ПРОБЛЕМА ДОВЕРИЯ ЧЕЛОВЕКА ИСКУССТВЕННОМУ ИНТЕЛЛЕКТУ КАК НОВЫЙ ЭТИЧЕСКИЙ ВЫЗОВ

В. Н. Сокольчик

Один самых сложных вызовов искусственного интеллекта (далее – ИИ) – можем ли мы доверять ИИ. По сути именно доверие отражает нашу готовность взаимодействовать с ИИ, устанавливать ограничения для его

использования, выстраивать механизмы контроля, и, в конечном итоге, принимать результаты и выводы, предложенные ИИ. В паттерне взаимоотношений человека и ИИ доверие становится основой нашего взаимодействия и основой для дальнейшего развития ИИ.

Проводимые автором в составе исследовательского коллектива инициативного проекта по этическому сопровождению ИИ пилотные опросы, интервьюирование ученых, специалистов сферы IT, представителей сферы здравоохранения, педагогов и обучающихся, а также изучение научных исследований по вопросу доверия общества к ИИ показывают, что уровень доверия ИИ среди пользователей достаточно невысок. Причиной этого наши респонденты называют ошибки ИИ (в том числе ошибки, которые в профессиональной сфере могут привести к трагическим последствиям), ложь и заблуждения, которые ИИ достаточно часто продуцирует в ответ на запросы человека, предубеждения ИИ (связанные с особенностями представлений разработчиков, некорректным использованием баз данных и отсутствием учета социокультурных особенностей среды, где будут применяться системы ИИ), непрозрачность предоставляемых ИИ выводов (результатов) и т. д.

Безусловно, многие основания недоверия к ИИ будут со временем «сниматься», вследствие развития последнего, однако, во-первых, с расширением нашего постоянного взаимодействия с ИИ будут появляться новые проблемы и вызовы доверия, и, во-вторых, в отсутствии внимания к проблеме и отсутствии потребности ее решать, феномен «зловещей долины» вероятно будет прогрессировать в общественном и индивидуальном сознании [1].

Огромную роль для построения сбалансированных взаимоотношений с ИИ и, в частности, для формирования доверия ИИ, играет этика. Этическое сопровождение развития и использования ИИ в условиях экспоненциального роста его значения для человека сегодня становится обязательным условием для существования социума и индивида.

Сегодня обществу необходимы этико-правовые стандарты и рекомендации, обеспечивающие взвешенное и взаимообусловленное сотрудничество в системе взаимоотношений «разработчик ИИ – ИИ – пользователь ИИ». Необходимо понимать, что такие стандарты могут опираться скорее на этические нормы регулирования взаимоотношений, нежели на жесткие нормы права. Это связано со скоростью развития ИИ и технологий его разработок и непредсказуемостью путей совершенствования и распространения ИИ. По мнению автора, наиболее приемлемый путь регулирования взаимодействия ИИ с человеком в контексте доверия последнего к искусственному интеллекту – этико-правовые рекомендации и общая регламентация принципов взаимодействия.

В ряду уже действующих примеров такой регламентации можно назвать много документов, имеющих разный статус и сферу деятельности, среди наиболее известных – кодекс этики искусственного интеллекта (Россия) [2], Билль о правах ИИ (США) [3] и утвержденный в мае 2024 г. Регламент использования ИИ (Artificial Intelligence Act). Европейского союза (далее – Регламент) [4]. Авторы регламента, определяя цели создания соответствующего документа, обозначают значимость ограничений ИИ в контексте его влияния на человека и общество и в контексте формирования доверия общества к ИИ. Акцентируется, что регламент нацелен на уважение свободы науки и развитие ИИ, «очерчивает» общие направления и этические рамки такого развития.

Так, в регламенте ИИ, обозначены три группы систем ИИ и общие установки их регулирования [4]. Первую группу составляют «запрещенные системы», которые воздействуют на подсознание человека, могут «вычислять» слабые стороны социальных групп, взаимоотношений, человеческой личности и так далее, а также системы ИИ, которые осуществляют разного рода социальную оценку, выработку критериев распределения и принимают соответствующие решения. Такие системы должны быть взяты под постоянный контроль общественных структур (включая и профессиональные службы), требуют применения достаточно жестких ограничений вплоть до полного запрещения (ограничения) соответствующих исследований.

Вторую группу составляют системы высокого риска, где вероятность причинения ущерба, несправедливого распределения и так далее достаточно велика. По мнению разработчиков документа, использование таких систем может привести к дискриминации и стигматизации в обществе, установлению несправедливых систем распределения и другим неблагоприятным последствиям для социума. С системами высокого риска мы сталкиваемся в сфере финансов, медицины, правосудия, при организации доступа к значимым ресурсам – таким, как электричество, телекоммуникации и так далее. Типичный пример – сортировка пациентов для неотложной медицинской помощи (например, в ситуации катастроф, пандемий и так далее), которая при отсутствии этического обучения ИИ (а также разработчика) может привести не только к «механистическому» подходу к человеку, но и к летальным исходам, которых можно было бы избежать, и даже к разрушению ценностей нашей цивилизации. По мнению автора, к таким системам сегодня уже может быть отнесены и некоторые системы генеративного ИИ, которые могут оказать значимое негативное влияние на образование, науку, искусство. В контексте изучения систем высокого риска (в плане их воздействия на человека) проблемы приоритетной социальной и личностной коммуникации с ИИ также могут формировать проблемы: эмоциональной зависимости от ИИ, разрушение социальной адаптации, потерю адекватного восприятия

реальности. Особенностями работы с системами ИИ высокого риска, реализация которых может обеспечить доверие общества, это:

– требование использования систем ИИ с учетом предполагаемой цели создания;

– непрерывная система управления рисками на протяжении всего жизненного цикла системы;

– постоянный мониторинг взаимодействия ИИ с человеком и со средой;

– адекватное обозначение (в том числе в качестве инструктажа) любых возможных рисков целевого и нецелевого использования ИИ;

– необходимость социально-этической экспертизы систем ИИ;

– обучение ИИ с использованием высококачественных данных, обязательное использование при обучении ИИ этических регулятивов (принципы, ценности, нормы).

Особенную значимость в контексте систем высокого риска приобретает постоянный контроль и экспертиза решений, принимаемых ИИ и, по возможности, использование этих решений только в статусе рекомендаций.

Третья группа систем ИИ, обозначенная в Регламенте, – это системы низкого риска, которые не оказывают значимого негативного воздействия на пользователей. Однако в свете решения проблемы доверия общества, разработчики Регламента акцентируют обязательность информирования пользователей о том, что он взаимодействуют с ИИ даже при отсутствии видимых рисков (например, при анонимном анкетировании или опросе). Своевременное получение адекватной правдивой информации о взаимодействии с ИИ, по мнению автора, – это также одно из условий формирования доверия общества к использованию ИИ.

Таким образом, реализация мер этико-правовой регламентации использования ИИ, создание экспертных и консультативных социально-этических структур, этическое обучение ИИ и формирование этических требований к взаимодействию с ИИ вкупе с всеобщим информированием, общественным обсуждением поставленного вопроса, образованием и просвещением общества – основные ступени формирования доверия общества и индивида к использованию ИИ.

Литература и источники

1. Fisher, R. Trust in artificial intelligence: Global study on the shifting public perceptions of AI / R. Fisher [et al.] // KPMG, 2023 [Электронный ресурс]. – Режим доступа: https://assets.kpmg.com/ content/ dam/ kpmgsites/ xx/ pdf/ 2023/ 09/ trust-in-ai-global-study-2023.pdf.c oredownload.inline.pdf. – Дата доступа: 10.01.2024.

2. Кодекс деятельности ИИ (Россия, 2022) [Электронный ресурс]. – Режим доступа: https://ethics.a-ai.ru. – Дата доступа: 01.02.2023.

3. Blueprint for an AI Bill of Rights / The White House, 2022 [Электронный ресурс]. – Режим доступа: https://www.whitehouse.gov/ ostp/ ai-bill-of-rights/. – Дата доступа: 01.02.2023.

4. Регламент Европейского союза об искусственном интеллекте. Анализ основных положений и принципов регулирования, 2024 [Электронный ресурс]. – Режим доступа: https://ai.gov.ru/ knowledgebase/ dokumenty-po-razvitiyu-ii-v-drugikh-stranakh/ 2024_ reglament_ evropeyskogo_ soyuza_ob_ iskusstvennom_ intellekte_ano_cifrovaya_ekonomika_/. – Дата доступа:15. 09.2024.

Национальная академия наук Беларуси
Институт философии НАН Беларуси

ИНТЕЛЛЕКТУАЛЬНАЯ КУЛЬТУРА БЕЛАРУСИ:
от Просвещения к Современности

Материалы
Восьмой международной научной конференции
(21–22 ноября 2024 года, г. Минск)

В двух томах
Том 2

МИНСК
ИЗДАТЕЛЬСТВО «ЧЕТЫРЕ ЧЕТВЕРТИ»
2024