УДК 519.816:004.85

ПОТЕНЦІЙНІ УПЕРЕДЖЕННЯ ТА ЕТИЧНІ ПИТАННЯ У ВИКОРИСТАННІ МАШИННОГО НАВЧАННЯ ДЛЯ ПРИЙНЯТТЯ РІШЕНЬ

Ярохно Д.В.
Науковий керівник – д.т.н., доцент Петрова Р.В.
Харківський національний університет радіоелектроніки, каф. СТ
м. Харків, Україна
тел.: +38(097) 422-00-52, e-mail: dmytro.iarokhno@nure.ua

This paper was created to explore the potential biases and ethical issues associated with the use of machine learning for decision-making. Examples of the use of machine learning in various fields are analyzed and the possible risks and consequences of misusing this technology are highlighted. Examples of possible solutions to the problem of bias in organizations are provided, and recommendations for the ethical use of this technology are given.

В останні роки машинне навчання стало надзвичайно популярним і широко використовується у багатьох галузях, включаючи медицину, фінанси, транспорт та інші. Однак, разом із зростанням використання машинного навчання, з’являються етичні питання та потенційні упередження, які потрібно враховувати при прийнятті рішень. Найбільш використовувані моделі які зараз застосовуються це моделі прогнозування, що містяться у бізнес-додатках, наприклад вони потрібні для автоматичного схвалення кредитів, надання рекомендацій клієнтам, відбору персоналу або розпізнавання різних речей.

Проблемою при прийнятті рішень за використання машинного навчання є упередженість. Наприклад, при оцінці кредитоспроможності клієнти, у багаторічній історії кредитного обслуговування яких не спостерігається правопорушень, зазвичай визначаються, як клієнти з низьким ризиком. Але уявіть, що іпотечні кредити цих клієнтів оплачувалися протягом багатьох років коштом суттєвих податкових пільг, які більше не надаються. Ступінь ризику змінюється, проте, якщо програма про це «не знає», вона не може дати адекватну оцінку. Машинне навчання може закріплювати й посилювати поведінкові упередження людей, і це проблема, з якою ми стикаємося в соціальних мережах. Наприклад, алгоритм фільтрування новин базується на особистих вподобаннях користувачів, що збільшує природні упередження аудиторії. Крім того, сайти можуть систематично приховувати протилежні точки зору [1].

Щоб зменшити упередженість у системах штучного інтелекту, організації можуть скористатися кількома процесами та засобами контролю. Такі засоби можна розділити на дві категорії:

- Контроль на рівні організації – організації можуть встановити засоби контролю на рівні підрозділу або на найвищому рівні, щоб створити ефективне середовище контролю, яке допоможе упоратися з упередженістю.
- Контроль на рівні процесу – для забезпечення вільної від упередженості розробки та функціонування систем штучного інтелекту слід впровадити внутрішні механізми контролю, такі як протоколи збору даних, розподіл обов'язків щодо систем штучного інтелекту та періодичні перевірки результатів цих систем [2].

Статистичне управління Великобританії розробило шість етичних принципів, які повинні бути враховані при застосуванні машинного навчання для забезпечення суспільного блага досліджень і статистики. Ці принципи стосуються збереження конфіденційності даних, розуміння можливих ризиків і обмежень нових методів та технологій, дотримання законодавчих вимог, врахування суспільної прийнятності проекту і забезпечення прозорості у зборі, використанні та обміні даними [3].

Існують різні ризики упередженості в системах штучного інтелекту, які залежать від їх призначення. Наприклад, системи, які рекомендують товари для покупок, мають менший ризик, ніж системи, які приймають рішення щодо кредитних заявок. Для кожної системи можуть бути потрібні різні засоби контролю, щоб запобігти упередженості. Крім того, існують інші ризики, такі як безпека моделі та конфіденційність даних, які необхідно враховувати. Проте, позитивним є той факт, що можна керувати упередженнями, якщо ми щиро про них говоримо. Люди повинні постійно усувати реальні обмеження машинного навчання. Для бізнесу це може означати створення додаткової цінності, заснованої на інсайтах, які отримані за допомогою добре контрольованих машин. Це є найреалістичнішим алгоритмом досягнення впливу при машинному навчанні.

Список використаних джерел:

1. Baer T. Controlling machine-learning algorithms and their biases [Електронний ресурс] / T. Baer, V. Kamalnath // McKinsey & Company. – 2017. – Режим доступу до ресурсу: https://is.gd/CVnwqV.

2. Sutaria N. Bias and Ethical Concerns in Machine Learning [Електронний ресурс] / Niral Sutaria // ISACA. – 2022. – Режим доступу до ресурсу: https://is.gd/wsI95O.

3. Ethical considerations in the use of Machine Learning for research and statistics [Електронний ресурс] // UK Statistics Authority. – 2021. – Режим доступу до ресурсу: https://is.gd/MpUyTX.