ДИСКУРС. 2022. Т. 8, № 6. С. 17–30
DISCOURSE. 2022, vol. 8, no. 6, pp. 17–30

Оригинальная статья
УДК 172:004.8
http://doi.org/10.32603/2412-8562-2022-8-6-17-30

Искусственный интеллект: в поисках формализации этических оснований

***Раиса Ильинична Мамина*[1]✉*, Анна Валерьевна Ильина*[2]**

[1, 2]*Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В. И. Ульянова (Ленина), Санкт-Петербург, Россия*

[1]*maminaraisa@yandex.ru, https://orcid.org/0000-0003-3301-636X*
[2]*a.ilyina2045@gmail.com, https://orcid.org/0000-0003-3002-1841*

Введение. В цифровую эпоху роль искусственного интеллекта (ИИ) в жизни социума во многом определяется технологическим прорывом, который произошел в области прикладного ИИ за последние два десятилетия. Однако инновации в области ИИ обозначили не только новые возможности для человека и общества в целом, но и целый ряд проблем, прежде всего социально-этического характера. В частности, проблема «moral machines» определяется мировым научным сообществом как одна из приоритетных. В статье рассматриваются вопросы создания этических инструментов, регулирующих коммуникацию человека с ИИ, достигнутые результаты, тенденции развития данной проблематики.

Методология и источники. В статье применяется методология культур-философского, аксиологического и междисциплинарного подходов. В качестве источников использованы научные исследования отечественных и зарубежных авторов, документы, публикации и сайты, посвященные современному состоянию ИИ и задачам, которые стоят перед специалистами в области этики ИИ.

Результаты и обсуждение. Актуальность проблем этики ИИ определила необходимость систематизации базовых этических понятий с целью привнесения этичности в технологию прикладного ИИ. Показано, что сегодня накоплена определенная нормативно-этическая база в области этики ИИ, однако проблема формализации этических норм остается все еще не решенной. Сложность в том, что этичность ИИ носит не смысловой, а формализованный характер.

Заключение. В настоящее время процесс формирования нормативно-этической базы ИИ находится в фокусе повышенного внимания науки и индустрии ИИ. Однако, чтобы коммуникация «человек – машина» развивалась с позиций человеко-ориентированной цифровой эпохи, решение проблем этичности ИИ дополняется созданием новых профессий в этой области и новыми требованиями к образовательным траекториям (в частности, к подготовке инженеров и разработчиков ИИ-систем), соответствующими специфике новых реалий.

Ключевые слова: цифровые технологии, прикладной ИИ, общий ИИ, этика ИИ, ценности, этические принципы, цифровая этика, коммуникация «человек – машина», новые ИИ-профессии, цифровая гуманитаристика

Для цитирования: Мамина Р. И., Ильина А. В. Искусственный интеллект: в поисках формализации этических оснований // ДИСКУРС. 2022. Т. 8, № 6. С. 17–30. DOI: 10.32603/2412-8562-2022-8-6-17-30.

ДИСКУРС. 2022. Т. 8, № 6. С. 17–30
DISCOURSE. 2022, vol. 8, no. 6, pp. 17–30

Original paper

Artificial Intelligence: in Search for Formalization of Ethical Foundations

***Raisa I. Mamina*[1]✉*, Anna V. Ilina*[2]**

[1, 2]*Saint Petersburg Electrotechnical University, St Petersburg, Russia*

[1]*maminaraisa@yandex.ru, https://orcid.org/0000-0003-3301-636X*
[2]*a.ilyina2045@gmail.com, https://orcid.org/0000-0003-3002-1841*

Introduction. In digital age the role of artificial intelligence (AI) in society is largely determined by the technological breakthrough in the field of applied AI over the past two decades. However the implementation of AI innovations not only opens up new opportunities for individuals and society as a whole, but also raises a number of problems, primarily of a socio-ethical focus. In particular, the scientific community considers the problem of "moral machines" to be of high research priority. The article deals with the problems of ethical regulation of AI in human-machine communication, latest research results, and trends in this field.

Methodology and sources. The article is based on methodology of cultural-philosophical, axiological and interdisciplinary approaches. There were also used the following sources: scientific research of Russian and foreign authors, documents, publications and websites dedicated to the current state of AI and the tasks to be solved by specialists in the field of AI ethics.

Results and discussion. The urgency of issues in the field of AI ethics determines the need for systematization of basic ethical concepts in order to integrate ethics into applied AI. It is argued that despite the accumulation of regulatory basis in the field of AI ethics, the problem of conclusive formalization of ethical norms in this field is still unresolved. The main difficulty with the aforementioned norms lies in the fact that ethics of AI is more dependent upon formalization than upon semantics.

Conclusion. Currently, the process of establishing a regulatory framework for ethics of AI is actively discussed in industry and science. However, if we want the human-machine communication to start its development from the standpoint of a human-centric digital age, it is important not only to solve ethical problems, but also to create new professions in the field of AI ethics, as well as to introduce new approaches towards the training of engineers and developers of AI systems that meet demands of the time.

Keywords: digital technologies, applied AI, general AI, AI ethics, values, ethical principles, digital ethics, human-machine communication, new AI professions, digital humanities

For citation: Mamina, R.I. and Ilina, A.V. (2022), "Artificial Intelligence: in Search for Formalization of Ethical Foundations", *DISCOURSE*, vol. 8, no. 6, pp. 17–30. DOI: 10.32603/2412-8562-2022-8-6-17–30 (Russia).

Введение. В условиях цифровых трансформаций технологический прогресс – это не только новые возможности, но и новые вызовы, породившие проблемы социального, психологического, этического и правового характера. При этом проблемы этики искусственного интеллекта (ИИ) определяются специалистами как ключевые проблемы современности. Главная сложность заключается в том, что этичность ИИ носит не смысловой, не ценностный, а формализованный характер. Отсюда необходимость перевода этических норм в нравственные стандарты ИИ. Решение этой проблемы, как подчеркивают специалисты

НИЦ «Курчатовский институт», включает в себя две основные задачи – создание форм представлений норм и выбор соответствующего математического аппарата для работы с этими формами [1, с. 94]. Как следствие, проблема этичности ИИ стала общим делом гуманитариев и инженерного сообщества. В рамках представленной статьи показано, что осмысление и решение проблемы формализации этики ИИ с позиций гуманитаризации технических знаний включила в себя разработку нормативно-этической базы ИИ-систем, а также формирование новых ИИ-профессий этической направленности и, соответственно, новых образовательных подходов к подготовке инженеров и разработчиков ИИ-систем, соответствующих специфике новых реалий.

Методология и источники. Представленные в работе выводы опираются на методологию культур-философского, аксиологического и междисциплинарного подходов. Проведенный анализ основывается на специальной литературе, в частности: коллективное монографическое исследование ведущих отечественных специалистов, занимающихся проблематикой ИИ «Сильный искусственный интеллект» (2021); концепция П. Доэрти и Дж. Уилсона, представленная в книге «Человек + машина. Новые принципы работы в эпоху искусственного интеллекта» (2019), аналитические доклады, разработанные под эгидой российского Центра подготовки руководителей цифровой трансформации «Этика и “цифра”: этические проблемы и цифровые технологии» (2020) и «Этика и “цифра”: от проблем к решениям» (2021), документы, научные публикации и сайты, посвященные современному состоянию ИИ и задачам, которые стоят перед специалистами в области этики ИИ.

Результаты и обсуждение. История современного ИИ начинается с 2010 г., когда мощность компьютеров позволила сочетать технологию больших данных (Big Data) с методами глубокого обучения (Deep Learning), которые основываются на использовании искусственных нейронных сетей. Следствием этого стали не только новые удивительные возможности прикладного ИИ, но и проблемы, с которыми столкнулись и человек, и общество в целом.

Прикладной ИИ и его проблематика. Прикладной, или слабый ИИ (Narrow AI) – это искусственный интеллект, предназначенный для решения какой-либо одной интеллектуальной задачи или их определенного множества. С его помощью решаются в основном прикладные задачи, поскольку, несмотря на использование нейронных сетей при создании ИИ-систем, речь идет о весьма упрощенном аналоге естественных нейронных сетей [2]. Слабому ИИ традиционно противопоставляется гипотетический ИИ, который американский философ Дж. Серл определил, как сильный ИИ (Strong AI).

Принято считать, что сильный ИИ будет обладать способностью мыслить, как человек, но принимать решения без участия человека. Сегодня сильный ИИ называют гипотетическим или суперинтеллектом (Artificial Super Intelligence, ASI), чтобы подчеркнуть, что он превосходит уровень интеллекта отдельного человека. Однако, поскольку до сих пор еще нет окончательного понимания, как работает естественный интеллект (ЕИ), в данное время сильный ИИ – это только абстрактное понятие, отражающее наши гипотетические представления о том, каким будет ИИ в будущем. Одна из целевых установок научного сообщества в плане дальнейшего развития проблематики ASI – это воспроизведение семантических ассоциаций между объектами и действиями в искусственных системах [2], что в реалиях Web 3.0 труднодостижимо. Предполагается, что решение этой задачи будет возможно только в условиях четвертого поколения Сети – Web 4.0 [3].

ДИСКУРС. 2022. Т. 8, № 6. С. 17–30
DISCOURSE. 2022, vol. 8, no. 6, pp. 17–30

В настоящих условиях узким методам прикладного ИИ противопоставляют не сильный ИИ, а общий ИИ (Artificial General Intelligence, AGI). В частности, монография «Сильный интеллект: на подступах к сверхразуму» (2021) посвящена научным подходам к созданию полноценного сильного ИИ и потенциалу его применения. Особое внимание авторы монографии уделяют общему ИИ, который рассматривается как альтернатива узким методам ИИ и определяется без отсылки к человеческому интеллекту, как это принято в случае с сильным ИИ. «Идея общего ИИ, – пишут авторы монографии, – предполагает, что компьютеры смогут самостоятельно решать как новые узкие, так и сложные задачи, чем будут заметно отличаться от критикуемых систем ИИ» [4, с. 33]. Однако общий ИИ находится только в процессе научно-исследовательских разработок, поэтому в центре внимания специалистов остается прикладной ИИ, его возможности, его проблематика.

Применение прикладного ИИ практически во всех сферах жизнедеятельности человека при всех его удивительных возможностях породило целый ряд социальных и социально-психологических проблем, в частности – это вытеснение человека машиной из производственных процессов, исчезновение целого ряда профессий, сокращение рабочих мест, безработица; разрыв между доходами от капитала и доходами от труда в результате развития и активного применения ИИ в бизнес-процессах и др. Однако особую значимость, как подчеркивают аналитики, составляют проблемы ИИ этического и правового характера. Одной из них стало ограничение прав и свобод личности на рабочем месте, которое приняло такие формы, как цифровая экономическая экспансия, цифровая диктатура, цифровое неравенство и дискриминация, правовой нигилизм в цифровом формате.

Например, такое явление, как цифровая финансовая экспансия, детерминировано ничем иным, как потребительским комфортом предоставляемых клиентам финансовых и банковских услуг. Однако за этим комфортом за счет колоссальных возможностей технологий Big Data стоит цифровой профиль клиента. На основе получаемых сведений ИИ-системы способны составить полный аналитический портрет клиента с точки зрения рисков, структуры потребления, доходов и предпочтений. Одно из следствий – цифровая дискриминация, которая проявляется, в частности, как ограничения при приеме на работу по половому, возрастному, расовому и другим признакам. В последнее время наиболее распространенными стали ограничения по социальному статусу и доходам при выдаче банками кредитов и погашении займов и др. В итоге, такая «необъективность ИИ»/«пристрастность ИИ» получила название AI bias [5].

Повышенный интерес научного сообщества к AI bias объясняется и тем, что результаты внедрения технологий ИИ все чаще наступают на основные ценности и ценностные ориентиры современного социума. Анализ феномена AI bias позволил исследователям утверждать, что главная причина некорректности ИИ – человеческий фактор, поскольку даже самые усовершенствованные программы созданы человеком и зависят от человека, его знаний, его ценностей и нравственных ориентиров, которые он вкладывает в технологические системы. Разработка базовых этических понятий, способных привнести этичность в саму технологию, стала в настоящих условиях одной из главных тем современности.

Этика ИИ: мировые стратегии. В обсуждении этического аспекта взаимодействия «человек – машина» приняли участие, прежде всего, ученые и практики с мировым именем

ДИСКУРС. 2022. Т. 8, № 6. С. 17–30
DISCOURSE. 2022, vol. 8, no. 6, pp. 17–30

(Н. Бостром, Р. Курцвейл, Г. Леонгард, Ст. Хокинг и др.), а также крупные международные корпорации (Apple, Google, Microsoft и др.). Одним из главных результатов данного дискурса стало появление ряда этических кодексов и корпоративных документов, регламентирующих коммуникацию человека с ИИ, в частности, это «Пять законов робототехники» (Google, 2016), «Десять законов для людей и ИИ» (CEO Microsoft Сати Наделлы, 2016) и др.

Однако первым международным документом нормативно-этического характера, направленным на решение проблемы регулирования отношений «человек – машина» стали «23 Азиломарских принципа искусственного интеллекта», принятые и опубликованные по итогам конференции разработчиков и исследователей в сфере ИИ (Азиломар, Калифорния, США, 2017). В частности, в разделе «Этика и ценности» среди выделенных принципов, таких как «Человеческие ценности», «Свобода и конфиденциальность», «Контроль ИИ человеком» и другие, особого внимания с позиций этики ИИ заслуживает принцип «Синхронизация ценностей», который отражает главный этический ИИ-норматив: «Системы ИИ с высокой степенью автономности должны быть разработаны таким образом, чтобы их цели и поведение были согласованы с человеческими ценностями на всем протяжении работы» [6]. В целом, Азиломарский документ призывает направить свои усилия на создание управляемого, надежного и полезного ИИ, отказаться от гонки вооружений на основе ИИ, а также подумать о безопасности разработок в области ИИ и об ответственности самих разработчиков.

Актуальность разработок в области этики ИИ в целях формализации этических норм привела к появлению и ряда других документов, среди которых особенно выделяются «Рекомендации для этически обоснованного проектирования. Концепция взаимодействия людей с искусственными интеллектом и автономными системами с приоритетом человеческих ценностей» («Глобальная инициатива» IEEE, 2017) [7]. Документ был представлен Институтом инженеров в области электротехники и электроники (IEEE), одним из основных мировых разработчиков стандартов. По оценкам специалистов, создание данного документа послужило хорошим примером того, как социальные и этические проблемы экспоненциальных технологий выносятся на общепланетарный уровень обсуждения.

Следующим важным шагом в направлении этики ИИ на международном уровне стал двухлетний проект разработки первого глобального нормативного документа об этических аспектах искусственного интеллекта, к которому приступила ЮНЕСКО в 2019 г. Главная идея – охватить все области, которые определяют развитие и применение ИИ в рамках подхода, ориентированного на человека, с целью уменьшение рисков и трудностей, связанных с ИИ, особенно с точки зрения усугубления существующего неравенства, а также последствий для прав человека. Итоговый документ «Рекомендация об этических аспектах искусственного интеллекта» был утвержден 16 ноября 2021 г. в Париже в ходе 41-й сессии Генеральной конференции ЮНЕСКО. 193 страны, которые дали мандат для его разработки, пришли к определенному согласию относительно целей, установок и принципов этики ИИ и приняли данный исторический документ – первое глобальное соглашение по этике ИИ. По оценкам аналитиков, несмотря на ряд критических замечаний, в документе собраны практически все возможные размышления и идеи о регулировании технологий ИИ, определяется практический характер этики ИИ. В частности, в документе утверждается, что «этические принципы выступают в качестве гибкой основы для нормативной оценки, а также методи-

ческого руководства в вопросах применения технологий на основе ИИ». При этом отдельно подчеркивается, что «человеческое достоинство, благополучие человека и недопущение нанесения вреда рассматриваются как целевой ориентир, уходящий корнями в этику науки и технологии» [8].

В рамках международного сотрудничества по вопросам этики ИИ особо выделяется и такая структура, как Глобальное партнерство по ИИ (Global Partnership on Artificial Intelligence, GPAI), основанное в июне 2020 г. странами Евросоюза и еще 14 государствами (Австралия, Великобритания, Канада, Мексика, Республика Корея и др.). Глобальное партнерство позиционируется как мультистейкхолдерная международная инициатива по разработке и использованию ИИ, функционирующего на базе таких ценностей, как инклюзивность, инновации, многообразие, соблюдение прав человека, стремление к экономическому росту» [9].

Еще одним важным направлением международных стратегий в этой области является проведение тематических конференций. К крупнейшим из них относят такие, как «AI Journey», «The International Conference on Artificial Intelligence (ICOAI)», «IEEE Conference» и др. В целом, как перечисленные, так и другие, принятые международным сообществом документы, посвященные этике ИИ, деятельность GPAI и других институциональных структур, включая международные научно-исследовательские центры и лаборатории по этике ИИ, проведение специализированных конференций, форумов и других мероприятий во многом способствуют активному обсуждению этической проблематики ИИ на глобальном уровне.

Этика ИИ: национальные стратегии. В настоящее время национальные документы стратегического развития этики ИИ имеются в большинстве стран – Великобритании, Канаде, Китае, Корее, России, США, Франции и др. По результатам исследования Microsoft и РАЭК, проведенного в конце 2020 г., национальными стратегиями развития ИИ располагают более 30 стран, еще порядка 20 занимаются их разработкой. Как правило, в этих документах содержится описание подходов к развитию технологий ИИ, сохраняющих автономию и свободу воли человека, а также ориентацию на потенциальные риски применения систем ИИ. Все кодексы этики систем ИИ (КЭСИИ), как правило, постулируют, что финальные решения принимает человек, он же несет ответственность за негативные последствия [10].

Сегодня в числе наиболее обсуждаемых национальных документов по этике ИИ, имеющих в том числе и международное значение, называют китайский закон о конфиденциальности персональных данных, вступивший в силу 1 ноября 2021 г. [11] и «Билль этичности искусственного интеллекта», который был опубликован 4 октября 2022 г. на официальном сайте Белого дома [12]. Билль описывает 5 основных принципов формирования ИИ-систем, защищающих конечных пользователей от возможной дискриминации и злоупотребления персональными данными со стороны ИИ. Аналитики отмечают, что данный документ является в некотором роде повторением закона, принятого в Китае, поскольку так же, как и китайский вариант, направлен на регулирование процесса взаимодействия крупных IT-корпораций и пользователей их продуктов.

В рамках отечественных стратегий речь, прежде всего, идет о таком официальном документе, как «Национальные стратегии развития искусственного интеллекта на период до

ДИСКУРС. 2022. Т. 8, № 6. С. 17–30
DISCOURSE. 2022, vol. 8, no. 6, pp. 17–30

2030 года» (Москва, 2019) [13], а также «Национальном кодексе этики ИИ» (Москва, 2021) [14], который построен исключительно на добровольной основе. По оценке аналитиков, оба документа в целом совпадают с Рекомендациями ЮНЕСКО. Однако есть и отличия, например, российский КЭСИИ представляет собой, прежде всего, практически применимый документ, поскольку содержит не только общие принципы, но и конкретные положения, базирующиеся на человекоориентированном и рискоориентированном подходах к пониманию перспектив развития ИИ.

Среди целого ряда других отечественных достижений в плане нацстратегий в вопросах этики ИИ следует также выделить аналитические доклады, разработанные под эгидой Центра подготовки руководителей цифровой трансформации: «Этика и “цифра”: этические проблемы и цифровые технологии» (2020) [15] и «Этика и “цифра”: от проблем к решениям» (2021) [16]. Основная цель первого доклада – обозначить этические проблемы, которые имеют большое значение для цифрового общества, а также представить подходы для их решения. В связи с тем, что цифровая этика только формируется как в России, так и в мире, в докладе речь идет только о подходах, а не о готовых решениях. Во втором докладе происходит переход к поиску вариантов для решения этических дилемм, а также изучению роли этики для цифровой трансформации социума.

Весомый вклад в развитие отечественных национальных стратегий в области этики ИИ вносят научно-исследовательские центры (НИЦ). Например, среди центральных направлений одного из крупнейших научных центров России НИЦ «Курчатовский институт» – создание нового поколения ИИ, в котором используются биоподобные технологии, основанные на принципах импульсных архитектур [2], что в будущем, хотя и весьма отдаленном, предполагает возможность самообучения и эмоционально-осмысленного поведения ИИ. Современные ИИ-системы – это не самообучаемые системы, для их обучения, как уже отмечалось, используются специально подготовленные данные. Сегодня к этим данным должны добавиться «нравственные стандарты» ИИ, в разработке которых активное участие принимают ученые НИЦ [1]. Большие надежды, включая новые решения в области этики ИИ, возлагаются и на новый Национальный центр развития искусственного интеллекта, открытый по инициативе Правительства страны на площадке Высшей школы экономики (НИУ ВШЭ).

Таковы некоторые основные достижения в создании нормативно-этической базы этики ИИ, которая предполагает «мягкое» регулирование проблемных зон этического характера и находится в фокусе повышенного внимания научного сообщества на международном и национальном уровнях. Вектор этой направленности, как подчеркивают специалисты, позволяет говорить о том, что мы входим в новую фазу наставничества по отношению к ИИ, которое рассматривает развитие и применение ИИ в рамках подхода, ориентированного на человека, его безопасность, достоинство и свободу. В качестве одной из важных форм такого нового наставничества можно рассматривать и создание новых профессий в области этики ИИ.

Этика ИИ: новые профессии. Разработка этических стандартов напрямую связана с созданием новых профессий в области этики ИИ. В контексте зарубежных и отечественных разработок в этой сфере высокую оценку специалистов получила концепция, разработанная руководителями компании Accenture П. Доэрти и Дж. Уилсоном и представленная в их

совместной книге «Человек + машина. Новые принципы работы в эпоху искусственного интеллекта» [17].

Специфика концепции заключается в том, что, авторы книги, помимо видов деятельности, доступных только человеку, и видов деятельности, доступных только машине, выделяют и анализируют смешанные виды деятельности человека и машины, которую П. Доэрти и Дж. Уилсон называют «недостающей серединой». При этом они объясняют: «недостающей», потому что «практически никто не говорит и лишь немногие работают над тем, чтобы заполнить эту лакуну» [17, с. 36]. В рамках «недостающей середины» люди разрабатывают, обучают ИИ-приложения и управляют ими, в свою очередь машины расширяют возможности человека и сотрудничают с ним, что позволяет прийти к результатам, которые раньше считались недостижимыми. Основная идея смены парадигмы ИИ, согласно П. Доэрти и Дж. Уилсону, заключается в трансформации управления персоналом и бизнес-процессами компании, базирующейся на понимании того, что люди и машины – это партнеры, а не противоборствующие стороны. В эпоху искусственного интеллекта, как особо подчеркивают авторы, только такой подход, где человек и машина – команда, гарантирует компаниям лидерство в своей отрасли, а значит и процветание.

В этой связи отмечается, что системы ИИ высокого уровня сложности требуют привлечения специалистов в области бизнеса и технологий, которые должны заниматься *обучением, разъяснением* и *обеспечением устойчивости* систем ИИ. При этом в рамках каждого из указанных направлений речь идет о совершенно новых профессиях и функциях, требующих новых этически обусловленных навыков, в которых еще никогда не возникала потребность. В частности, в рамках направления «обучение» – речь идет об обучении ИИ умению взаимодействовать с людьми и этичности поведения, отсюда потребность в таких новых профессиях, как специалист по обучению эмпатии, специалист по обучению личностным качествам, специалист по обучению мировоззрению и локализации и др. [17, с. 151–159].

Так, например, специалист по обучению эмпатии, согласно авторам концепции, – это человек, который учит системы ИИ демонстрировать сочувствие во взаимоотношениях с людьми. Главная цель – добиться того, чтобы система обсуждала с человеком проблему или сложную ситуацию, в которую он попал, и проявляла сочувствие, сострадание или даже юмор, т. е. необходимую эмоциональную поддержку. В качестве конкретного примера такого поведения Доэрти и Уилсон приводят стартап Koko, который разработал систему машинного обучения, призванную помочь таким чат-ботам, как Siri компании Apple и Alexa компании Amazon реагировать на вопросы пользователей с сочувствием и пониманием. Предполагается, что со временем благодаря потенциалу, заложенному в современных системах ИИ, данный алгоритм Koko поможет ботам Siri и Alexa оказывать серьезную эмоциональную поддержку людям, которые в ней нуждаются [17, с. 153–154].

Помимо новых этически обусловленных ИИ-профессий, применительно к смешанным видам деятельности П. Доэрти и Дж. Уилсон вводят также понятие «интегрированные компетенции» и предлагают 8 таких компетенций. Среди них особое внимание вызывают компетенции *взаимное обучение* и *неустанное переосмысление.* Например, суть навыка «взаимное обучение» заключается в том, что «в эпоху слияния человека и машины: одна из самых важных характеристик, будь то человек или машины – не просто обладать необходимыми

ДИСКУРС. 2022. Т. 8, № 6. С. 17–30
DISCOURSE. 2022, vol. 8, no. 6, pp. 17–30

навыками, а уметь учиться» [17, с. 254–255]. «Неустанное переосмысление» рассматривается как базовый навык, который служит основой для всех других: «Именно способность к переосмыслению позволяет людям легче адаптироваться к меняющему миру, в котором передовые технологии и искусственный интеллект непрерывно преобразуют рабочие процессы, бизнес-модели и целые отрасли» [17, с. 257].

Однако следует отметить, что концепция П. Доэрти и Дж. Уилсона имеет прогностический характер, поэтому открывает широкие возможности для обсуждения, рекомендаций, предложений, замечаний в целях ее совершенствования.

В рамках такого обсуждения обратимся еще раз к содержательным характеристикам профессии «специалист по обучению эмпатии», представленной в направлении «обучение систем ИИ». Авторы концепции считают, что такому специалисту не обязательно обладать традиционным дипломом о высшем образовании, поскольку людей, наделенных от природы состраданием, можно обучить необходимым навыкам в рамках корпоративной программы профессиональной подготовки специалиста по психологии [17]. Однако в настоящих условиях такого уровня подготовки уже недостаточно для новых профессий, тем более что в концепции речь идет о взаимном обучении и совместном создании ценностей в процессе взаимодействия машин и обучающих их людей. Такое взаимное создание ценностей предполагает и необходимое понимание механизма работы создаваемого алгоритма, и владение не только корпоративной системой ценностных установок, но и более широкими и глубокими познаниями в этой области.

Или, например, описание профессии «менеджер по соблюдению этических норм», представленной в направлении «обеспечение устойчивости систем ИИ». Доэрти и Уилсон, в частности, считают, что в случаях предвзятого поведения ИИ специалист по этике в сотрудничестве с экспертом по алгоритмам должен раскрыть причины такого поведения и принять надлежащие меры по их устранению [17]. Представляется, что на сегодняшний день такой рабочий союз специалиста по этике с экспертом по алгоритмам более чем актуален, но в перспективе речь должна идти все-таки о специалистах, обладающих профессиональными, сквозными междисциплинарными, IT- и гуманитарными компетенциями, в частности, в области этики ИИ.

Таковы некоторые из возражений авторов данной статьи в рамках обсуждения особенностей отдельных новых профессий, предлагаемых Доэрти и Уилсоном. Однако, подчеркнем еще раз, что разработанная ими новая парадигма «человек + машина – это партнеры» оценивается аналитиками как особо значимая заслуга, отразившая вызовы времени.

Ценности и этика ИИ. Необходимым фоном нормативно-этических документов, этических кодексов, а также новых ИИ-профессий в коммуникации «человек – машина» выступает главный мировоззренческий вопрос, который беспокоит специалистов и всю прогрессивную общественность: «Какие ценности мы хотим передать машинам?».

Традиционные ценности как основополагающие регулятивы в практиках реального бытия социума всегда определяли этические принципы, смыслы которых конкретизировались в поведенческих актах и практиках человеческого общежития. В условиях новых реалий произошла определенная девальвация целого ряда ценностей и ценностных ориентиров, что, в первую очередь, коснулось поколения Z. По свидетельству аналитиков, значительное

ДИСКУРС. 2022. Т. 8, № 6. С. 17–30
DISCOURSE. 2022, vol. 8, no. 6, pp. 17–30

отличие зетов даже от предшествующего поколения Y, заключается в том, что новое поколение не видит различий между виртуальным и реальным [18, с. 81]. Изменение шкалы ценностных координат и особая направленность на себя, обусловленные цифровым технологическим детерминизмом, определили свободу как главную поведенческую ценность поколения Z в реальных и виртуальных практиках его бытия и стремления к самореализации. В новой реальности тема воспитания и образования молодого поколения приобретает особые смыслы и значения для прогрессивного развития современного социума. Сложившаяся ситуация требует аудита и синхронизации традиционных ценностей и ценностей цифровой эпохи как детерминирующих оснований, на которых выстраивается вся архитектура этических принципов, норм и правил, регулирующих отношения и поведение людей на всех уровнях коммуникативного взаимодействия в условиях цифрового общества. При этом, чтобы поведение ИИ было этичным по отношению к человеку в коммуникации с ИИ, как уже отмечалось, специалисты говорят о необходимости перевести этические принципы в базовые этические понятия, усваиваемые машинами, поскольку этичность ИИ носит не смысловой, а формализованный характер.

К настоящему времени уже разработаны специальные модели, предназначенные для распознавания эмоций, системы компьютерного зрения, обработки естественного языка, анализа данных (машинного обучения), обработки символьной информации (рассуждений на основе знаний) и т. д. Существует также обширный математический инструментарий, который уже используется для формализации понятий этики [1]. Однако эксперты считают, что проблема формализации этических норм остается все еще не решенной, поскольку тесно связана и с более общей задачей, а именно – с формализацией гуманитарного знания. Фактически объективируется роль и значение цифрового гуманитарного знания (e-Humanities/ Digital Humanities или DH), которое оценивается специалистами не как замена или отказ от традиционных гуманитарных запросов, а как естественное продолжение и расширение традиционной сферы гуманитарного знания, базирующегося на информационной методологии и новой «междициплинарности» [19, с. 10].

Целевые установки e-Humanities направлены на взаимодействие техногенных и гуманитарных ценностей и идеалов с позиций неотделимости духовной составляющей современного социума от его технологической и материальной составляющих. При этом цифровые гуманитарные науки не являются новой наукой, а представляют собой междисциплинарное направление социально-гуманитарных наук, основанное на применении цифровых технологий, выполняющих инструментальную роль в достижении целей каждого конкретного гуманитарного знания [20]. Применительно к ИИ речь идет, в частности, о цифровой этике и цифровом этикете как относительно самостоятельных предметных знаниях DH, детерминированных ценностями и идеалами информационного общества и включающих в себя интерфейсы «человек – человек» и «человек – машина». Однако, представляя собой синергетический эффект объединения гуманитарных и технологических наук, DH в настоящее время находится в стадии своего становления, включая цифровую этику и цифровой этикет, а это еще одна большая тема, которая напрямую связана с обучением машины взаимодействию с человеком.

Проведенный анализ в целом показал, что решение проблемы этичности искусственного интеллекта обусловило необходимость формирования нормативно-этической базы ИИ

ДИСКУРС. 2022. Т. 8, № 6. С. 17–30
DISCOURSE. 2022, vol. 8, no. 6, pp. 17–30

как процесса создания этических инструментов «мягкого» регулирования проблемных зон этического характера. Создание таких инструментов является причиной и следствием разработки новых этически обусловленных ИИ-профессий, их специфики, содержания и функционала, а также новых образовательных траекторий: в частности, оформление цифровой этики и цифрового этикета как предметных знаний DH, базирующихся на ценностях и идеалах современного социума, а также совершенствование образовательной системы, ориентированной в том числе на новые подходы к получению знаний при подготовке ИИ-специалистов, соответствующие вызовам времени. В этой связи представляется весьма актуальным обращение к STEM – первой образовательной модели мирового значения, которая реализует обеспечение сквозного взаимодействия между прикладными задачами, фундаментальными исследованиями и системой образования, но уже в ее третьей, новой модификации – к модели i-STEAM образования, разработанной в Израиле. Новая модель включила в себя не только гуманитарную компоненту (STEAM), но и инновационную (i-STEAM), что отражает запросы цифровой цивилизации [21].

Заключение. Развитие цифровых коммуникаций, в частности коммуникации «человек – машина», с позиций человекоориентированной цифровой эпохи (Human-centric Digital Age) напрямую зависит как непосредственно от личности самого человека, его воспитания, образования, системы его ценностных установок, идеалов, которые он несет с собой в мир ИИ, так и от безусловного понимании на уровне всего социума, что взаимодействие человека и машины должно развиваться исключительно на принципах партнерства как необходимого условия его прогрессивного развития. Однако речь идет о том, что принципам партнерства должны обучаться не только машины, но и сам человек, вступающий в диалог с машиной, в первую очередь это касается инженеров и разработчиков ИИ.

Если, как подчеркивают специалисты, видеть в этичности ИИ как конкурентные преимущества, так и важные цивилизационные смыслы, мы сможем не только выстроить партнерские отношения с ним, но и отказаться от страха перед восстанием машин. «Четвертая промышленная революция, – пишет Карл Шваб, известный немецкий экономист, создатель и руководитель ежегодного Всемирного экономического форума в Давосе, – обладает потенциалом роботизировать человечество и поставить под угрозу наши традиционные источники смыслов, таких как работа, общество, семья, личность. В наших силах не допустить развитие такого сценария, а использовать четвертую промышленную революцию для движения человечества вверх к новому коллективному и моральному сознанию, основанному на едином представлении о судьбе. Всем нам надлежит постараться, чтобы произошло именно так» [22, с. 24].

СПИСОК ЛИТЕРАТУРЫ

1. Карпов В. Э., Готовцев П. М., Ройзензон Г. В. К вопросу об этике и системах искусственного интеллекта // Философия и общество. 2018. № 2 (87). С. 84–105. DOI: 10.30884/jfio/2018.02.07.

2. Лескова Н. Л. Искусственный интеллект обучит себя сам // В мире науки. 2019. № 11. С. 92–97.

3. Almeida F. L. Concept and Dimensions of Web 4.0 // International journal of computers & technology. 2017. Vol. 16 (7). P. 7040–7046. DOI: https://doi.org/10.24297/ijct.v16i7.6446.

4. Сильный искусственный интеллект: на подступах к сверхразуму / М. С. Бурцев, О. Л. Бухвалов, А. А. Ведяхин и др. М.: Интеллектуальная литература, 2021.

5. Bias and discrimination in AI: a cross-disciplinary perspective / X. Ferrer, T. van Nuenen, J. M. Such et al. // IEEE Technology and Society Magazine. 2021. Vol. 40 (2). P. 72–80. DOI: 10.1109/MTS.2021.3056293.

6. AI Principles: Open Letter // Future of Life Institute. URL: https://futureoflife.org/ai-principles/ (дата обращения: 13.08.2022).

7. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems // Institute of Electrical and Electronics Engineers. URL: http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html (дата обращения: 23.09.2022).

8. Рекомендация об этических аспектах искусственного интеллекта. 2021 // ЮНЕСКО. URL: https://unesdoc.unesco.org/ark:/48223/pf0000380455_rus (дата обращения: 17.10.2022).

9. About GPAI // The Global Partnership on Artificial Intelligence. URL: https://gpai.ai/about/ (дата обращения: 09.10.2022).

10. Плуготаренко С. А. Зачем нужны кодексы этики для искусственного интеллекта // Инвест-форсайт. 15.11.2021. URL: https://www.if24.ru/etika-dlya-ai/ (дата обращения: 18.10.2022).

11. Translation: Personal Information Protection Law of the People's Republic of China // DigiChina. URL: https://digichina.stanford.edu/work/translation-personal-information-protection-law-of-the-peoples-republic-of-china-effective-nov-1-2021/ (дата обращения: 19.10.2022).

12. Blueprint for an AI Bill of Rights: A Vision for Protecting Our Civil Rights in the Algorithmic Age // The White House. URL: https://www.whitehouse.gov/ostp/news-updates/2022/10/04/blueprint-for-an-ai-bill-of-rightsa-vision-for-protecting-our-civil-rights-in-the-algorithmic-age/ (дата обращения: 19.10.2022).

13. Национальная стратегия развития искусственного интеллекта на период до 2030 г. // Гарант.ру. URL: https://www.garant.ru/products/ipo/prime/doc/72738946/#1000 (дата обращения: 15.10.2022).

14. Кодекс этики в сфере ИИ // Альянс в сфере искусственного интеллекта. URL: https://a-ai.ru/ethics/index.html (дата обращения: 15.10.2022).

15. Ткачева К. А., Шепелева О. С. Этика и «цифра»: этические проблемы цифровых технологий. М.: РАНХиГС, 2020.

16. Этика и «цифра»: от проблем к решениям / под ред. Е. Г. Потаповой, М. С. Шклярук. М.: РАНХиГС, 2021.

17. Доэрти П., Уилсон Дж. Человек + машина. Новые принципы работы в эпоху искусственного интеллекта / пер. О. Сивченко, Н. Яцюк. М.: Манн, Иванов и Фербер, 2019.

18. Стиллман Д., Стиллман И. Поколение Z на работе. Как его понять и найти с ним общий язык / пер. с англ. Ю. Кондукова. М.: Манн, Иванов и Фербер, 2018.

19. Можаева Г. В. Digital Humanities: цифровой поворот в гуманитарных науках // Гуманитарная информатика. 2015. Вып. 9. С. 8–23. DOI: 10.17223/23046082/9/1.

20. Мамина Р. И., Елькина Е. Е. Digital Humanities: новая наука или конвергентные модели и практики глобального сетевого проекта? // ДИСКУРС. 2020. Т. 6, № 4. С. 22–38. DOI: https://doi.org/10.32603/2412-8562-2020-6-4-22-38.

21. Международный опыт развития предпринимательского и STEAM-образования в странах ОЭСР и в мире: аналитический отчет / Б. А. Газдиева, А. А. Ахметжанова, Ж. О. Сагындыкова и др. Кокшетау: Изд-во КГУ им. Ш. Уалиханова, 2018.

22. Шваб К., Девис Н. Технологии четвертой промышленной революции / пер. с англ. К. Ахметова, А. Врублевского, В. Карпюка, А. Козлова. М.: Бомбора, 2018.

Информация об авторах.

Мамина Раиса Ильинична – доктор философских наук (2007), профессор кафедры философии Санкт-Петербургского государственного электротехнического университета «ЛЭТИ» им. В. И. Ульянова (Ленина), ул. Профессора Попова, д. 5Ф, Санкт-Петербург,

ДИСКУРС. 2022. Т. 8, № 6. С. 17–30
DISCOURSE. 2022, vol. 8, no. 6, pp. 17–30

197022, Россия. Автор более 100 научных публикаций. Сфера научных интересов: аксиосфера современного социума, коммуникативные практики, кросскультурное сотрудничество, цифровые коммуникации, цифровой этикет, цифровая самопрезентация, инновационные образовательные траектории.

Ильина Анна Валерьевна – аспирант кафедры философии Санкт-Петербургского государственного электротехнического университета «ЛЭТИ» им. В. И. Ульянова (Ленина), ул. Профессора Попова, д. 5Ф, Санкт-Петербург, 197022, Россия. Автор 2 научных публикаций. Сфера научных интересов: этика искусственного интеллекта.

О конфликте интересов, связанном с данной публикацией, не сообщалось.
Поступила 14.10.2022; принята после рецензирования 11.11.2022; опубликована онлайн 23.12.2022.

REFERENCES

1. Karpov, V.E., Gotovtsev, P.M. and Roizenzon, G.V. (2018), "On the issue of ethics and artificial intelligence systems", *Philosophy and Society*, no. 2 (87), pp. 84–105. DOI: 10.30884/jfio/2018.02.07.

2. Leskova, N.L. (2019),"Artificial intelligence will learn by itself", *V mire nauki*, no. 11, pp. 92–97.

3. Almeida, F.L. (2017), "Concept and Dimensions of Web 4.0", *International J. of computers & technology*, vol. 16 (7), pp. 7040–7046. DOI: https://doi.org/10.24297/ijct.v16i7.6446.

4. Burtsev, M.S., Bukhvalov, O.L., Vedyakhin, A.A. et al. (2021), *Sil'nyi iskusstvennyi intellekt: na podstupakh k sverkhrazumu* [Strong artificial intelligence: approaching the superintelligence], Intellektual'naya literatura, Moscow, RUS.

5. Ferrer, X., van Nuenen, T., Such, J.M., Coté, M. and Criado, N. (2021), "Bias and discrimination in AI: a cross-disciplinary perspective", *IEEE Technology and Society Magazine*, vol. 40 (2), pp. 72–80. DOI: 10.1109/MTS.2021.3056293.

6. "AI Principles: Open Letter" (2017), *Future of Life Institute*, available at: https://futureoflife.org/ai-principles/ (accessed 13.08.2022).

7. "The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems" (2022), *Institute of Electrical and Electronics Engineers*, available at: http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html (accessed 23.09.2022).

8. "Recommendation on the ethics of artificial intelligence" (2021), *UNESCO*, available at: https://unesdoc.unesco.org/ark:/48223/pf0000380455 (accessed 17.10.2022).

9. "About GPAI", *The Global Partnership on Artificial Intelligence*, available at: https://gpai.ai/about/ (accessed 09.10.2022).

10. Plugotarenko, S.A. (2021), "Why do we need codes of ethics for artificial intelligence", *Invest-Forsait*, available at: https://www.if24.ru/etika-dlya-ai/ (accessed 18.10.2022).

11. "Translation: Personal Information Protection Law of the People's Republic of China" (2021), *DigiChina*, available at: https://digichina.stanford.edu/work/translation-personal-information-protection-law-of-the-peoples-republic-of-china-effective-nov-1-2021/ (accessed 19.10.2022).

12. "Blueprint for an AI Bill of Rights: A Vision for Protecting Our Civil Rights in the Algorithmic Age" (2022), *The White House*, available at: https://www.whitehouse.gov/ostp/news-updates/2022/10/04/blueprint-for-an-ai-bill-of-rightsa-vision-for-protecting-our-civil-rights-in-the-algorithmic-age/ (accessed 19.10.2022).

13. "National strategy for the development of artificial intelligence for the period up to 2030" (2019), *Garant.ru*, available at: https://www.garant.ru/products/ipo/prime/doc/72738946/#1000 (accessed 15.10.2022).

14. "AI Ethics Code", *AI Alliance Russia*, available at: https://a-ai.ru/ethics/index-en.html (accessed 15.10.2022).

15. Tkacheva, K.A. and Shepeleva, O.S. (2020), *Etika i "tsifra": eticheskie problemy tsifrovykh tekhnologii* [Ethics and digital: ethical is-sues of digital technologies], RANHiGS, Moscow, RUS.

ДИСКУРС. 2022. Т. 8, № 6. С. 17–30
DISCOURSE. 2022, vol. 8, no. 6, pp. 17–30

16. *Etika i "tsifra": ot problem k resheniyam* [Ethics and digital: from problems to solutions] (2021), in Potapova, E.G. and Shklyaruk, M.S. (eds.), RANHiGS, Moscow, RUS.

17. Daugherty, P. and Wilson, J. (2019), *Human + Machine: Reimagining Work in the Age of AI*, Transl. by Sivchenko, O. and Yatsyuk, N., Mann, Ivanov i Ferber, Mocow, RUS.

18. Stillman, D., Stillman, I. (2018), *Gen Z @ Work: How the Next Generation Is Trans-forming the Workplace*, Transl. by Kondukov, Yu., Mann, Ivanov i Ferber, Moscow, RUS.

19. Mozhaeva, G.V. (2015), "Digital humanities: digital turn in the humanities", *Humanitarian informatics*, no. 9, pp. 8–23. DOI: 10.17223/23046082/9/1.

20. Mamina, R.I. and Yelkina, E.E. (2020), "Digital Humanities: Is it a New Science or a Set of Models and Practices of the Global Network Project?", *DISCOURSE*, vol. 6, iss. 4, pp. 22–38. DOI: https://doi.org/10.32603/2412-8562-2020-6-4-22-38.

21. Gazdieva, B.A., Akhmetzhanova, A.A., Sagyndykova, Zh.O. et al. (2018), *Mezhdunarodnyi opyt razvitiya predprinimatel'skogo i STEAM-obrazovaniya v stranakh OESR i v mire: analiticheskii otchet* [International experience in the development of entrepreneurial and STEAM education in OECD countries and in the world: analytical report], Izd-vo KGU im. Sh. Ualikhanova, Kokshetau, KAZ.

22. Schwab, K. and Davis, N. (2018), *Shaping the Fourth Industrial Revolution*, Transl. by Akhmetov, K., Vrublevskii, A., Karpyuk, V. and Kozlov, A., Bombora, Moscow, RUS.

Information about the authors.

Raisa I. Mamina – Dr. Sci. (Philosophy) (2007), Professor at the Department of Philosophy, Saint Petersburg Electrotechnical University, 5F Professor Popov str., St Petersburg 197022, Russia. The author more than 100 scientific publications. Area of expertise: axiosphere of modern society, communication practices, cross-cultural cooperation, digital communications, digital etiquette, digital self-presentation, innovative educational trajectories.

Anna V. Ilina – Postgraduate at the Department of Philosophy, Saint Petersburg Electrotechnical University, 5F Professor Popov str., St Petersburg 197022, Russia. The author 2 scientific publications. Area of expertise: ethics of artificial intelligence.

No conflicts of interest related to this publication were reported.
Received 14.10.2022; adopted after review 11.11.2022; published online 23.12.2022.