УДК 341

РИСКИ, СВЯЗАННЫЕ С УПРАВЛЕНИЕМ ИСКУССТВЕННЫМ ИНТЕЛЛЕКТОМ: ВЗГЛЯД СО СТОРОНЫ

Н. М. Шевко

Белорусский государственный университет,
ул. Ленинградская 8, 220030, г. Минск, Беларусь, ms.shevko@inbox.ru

Пристальное внимание к потенциальным недостаткам, связанным с использованием технологий искусственного интеллекта (далее – ИИ), обусловлено тем, что они способны вызвать риски и вред. С одной стороны, эффективное управление ИИ будет способствовать смягчению рисков. С другой стороны, управление ИИ уже представляет существенную проблему, в разрешении которой заинтересованы: государственные организации и правительства; частный сектор и общество.

Ключевые слова: искусственный интеллект; управление искусственным интеллектом; дискриминация; угроза правам человека; технология «дипфейк».

RISKS IN MANAGING ARTIFICIAL INTELLIGENCE: A VIEW FROM THE OUTSIDE

N. M. Shevko

Belarusian State University, Leningradskaya st. 8, 220030
Minsk, Belarus, ms.shevko@inbox.ru

The focus on potential downsides associated with the use of artificial intelligence (hereinafter referred to as AI) is due to the fact that they can cause risks and harm. On the one hand, effective governance of AI will help mitigate the risks. On the other hand, governance of AI already represents a significant challenge in the resolution of which: public organizations and governments; the private sector and society are interested.

***Keywords*:** artificial intelligence; artificial intelligence governance; discrimination; human rights threat; deepfake technology.

Как верно отмечает Генеральный Секретарь Организации Объединенных Наций (далее ООН) Антониу Гутерриш: «Если мы хотим использовать преимущества искусственного интеллекта и устранить риски, мы все должны работать вместе — правительства, промышленность, академические круги и гражданское общество — над разработкой структур и систем, которые обеспечивают ответственные инновации…» [1]. Стоит согласиться с тем, что искусственный интеллект (далее ИИ) и другие пе-

редовые технологии обладают значительным потенциалом для поддержки инклюзивности, сокращения неравенства, спасения Целей устойчивого развития (ЦУР) и укрепления деятельности системы ООН [2]. Однако использование положительного влияния ИИ требует пристального внимания к его потенциальным недостаткам, в том числе путем защиты конфиденциальности данных, смягчения предубеждений и обеспечения прозрачных процессов принятия решений. В связи с этим, особо важно использовать возможности ИИ, одновременно устраняя риски и вред, которые составляют его темную сторону.

Так, группа FutureTech из Массачусетского технологического института (MIT) в сотрудничестве с другими экспертами составила базу данных, включающую более 700 потенциальных рисков, классифицированных по причинам возникновения и разделенных на семь основных типов [3]. При этом важно отметить, что основные опасения от использования технологий ИИ связаны с безопасностью, предвзятостью и дискриминацией, а также вопросами конфиденциальности. Среди основных рисков, способных дать сбой и нанести вред человечеству особого внимания заслуживают следующие:

1. ИИ становится разумным, что приведет к тому, что ИИ достигнет разумности и будет способен воспринимать эмоции или ощущения, а также приобретет субъективный опыт, включая удовольствие и боль. В сложившейся ситуации, разумный ИИ может столкнуться с жестоким обращением или пострадать, если не будут реализованы его права. В этом случае перед учеными и регулирующими органами может встать вопрос: «Заслуживают ли системы ИИ аналогичного к себе отношения, как к людям, животным или окружающей среде?».

2. ИИ может преследовать цели, противоречащие интересам человека, что потенциально может привести к тому, что неправильно настроенный ИИ выйдет из-под контроля и причинит серьезный вред человеку и обществу. Данную проблему обычно называют проблемой «черного ящика» ИИ, которая связана с тем, что решения, принимаемые ИИ, больше не могут или только частично могут быть отслежены людьми. Стоит согласиться с Bernd W. Wirtz, Jan C. Weyerer, Ines Kehl которые указывают на то, что потеря контроля над ИИ или неправильное использование ИИ, в частности, в чувствительных областях применения, таких как правовая система, армия, политические коммуникации, транспорт, гражданская оборона или ядерная энергетика могут представлять серьезную угрозу правам человека, общественной безопасности и порядку, а также легитимности предоставления государственных услуг, тем самым препятствуя созданию общественной ценности с помощью ИИ [4], что представляется одной из доминирующих проблем в управлении ИИ.

3. ИИ может лишить людей свободы воли, что, с одной стороны, приведет к снижению уровня критического мышления и навыков решения проблем у людей. С другой стороны, широкое внедрение ИИ для выполнения человеческих задач приведет к значительному сокращению рабочих мест и растущему чувству беспомощности среди населения.

4. Возникновение неадекватной привязанности к ИИ, а именно, переоценивание способностей ИИ и занижение собственных способностей. Как результат, возникновение чрезмерной зависимости от технологий ИИ, изолированность от человеческих отношений, развитие эмоциональной зависимости и росту доверия к возможностям ИИ.

5. Технология «дипфейк» может облегчить искажение реальности и привести к распространению пропаганды или дезинформации, а также к манипулированию общественным мнением. В результате может увеличиться число сложных фишинговых схем, использующих созданные ИИ изображения, видео и аудиосообщения [4].

Таким образом, негативное влияние вышеуказанных рисков и их последствия требуют эффективного управления, что, по мнению автора, позволит не только смягчить эти риски, но и выработать широкое понимание управления ИИ, основанное на международно-правовом подходе. Более того, управление ИИ особо нуждается в дельнейших исследованиях в силу того, что управление ИИ уже представляет фундаментальную проблему в разрешении которой заинтересованы как государственные организации и правительства, так и частный сектор.

Библиографические ссылки

1. Secretary-General's message for Third Artificial Intelligence for Good Summit [Electronic resource] // URL: https://www.un.org/sg/en/content/sg/statement/2019-05-28/secretary-generals-message-for-third-artificial-intelligence-for-good-summit (date of access: 24.10.2024).

2. Artificial Intelligence [Electronic resource] // URL: https://unsceb.org/topics/artificial-intelligence (date of access: 24.10.2024).

3. What are the risks from Artificial Intelligence? [Electronic resource] // URL: https://airisk.mit.edu/ (date of access: 24.10.2024).

4. Bernd W. Wirtz, Jan C. Weyerer, Ines Kehl Governance of artificial intelligence: A risk and guideline-based integrative framework [Electronic resource] / URL: https://www.sciencedirect.com/science/article/abs/pii/S0740624X22000181 (date of access: 24.10.2024).