УДК 165.41+177(5)

КОЗАЧЕНКО Надія *– кандидат філософських наук, доцент кафедри філософії, Криворізький державний педагогічний університет, 54, пр. Університетський, м. Кривий Ріг, Україна, індекс 50086 (N.P.Kozachenko@gmail.com)*

ORCID: https://orcid.org/0000-0003-2358-9076

Scopus Author ID: 57214946191

DOI: https://doi.org/10.24919/2522-4700.49.6

Бібліографічний опис статті: Козаченко, Н. (2024). Проблема епістемічної надійності знання в контексті штучного інтелекту. *Людинознавчі студії: збірник наукових праць Дрогобицького державного педагогічного університету імені Івана Франка. Серія «Філософія»*, № 49, 94–109, doi: https://doi.org/10.24919/2522-4700.49.6

ПРОБЛЕМА ЕПІСТЕМІЧНОЇ НАДІЙНОСТІ ЗНАННЯ В КОНТЕКСТІ ШТУЧНОГО ІНТЕЛЕКТУ

Анотація. *У статті здійснюється аналіз проблеми епістемічної надійності знання, отриманого за допомогою штучного інтелекту, з урахуванням аспектів прозорості, обґрунтованості, а також впливу соціокультурних факторів на процес навчання ШІ. Основна* ***мета*** *роботи – виявлення чинників, що впливають на надійність знання ШІ, і розгляд шляхів покращення прозорості його моделей для підвищення довіри користувачів.* ***Методологія*** *дослідження ґрунтується на аналітичному огляді сучасних підходів у сфері епістемології штучного інтелекту та філософії науки. Використано методи порівняльного аналізу для виявлення подібностей і відмінностей між людським та машинним пізнанням, а також розглянуто концепції, що розкривають природу обґрунтованості знання в умовах недостатньої прозорості алгоритмів глибокого навчання. Задіяно метод концептуального аналізу до проблеми створення «пояснювального ШІ» та «вирівнювання ШІ» в контексті порівняння змістовного наповнення термінів людського пізнання та моральності з відповідними термінами, які ми застосовуємо до ШІ. Застосовано метод уявного експерименту для*

моделювання ситуацій упередженості та непрозорості рішень ШІ. ***Наукова новизна*** *статті полягає у систематичному підході до дослідження проблеми епістемічної надійності ШІ у контексті його здатності пояснювати свої рішення. Зокрема, у статті розкрито, як соціальні, культурні та етичні аспекти впливають на процес навчання ШІ, що в свою чергу призводить до упереджень у прийнятті рішень. Вперше досліджено питання теоретичної навантаженості фактів, коли дані для навчання ШІ розглядаються в межах концептуальної рамки, що впливає на їх об'єктивність. Вперше застосовано проблему трилеми Мюнхгаузена для характеристики пояснювальних можливостей ШІ.* ***Висновки****. У статті обґрунтовано необхідність покращення прозорості алгоритмів ШІ з метою підвищення їх надійності. Основними висновками є важливість інтеграції етичних принципів у процес розробки ШІ для зниження рівня упередженості, а також розробка моделей пояснювального ШІ, здатного прозоро обґрунтовувати свої рішення. Це дозволить підвищити довіру користувачів та забезпечити ефективніше використання ШІ у критично важливих сферах. Перспективи подальших досліджень включають розробку інструментів для врахування епістемологічних характеристик надійності знання, здобутого ШІ.*

Ключові слова: *філософія штучного інтелекту, етика штучного інтелекту, епістемологія штучного інтелекту, методологія наукового пізнання, надійність знання, упередженість, вирівнювання ШІ, пояснювальний ШІ.*

KOZACHENKO Nadiia *– Candidate of Philosophy Sciences, Associate Professor at the Philosophy Department, Kryvyi Rih State Pedagogical University, 54, Universytetsky ave., Kryvyi Rih, Ukraine, postal code 50086 (N.P.Kozachenko@gmail.com)*

ORCID: https://orcid.org/0000-0003-2358-9076

Scopus Author ID: 57214946191

DOI: https://doi.org/10.24919/2522-4700.49.6

To cite this article: Kozachenko, N. (2024). Problema epistemichnoi nadiinosti znannia v konteksti shtuchnogo intelektu [The challenge of epistemic reliability of knowledge in the context of artificial intelligence]. *Liudynoznavchi studii: zbirnyk naukovykh prats Drohobytskoho derzhavnoho pedahohichnoho universytetu imeni Ivana Franka. Seriia "Filosofiia" –*

Human Studies. Series of "Philosophy": a collection of scientific articles of the Drohobych Ivan Franko State Pedagogical University, № 49, 94–109, doi: https://doi.org/10.24919/2522-4700.49.6

THE CHALLENGE OF EPISTEMIC RELIABILITY OF KNOWLEDGE IN THE CONTEXT OF ARTIFICIAL INTELLIGENCE

Summary. *The article analyzes the problem of epistemic reliability of knowledge obtained through artificial intelligence, considering aspects of transparency, justification, and the impact of sociocultural factors on AI training processes.* ***Aim***. *The main aim of the study is to identify the factors affecting the reliability of AI-generated knowledge and to explore ways of improving the transparency of AI models to enhance user trust.* ***Methodology.*** *The research methodology is based on an analytical review of current approaches in the field of epistemology of artificial intelligence and philosophy of science. Methods of comparative analysis are employed to identify similarities and differences between human and machine cognition, as well as to examine concepts that reveal the nature of the justification of knowledge under conditions of insufficient transparency of deep learning algorithms. The conceptual analysis method is applied to the problem of creating «explainable AI» and «AI alignment» in the context of comparing the content of terms related to human cognition and morality with corresponding terms applied to AI. The method of thought experiment is used to model situations involving bias and the opacity of AI decisions.* ***Scientific Novelty.*** *The scientific novelty of the article lies in a systematic approach to studying the problem of epistemic reliability of AI in the context of its ability to explain its decisions. In particular, the article reveals how social, cultural, and ethical aspects influence the AI training process, which in turn leads to biases in decision-making. For the first time, the issue of theory-laden facts is investigated, where training data for AI is considered within a conceptual framework that affects their objectivity. The Münchhausen trilemma problem is also applied for the first time to characterize the explanatory capabilities of AI.* ***Conclusions.*** *The article substantiates the need to improve the transparency of AI algorithms to enhance their reliability. The main conclusions emphasize the importance of integrating*

ethical principles into the development process of AI to reduce bias and the development of explainable AI models capable of transparently justifying their decisions. This will increase user trust and ensure more effective use of AI in critical domains. Prospects for future research include the development of tools to account for the epistemological characteristics of the reliability of AI-generated knowledge.

Key words: *philosophy of artificial intelligence, ethics of artificial intelligence, epistemology of artificial intelligence, methodology of science, reliability of knowledge, bias, AI alignment, explainable AI.*

В рамках філософських досліджень штучного інтелекту дослідники звертають особливу увагу на питання щодо прозорості прийняття рішень штучним інтелектом, справедливості його суджень та можливості користувачів довіряти результатам, отриманих за допомогою ШІ. Зазвичай ця тематика поширена серед дослідників етики штучного інтелекту і це не випадково, оскільки подібні питання стосуються правомірності застосування ШІ для прийняття досить важливих рішень. Ми вважаємо, що ці питання також безпосередньо стосуються епістемологічної проблематики ШІ, а не лише моральних його аспектів. По суті, це проблема достовірності або, в контексті епістемології ШІ, – проблема надійності знання.

Актуальність проблеми зумовлена тим, що останніми роками значно підвищилася точність моделей, заснованих на глибокому навчанні, але це зростання було досягнуте за рахунок збільшення складності моделі, що в свою чергу привело до виникнення низки помилок (в тому числі й «галюцинацій ШІ»), коли інформація, рішення або висновок, який робить ШІ, суперечить здоровому глузду, практиці або взагалі наявній реальності. При цьому ШІ припускається помилок без пояснення їхніх причин, що робить неможливим повністю довіряти наданій ним інформації (Trotta, 2023, p. 727). Таким чином, виникає необхідність перегляду підходу до розробки моделей ШІ, який би включав прозорість, зрозумілість і підтвердженість як невід'ємні ознаки надійності інформації.

Проблема епістемічної надійності знання в контексті штучного інтелекту (ШІ) є особливо актуальною у сучасній філософії

ШІ, зокрема через збільшення використання ШІ для прийняття важливих рішень. Таким чином, *постановка проблеми* нашого дослідження полягає в наступному: підвищення точності моделей, заснованих на глибокому навчанні, супроводжується зростанням складності цих моделей, що призводить до проблеми прозорості та обґрунтованості їх висновків, це ставить під сумнів можливість користувачів повністю довіряти інформації, наданій ШІ, оскільки відсутність пояснень щодо процесу прийняття рішень знижує рівень довіри до них. Але дослідження епістемологічних проблем ШІ можуть надати натхнення щодо розвідок людського процесу пізнання, який наразі також не є до кінця прозорим. *Метою цього дослідження* є проблематизація епістемологічної надійності знання, отриманого за допомогою штучного інтелекту, в контексті класичних епістемологічних проблем. Таким чином, основними завданнями нашого дослідження є виявлення факторів, що впливають на епістемічну надійність знання ШІ, аналіз впливу упередженості даних на результати, отримані за допомогою ШІ та розгляд шляхів покращення прозорості моделей ШІ для підвищення рівня довіри користувачів.

Аналіз останніх досліджень та публікацій показує, що основні напрямки, в яких рухаються філософи штучного інтелекту є переважно прикладними і тісно пов'язані з питаннями етики ШІ. Так, А. Касірзаде і І. Габрієль зазначають, що великі мовні моделі викликають низку ризиків, включаючи виробництво неправдивої, образливої або нерелевантної інформації, яка може призвести до цілого ряду шкоди. Таким чином, ключове соціальне та етичне питання стосується узгодження спілкування з ШІ з відповідними нормами та цінностями (Kasirzadeh, 2023, p. 27). Йюхень Цанг наголошує, шо ШІ отримує цінності й настанови тільки від людини, тому його розробники несуть повну відповідальність за етичні аспекти роботи ШІ: «AI is still simply a mimic of BI (Biological Intelligence)» (Youheng, 2023, p. 30). Л. Беленгер розробляє цілісну концептуальну схему, яка має поєднати автоматичні міркування з людиноцентричним підходом і в рамках моделювання ШІ намагається дати відповідь на своє ж питання: «Як абстрактні ідеї, такі як справедливість чи соціальна справедливість, можна перевести у відповідні етичні рамки?» (Belenguer, 2022).

Основний матеріал. Ми цілком погоджуємося з тим, що етична складова проблеми надійності знання та довіри до ШІ дуже важлива, але ці питання також належать до царини епістемології. Їхній аналіз в контексті ШІ може, зокрема, пролити світло на деякі невирішені питання і людського пізнання. Проблема надійності знання – це класична епістемологічна проблема і її постановка в контексті ШІ в деякому сенсі наближає його до людського пізнання: по суті – це ті самі проблеми, доповнені специфікою ШІ. Для людини важливо визначити ознаки надійності інформації, але для штучного інтелекту ця проблема постає ще складнішою, оскільки він не є самостійним епістемічним агентом і не здійснює рефлексію над власним знанням. У контексті ШІ це означає, що, навіть якщо алгоритм здається обґрунтованим, залишається питання, на основі чого він приймає рішення і наскільки виправдані його висновки; це питання не здатний поставити сам ШІ, але таке питання може ставити людина, ґрунтуючись на реальності, практиці, здоровому глузді та інтуїції.

Одразу розглянемо приклад непрозорості висновків ШІ, які дають підставу ставити під сумнів його судження. Припустімо, що алгоритм глибокого навчання використовується для діагностики хвороби за знімками рентгену. Модель може показувати високий рівень точності у статистиці, але залишається питання: на основі яких саме параметрів вона приймає рішення і наскільки виправдані її висновки? Адже саме це питання ставить собі лікар для отримання впевненості у діагнозі.

Наприклад, під час аналізу рентгенівських знімків ШІ може відмітити певні патерни, які він вважає ознакою відповідної хвороби. Лікарі так само приймають рішення на основі тих самих ознак, але в цьому випадку виникає питання довіри, яке вимагає підтвердження: це вже питання дворівневої довіри – довіри собі й довіри вторинному або допоміжному джерелу. Лікарі мають знати, чому саме ШІ прийшов до такого висновку – це може бути через те, що він виявив ущільнення в певному місці, або ж певний рівень інтенсивності кольору на знімку, або через дефект знімку, або просто тому, що йому задали шукати ознаки цієї хвороби. Останній варіант найстрашніший: якщо ми задаємо моделі ШІ шукати відповідні ознаки, вона їх знайде. Те

саме стосується фахівців: якщо кардіолог шукає ознаки серцевої недостатності – він їх знайде. Однак, сам алгоритм ШІ не може поставити під сумнів свої висновки. Це можуть зробити лише фахівці, в даному випадку лікарі, базуючись на своєму досвіді, інтуїції чи здоровому глузді. Так, на знімку може бути морфологічна особливість тіла, похибка знімку чи інше захворювання в агресивній стадії тощо, але ШІ досі не може працювати з виключеннями і з розширеним контекстом. Лише лікарі можуть запідозрити, що ШІ зробив помилку, наприклад, через те, що в навчальних даних ШІ могли бути присутні упередження або через можливе неправильне трактування зображень, або це була унікальна особливість організму. Зважаючи на те, що лікарські помилки так само мають місце і що ШІ дійсно економить час і уважність, обробляючи великі набори даних, тим не менш, маємо констатувати, що узагальнення та інтерпретація досі належать людям. Особливо, якщо це стосується людського життя і важливих рішень щодо нього.

Незважаючи на величезний спектр практичних застосувань, проблема епістемічної надійності ШІ пов'язана з низкою суто теоретичних проблем епістемології. Перш за все йдеться про **проблему обґрунтування**: якщо вважати надійним знанням с*а*ме обґрунтоване знання, виникає необхідність врахувати традиційні епістемологічні питання щодо обґрунтування. Згідно трилеми Мюнхгаузена, сформульованої Г. Альбертом: будь-яке обґрунтування або впадає в нескінченний регрес – а тому нездійсненне, або впадає в логічне коло – а тому невиправдане, або передбачає довільну зупинку в обґрунтуванні – а тому догматичне (Albert, 1985, p. 13).

Прикладом застосування штучного інтелекту, де виникає питання щодо обґрунтованості його рішень, є кредитна оцінка, коли моделі ШІ, що базуються на алгоритмах глибокого навчання, аналізують параметри: доходи, кредитну історію, використання кредитних лімітів, соціальні фактори тощо, і на основі цього визначають, чи має заявник право на отримання кредиту. У випадку роботи ШІ неможливо з'ясувати, які конкретні фактори найбільше вплинули на оцінку, через складність алгоритмів глибокого навчання. Наприклад, клієнт може отримати відмову в кредиті, але не отримати чіткого пояснення,

які конкретні причини були ключовими для цього рішення. Це породжує проблеми з прозорістю та етикою: людина, яка подає заявку на кредит, може не знати, чому її відмовили, і це може здаватися їй необґрунтованим чи навіть дискримінаційним. У фінансових установах така непрозорість може викликати сумніви у менеджерів, які хочуть бути впевнені у справедливості алгоритмів, що приймають рішення. Відтак, сама відсутність можливості пояснення від того, хто вирішував, ставить під питання надійність знання, здобутого за допомогою ШІ, який надав рішення. Висновок ШІ може виглядати як догматичний, оскільки його підгрунтя не зрозумілі. Або ж виглядати як коло в обґрунтуванні: вам відмовлено в кредиті, оскільки у вас недостатньо позитивна кредитна історія, підставою для якої є те, що вам до цього часу не давали кредити; або ж впадати в нескінченний регрес і не давати відповіді.

Наступна проблема – це проблема **теоретичної навантаженості фактів**: якщо фактичність вважати підґрунтям обґрунтованості, то потрібно враховувати, що будь-який факт існує не сам по собі, а в системі теоретичних уявлень тієї концепції, в рамках якої він виявлений. Факти, що використовуються для навчання алгоритмів, завжди вбудовані в певну теоретичну або концептуальну рамку (Franklin, 2015, p. 157–159). Це означає, що інформація, яку вважають об'єктивною, завжди піддається впливу теорії або підходу, в межах якого вона отримана. В контексті ШІ це, зокрема, проявляється через можливу упередженість даних, на яких алгоритм навчався, що впливає на його рішення і результати (Baker, 2022).

Упередженість фактів проявляється не лише в науково-теоретичному сенсі, але й на буденному рівні, коли автори даних для ШІ несвідомо демонструють расові, гендерні чи вікові стереотипи. Упередженість у даних є значною проблемою для епістемології ШІ, адже ця упередженість є саме тим, що люди передають ШІ. До того ж упередженість викликає додаткові запитання: Що означає неупередженість мовних моделей? Яка концепція упередженості, серед множини варіантів, повинна слугувати центром для коригувальних дій? (Kasirzadeh, 2023, c. 29). Очевидно, що упереджені дані можуть призвести до упереджених рішень, що особливо небезпечно у сферах, які впливають на

життя людей (наприклад, прийняття рішень щодо надання кредиту, оцінки кандидатів на роботу тощо) (Trotta, 2023; Belenguer, 2022). Упередженість може стосуватися не тільки стереотипів, але й зумовлюватися ціннісними перевагами авторів навчальних даних: етичними, естетичними чи загальносоціальними й культурними особливостями, які стосуються традиції, прийнятності, поняття гарного, допустимого чи бажаного (Belenguer, 2022).

Проблема фактичності також пов'язана з питанням об'єктивності і рівня опрацьовуваних ШІ даних. ШІ працює з великим обсягом даних, які можна розглядати як об'єктивні в силу їхнього обсягу, але питання інтеграції суб'єктивного досвіду людини в цей процес є досить складним. ШІ має справу лише з числовими або категорійними даними, тоді як знання людини включає суб'єктивні аспекти: інтуїція, переживання, емоції, ситуативний контекст та особистісний стан. Це обмежує можливість створення «усвідомленого» знання, яке б відповідало людському розумінню і було б доступним для людини (Trotta, 2023).

Крім того, навіть після налаштування алгоритмів з урахуванням расових та гендерних відмінностей, вони все ще демонструють упередженість. Наприклад, алгоритм, який використовувався для визначення кандидатів на роботу в галузях, де переважають чоловіки, мав тенденцію віддавати перевагу кандидатам, які більше відповідали стереотипам про маскулінність. Наприклад, алгоритм асоціював слово «empowerment» з жінками, і жінки-хірурги, які використовували це слово у своїх біографіях, мали менше шансів бути правильно ідентифікованими як хірурги (Johnson, 2021). Коли алгоритми навчалися на біографічних текстах, вони використовували контекстуальні патерни, але в них слово «empowerment» частіше зустрічалося в текстах, де йшлося про жінок і в описах, пов'язаних із традиційно жіночими професіями та соціальними ініціативами. Це призвело до того, що на основі тренувальної вибірки алгоритм почав автоматично асоціювати цей термін саме з жінками, зменшуючи ймовірність ідентифікації жінок, які використовували це слово, як представників «чоловічих» професій, наприклад хірургів. Така упередженість вказує на те, що алгоритми схильні підтримувати соціальні стереотипи, а не об'єктивно оцінювати

компетенцію кандидатів. У цьому випадку алгоритми, розроблені для боротьби з дискримінацією, насправді продовжують існуючі стереотипи, підкріплені способом представлення даних (Köchling, 2023).

Крім того, проблема упередженості в алгоритмах часто є наслідком використання похідних (неявних) властивостей, які явно не корелюють із соціальними характеристиками. Алгоритм міг віддавати перевагу кандидатам, які закінчили програми або навчальні заклади, де основний акцент робився на підтримці жінок у технічних спеціальностях, наприклад, спеціальні STEM-програми для жінок. Таким чином, навіть якщо гендер явно не враховувався, назва навчального закладу або участь у програмі могла виступати проксі-атрибутом для визначення гендеру. Це ще більше ускладнює розуміння, адже навіть при видаленні явних ознак дискримінації, алгоритм може використовувати непрямі ознаки, що продовжують впливати на прийняті рішення.

Ще одна проблема – **проблема контекстуальності знання** – ніколи повністю невідомо, наскільки широкий чи наскільки вузький контекст виявиться релевантним для обґрунтування знання. У випадку ШІ контекстуальність знання часто стає проблемою через обмеження моделі розуміти та інтерпретувати широкий спектр реальних ситуацій. Уявімо, що ви користуєтеся розумною системою управління розкладом, яка допомагає вам планувати зустрічі, нагадує про важливі події і координує ваші щоденні активності. Ви просите цю систему запланувати зустріч із колегою на наступний тиждень. Система переглядає ваш календар, визначає вільний час і автоматично планує зустріч, обираючи найближчий вільний слот. Однак, проблема контекстуальності полягає в тому, що система не розуміє деталей, які важливі для цієї зустрічі. Наприклад, вона не враховує, що ви надаєте перевагу проведенню важливих зустрічей на початку тижня, щоб залишити достатньо часу для виконання наступних завдань. І тут постає питання: якою мірою ми маємо уточнювати контекст релевантності для прийняття рішення? Чи не втрапимо ми тут в нескінченність? І як вирішити, які умови будуть релевантними?

Дійсно, якщо ми продовжимо уточнювати контекст, намагаючись охопити всі можливі варіанти, ми можемо ніколи

не досягти остаточного набору релевантних умов. Реальний світ є надзвичайно складним і контексти можуть змінюватися нескінченно, але, з одного боку, ми хочемо, щоб система була досить адаптивною та гнучкою, а з іншого боку, спроба охопити всі можливі варіанти контексту може зробити алгоритм занадто складним для практичного застосування. Щоб уникнути нескінченного регресу, в епістемології та в розробці ШІ часто застосовується принцип достатньої релевантності. Він передбачає, що ми маємо обмежити контекст до тих параметрів, які є достатньо релевантними для конкретного рішення або дії. Це означає, що замість спроби охопити весь можливий контекст, ми виділяємо лише ті аспекти, які мають найбільший вплив на кінцеве рішення. Наприклад, якщо алгоритм планує зустріч, то достатньо знати основні уподобання часу для всіх учасників, не обов'язково враховувати їхній настрій або плани на відпочинок. Очевидно, що обмежити контекст досить проблематично.

Крім того, алгоритм ШІ може добре працювати в лабораторних умовах або при конкретному сценарії, але давати похибки при перенесенні в інший контекст, в той час як людина має здатність «орієнтуватися у ситуації». Це знову ж піднімає питання про обмеженість та релевантність знання, яке генерує ШІ. Варто зазначити, що алгоритми ШІ обробляють дані, не маючи по суті ніякого «розуміння» контексту в людському сенсі (як наприклад описує Сьорл в «китайській кімнаті»). Так, глибокі нейронні мережі можуть виконувати класифікацію з високою точністю, але залишаються чорною скринькою, і ми не завжди розуміємо, як саме алгоритм прийшов до певного висновку. Це ставить під сумнів обґрунтованість і правильність знання, яке виробляє ШІ.

У пізнанні штучного інтелекту також актуально постає **проблема джерела знання**. І в нашому контексті йдеться навіть не про загальні філософські підходи, а, скоріш, про їхні наслідки, зумовлені тим, що ШІ отримує вторинні дані, надані людиною безпосередньо або опосередковано. ШІ здобуває знання на основі навчальних даних, але в багатьох випадках ці дані є неповними або мають упередження. Таким чином, знання, отримане ШІ, не можна вважати об'єктивним або надійним, якщо воно було здобуте на основі упереджених даних. Але виникає питання, як

перевірити упередженість даних? Чи здатні користувачі або розробники ШІ перевірити свої дані на упередженість? Це ставить під сумнів надійність знання, яке ШІ генерує. У той час як люди можуть оцінити джерело інформації та його надійність, інтуїтивно обмеживши контекст необхідними даними, алгоритми ШІ зазвичай не мають обмежувати джерела, окрім наданого людиною контексту. В цьому сенсі ШІ незначні помічники. Більш того, люди, як носії когнітивних упереджень, можуть цілком передати їх штучному інтелекту. Алгоритми можуть неусвідомлено використовувати когнітивні упередження, які були закладені у вихідних даних або в запрограмованих алгоритмах. Це може призвести до систематичних помилок, що не завжди легко ідентифікувати і виправити.

Якщо переформулювати аристотелівський вислів «Людина від природи прагне пізнання» в контексті штучного інтелекту, то ми б запропонували таку його модифікацію: «Штучний інтелект здійснює пізнання, якого прагне людина». В цьому виразі досить явно прослідковується наша дослідницька настанова на розуміння ШІ як інструменту, який відрізняється від людини відсутністю цінностей, мотивації, прагнень. Відповідь на питання – «чому ШІ робить, те що він робить?» – дуже проста – тому що йому надає запит людина, прямо чи опосередковано. Дійсно, ШІ не має «внутрішнього прагнення» до пізнання, а натомість реалізує завдання, які йому надає людина. Це є фундаментальною відмінністю між людським та штучним інтелектом, що підкреслює їхні різні епістемічні властивості. Відтак, маємо визнати, що досвід міркування ШІ отримує від людини. Так, ШІ може знайти невідому людині закономірність в даних, але саме поняття закономірності, завдання її знаходження і дані він отримав від людини. Разом з цим ШІ отримує цілу низку недоліків людського пізнання – і перш за все, невпевненість. Причому ця невпевненість стосується не ШІ: він будь-яку інформацію сприймає як факт, допоки не отримає від користувача завдання її перевірити, – ця невпевненість стосується саме користувача, який не є переконаним у відповідях штучного інтелекту. Невпевненість породжується кількома параметрами: непрозорістю міркувань ШІ, недовірою до використовуваних ним даних, незрозумілістю алгоритмів його

роботи. Сюди можна додати неочевидні параметри: неясність мотивів розробників ШІ, незрозумілість цінностей і принципів, якими керується ШІ, підозра щодо комерціалізації моделей ШІ.

У класичній епістемології проблема обґрунтування є основною для визначення надійності знання. Якщо для людського пізнання надійність знання пов'язана з обґрунтуванням через різні типи доказів, то в ШІ ця проблема постає ще складнішою через необхідність об'єктивізації таких процесів, які є не лише формально правильними, а й доступними для пояснення користувачам. Одним із важливих кроків у цьому напрямку є розробка моделі пояснювального ШІ, який має бути прозорим і зрозумілим для людей, щоб забезпечити довіру між людиною і машиною, внаслідок якої людина могла б покладатися на отриману від ШІ інформацію. Прозорість є необхідною для створення довіри між людиною та ШІ, особливо коли йдеться про критичні сфери, такі як освіта, медицина та правосуддя. Дослідники розглядають концепцію «пояснювального ШІ», що забезпечує можливість зрозуміти, як і чому ШІ приймає певні рішення (Trotta, 2024). Ці дослідження зосереджені на розробці методів, які роблять процес прийняття рішень алгоритмами ШІ зрозумілим для користувачів. Так, А. Тротта та його колеги, досліджують пояснювальний ШІ у контексті соціальних інтересів та впливу на громадськість. Вони працюють над розробкою «простору пояснень», де користувачі можуть взаємодіяти з ШІ і надавати зворотний зв'язок щодо рішень, які приймає алгоритм. Їхні дослідження спрямовані на те, щоб зробити ШІ більш «людиноцентричним», де пояснення є не лише технічно коректними, але й зрозумілими для людей без спеціальних знань. В. Ломонако зосереджує свої дослідження на розробці пояснювальних моделей, які враховують як технічну коректність, так і потреби користувачів. Його роботи стосуються пояснювального ШІ у контексті навчання з підкріпленням і того, як пояснення можуть бути інтегровані у процес прийняття рішень ШІ для забезпечення більшої прозорості та довіри (Trotta, 2024). Проблема «вирівнювання ШІ» (AI alignment) є однією з ключових сучасних філософських проблем штучного інтелекту.

Йдеться про те, як зробити так, щоб ШІ діяв згідно з людськими цінностями, включаючи не тільки утилітарні аспекти, такі як максимізація загального блага, але й дотримання прав та етичних обмежень. Виникає питання, як навчити ШІ розпізнавати та поважати етичні принципи, включаючи ідеї прав, які діють як «обмеження на дії» і мають пріоритет над загальною вигодою (McDonald, 2023).

*Висновки***.** Проблема епістемічної надійності знання, отриманого за допомогою ШІ, є складною і багатогранною, що вимагає врахування різних аспектів – від прозорості моделей ШІ до соціальних та етичних наслідків. Очевидно, що для забезпечення надійності знання необхідно покращити прозорість алгоритмів та підвищити рівень довіри користувачів, що можна досягти через пояснювальний ШІ, здатний зрозуміло обґрунтовувати свої рішення. Подальші дослідження можуть зосереджуватися на розробці підходів до підвищення прозорості та інтеграції етичних принципів у розробку моделей штучного інтелекту, що дозволить ефективно знизити рівень упередженості та підвищити довіру користувачів до результатів, які надає ШІ. Так само важливим залишається питання людського «самопізнання» крізь пізнання епістемічних процесів ШІ, оскільки вони зумовлені діяльністю людини як основного епістемічного агента.

ЛІТЕРАТУРА

1. Albert H., Rorty M.V. The Problem of Foundation. *Treatise on Critical Reason*. Princeton University Press, 1985.

2. Baker R.S., Hawn A. Algorithmic Bias in Education. *International Journal of Artificial Intelligence in Education*. Vol. 32. 2022. P. 1052–1092.

3. Belenguer L. AI Bias: Exploring Discriminatory Algorithmic Decision-Making Models and the Application of Possible Machine-Centric Solutions Adapted from the Pharmaceutical Industry. *AI and Ethics*. Vol. 2. № 4. 2022. P. 771–787.

4. Donald A., et al. Bias Detection for Customer Interaction Data: A Survey on Datasets, Methods, and Tools. *IEEE Access*. Vol. 11. 2023. P. 53703–53715, 2023.

5. Franklin A. The Theory-Ladenness of Experiment. *Journal for General Philosophy of Science*. Vol. 46. 2015. P. 155–166.

6. Johnson G.M. Algorithmic Bias: On the Implicit Biases of Social Technology. *Synthese*. Vol. 198. 2021. P. 9941–9961.

7. Kasirzadeh A., Gabriel I. In Conversation with Artificial Intelligence: Aligning Language Models with Human Values. *Philosophy & Technology*. Vol. 36. 2023.

8. Köchling, A., Wehner, M.C. Discriminated by an Algorithm. *Business Research*. Vol. 13. 2023. P. 795–848.

9. McDonald F.J. AI, Alignment, and the Categorical Imperative. *AI and Ethics*. Vol. 3. 2023. P. 337–344.

10. Trotta A., Ziosi M., Lomonaco V. The Future of Ethics in AI: Challenges and Opportunities. *AI & Society*. Vol. 38. 2023. P. 439–441.

11. Yang W., Wei Y., Wei H., et al. Survey on Explainable AI: From Approaches, Limitations, and Applications Aspects. *Human-Centric Intelligent Systems*. Vol. 3. 2023. P. 161–188.

12. Zhang Y., Tiňo P., Leonardis A., Tang K. A Survey on Neural Network Interpretability. *Ithaca: Cornell University Library. arXiv.org*. Vol. 5. №. 5. 2021. P. 726–742.

REFERENCES

1. Albert, H., Rorty, M.V. (1985) The Problem of Foundation. *Treatise on Critical Reason*. Princeton University Press.

2. Baker, R.S., Hawn A. (2022) Algorithmic Bias in Education. *International Journal of Artificial Intelligence in Education*. Vol. 32. pp. 1052–1092. https://doi.org/10.1007/s40593-021-00285-9.

3. Belenguer, L. (2022) AI Bias: Exploring Discriminatory Algorithmic Decision-Making Models and the Application of Possible Machine-Centric Solutions Adapted from the Pharmaceutical Industry. *AI and Ethics*. Vol. 2, no. 4. pp. 771–787. doi: 10.1007/s43681-022-00138-8.

4. Donald, A., et al. (2023) Bias Detection for Customer Interaction Data: A Survey on Datasets, Methods, and Tools. *IEEE Access*. Vol. 11. pp. 53703–53715. doi: 10.1109/ACCESS.2023.3276757.

5. Franklin, A. (2015) The Theory-Ladenness of Experiment. *Journal for General Philosophy of Science*. Vol. 46. pp. 155–166. https://doi.org/10.1007/s10838-015-9285-9.

6. Johnson, G.M. (2021) Algorithmic Bias: On the Implicit Biases of Social Technology. *Synthese*. Vol. 198. pp. 9941–9961. https://doi.org/10.1007/s11229-020-02696-y.

7. Kasirzadeh, A., Gabriel, I. (2023) In Conversation with Artificial Intelligence: Aligning Language Models with Human Values. *Philosophy & Technology*. Vol. 36. p. 27. https://doi.org/10.1007/s13347-023-00606-x.

8. Köchling, A., Wehner, M.C. (2023) Discriminated by an Algorithm. *Business Research*. Vol. 13. pp. 795–848. https://doi.org/10.1007/s40685-020-00134-w/’

9. McDonald, F.J. (2023) AI, Alignment, and the Categorical Imperative. *AI and Ethics*. Vol. 3. pp. 337–344. https://doi.org/10.1007/s43681-022-00160-w.

10. Trotta, A., Ziosi, M., Lomonaco, V. (2023) The Future of Ethics in AI: Challenges and Opportunities. *AI & Society*. Vol. 38. pp. 439–441. https://doi.org/10.1007/s00146-023-01644-x.

11. Yang, W., Wei, Y., Wei, H., et al. (2023) Survey on Explainable AI: From Approaches, Limitations, and Applications Aspects. *Human-Centric Intelligent Systems*. Vol. 3. pp. 161–188. https://doi.org/10.1007/s44230-023-00038-y.

12. Zhang, Y., Tiňo, P., Leonardis, A., Tang, K. (2021) A Survey on Neural Network Interpretability. *Ithaca: Cornell University Library. arXiv.org*. Vol. 5, no. 5. pp. 726–742.