

УДК 171:004

DOI <https://doi.org/10.30970/PPS.2023.46.18>

ЕТИЧНІ ПИТАННЯ ВИКОРИСТАННЯ ТЕХНОЛОГІЙ МАШИННОГО НАВЧАННЯ

Олеся Паньків

*Національний університет «Львівська політехніка»
Інститут гуманітарних та соціальних наук, кафедра філософії
вул. С. Бандери, 12, 79000, м. Львів, Україна*

Стрімкий розвиток сучасних технологій демонструє людині нові знання, ідеї та можливості. В останні десятиліття значного прогресу в багатьох сферах життя досягнув штучний інтелект (ШІ). В даній статті розглядається машинне навчання – галузь штучного інтелекту, метою якого є повне або часткове автоматизування складних аналітичних завдань, що дає можливість забезпечити максимально точні прогнози на основі початкових даних та виконувати завдання, які раніше вимагали людської інтелектуальної діяльності. Сьогодні машинне навчання застосовують в численних галузях – від фіксації ризиків серцевої недостатності до подолання наслідків зміни клімату. Однак, у певних програмах, побудованих на машинному навчанні, дослідники вбачають порушення прав людини та конфіденційності. Зокрема, в статті наголошується на важливості звернення до етичних питань, які стосуються використання технологій машинного навчання: програми з розпізнавання обличчя (DeepFace) та технології GPT. Найчисельнішими є дискусії, пов'язані з наслідками використання ChatGPT. В статті відзначається можливість зловживання GPT, як і будь-яким іншим технологічним інструментом, для створення фальшивих новин, генерування тексту, копіювання тексту з порушенням авторських прав тощо. Відзначено, що розвиток технологій машинного навчання не лише впливає на людську поведінку, вносить зміни в її ціннісні орієнтири, а й створює нові форми зла: «штучне зло» (Л. Флоріді), «банальне (автоматизоване) зло» (А. Чан). Це вказує на те, що не обов'язково бути людиною аби стати жертвою злих дій. Наголошено на необхідності розгляду та аналізу потенціалу машинного навчання для визначення перспектив його подальшого розвитку та вирішення етичних питань, що виникають при його використанні.

Ключові слова: штучний інтелект, машинне навчання, «банальне (автоматизоване) зло», GPT.

Сьогодні ми живемо у технологічному світі і більшість наших повсякденних дій відбувається завдяки технічним та технологічним можливостям – від користування мобільним телефоном до більш складніших технологічних рішень в сфері медицини, транспорту, бізнесу тощо. Завдяки розвитку штучного інтелекту ми рухаємося до повністю автоматизованого світу, де штучні машини будуть самостійно діяти та приймати рішення. З цього приводу виникає подвійне відчуття: з однієї сторони, це допоможе у вирішенні складних завдань та зробить наше життя більш зручним та продуктивнішим, з іншого, породжує ряд занепокоєнь щодо потенційного негативного впливу на людей та суспільство в цілому.

Метою та завданням даної статті є виявлення етичних проблем розвитку штучного інтелекту, зокрема, оцінка викликів та ризиків, які з'являються з новими рішеннями та можливостями машинного навчання.

Машинне навчання є однією з найбільш розвинених галузей штучного інтелекту, а його філософські аспекти стали предметом обговорення серед багатьох впливових філософів, етиків та інших дослідників. Провідні дослідники проблем машинного навчання Р. Курцвейл [1], Н. Бостром [2], П. Домінгос [3] схиляються до думки, що дана технологія

може стати найбільш прогресивним та корисним винаходом сучасної цифрової епохи. Можливості, які відкриває машинне навчання, виглядають доволі перспективно, однак, впливаючи та трансформуючи різні сфери людської діяльності, породжують ряд етичних проблем: довіри, відповідальності та безпеки. Саме тому з'являється необхідність в розгляді та аналізі даних проблем, щоб зрозуміти прогностичний потенціал машинного навчання для визначення перспектив розвитку сучасного суспільства.

Машинне навчання (machine learning) – галузь штучного інтелекту, в якій комп'ютери використовують величезні обсяги даних, щоб навчитися виконувати завдання, а не бути запрограмованими для їх виконання [4]. Метою машинного навчання є повне або часткове автоматизування складних аналітичних завдань, що дає можливість забезпечити максимально точні прогнози на основі початкових даних. «Доступність великої кількості даних, поєднаних поточною обчислювальною потужністю є важливим фактором в основі революції машин» [5]. На даний момент машинне навчання охоплює широкий спектр додатків: від банків, ресторанів, заправок до роботів на виробництві. Нові завдання, що виникають практично щодня, призводять до появи нових напрямків машинного навчання. Хоча технології штучного інтелекту забезпечують вражаючі результати (наприклад, скринінг раку та створення інклюзивних середовищ для людей з обмеженими можливостями), вони також ставлять перед суспільством нові, небачені раніше виклики. Наприклад, використання штучного інтелекту збільшує ризик посилення гендерних та етнічних упереджень, порушує приватність та гідність людей, створює загрози масового стеження та сприяє використанню ненадійних технологій у правоохоронних органах.

За останні кілька десятиліть машинне навчання стало значно прогресивнішим завдяки систематичному збору великих обсягів даних за допомогою інтерактивним технологіям, які доволі часто порушують конфіденційність, мова йде про платформи соціальних мереж, програми для мобільних телефонів тощо. Хоча доволі поширеною є думка про те, що інформатика є нейтральною, оскільки спирається на «штучні інтелектуальні алгоритми», які нібито сліпі до соціального контексту, Гебру вважає лише розширенням «міфу наукової об'єктивності» [6]. Уявна нейтральність алгоритмів сама по собі може вважатися небезпечною, оскільки завдяки їй може здійснюватись дискримінація. Звертаючись до етичних проблем породжених розвитком машинного навчання, Сеннеті використовує поняття «банального (автоматизованого) зла» [5]. Покликаючись на концепцію «банального зла», яка була розроблена Ханною Арент [7], прагне пролити світло на питання нейтральності алгоритмів. Згадуючи випадок Адольфа Ейхмана, який просто виконував накази. Х. Арентд стверджувала, що зло може бути спричинене відсутністю критичного мислення, а не індивідуальною схильністю до зла. Поняття «банальне зло» було використане для позначення випадків, коли люди відмовляються критично мислити, виконуючи накази, що призводять до жахливих наслідків. Зло є наслідком відмови багатьох людей мислити, виконуючи накази. Цей приклад Сеннет приводить не випадково, оскільки вважає, що його можна застосовувати в контексті «дотримання правил», «конформістських алгоритмів» які сліпо виконують накази своїх потенційно упереджених програмістів або самі приходять до морально проблемних рішень через відсутність людського критичного мислення. «Алгоритм підкоряється наказам, наданим кодексом, не думаючи і не сумніваючись у соціальних і етичних наслідках дій» [5]. Це, в свою чергу, може завдати шкоди вразливим верствам населення. Крім того, особи з морально сумнівними цілями можуть використовувати інструменти машинного навчання у власних цілях. Банальність автоматизованого зла може мати небезпечні наслідки, до яких етикам та філософам необхідно підійти критично, а науковцям які займаються даними та комп'ютерними технологіями – запобігти цьому.

Сьогодні машинне навчання застосовують в багатьох сферах. Успішними прикладами є діагностика медичних захворювань, автономні автомобілі, робототехніка тощо. Проте, у певних програмах побудованих на машинному навчанні, дослідники К. Вонг, А. Добсон [8], Р. Алєн, Д. Мєстерс [9] вбачають певні виклики та ризики, які порушують права людини та конфіденційність. Зокрема, викликає доволі серйозні етичні питання програма з розпізнавання облич DeerFace розроблена компанією Facebook. Одним з основних ризиків використання DeerFace та інших алгоритмів розпізнавання облич є порушення приватності. Ці алгоритми можуть зібрати та зберігати велику кількість особистих даних про людей, використовувати їх без згоди власників цих даних. Це може призвести до можливого зловживання та навіть стеження за людьми.

Іншим ризиком є можливість помилкового розпізнавання. Це може призвести до помилкових ідентифікацій та неправомірних звинувачень. Крім того, використання DeerFace може призвести до зростання расової та гендерної дискримінації. Алгоритми розпізнавання облич можуть містити певні спотворення та помилки, що можуть призвести до неправомірних рішень та дискримінації проти окремих груп людей. Китайський уряд використовує дану технологію для моніторингу переміщень громадян, оцінки їх поведінки, що в подальшому може впливати на результати співбесіди, навіть надання «соціальних кредитів» [8]. Це може, в свою чергу, може лише сприяти дискримінації [9].

Технологія, яка є невіддільною людському контролю, «автономна та некерована сила, яка дегуманізує все, до чого торкається» [10], ця позиція стала сьогодні основою багатьох етичних дискусій, пов'язаних з можливими наслідками використання GPT-3. Ці дискусії виникають у зв'язку зі збільшенням занепокоєнням щодо потенційної шкоди, яку може завдати дана технологія в майбутньому. GPT-3 (Generative Pre-trained Transformer) – авторегресійна модель мови третього покоління, яка використовує глибоке навчання для створення тексту та навчається за допомогою безлічі інтернет-даних. Кількість активних користувачів ChatGPT (чат-боту) із штучним інтелектом розроблений лабораторією OpenAI, лише за 2 місяці перевищила 100 млн, що стало історичним рекордом. Від 18 лютого 2023 року ChatGPT став доступним і для українських користувачів. Цей чат може бути використаний для багатьох різних цілей, таких як написання статей, відповідей на запитання, письмової комунікації, написання рядків програмного коду тощо. Однак, можливе зловживлення GPT, як і будь-яким іншим технологічним інструментом. Наприклад, GPT може бути використаний для створення фальшивих новин, оскільки він здатен генерувати вигадані факти і події, що можуть бути неправдивими, генерувати високоякісний текст, який можна використовувати для створення спаму, надсилати рекламу, або ж копіювати тексти без дозволу авторів. Застереження висловлюють Л. Флоріді та М. Чіріатті [11]. Дослідники попереджають, що використання GPT-3 призведе до реформування ринку праці, онлайн-маркетинг стане керованим штучним інтелектом, а читачі тексту повинні будуть «звикнути не знати, чи є джерело штучним чи людським» [11]. Вони наголошують, на потребі у створенні цифрової культури: «щоб теперішні та майбутні громадяни, користувачі та споживачі знали про нову *інфосферу*, в якій вони живуть і працюють». Крім того, людство повинно стати більш критичним в новій інфосфері [11]. Це міркування знову застерігає від бездумності та некритичності, яка може стати джерелом виникнення зла. Примітно, що хоча автори визнають необхідність законодавчих змін (таких як поправки до закону про авторське право) та певних рішень для громадської обізнаності, вони ж в свою чергу припускають, що GPT-3 призведе до майбутнього, наповненого «семантичним сміттям» [11]. Слід відзначити, що Л. Флоріді та М. Сандерс ще у 2001 році, досліджуючи

розвиток інформаційних технологій у статті «Штучне зло і основи комп'ютерної етики» (Artificial evil and the foundation of computer ethics) [12], зосередили увагу на нових формах зла. Мова йде про поняття «штучного зла» (artificial evil) як можливу властивість систем штучного інтелекту, яка полягає в навмисному виконанні завдань, які шкодять людям або суспільству в цілому. Ця властивість може виникнути внаслідок недосконалої розробки системи, помилкового налаштування алгоритмів, а також в результаті зловживання системою з боку користувача. Флоріді та Сандерс зазначають, що потенційне «штучне зло» може бути відображене у випадках, коли система штучного інтелекту допускає помилку, або в разі зловживання користувачем, спрямовує роботу системи на шкоду іншим людям або суспільству в цілому.

Попри песимістичні погляди, незадовільних рішеннях щодо нової технології, А. Чан вважає, що в підході Л. Флоріді та М. Чіріатті спостерігається перенесення свободи дій від людей до машини. Їхня точка зору, на думку дослідниці, ігнорує людей, а саме: розробників ШІ, політиків, громадянське суспільство, навіть зловмисників, які безпосередньо відповідальні за використання GPT-3 або інших паралельних мовних моделей як інструменту дезінформації. Це вимагає більш детального аналізу сфери людської діяльності та прийняття рішень в межах розвитку ШІ та, зокрема, машинного навчання. «Замість того, щоб прийняти GPT-3 та його неправильне використання, ці автори наголошують на необхідності попередніх регулятивних заходів, щоб запобігти можливому надходженню дезінформації» [10]. Зокрема, звертаючись до нової технологічної можливості GPT-3, А. Чан відзначає, що GPT-3 і мовні моделі знаходяться в зародковому стані, і це дозволяє запроваджувати відповідну політику та закони щодо безпеки.

Гадаю, що сьогодні в інтелектуальних колах з великою пересторогою сприймають можливість ChatGPT, стверджуючи, що GPT-3 означає кінець традиційної журналістики, поезії, літератури тощо. І, звичайно, є підставити так міркувати. Це питання хвилює і викладачів філософії. Чи можемо отримати від студентів філософські есе, створені GPT-3? Роман Кшановський, доктор філософії Папського університету Івана Павла II у Кракові, провів дослідження «GPT-3 та філософія» яке дало можливість проаналізувати особливості виконання завдань з філософії. «Результати вказують на те, що GPT-3 у досліджуваній галузі філософії працював непередбачувано, майже випадково. На деякі запитання GPT-3 давав відповіді, які можна вважати студентською роботою, але на інші запитання GPT-3 просто генерував суміш зв'язного та невідповідного тексту, хоча все ще граматично правильний текст. Ми не знаємо, чому на одні запитання дається правильна відповідь, а на інші – ні» [13], відзначено у результатах дослідження. На думку дослідника, очікування, що GPT-3 є зорею нової ери синтетичної філософії або сутінком традиційної, не видається виправданим. «Філософія, як ми її розуміємо, не є комбінаторним повторенням речей, як вже було сказано. Філософія – це відкриття прихованих, неявних смислів, неочевидних речей, прихованих за стінами» [13]. Тому, як підкреслює Р. Кшановський, ми не повинні очікувати від GPT-3 потоку синтетичної філософії з блискучими новими філософськими ідеями.

В якості невеличкого експерименту, користуючись можливостями ChatGPT, було цікаво отримати відповідь на питання про можливі ризики користування даним чатом. Відповідь не забарилася:

«ChatGPT – це інструмент, розроблений, щоб надавати корисні відповіді на запитання, однак, як і у будь-якого технологічного інструменту, є деякі потенційні ризики, які можуть виникнути при користуванні ним.

Ось декілька можливих ризиків користування ChatGPT:

Некоректні відповіді: ChatGPT є інструментом машинного навчання, тому іноді можуть виникати некоректні або неправильні відповіді на запитання, особливо якщо запитання містить неточну інформацію або якщо в ньому міститься багатозначність.

Конфіденційність даних: при користуванні ChatGPT можуть передаватися конфіденційні дані, такі як імена, адреси електронної пошти або номери телефонів. Хоча ChatGPT не зберігає жодних конфіденційних даних, існує ризик, що дані можуть бути викрадені під час передачі.

Залежність від технології: користувачі можуть стати залежними від ChatGPT або інших подібних технологій для відповідей на запитання і, отже, втратити здатність до самостійного пошуку інформації або розв'язання проблем.

Вплив на мовну культуру: через те, що ChatGPT може використовуватися для зв'язку з людьми, є ризик, що він може впливати на мовну культуру і сприяти використанню некоректних або неприйнятних виразів.

Важливо зазначити, що ChatGPT зазвичай намагається давати точні та корисні відповіді, і більшість користувачів використовують його без проблем».

Слід відзначити, що зазвичай існує тенденція переоцінювати вплив та загрози від нової технології, а також недооцінювати, яких наслідків можна очікувати. Безумовно, машинне навчання відкриває для людства великі перспективи, однак, впливаючи та трансформуючи різні сфери людської діяльності, слід не забувати про ризики які несуть в собі дані технології: від зловживання конфіденційними даними, порушення авторських прав до ряду проблем які стосується безпеки. У випадку з технологією GPT, важливою умовою є усунення шкідливих вихідних даних мовної моделі, задля сприяння якості, справедливості та прозорості. Це, звісно, не просте завдання, оскільки мовна модель, навчена величезною кількістю Інтернет-контенту, обов'язково міститиме токсичну, упереджену, сексистську та насильницьку мову. Тому законодавчо слід вимагати жорсткіших стандартів збору даних, особливо для чутливих контекстів, таких як освіта, охорона здоров'я чи кримінальне правосуддя. Відповідальними за додавання важливої соціотехнічної інформації мають бути також розробники. Важливо, щоб етичні рішення для боротьби з маніпуляціями та упередженнями були також спрямовані на підвищення рівня цифрової грамотності суспільства. Сфера штучного інтелекту є доволі динамічною та стрімко розвивається, тому при розробці програмних застосувань, що задіюють підходи машинного навчання, критично важливим є спрямованість до людини, а не проти людини. Властиво, це положення було відзначено як одне з основних у маніфесті цифрового гуманізму [14] та у першому глобальному стандарті етики штучного інтелекту 25 листопада 2021 року на Генеральній конференції ЮНЕСКО [15].

Дана стаття залишає відкритими ряд етичних питань щодо того як можна зменшити упередженість і водночас запобігти зловмисному використанню технологій машинного навчання.

Список використаної літератури

1. Kurzweil R. How to create a mind: The secret of human thought revealed. Washington: Viking Penguin, 2012. 352 p.
2. Bostrom N. Super intelligence: Paths, Dangers, Strategies. Oxford: Oxford University Press, 2013. 352 p.
3. Domingis P. The master algorithm: How the quest for the ultimate learning machine will remake our world. Washington: MT Press, 2015. 356 p.

4. Oxford Learner's Dictionaries API. URL: <https://www.oxfordlearnersdictionaries.com/us/definition/english/machine-learnin/g>
5. Senent R. M. Bueso D. The Banality of (Automated) Evil: Critical Reflections on the Concept of Forbidden Knowledge in Machine Learning Research. *Recerca. Revista de Pensament i Anàlisi*. 2022. Vol. 27(2), P. 1–26.
6. Gebru T. Race and Gender. In *Oxford Handbooks of AI Ethics*. Oxford: Oxford Handbooks. 2019. P. 251–259.
7. Арендт Х. Банальність зла. Суд над Айхманом в Єрусалимі. Пер. з англ. А. Котенко. К.: Дух і Літера. 2013. 367 с.
8. Wong K. Dobson A. We're just data: Exploring China's social credit system in relation to digital platform ratings cultures in Westernised democracies. *Global Media and China*. 2019. Vol. 4(2). P. 220–232.
9. Allen R. Masters D. Artificial Intelligence: the right to protection from discrimination caused by algorithms, machine learning and automated decision-making. *ERA Forum 20*. Springer. 2020. P. 585–598.
10. Chan, A. GPT-3 and InstructGPT: technological dystopianism, utopianism, and “Contextual” perspectives in AI ethics and industry. *AI Ethics*. 2023. Vol. 3. P. 53–64.
11. Floridi L., Chiriatti, M. GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds & Machines*. 2020. Vol. 30. P. 681–694. URL: <https://link.springer.com/article/10.1007/s11023-020-09548-1#citeas>
12. Floridi L., Sanders J.W. Artificial evil and the foundation of computer ethics. *Ethics and Information Technology*. 2001. Vol.3, P. 55–66. URL: <https://link.springer.com/article/10.1023/A:1011440125207>
13. Krzanowski Roman. GPT-3 and Philosophy. Preliminary Tests Results. 2022. (hal-03640357) URL: <https://hal.science/hal-03640357#>
14. Vienna Manifesto on Digital Humanism. URL: <https://dighum.ec.tuwien.ac.at/dighum-manifesto/May, 2019>
15. Unesco. URL: <https://www.unesco.org/en/articles/unesco-member-states-adopt-first-ever-global-agreement-ethics-artificial-intelligence>

ETHICAL ISSUES OF MACHINE LEARNING TECHNOLOGY USAGE

Olesia Pankiv

*Lviv Polytechnic National University,
Institute of Humanities and Social Sciences, Department of Philosophy
Bandera str., 12, 79000, Lviv, Ukraine*

The rapid development of modern technology brings humans new knowledges, ideas, and opportunities. Over the past decade, significant progress has been made in many areas of life, including artificial intelligence (AI). This article focuses on machine learning, a branch of AI whose goal is to fully or partially automate complex analytical tasks, providing highly accurate predictions based on initial data and performing tasks that previously required human intellectual activity. Today, machine learning is applied in numerous fields, from heart failure risk monitoring to mitigating the effects of climate change. However, in certain programs based on machine learning, researchers see violations of human rights and confidentiality. In particular, the article emphasizes the importance of addressing ethical issues related to the use of machine learning technologies, such as the DeepFace facial recognition program and GPT technology. The most frequent discussions are related to the consequences of using ChatGPT. The article notes the possibility

of misuse of GPT, like any other technological tool, for creating fake news, generating text, copying text while violating copyright, and so on. It is noted that the development of machine learning technologies not only affects human behavior, but also changes its value orientations and creates new forms of evil, such as "artificial evil" (L. Floridi) and "banal (automated) evil" (A. Chan). This indicates that it is not necessary to be a human to become a victim of evil actions. The article emphasizes the need to examine and analyze the potential of machine learning to determine the prospects for its further development and address the ethical issues that arise in its use.

Key words: artificial intelligence, machine learning, "banal (automated) evil", GPT.