Савельева Анастасия

Этические вопросы искусственного интеллекта

«…над чем бы ученые ни работали,
у них все равно получается оружие»

К. Воннегут.

Этика относится к сфере философского знания, предметы которого – мораль и нравственность. При первом взгляде на это определение этики кажется, что не может быть связи современных технологий и социально-философских исследований. Соответственно, все доводы в сторону необходимости этики в искусственном интеллекте (ИИ), кажущиеся очевидными, касаются области научной фантастики и навеяны популярной культурой.

Однако, на самом деле, в современных реалиях вопросы этики ИИ стоят достаточно остро и даже не в перспективе. Сегодня ИИ решает множество практических задач, часто пользователи не задумываются, в каком окружении технологий находятся, потому что это уже привычный уклад жизни.

Существует более подходящее для понимания данного вопроса понятие этики – практика применения стандартов наилучшего поведения в различных областях деятельности. При этом важно также понимать, что такая практика и стандарты не могут быть в полной мере урегулированы правовыми вопросами относительно сферы человеческой деятельности. Философия предполагает, что право есть определенный минимум нравственности, поскольку нравственное требование предполагает свободное или добровольное исполнение, тогда как правовое, напротив, допускает прямое или косвенное принуждение [1]. Сформулированы и определения этики ИИ, одно из которых – проектирование машин, совершающих «поступки», которые, когда они совершены людьми, являются критериями обладания «этическим статусом» у этих людей [2]. Другое – определяет этику ИИ, как моральное поведение машин с ИИ [3]. Очевидны различные подходы к этике ИИ.

В любом случае, на данном этапе развития технологии, ИИ усиливает важность человеческой этики, поскольку на ней базируется – ИИ обучается на данных, которые ему предоставляет разработчик, который в свою очередь должен придерживаться определенных моральных и этических принципов при их отборе.

Примечательным является тот факт, что нормы морали для каждого человека разные, а особенностям этических норм в различных культурах посвящено немало исследований, что обуславливает новые барьеры в вопросах смежного характера – человеческой этики и этики ИИ. Ярким примером может служить исследование MIT, в основу которого легла известная классическая ультимативная дилемма – «проблема вагонетки». Ученые MIT разработали платформу для сбора человеческих мнений о нравственном выборе в различных сценариях, который человеку необходимо сделать «за» самоуправляемый автомобиль, находящийся в движении. Эта работа направлена не только на выявление наиболее распространенного мнения относительно моральных дилемм, но и на достижение более четкого понимания, как люди воспринимают машинный интеллект, который делает сложный выбор. Составленные картины мнений общественности зачастую

радикально различны [4]. На основе ответов исследователи разделили страны на несколько групп: «восточные» (в основном страны исламской и конфуцианской культур), «западные» (самая большая группа, включающая в том числе страны Европы) и «южные» (в основном страны Латинской Америки). Например, южные страны чаще всего «спасали» молодых людей, женщин или людей с более высоким статусом, особенно в сравнении с восточными – жители этих стран часто выбирали «спасение» для пожилых или бездействие.

«Зеркало общества» – с таким заголовком в СМИ была опубликована история бота-мизантропа компании Microsoft – твиттер чат-бота по имени Тай, который уже через один день после запуска перестал давать дружелюбные ответы на вопросы подписчиков, а затем и вовсе стал публиковать расистские высказывания и т.д. Проблема возникла по следующим причинам – разработчики в качестве одного из показателей эффективности бота задали «уровень вовлеченности пользователей», а как известно негативные комментарии не только преобладают в чатах, но и вызывают всплеск активности и дискуссий. Более того, часть «особенных» слов и сокращений была внесена разработчиками в базу данных бота сразу, часть – попала в ходе обучения, поскольку сами пользователи стали задавать боту неэтичные вопросы, на которых бот «дообучился».

Таким образом, очевидна важность информации, заложенной разработчиками в качестве основы для обучения. В довершение изложенного выше, возможно привести гипертрофированный пример из популярной культуры – персонажа комиксов и фильмов Альтрона, который изначально был создан как робот-ассистент с ИИ, но поскольку обучался на данных, находящихся в Интернет-сети, усвоил информацию о современном мире и истории человечества (о войнах, геноцидах, смертоносных событиях, агрессии, в том числе в постах и комментариях) и заключил, что глобальный мир на Земле возможен только в случае истребления человечества. Конечно, на сегодняшний день подобного, так называемого сильного ИИ, еще не создано, а обучение, в том числе и обучение с подкреплением (поход, предполагающий обучение в процессе работы системы с ИИ, а не на основе исходных данных) слабого ИИ нестабильно – при наличии явных успехов отдельных систем с ИИ в большинстве случаев нельзя утверждать о стабильно высоком уровне результатов работы всех систем с ИИ в настоящее время.

Даже при определенной общей нестабильности, качественно разработанный слабый ИИ показывает внушительные результаты. Характерным примером служат победы системы с ИИ AlphaGo в партиях игры ГО, которая считается наиболее сложной для ИИ из-за большого размера доски, фактически бесконечного количества ходов и трудностей в их прогнозировании [5]. Также примечательны успехи веб-приложения Cleverbot и чат-бота «Женя Густман», прошедших тест Тьюринга [6].

Очевидно, вопрос этики в ИИ неразрывно связан с ответственным использованием данных. Важными являются не только исходные и «подкрепляемые» данные, на которых ИИ обучается, но и те результаты и решения, которые выдаются на выходе систем, а также каким образом эти данные будут использоваться в дальнейшем. Так существуют документальные подтверждения, что медицинский ИИ от IBM – Watson совершает врачебные ошибки [7]. Онкологи раскритиковали некорректные методы лечения, которые алгоритм предлагает пациентам. В частности, ИИ предложил одному из пациентов с установленным диагнозом рак легкого с кровотечением назначение полностью

соответствующие протоколу лечения по диагнозу, однако неприменимое в случае осложнения в виде кровотечения. В свою очередь IBM объявили, что описанные случаи – просто часть процесса обучения. Отличительным является тот факт, что ошибки были выявлены пользователями, непосредственно применявшими данную систему. Чаще всего выработанное безусловное доверие к компьютерным системам и их возможностям препятствует оспариванию решений ИИ, что является особенно актуальной проблемой, связанной в том числе с вопросами биомедицинской этики как в данном примере.

Другим примером в доказательство тезиса о нестабильности результатов работы ИИ – применение систем распознавания лиц. В Китае распространена подобная система, результаты деятельности которой позволяют штрафовать за переход дороги в неположенном месте, подтверждать банковские операции и т.д. Зафиксированы случаи сбоев, когда система по ошибке распознавала лица на рекламных плакатах, а также не идеальность системы в области безопасности, система уязвима для взломов и мошенничества [8, 9]. Спорными относительно этичности являются приложения подобные Gradient, которые в качестве развлекательной функции «определяют» национальность по фотографии пользователя.

Относительно технологии распознавания лиц стоит также упомянуть вопрос защиты данных. В частности, во Франции предлагаемый к внедрению на платформы госуслуг сервис Alicem, предоставляющий удаленный доступ к услугам с помощью распознавания лиц, столкнулся с судебными исками активистов, поскольку такая система нарушает европейский регламент по защите данных, который запрещает корпорациям и государствам собирать конфиденциальную информацию о гражданах без их разрешения [10]. Пользователям такого сервиса не предоставляется возможность отказаться от сбора биометрических данных, а, следовательно, отказывается в удаленном доступе к услугам. Власти ряда американских городов также запретили применение технологии распознавания лиц.

Противоречивой является и технология Deepfake (дипфейк), синтезирующая изображение на основе ИИ. Существуют конструктивные варианты использования данной технологии, однако споры вызывают решения, которые относятся не только к такой киберугрозе, как подмена цифровой личности, но и в целом к этическому вопросу такой реализации – возмущает манипулятивность технологии и искажение реальности. Созданные бельгийским специалистом по визуальным эффектам и опубликованные на платформе TikTok видео с Томом Крузом (на которые актер не давал своего согласия) отличаются высокой реалистичностью [11].

Итак, можно выделить пять больших групп этических вопросов относительно ИИ: прозрачность (понимание алгоритмов, этика исходных данных, автономия), внешнее управление (элементы и операторы), доверие (безопасность, риски, ответственность), индикаторы (метрики, показатели эффективности), цели (оправданность применения). Остро стоит сквозной вопрос – кто может принимать решение по всем приведенным аспектам в момент возникновения этической дилеммы.

Женевьева Белл – австралийский антрополог, наиболее известная своей работой на стыке культурной практики и технологического развития, является одним из агитаторов инвестирования в обсуждение и поиск решений в области этики, морали, культурных и

философских базисов в новом цифровом мире, однако вопрос создания единой платформы для обсуждения остается открытым [12]. Одно остается очевидным – политика и соответствующие нормативные документы должны быть разработаны. Безусловно, вопросы регулирования, надзора, подотчётности и открытости являются труднореализуемыми.

В качестве подхода к решению проблем ошибок при принятии решений (и смежной проблемы ответственности за действия ИИ), алгоритмической предвзятости, угроз нарушения конфиденциальности и приватности, а также риска внешнего вмешательства, в рамках настоящего эссе возможно предложить следующие направления:

- Обеспечение валидации, сертификации и регулирования прозрачности систем с ИИ
- Исключение предвзятости и предубеждений в исходных данных, на которых ИИ обучается – создание процессов и указаний для их проверки
- Социально-гуманитарная экспертиза этики полного цикла технологии ИИ (применение ИИ должно быть оправданным)
- Создание стандартов на разработку этически ориентированных систем с ИИ (должен быть сформирован уровень допустимой обработки метаданных, приоритет – обезличенные данные, для исключения предвзятости системы и любого вида дискриминации)
- Повышение квалификации операторов систем с ИИ
- Развитие систем информационной безопасности

Для реализации указанных решений возможно признать деятельность в области создания систем с ИИ лицензируемым видом деятельности, причем такая лицензия может выдаваться уполномоченным органом ФОИВ в сфере ИКТ в соответствии с национальным законодательством.

Для исключения предвзятости и предубеждений в исходных данных разработка и использование ИС должны вестись на недискриминационной основе, исключающей ограничения по национальному, половому, возрастному, социальному признакам, с полным соблюдением конституционных прав и свобод личности, общепризнанных этических и социальных норм.

Для обеспечения контроля решений системы с ИИ необходимо участие оператора, а к осуществлению эксплуатации систем стоит допускать персонал прошедший повышение квалификации в соответствии со стандартами.

Относительно этики полного цикла технологии ИИ стоит выделить этапы такого цикла: этика высших принципов функционирования (например, ответственности и предосторожности), этика подготовки и деятельности специалистов-разработчиков, решение этических проблем внедрения в отдельных секторах. Социально-гуманитарная, в частности – этическая, экспертиза как исследовательская практика может занять важную роль в решении этических проблем систем ИИ при сопровождении сложных инновационных проектов. С помощью такой экспертизы возможно исследовать целевые и ценностные установки проекта, проанализировать этичность средств достижения целей, а также определить возможные последствия и риски. Так возможно утвердить, что использование ИИ должно допускаться исключительно после проведения социально-

гуманитарной экспертизы этики полного цикла ИИ в такого рода проектах. При этом экспертиза может осуществляться в разных форматах в зависимости от масштабов проектов и сферы применения, это могут быть – комитеты и комиссии, группы и департаменты внутри проекта, а также индивидуальные эксперты.

В части повышения безопасности использования ИИ необходимо совершенствование регуляторной базы в области информационной безопасности и обработки персональных данных. Однако, стоит отметить, что приоритетно применение обезличенных данных в целях формирования комфортной городской среды и улучшения благосостояния общества.

Деятельность по регулированию как технологии ИИ, так и этических аспектов, повсеместно признана одной из приоритетных. В 2020 году Ватикан представил проект доктрины «Об этике искусственного интеллекта», в которой выработаны фундаментальные элементы полезных инноваций, предложены принципы реализации систем с ИИ [13]. С 2020 года ЮНЕСКО также занимается разработкой рекомендации об этических аспектах ИИ [14].

В 2021 году Еврокоммисия представила комплексный законопроект, посвященный детальному регулированию систем ИИ [15]. В документе разделяются технологии ИИ по степени риска для безопасности, жизни и прав граждан. Возможно заключить, что положения законопроекта руководствуются двумя важными этическими принципами – принципом ответственности и принципом предосторожности.

В заключении хотелось бы отметить, что в современной науке исчезает разграничение опасности и безопасности внутри исследовательского процесса. Цитата, размещенная в эпиграфе к настоящему эссе, заставляет задуматься о деятельности ученных, которая в своей основе направлена на достижение научно-технологического прогресса, но результаты такой деятельности могут использоваться в решениях совершенно далеких от этических норм. Так в открытом письме [16], подписанным более чем 110 представителями различных компаний-разработчиков в области ИИ и робототехники (в том числе Илоном Маском), признается особая ответственность за деятельность, а также содержится призыв – упорно трудиться над поиском средств предотвращения гонки вооружений в отношении оружия, основанном на технологии ИИ, защиты гражданского населения от неправомерного использования и предотвращения дестабилизирующего воздействия таких технологий.

[1] Соловьев В.С. Оправдание добра. Нравственная философия: Сочинения в 2 т. Т1: Мысль, 1990, С.450, С.453.

[2] Steve Torrance. Artificial agents and the expanding ethical circle. AI & SOCIETY. № 28 (4), 2012

[3] James Moor. What Is Computer Ethics? Metaphilosophy. № 16 (4). С. 266 – 275, 2007

[4] Портал проекта MIT «Moral Machine». URL: https://www.moralmachine.net/hl/ru

[5] ТАСС. Новостная статья. Программа AlphaGo обыграла признанного чемпиона. URL: https://nauka.tass.ru/sci/6822187

[6] РИА Новости. Новостная статья. Российская компьютерная программа впервые прошла тест Тьюринга. URL: https://ria.ru/20140609/1011250142.html

[7] Jennings Brown. IBM Watson Reportedly Recommended Cancer Treatments That Were 'Unsafe and Incorrect'. URL: https://gizmodo.com/ibm-watson-reportedly-recommended-cancer-treatments-tha-1827868882

[8] Известия. Новостная статья. В Китае система распознавания лиц выписала штраф фотографии на автобусе. URL: https://iz.ru/815961/video/v-kitae-sistema-raspoznavaniia-litc-vypisala-shtraf-fotografii-na-avtobuse

[9] Первый канал. Новостная статья. Взломана китайская система распознавания лиц. URL: https://1tv.ru/s/news/2021-04-07/404459-vzlomana_znamenitaya_kitayskaya_sistema_raspoznavaniya_lits_zayavlennaya_kak_nepristupnaya

[10] Deutsche Welle. Новостная статья. Лиза Луи, Александра Елкина. Система распознавания лиц в приложении Alicem. URL: https://p.dw.com/p/3SXju

[11] Esguire. Новостная статья. Автор роликов с Томом Крузом создал дипфейк-студию. URL: https://esquire.ru/articles/280473-pomnite-bezumno-realistichnye-dipfeyki-s-tomom-kruzom-v-tiktok-avtor-etih-rolikov-sozdal-sobstvennuyu-studiyu-dipfeyk-tehnologiy/#part1

[12] The Guardian. Экспертное мнение. Genevieve Bell. In our focus on the digital, have we lost our sense of what being human means? URL: https://www.theguardian.com/commentisfree/2017/oct/24/in-our-focus-on-the-digital-have-we-lost-our-sense-of-what-being-human-means

[13] Портал Папской академии жизни. Новостная статья. Rome Call for AI Ethics. URL: http://www.academyforlife.va/content/pav/en/events/intelligenza-artificiale.html

[14] Портал ЮНЭСКО. Доклад. Разработка рекомендации об этических аспектах искусственного интеллекта. URL: https://ru.unesco.org/artificial-intelligence/ethics

[15] Предложение в Регламент Европейского парламента и Совета. Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts - 2021. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN

[16] Открытое письмо. An Open Letter to the United Nations Convention on Certain Conventional Weapons. URL: https://www.cse.unsw.edu.au/~tw/ciair/open.pdf