УДК 165.1:004.8

ЕПІСТЕМОЛОГІЧНІ МЕЖІ ДОВІРИ ТА ОБҐРУНТУВАННЯ ЗНАННЯ В НЕІНТЕРПРЕТОВАНИХ СИСТЕМАХ ЕКСТРАЦІЇ ДАНИХ

Шатило І.Ю.
e-mail: ihor.shatylo@nure.ua
Науковий керівник – к. філос. н., доц. Старікова Г.Г.
Харківський національний університет радіоелектроніки, каф. СГН
м. Харків, Україна

This paper examines non-interpretable «black box» data extraction systems from an epistemological perspective, arguing that in a digital society such systems, though fundamental to knowledge production, challenge the classical definition of knowledge as «justified true belief» (JTB). While a «black box» may provide outputs that are true (accurate) and generate belief (the user trusts the data), it fails to supply the third crucial component: justification (logos). Without access to the system's reasoning, the user lacks a rational basis for belief and must rely on faith in the algorithm. We conclude that interpretability is not merely a technical feature but an epistemological necessity if AI outputs are to be legitimately considered knowledge.

Сучасне цифрове суспільство все більше покладається на автоматизовані системи для формування своєї бази знань. Системи екстракції даних (веб-скрейпінгу, аналізу документів) відіграють у цьому ключову роль, автоматично збираючи «факти» про світ, які стають основою для бізнес-аналітики, наукових досліджень та формування громадської думки. Дані, видобуті цими системами, часто сприймаються як об'єктивна істина, як сировина для знання.

Проте технологічний розвиток цих систем, особливо перехід до складних моделей глибокого навчання, створює фундаментальну проблему «чорної скриньки». Сама їхня архітектура – мільярди параметрів у багатовимірних векторних просторах та нелінійні перетворення – робить їхні внутрішні процеси практично незбагненними для людського розуму. Система надає результат (видобуті дані), але не надає раціонального, доступного для людини пояснення, чому саме ці дані були видобуті і як був зроблений висновок. Ця непрозорість створює не просто технічну чи етичну проблему, а й глибоку епістемологічну дилему.

Класична епістемологія, починаючи з платонівського діалогу «Теетет», визначає знання (episteme) як «обґрунтоване істинне переконання» (англ. Justified True Belief, JTB). Щоб стверджувати, що ми щось знаємо (наприклад, «ціна товару X становить Y»), ми повинні: а) вірити в це, б) це має бути істиною, і в) ми повинні мати обґрунтування (logos) для нашої віри [1].

Традиційна, або «інтерналістська», теорія обґрунтування вимагає

свідомого доступу до підстав. Проте навіть «екстерналістські» теорії, що покладаються на надійність (reliability) процесу, стикаються з проблемою «чорних скриньок». Система може бути надійною, але без обґрунтування ми не знаємо, чому вона є правильною. Ця проблема загострюється у динамічних середовищах, де відбувається «concept drift»: надійність, продемонстрована вчора, може непомітно деградувати, і без інтерпретації ми не можемо це перевірити. Саме компонент обґрунтування відрізняє раціональне знання від випадкового вгадування чи сліпої віри.

Застосовуючи цю тріаду до «чорної скриньки», ми бачимо кризу обґрунтування. Неінтерпретована система екстракції даних може надавати істинні (тобто точні) дані. Користувач, довіряючи системі, формує переконання. Однак, у нього повністю відсутній третій, ключовий компонент – обґрунтування. Користувач не може відповісти на питання: «Звідки ви знаєте, що ці дані правильні?». Єдина можлива відповідь – «Тому що так сказала система», – що є апеляцією до авторитету, а не раціональним обґрунтуванням [2].

Таким чином, неінтерпретована система ШІ перестає бути інструментом пізнання і перетворюється на сучасного «оракула». Вона надає пропозиції про світ («факт А є істинним»), які вимагають віри, а не раціонального прийняття. Довіра до такого оракула є не епістемологічною, а радше фідеїстичною, що фундаментально суперечить ідеалам Просвітництва, на яких побудована сучасна наука та суспільство і які вимагають доказовості, верифікації та раціонального сумніву. Це входить у прямий конфлікт з ідеалом, вираженим у Кантівському «Sapere aude» («Май мужність користуватися власним розумом»), оскільки користувач змушений відмовитись від власного розуміння на користь сліпої довіри. Це також ілюструє феномен, який Ленгдон Віннер назвав «технологічним сомнамбулізмом» – ми несвідомо приймаємо технологічні артефакти, не осмислюючи їхні глибокі філософські та соціальні наслідки [3].

Ця ситуація також підважує концепцію епістемічної відповідальності – обов’язку раціонального агента (людини) формувати свої переконання на основі достатніх доказів та обґрунтувань [4]. Коли аналітик чи науковець покладається на дані з «чорної скриньки», він, по суті, делегує свою епістемічну відповідальність непрозорому механізму. Це стає неприйнятним у сферах з високою ціною помилки (медична діагностика, юридичні рішення), де «система так вирішила» не може бути прийнятним обґрунтуванням для дії [5].

Більше того, відсутність обґрунтування робить нас вразливими до «тихих помилок». У традиційному процесі пізнання обґрунтування дозволяє нам виявляти хиби в міркуваннях. Неінтерпретована система може бути «тихо неправою» – генерувати помилкові, але правдоподібні дані (подібно до «галюцинацій» у великих мовних моделях). Оскільки процес міркування прихований, у нас немає механізму для аудиту чи ефективної

фальсифікації результату, окрім дорогої ручної перевірки.

Ця криза обґрунтування не є неминучою, вона є наслідком архітектурного вибору на користь непрозорих моделей. Альтернативою є створення засадничо інтерпретованих (англ. interpretable by design) систем. Розробка гібридних моделей, що поєднують нейронні методи з прозорими символьними правилами або нечіткою логікою, є прямим вирішенням цієї епістемологічної проблеми.

Такі системи можуть бути спроєктовані для того, щоб надавати не лише результат (дані), але й обґрунтування (наприклад, «дані видобуто, бо вони відповідали правилу А та візуальному шаблону Б»). Пояснення, згенероване такою гібридною системою, відновлює розірваний епістемологічний ланцюг. Воно перетворює висновок ШІ з догматичного твердження оракула на фальсифіковану гіпотезу з доказовою базою. Користувач отримує не лише «істинне переконання», але й «обґрунтування», що дозволяє йому виконати свою епістемічну відповідальність.

Таким чином, інтерпретованість – це не просто бажана технічна опція чи «nice-to-have» для зручності. Це фундаментальна епістемологічна вимога для того, щоб дані, згенеровані ШІ, могли легітимно вважатися знанням у строгому сенсі цього слова.

Список використаних джерел:

1. Rodrigues L. E. Epistemology: A Contemporary Introduction to The Theory of Knowledge. Disputatio. 2012. Т. 4, № 33. С. 540–545. URL: https://doi.org/10.2478/disp-2012-0017 (дата звернення: 07.11.2025).

2. Ribeiro M., Singh S., Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, м. San Diego, California. Stroudsburg, PA, USA, 2016. URL: https://doi.org/10.18653/v1/n16-3020 (дата звернення: 10.11.2025).

3. Winner L. Technologies as Forms of Life. Ethics and Emerging Technologies. London, 2014. С. 48–60. URL: https://doi.org/10.1057/9781137349088_4 (дата звернення: 12.11.2025).

4. Bonjour L., Code L. Epistemic Responsibility. The Philosophical Review. 1990. Т. 99, № 1. С. 123. URL: https://doi.org/10.2307/2185214 (дата звернення: 13.11.2025).

5. Doshi-Velez F., Kim B. Considerations for Evaluation and Generalization in Interpretable Machine Learning. The Springer Series on Challenges in Machine Learning. Cham, 2018. С. 3–17. URL: https://doi.org/10.1007/978-3-319-98131-4_1 (дата звернення: 15.11.2025).