ОСНОВНЫЕ НАПРАВЛЕНИЯ И ДИСКУССИОННОЕ ПОЛЕ ЭТИКИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

И. Г. Красникова

Этические аспекты создания и последствий использования искусственного интеллекта (ИИ), его влияния на жизнь человека и общество в целом в настоящее время становятся предметом широкого обсуждения в философии, науке и публичном пространстве. Этика искусственного интеллекта прежде всего связана с исследованием глобальных (экзистенциальных, социальных) рисков, которые возникают или могут возникнуть в будущем в связи с использованием ИИ и призвана очертить границы допустимого и определить должное в данной сфере. Специфика этики ИИ определяется ее *публичностью* – участием в этических дискуссиях философов, юристов, журналистов, политиков, экономистов и других представителей общественности и *нормативностью* – формированием на основе общечеловеческих ценностей кодексов, рекомендаций, принципов, регулирующих сферу ИИ.

На сегодняшний день этика ИИ представлена двумя направлениями:

во-первых, это этические исследования, связанные с разработкой принципов и правил, которыми необходимо руководствоваться при проектировании, конструировании, использовании ИИ и оценкой моральных рисков применения ИИ (*этика создания и использования ИИ*); во-вторых, это этика морального поведения систем ИИ (*машинная этика*).

Этика создания и использования ИИ носит нормативный характер и предполагает закрепление этических принципов и стандартов в различных документах (кодексах, правилах, рекомендациях) как на уровне государства, так и на уровне крупных корпораций (например, принципы Google, Microsoft и др. в области ИИ), связанных с созданием ИИ, а также в конкретных сферах человеческой деятельности (например, руководство Всемирной организации здравоохранения о применении ИИ в медицине). Так, в российском Кодексе этики в сфере искусственного интеллекта (2021 г.) подчеркивается, что при развитии технологий ИИ *человек, его права и свободы должны рассматриваться как наивысшая ценность*: разработчики ИИ не должны допускать создание ИИ, который может угрожать автономии и свободе воли человека в принятии им решений; обязаны обеспечить справедливость и не допускать дискриминации отдельных лиц или групп лиц по признакам расовой, национальной, половой принадлежности, политических взглядов, религиозных убеждений, возраста, социального и экономического статуса; обязаны проводить оценку потенциальных рисков применения ИИ, включая социальные последствия для человека, общества и государства; не должны допускать использование технологий ИИ в целях причинения вреда жизни или здоровью человека, имуществу граждан, окружающей среде; должны информировать пользователей об их взаимодействии с ИИ; обеспечить конфиденциальность и защиту персональных данных, обработка которых осуществляется ИИ [1]. В Регламенте об искусственном интеллекте, принятом в 2024 г. Европейским парламентом, создание и использование ИИ регулируется на основе оценки риска для человека. Так, в данном документе к *недопустимому риску* относится ИИ, который предполагает биометрическую идентификацию и категоризацию людей, распознавание человеческих эмоций на рабочем месте или систему социального рейтинга (social scoring) и контроля. К *высокому риску* относятся системы ИИ, используемые для работы инфраструктуры (в частности, транспорта), если такое использование может поставить под угрозу жизнь и здоровье граждан; ИИ в образовании в случае, если он может повлиять на доступ к образованию, например, через оценку экзаменов; ИИ, применяемый в роботизированной хирургии. *Ограниченный риск* связан с несоблюдением требований прозрачности. ИИ. Например, при использовании чат-ботов необходимо сообщать, что человек общается с машиной. Контент, созданный ИИ, должен быть соответствующим образом маркирован как искусственно сгенерированный. К *минимальному (или отсутствующему)*

риску относятся, например, видеоигры с поддержкой ИИ или фильтры для спама. В целом, для этики создания и использования ИИ характерен человеко-ориентированный подход, согласно которому автономия человека, ценность его жизни, здоровья рассматриваются в качестве фундаментальных оснований этики ИИ, и риско-ориентированный подход, который определяет границы допустимого в сфере ИИ.

В этике ИИ также обсуждаются проблемы, связанные с разработкой этических компонентов и программированием этических решений и ограничений в системах ИИ [2]. Одними из первых тему машинной этики (робоэтики) обозначили писатели, работающие в жанре научной фантастики (например, три закона робототехники А. Азимова) и во многом определили ее дискуссионные вопросы: Может ли ИИ в полной мере считаться моральным агентом или стать таковым в будущем? Как возможна формализация этических норм при программировании систем ИИ? Какими этическими нормами и критериями необходимо руководствоваться в случае возникновения этических дилемм (ситуации выбора, в которой ни одно решение не является абсолютно правильным)? На сегодняшний день эти вопросы становятся все более актуальными. Так, перед разработчиками беспилотного транспорта остро строит проблема: какие этические правила закладывать в программу? В 2016 г. была запущена платформа Moral Machine (Режим доступа: https://www.moralmachine.net/hl/ru), разработанная Массачусетским технологическим институтом, где каждый желающий мог пройти тест из смоделированных ситуаций и выбрать, как стоит вести себя беспилотному автомобилю и кем предпочтительней пожертвовать, когда потери в результате аварии неизбежны (Например, в случае отказа тормозов беспилотного автомобиля, что предпочтительнее: сохранять курс (тогда погибнут двое пожилых людей, переходящих дорогу на зеленый свет) – или свернуть и врезаться в ограждение (погибнет один взрослый и ребенок, находящиеся в машине)?). Результаты исследования должны были помочь разработать *универсальную этику* беспилотных автомобилей и понять, должна ли она отличаться в зависимости от конкретного культурного региона. Анализ ответов показал, что у представителей разных культур есть общие представления – большинство жертвовали животными в пользу людей, преступниками – в пользу законопослушных граждан, а также стремились спасти больше жизней. Однако были выявлены и кросс-культурные различия в моральном выборе. Так, в странах с высоким уровнем коллективизма (восточные страны) испытуемые предпочитали спасать пожилых людей ценой жизни молодых. В странах Северной Америки и Европы чаще предпочитали не вмешиваться в действия машины и жертвовать пешеходами, а в странах Латинской Америки и Южной Африки чаще жертвовали пожилыми в пользу молодых, низкостатусными в пользу людей высокого социального

статуса [3].

Возможно ли создать универсальную машинную этику? Необходимо ли учитывать культурные предпочтения при программировании этических решений? Какой этический подход закладывать в конкретную систему ИИ? Кто будет нести ответственность за принятые ИИ решения в случае сложного морального выбора? Способен ли ИИ к развитию в нравственном плане? Какими этическими принципами необходимо руководствоваться, создавая ИИ? Каковы экзистенциальные риски использования ИИ? Поиск ответов на эти вопросы во многом задает проблемное поле этики ИИ, развитие которой невозможно без обращения к данным нейронаук, эмпирическим исследованиям в социологии и психологии ИИ, а также анализа сложных моральных ситуаций и коллизий, которые уже сегодня возникают в сфере ИИ.

Литература и источники

1. Кодекс этики в сфере ИИ [Электронный ресурс]. – Режим доступа: https://ethics.a-ai.ru. – Дата доступа: 11.09.2024.

2. Разин, А. В. Этика искусственного интеллекта / А. В. Разин // Философия и общество. – 2019. – Вып. 1 (90). – С. 57–73.

3. The Moral Machine experiment / E. Awad [et al.] // Nature – 24 October 2018 [Electronic resource]. – Mode of access: https://www.nature.com/ articles/ s41586–018–0637–6. – Date of access: 11.09.2024.

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