УДК 004.8:17
DOI: https://doi.org/ 10.17721/3041-2323.2024.121-126

Ілля ДІДИК, студ.
e-mail: ilya.didyk@gmail.com
Запорізький національний університет,
Запоріжжя, Україна

РОЗВ'ЯЗАННЯ МОРАЛЬНО-ЕТИЧНИХ ЗАДАЧ ЗА ДОПОМОГОЮ ШТУЧНОГО ІНТЕЛЕКТУ

Розглянуто проблематику морально-етичних задач і розв'язання їх за допомогою штучного інтелекту, увага акцентовано на сферах застосування, таких як автономні транспортні засоби, медичні рішення та фінансові системи. Висвітлено приклади моральних дилем, що виникають у цих контекстах, а також нейромережі, які використовують для розв'язання цих задач.

Ключові слова: ***штучний інтелект, морально-етичні задачі, дилема трамвая, нейромережі.***

Вступ

Штучний інтелект (ШІ) стрімко розвивається та щільно інтегрується у різні сфери нашого життя, він нині є одним із ключових інструментів як повсякдення, так і професійної діяльності. Одним із можливих напрямів його застосування є розв'язання морально-етичних задач, які виникають у процесі взаємодії з людьми та світом (Russell, & Norvig, 2020). Такі задачі виникають у багатьох сферах, починаючи з керування автономними транспортними засобами і закінчуючи медичними рішеннями, які приймаються на основі аналізу великих даних.

Результати

Однією з найвідоміших морально-етичних задач у сфері ШІ є "дилема трамвая" ("trolley problem", також її називають "проблемою вагонетки"), яка адаптується для автономних транспортних засобів (Goodall, 2014). Дилема полягає в тому, щоб визначити, як автомобіль / вантажівка / автобус має поводитися в ситуаціях, де неминуче трапляється аварія, і потрібно вибрати між більшим і меншим злом. Наприклад, коли треба обрати ситуацію між наїз-

дом на пішоходів і ризиком для життя пасажирів, що перебувають у машині. Різні компанії використовують унікальні підходи до цього питання за навчання своїх ШІ, наприклад Tesla Autopilot першим пріоритетом має мінімізацію шкоди, а другим – безпеку самого водія, а Mercedes-Benz Intelligent Drive навпаки в першу чергу має намір захистити водія та пасажирів авто. Uber Advanced Technologies Group зробили крок далі і навчають ШІ саме на етичних дилемах із моделюванням різних ситуацій, максимально зменшуючи ризик для вразливих учасників дорожнього руху.

ШІ також використовують для прийняття медичних рішень, таких як діагностика захворювань або вибір оптимального лікування, профілактики, реабілітації тощо (Topol, 2019). Основними викликами перед розробниками є етична стандартизація факторів конфіденційності, прозорості, відповідальності, безпеки та справедливості (Floridi, & Cowls, 2019). У цьому контексті виникає морально-етична задача – довірити автоматизованим системам прийняття рішень, особливо, коли йдеться про життя та здоров'я людини, з одного боку – виникає мінімізація людського фактора, з іншого – невизначеність покладання відповідальності у разі невдалого лікування. Прикладами компанії, що займаються такими питаннями, можуть слугувати Deep Mind Health від компанії Google. Їхній ШІ ставить у пріоритет прозорість і відповідальність, оброблення медичних даних відбувається з жорстким шифруванням й акцентуванням уваги на мінімізації використання даних про пацієнта. Водночас запускаються прозорі процеси звітності, що дозволяють пацієнтам контролювати, як і задля яких цілей використовують їхні дані. Компанія IBM зі своїм інструментом Watson Health обрала підхід довіри – рішення фокусується на довірі між лікарями, пацієнтами та технологіями, використовуються механізми анонімізації даних для конфіденційності. Також компанія IBM підтримує активний діалог з етичними комітетами, такими як IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, та навчає свою мережу відповідно до прийнятих етичних стандартів. Важливо зазначити, що всі компанії, які займаються інтеграцією ШІ у сферу медицини, наголошують на використанні своїх інструментів у контексті підтримки прийняття рішень, а не як остаточних авторитетів.

Використання ШІ у фінансових системах для оцінювання кредитоспроможності також піднімає питання етики (Pasquale, 2015). Компанії що займаються розробленням і впровадженням ШІ для розрахунку кредитних рейтингів, такі як Fair Isaac Corporation – зазначають, що ключовими морально-етичними дилемами є такі: прозорість, конфіденційність, справедливість / дискримінація. Для розв'язання цих проблем застосовують big data моделі, що за допомогою виведених моделей оцінювання розраховують рейтинг і прогнозують вдале повернення кредиту. Коли навіть таких вузьконаправлених моделей недостатньо – компанії звертаються до сторонніх сервісів, наприклад Zest AI та Upstart, що пропонують знижувати ризики з уникненням упереджень й аналізувати нетрадиційні дані, такі як освіта та досвід клієнта у різних сферах життя.

Методи розв'язання морально-етичних задач. Штучні нейронні мережі (нейромережі) є однією з ключових технологій, що використовується для розв'язання морально-етичних задач. Вони можуть навчатися на великих обсягах даних і приймати рішення на основі виявлених патернів (Russell, & Norvig, 2020; Moor, 2006). Нині пропонується розглянути кілька методів і підходів до розв'язання цих задач і проблем, що виникають.

"Навчання з підкріпленням" ("Reinforcement Learning"), або як його ще називають "Навчання з учителем" або "Навчання на основі винагороди" – цей метод використовують для навчання агентів ШІ, щоб вони приймали рішення на основі досвіду вчителя, власного досвіду, а також на підставі моделювання наслідків дій. Під час розв'язання морально-етичних задач навчені агенти ШІ можуть уникати певних ускладнених ситуацій або мінімізувати ризик.

"Глибинне навчання" ("Deep Learning") також має інтерпретації, такі як "Багатошарове навчання" або "Ієрархічне навчання", що використовуються для розпізнавання складних патернів у великих даних, наприклад зображення або текстові дані. Під час навчання нейромережа обирає ту модель поведінки, що є найбільш наближеною до поставленої проблеми, і застосовує набутий досвід. Такий метод може бути корисним для виявлення упереджень під час прийняття рішень, а також забезпечить етичність класифікації контенту.

Коли метод розв'язання задачі було обрано, треба забезпечити йому сприятливий рівень довіри (Bostrom, 2014), із цим допомагають такі методи:

а) *візуалізація активацій і фільтрів* дозволяє зрозуміти, як нейромережа обробляє інформацію, шляхом наочного відображення нейронів. Прикладом може слугувати модель, що класифікує зображення і під час роботи виділяє фільтри на шарах, щоб побачити, які саме зображення викликають найбільшу активацію;

б) *метод оберненої проєкції* передбачає зворотне проходження сигналу, щоб побачити які частини вхідного зображення найбільше вплинули на прийняте рішення. Як приклад – під час розпізнавання кішки, метод показує, що на прийняте рішення найбільш вплинули вуха та хвіст. Принципово відрізняється від попереднього методу тим, що працює постфактум;

в) *Grad-CAM або метод градієнтної камери* передбачає аналіз великої кількості об'єктів, обраховуючи ступінь значущості, будує градієнтну карту, або карту нормалей для візуалізації більш та менш важливих активаторів. Такий самий принцип роботи має і метод Attention Mechanisms;

г) *LIME або метод незалежних інтерпретованих пояснень* апроксимує рішення складних моделей на простіші моделі, які легше інтерпретувати. Прикладом може виступати класифікація тексту, де модель виокремлює певні слова-тригери, замість глибинного аналізу, і на їхній основі виробляє рішення – текст є офіційним чи неофіційним. Так само працює також метод Interpretable Surrogate Models;

д) *SHAP або пояснення застосунків Шеплі* – метод, заснований на концепції варіанта теорії ігор, що надає кількісні оцінки важливості кожної ознаки об'єкта під час прийняття рішень. Найпоширеніше застосування здобув цей метод у фінансовій індустрії, де він враховує фактори рівня доходу, кредитну історію, вік, та їхній ступінь впливу;

є) найбільш просунутий і надійний метод – XAI (Explainable AI Frameworks), який включає комплексні підходи й інструменти, розроблені для підвищення прозорості й інтерпретованості моделей ШІ. Компанії IBM і Google власноруч розробляють XAI-інструменти для пояснення рішень, що приймаються склад-

ними моделями, в них враховують не лише перевірку, а й регулювання в мануальному й автоматичному режимах.

Дискусія і висновки

Нині штучний інтелект поступово стає невід'ємною частиною багатьох сфер нашого життя, породжуючи нові морально-етичні виклики. Попередньо розглянуті приклади демонструють, що ШІ має потенціал стати ефективним допоміжним інструментом у повсякденній і професійній діяльності. Проте потрібно не лише отримувати відповіді, але й бачити що саме до них призвело, тому що штучний інтелект, звичайно може викреслити людський фактор певної діяльності, але розробляється він теж людиною, котра може помилитись. Важливо продовжувати вдосконалювати технології та підходи, що забезпечують прозорість, відповідальність і справедливість у процесах прийняття рішень.

Список використаних джерел

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press. https://nickbostrom.com/superintelligence

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1). https://hdsr.mitpress.mit.edu/pub/l0jsh9d1

Goodall, N. J. (2014). Machine ethics and automated vehicles. In A. Salerno et al. (Eds.), *Autonomous driving* (pp. 93–102). Springer. https://doi.org/10.1007/978-3-319-05990-7_10

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21(4), 18–21. https://doi.org/10.1109/MIS.2006.80

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press. https://www.hup.harvard.edu/catalog.php?isbn=9780674970847

Russell, S., & Norvig, P. (2020). *Artificial intelligence: A modern approach* (4th ed.). Pearson. https://aima.cs.berkeley.edu

Topol, E. (2019). *Deep medicine: How artificial intelligence can make healthcare human again*. Basic Books. https://www.basicbooks.com/titles/eric-topol/deep-medicine/9781541644632/

References

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press. https://nickbostrom.com/superintelligence

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1). https://hdsr.mitpress.mit.edu/pub/l0jsh9d1

Goodall, N. J. (2014). Machine ethics and automated vehicles. In A. Salerno et al. (Eds.), *Autonomous driving* (pp. 93–102). Springer. https://doi.org/10.1007/978-3-319-05990-7_10

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21(4), 18–21. https://doi.org/10.1109/MIS.2006.80

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press. https://www.hup.harvard.edu/catalog.php?isbn=9780674970847

Russell, S., & Norvig, P. (2020). *Artificial intelligence: A modern approach* (4th ed.). Pearson. https://aima.cs.berkeley.edu

Topol, E. (2019). *Deep medicine: How artificial intelligence can make healthcare human again*. Basic Books. https://www.basicbooks.com/titles/eric-topol/deep-medicine/9781541644632/

Отримано редакцією журналу / Received: 16.09.24
Прорецензовано / Revised: 25.09.24
Схвалено до друку / Accepted: 01.10.24

Ilya DIDYK, Student
e-mail: ilya.didyk@gmail.com
Zaporizhzhia National University, Zaporizhzhia, Ukraine

SOLVING MORAL AND ETHICAL PROBLEMS USING ARTIFICIAL INTELLIGENCE

The article examines the issues of moral and ethical challenges and their resolution using artificial intelligence, with a focus on applications such as autonomous vehicles, medical decisions, and financial systems. It highlights examples of moral dilemmas that arise in these contexts, as well as the neural networks used to address these challenges.

Keywords: ***artificial intelligence, moral and ethical challenges, trolley dilemma, neural networks.***

Автор заявляє про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The author declares no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.