Опубліковано: 08.12.2025