Этическая мысль
2024. Т. 24. № 1. С. 144–159
УДК 17.01

Ethical Thought
2024, Vol. 24, No. 1, pp. 144–159
DOI: 10.21146/2074-4870-2024-24-1-144-159

А.В. Углева, В.А. Шилова, Е.А. Карпова

Индекс «этичности» систем искусственного интеллекта в медицине: от теории к практике*

Углева Анастасия Валерьевна – кандидат философских наук, PhD, профессор. Национальный исследовательский университет «Высшая школа экономики». Российская Федерация, 109028, г. Москва, Покровский бульвар, д. 11.

ORCID: 0000-0002-9146-1026
e-mail: augleva@hse.ru

Шилова Валентина Александровна – кандидат социологических наук. Национальный исследовательский университет «Высшая школа экономики». Российская Федерация, 109028, г. Москва, Покровский бульвар, д. 11; Институт социологии ФНИСЦ РАН. Российская Федерация, 117218, Москва, ул. Кржижановского, дом 24/35, корпус 5.

ORCID: 0000-0002-8899-2707
e-mail: vshilova@yandex.ru

Карпова Елизавета Александровна – стажер-исследователь, аспирант. Национальный исследовательский университет «Высшая школа экономики». Российская Федерация, 109028, г. Москва, Покровский бульвар, д. 11.

ORCID: 0009-0005-0499-7930
e-mail: ea.karpova@hse.ru

* Публикация подготовлена за счет средств гранта на поддержку исследовательских центров в сфере искусственного интеллекта, в том числе в области «сильного» искусственного интеллекта, систем доверенного искусственного интеллекта и этических аспектов применения искусственного интеллекта, предоставленного АНО «Аналитический центр при Правительстве Российской Федерации» в соответствии с соглашением о предоставлении субсидии (идентификатор соглашения о предоставлении субсидии 000000D730321P5Q0002) и договором с ФГАОУ ВО «Национальный исследовательский университет «Высшая школа экономики» от 2 ноября 2021 г. № 70-2021-00139. Funding: The report study was funded by ANO Analytical Center under the Government of the Russian Federation in accordance with the agreement on granting the subsidy (identifier of the agreement 000000D730321P5Q0002) and the contract with National Research University Higher School of Economics, November 2, 2021. No. 70-2021-00139.

В статье представлена методика – «Индекс “этичности” систем искусственного интеллекта», задача разработки которой заключалась в оценке реальных и потенциальных рисков этического характера, возникающих на всех этапах жизненного цикла ИИ-систем. Сама система никакой «этичностью» не обладает, тогда как социально приемлемыми, с моральной точки зрения допустимыми и необходимыми, могут быть действия разработчиков и поставщиков данных в процессе ее проектирования и обучения, а также операторов и потребителей в ходе ее пилотирования и внедрения. Ряд вопросов, связанных, например, с конфиденциальностью персональных данных, отчасти регулируется актуальным законодательством, тогда как большая часть этических рисков лишь прогнозируется. Представленная методика разработана: для целей медицинского сообщества, нуждающегося в подтверждении моральной состоятельности и соответствия профессиональным стандартам внедрения новейших технологий в клиническую практику; в качестве инструмента, который может быть использован в деятельности этических комиссий по аналогии с биоэтическими комитетами при согласовании соответствующих медицинских исследований с участием ИИ; как дополнение к процедурам добровольной технической сертификации, а также в судебно-медицинской экспертизе. В «Индексе» отражены наиболее острые этические проблемы – доверия, распределенной ответственности, конфиденциальности данных, прозрачности и объяснимости ИИ-моделей, справедливости и недискриминации, обсуждаемые в современной этике в связи с рисками развития ИИ. Уникальность «Индекса» заключается в лежащем в его основе междисциплинарном подходе, включающем в себя методологию «дизайна ценностей» и «полевое» социологическое исследование, позволившее совместить теоретические гуманитарные подходы к пониманию содержания ключевых моральных понятий с их трактовкой в профессиональной медицинской этике. Это делает «Индекс» не только интересным с теоретической точки зрения как способ формализации этических категорий, но и полезным с практической точки зрения – как пример прикладного значения этики и использования ее инструментария для решения социально значимых задач.

Ключевые слова: искусственный интеллект в медицине, индекс этичности, этика в сфере искусственного интеллекта, ответственный искусственный интеллект, доверенный искусственный интеллект, недискриминация, конфиденциальность данных, объяснимость, судебно-медицинская экспертиза, этическая экспертиза

Значимость этической экспертизы в сфере искусственного интеллекта

В последние несколько лет в мире с целью выработки принципов и стандартов оценки для проведения гуманитарной экспертизы на всем жизненном цикле технологии было разработано множество стратегий, начиная с Глобальной рекомендации ЮНЕСКО по этике в сфере ИИ, общего регламента защиты данных (GDPR), предложений ЕС по идентификации систем повышенного риска, и вплоть до различных национальных документов, регулирующих сферу искусственного интеллекта (ИИ), а также кодексов, уставов, хартий, обеспечивающих профессиональное регулирование ИИ в конкретных отраслях, например, документы Института инженеров электротехники и электроники (IEEE) и проч. Что касается медицины, то в самых разных

научных и образовательных организациях как медицинского, так и не медицинского профиля сегодня предпринимается немало усилий по разработке отраслевых рекомендаций к Национальному кодексу этики в сфере ИИ РФ [Кодекс этики в сфере искусственного интеллекта, 2021], а также всевозможных приложений, регламентов и тому подобных национальных и международных документов (см., напр.: [Этические принципы, 2021])[1]. Однако, как правило, сами принципы и нормы формулируются в них не с позиции этики как философской субдисциплины, а с этико-правовой точки зрения. Они лишены нередко терминологической ясности и не уделяют должного внимания именно этическим и экзистенциальным рискам внедрения ИИ-технологий. Вместо этого в таких документах превалирует правовой подход и стремление весь процесс этической оценки свести к строгим алгоритмизированным процедурам.

Правовой подход позволяет фиксировать лишь принципиально поддающиеся кодификации и формализации морально-правовые понятия и нормы, а формализация этических понятий в алгоритмических процедурах лишает какой-либо «этичности» сами попытки свести оценку социальной приемлемости ИИ-систем и технологий к набору «закрытых» вопросов и простых ответов с арифметически просчитываемым результатом, как если бы существовали однозначно толкуемые нормы и правила поведения как в профессиональном медицинском сообществе, так и в обществе в целом. При этом в известных нам случаях[2] разрабатываемые способы оценки «этичности» ИИ до сих пор не опирались на сколько-нибудь значимые философские и социологические исследования и потому вряд ли могут отражать объективное положение дел в конкретных отраслях.

Такой исключительно «правовой» и отчасти «технологический» подходы страдают, во-первых, редукционизмом в понимании «этики», а во-вторых, проблематичностью в их рамках обнаружить и акцентировать внимание на потенциально конфликтных сферах применения новейших технологий, а также на сохранении общественного доверия к ним, что возможно при условии анализа зоны психологического дискомфорта и экзистенциальной резистентности к тому, что порождает обоснованные или необоснованные социальные страхи, например, связанные с нарушением принципа конфиденциальности данных, неприкосновенности частной жизни, информационной безопасности, психологического или физического вреда здоровью, репутационного ущерба и т.д. Они подогреваются множащимися случаями несанкционированного использования персональных данных, а также возникновением новых потенциально рискованных способов сбора, хранения, распространения, утилизации, предоставления (отражения) и использования данных в процессе принятия решений и организационных процедур. Особого внимания заслуживает использование необъективных, некорректных и нерепрезентативных данных для обучения

---

[1] О все возрастающей практике внедрения и инвестирования в ИИ-системы в здравоохранении см., напр., [Гусев, Шарова, 2023].

[2] В том числе в рамках работы по международной стандартизации Объединенного технического комитета ISO/IEC JTC 1, SC 42.

систем, непрозрачное принятие решений в области машинного обучения которых провоцирует недоверие к ним как со стороны оператора, так и со стороны потенциального потребителя. Будучи лишено возможности контроля и не обладая достаточным пониманием общественного воздействия ИИ-технологий после их внедрения, общество настороженно относится к разного рода инновациям вплоть до их полного отрицания. Для снятия данного напряжения необходима разработка системы общественного контроля и гуманитарной экспертизы (см.: [Брызгалина и др., 2023]), в основе которой должно лежать основанное на междисциплинарном и инклюзивном подходе фундаментальное научное исследование.

«Этичность» ИИ-систем

Следует заметить, что ИИ-системы, как и любые другие технические системы, не имеют внутренне присущих им свойств в области морали – эти свойства характерны для процессов жизненного цикла систем. Можно судить об этичности процедуры сбора данных при создании ИИ-системы, этичности применения систем для решения той или иной прикладной задачи, этичности интерпретации и использования результатов обработки данных системой и т.п. При этом неприемлемые социальные последствия использования ИИ-системы могут быть обусловлены непониманием или игнорированием потребителем ее свойств, а не свидетельством «неэтичности» самой системы. Важно отметить также, что для любых видов угроз и любых заинтересованных сторон степень соответствия требованиям непосредственно ИИ-систем корректно рассматривать в контексте конкретной прикладной задачи и в определенных (предусмотренных) условиях эксплуатации.

Важнейшее значение здесь приобретает доверие к ИИ. В Национальном стандарте «Способы обеспечения доверия к ИИ» [ГОСТ Р 59276-2020][3] доверие к ИИ-системам фактически напрямую связано с качеством, эффективностью ИИ-систем и удовлетворением потребительских ожиданий. Разработанная в соответствии со стандартом методика оценки доверия к ИИ [Алексеев, 2022], предлагает оценивать его как сочетание (произведения показателей) надежности и безопасности ИИ-системы. Для задач стандартизации эта методика – вполне работающий инструмент. Однако в медицине эффективность технологии в конечном счете оценивается как ее вклад в поддержание здоровья пациента, в повышение качества и увеличение продолжительности его жизни. А значит, уровень безопасности применения такого инструмента должен быть установлен до начала его проверки на эффективность. То есть оценка уровня доверия к ИИ-системе в медицине как произведения ее качества и безопасности этому требованию соответствует не в полной мере. Конечно, некоторые непосредственные этические риски (такие как ограничение свободы выбора, вмешательство в частную жизнь) при проведении проверки системы

[3] См. также: ISO/IEC JTC 1/SC 42/AHG 7 Convenorship: SCC CEN-CLC-JTC21 N216 AHG7 Proposal for NWI AI trustworthiness characterization. 2022-12-28.

на безопасность и эффективность могут быть в значительной степени формализованы – в том числе через совершенствование формы информированного добровольного согласия на участие в таком исследовании. Как и социальные риски от внедрения ИИ-системы в клинических целях, некоторые медицинские риски также могут быть измерены при проведении соответствующих испытаний или в рамках мониторинга ее рутинного применения. Однако это лишь часть рисков, которые возможно предвидеть, тогда как существуют этические риски, требующие уточнения базовых этических понятий. К таким рискам относятся, например, деградация фундаментальных эмпирических знаний и навыков, непосредственно связанных с элементарными рутинными навыками и процедурами; неспособность персонала решать задачи управления в случае выхода из строя ИИ-систем на длительное время; размывание ответственности за решения, принятые с участием экспертных систем на основе ИИ; десубъективация преимущественно человеческих видов деятельности, таких как врачебная деятельность, и др. В этой связи возрастает важность осуществления регулярного социального и этического мониторинга клинических практик, осуществляемых при поддержке ИИ-системы, а также социологических исследований общественного ожидания и восприятия таких практик.

«Доверенный» ИИ требует строгого соблюдения пяти принципов этики [Floridi, Cowls, 2022]: 1. Непричинение вреда; 2. Благодеяние, т.е. действие с максимальной пользой и минимизацией вреда; 3. Автономия – принцип уважения права человека на принятие собственных осознанных решений, в случае медицины – касающихся его медицинского обслуживания и конфиденциальности его личных данных. Решение является автономным, если оно (1) интенционально – целенаправленно; (2) сопровождается пониманием, т.е. оператор и потребитель имеют полную информацию о работе ИИ-системы и возможных прогнозируемых рисках ее применения; (3) свободно – исходит из воли человека, а не под давлением чужой воли [Gibson, 2007]. Принцип автономии на примере медицины подчеркивает важность получения информированного согласия[4] от пациентов на любые виды вмешательства в состояние его здоровья, а также сохранение права пациента на удаление нежелательной личной информации из базы данных и на прекращение медицинского обслуживания с помощью ИИ-системы, если прерывание лечения не повлечет за собой негативных последствий для здоровья и жизни пациента. Принцип автономии предполагает также автономию врача, который, будучи оператором ИИ-системы, должен обладать формальным правом и реальной возможностью осуществлять критическую оценку принятого ИИ-системой решения, чтобы оно было осмысленным и понятным; 4. Справедливость – противодействие разнообразным дискриминационным эффектам от решений ИИ-системы по причине, как правило, необъективных данных, которые могут использоваться для обучения ИИ-моделей и оценки их прогностической эффективности

[4] Благодаря информированному согласию пользователь получит уведомление о характере и назначении его взаимодействия с ИИ-системой, о конкретных целях ее использования и основаниях в данном конкретном случае, а также гарантию недопустимости ее использования для иных целей, не соответствующих условиям ее эксплуатации.

(см., напр.: [Anderson, Sutherland, 2024); 5. Принцип объяснимости, включающий в себя и эпистемологический смысл понятности (прозрачности[5]), и этический смысл подотчетности (ответственности).

Для всех, кто так или иначе соприкасается с этическими дилеммами в сфере медицины, нетривиальными и крайне актуальными этическими проблемами в условиях все более интенсивного внедрения ИИ-технологий являются не только проблема доверия к ИИ, но и проблема справедливости распределения ответственности в случае допущения медицинских ошибок (в том числе по причине несоответствия ИИ-системы требованию эксплуатации), а также искоренения предвзятостей, которые, как правило, сопровождают человеческие решения, но тем более сопровождают решения, принимаемые ИИ или с его участием, потому как пишущие алгоритмы люди вольно или невольно прописывают их в цифровых кодах на основании присущих им когнитивных предубеждений и предвзятостей. Под предвзятостью имеется в виду свойство ИИ-системы, заключающееся в принятии ошибочных решений, связанных со статистической смещенностью обучающей выборки исходных данных. Когнитивная предвзятость представляет собой предубеждение, возникающее при обработке и интерпретации информации. В итоге неоднозначным и слишком подвижным метрикам этической валидации ИИ в медицине могут противостоять плотно укорененные предвзятости – например, в отношении качества жизни людей с ограниченными возможностями, понимания медицинской нормы и т.п. – неявная (часто – неумышленная) дискриминация.

Методология оценки «этичности» ИИ

Для решения задачи недопущения неэтичных практик внедрения ИИ исследовательской группой «Этическая экспертиза в сфере ИИ» Центра ИИ НИУ ВШЭ был разработан «Индекс этичности систем искусственного интеллекта в медицине» (свидетельство о регистрации произведения № 8.0176-2023). Его предваряла теоретическая и практическая части исследования. Теоретический этап состоял, во-первых, в общей классификации источников ценностей (ценности по отношению к другим, права человека, утилитарные, перфекционистские, экзистенциальные ценности) при признании амбивалентности ценностных убеждений [Wong, 2023]. И, во-вторых, в выделении ключевых ценностей медицинской этики – таких как автономия, благодеяние, справедливость, ненанесение вреда и забота [Smajdor et al., 2022]. В соответствии с этими двумя типами классификаций были проанализированы ключевые источники по этике в сфере ИИ и по биоэтике, которые являются наиболее дискутируемыми и референтными в современной медицинской этике: Асиломарские

[5] Объяснимость – наличие четко выраженных причинно-следственных связей, на основании которых действует и принимает решение ИИ. Однако глубинное обучение устроено так, что противится интерпретации и объяснимости своих процессов, а потому создает угрозу снижения уровня доверия к ИИ. Поскольку в медицине, в частности, в рентгенологии, широко используется т.н. дискриминативный ИИ, основанный именно на глубинном обучении, то проблема доверия к ИИ здесь стоит особенно остро.

принципы искусственного интеллекта [Asilomar AI Principles, 2017]; Монреальская декларация об ответственном искусственном интеллекте, разработанная под эгидой Монреальского университета (2017) [Montreal Declaration, 2017]; Этически согласованный дизайн: концепция приоритета человеческого благополучия с помощью автономных и интеллектуальных систем [Ethically Aligned Design, 2019]; Заявление об искусственном интеллекте, роботических и автономных системах Европейской комиссии [Statement on artificial intelligence, 2018]; Доклад комитета по искусственному интеллекту Палаты лордов Великобритании [Select Committee on Artificial Intelligence, 2018]; Принципы партнерства в сфере ИИ [Partnership on AI, 2018]; Нюрнбергский кодекс [The Nuremberg Code, 1949, 181–182]; Женевская декларация Всемирной Медицинской Ассоциации [Declaration of Geneva, 1948]; Международный кодекс медицинской этики [International Code of Medical Ethics, 1949]; Хельсинкская декларация [Declaration of Helsinki, 1964]; Руководство по надлежащей клинической практике [Baber, 1994]; Конвенция о защите прав и достоинства человека в связи с применением достижений биологии и медицины: конвенция о правах человека и биомедицине [Convention on Human Rights and Biomedicine, 1997]; Всеобщая декларация о биоэтике и правах человека [ЮНЕСКО, 2005].

В основу эмпирической части пилотажного исследования с участием представителей врачебного сообщества и разработчиков ИИ-систем (IT-специалисты, предприниматели) легла качественная методология сбора и обработки данных [Полухина, 2023]. Сбор информации осуществлялся в рамках глубинного полуструктурированного экспертного интервью с врачами различных специальностей (онкологи, рентгенологи, сурдологи), имеющими опыт работы с ИИ-системами и/или на базе ИИ, и разработчиками таких систем для целей здравоохранения[6].

Гайд глубинного экспертного опроса был основан на позиции Ю. Хабермаса[7], заложившего традицию поиска истины в процессе «рационального обсуждения, направленного на достижение согласия» [Квале, 2003, 50]. Основной интерес представляла структура восприятия респондентами этических аспектов использования ИИ в медицинской сфере. А цель – выявление и систематизация основных критериев, по которым оценивается этическая составляющая при разработке, внедрении и применении ИИ в медицине с опорой как на общечеловеческие ценности, так и на морально-этические нормы и принципы, принятые в профессиональной медицинской среде.

Структура гайда включала в себя несколько блоков: 1. Общая информированность и опыт использования ИИ (для медицинских работников) и опыт разработки ИИ (для разработчиков); 2. Обусловленность доверия к ИИ-системам;

---

6 Всего было опрошено 45 человек: 23 врача и 22 разработчика ИИ в возрасте от 20 до 50 лет.

7 В основе эмпирического исследования лежит методологический подход Ю. Хабермаса к изучению проблем дискурса, позволивший выстроить анализ по принципам социальной обусловленности речевых высказываний. Интерес в данном случае представляют обоснования диалогической формы коммуникации, требовавшие особого внимания к языку, вне которого не может быть в должной мере изучена специфика коммуникативного действия.

3. Положительные и отрицательные аспекты применения ИИ для врача и для пациента; 4. Важные общечеловеческие ценности и их трактовка, которые необходимо соблюдать при разработке и внедрении ИИ; 5. Ответственность за медицинские ошибки, допущенные в ходе принятия клинического решения с помощью ИИ; 6. Необходимость в этической экспертизе (ее целесообразность на различных этапах жизненного цикла ИИ-системы); 7. Прогнозируемые риски, связанные с внедрением ИИ.

При анализе полученных данных использовались методы систематизации, конденсации смысла, осевого кодирования, категоризации значений, структурирования смыслов и интерпретации значений. При обработке полученных результатов использовалась концепция «дизайна ценностей» Е. Айзенберга и Й. Ван ден Ховена [Aizenberg, Van Den Hoven, 2020]. Ее значимость определяется лежащими в ее основе двумя взаимодополняющими подходами: 1. Дизайн информационных систем (Value Sensitive Design), предписывающий изучение и включение в процесс проектирования этических ценностей прямых и косвенных заинтересованных сторон; 2. Разработка ИИ-систем с активным участием основных заинтересованных сторон в процессе проектирования дизайна с целью удовлетворения их ценностных потребностей и этических требований (Participatory Design). Таким образом, итоговый набор толкований конкретных ценностных установок, используемых в «Индексе», является результатом «ценностного» дизайна, основанного на синтезе их идейно-философского содержания и особенностей профессиональных интерпретаций в медицине.

Индекс «этичности» ИИ-системы рассчитывается по результатам теста, состоящего из 20 вопросов разных типов: 1 – вопросы, представляющие наборы суждений без ограничения выбора вариантов ответа испытуемыми; 2 – вопросы, представляющие наборы суждений с ограничениями выбора вариантов ответа; 3 – дихотомические вопросы, предполагающие вариант ответа «да» или «нет». Также в тесте использованы вопросы-фильтры, при правильном ответе на которые испытуемый переходит к следующему содержательному вопросу. Вопросы теста сформулированы таким образом, чтобы можно было оценить когнитивные, эмоциональные (оценочные) и поведенческие факторы, влияющие на восприятие «этичности» или «неэтичности» той или иной ИИ-системы.

Индекс рассчитывается по матрице, которая представляет собой правила начисления баллов за варианты ответов, выбранных испытуемым. За каждый ответ можно получить 0 баллов или от одного до пяти баллов (в зависимости от характера вопроса). В расчете индекса учтено, что разные вопросы отличаются друг от друга по степени важности, соответственно, каждому вопросу приписывался определенный вес. Индекс состоит из суммы баллов по каждому блоку вопросов (важных, средних по значимости и менее значимых), деленной на максимально возможное количество баллов, которое может набрать испытуемый, выбрав все правильные ответы, соответственно, 29, 13 и 8. Далее каждый показатель умножается на его вклад в индексе, соответственно, 0.5, 0.3 и 0.2. В зависимости от успешности прохождения теста испытуемым, полученный результат приобретает числовое значение от 0 до 1,

соответствующее различным уровням этичности (от 0,8 до 1 – «система этична»; от 0,6 до 0,7 – «система скорее этична»; от 0 до 0,5 – «система неэтична»).

Присутствие в тесте возможности оставить комментарий с последующей обратной связью позволяет, с одной стороны, зафиксировать тонкости различий в толковании отдельных этических понятий, используемых в тесте, разными категориями респондентов; с другой стороны, – оставляет за респондентом право на получение квалифицированного подробного экспертного заключения как по пройденному тесту, так и по вопросу достижения более высокого уровня «этичности» внедряемой им ИИ-системы. Экспертная оценка уровня «этичности» системы формируется и определяется профессиональными знаниями, опытом и интуицией эксперта[8].

Моральные теории как основа экспертной оценки

В целом экспертные процедуры в рамках этики в сфере ИИ в медицине нередко обращены ко всем трем основным типам этических теорий: консеквенциализму (ориентируясь на возможные последствия внедрения ИИ-системы), этике долга (ориентируясь на базовые права и свободы человека, а также сопряженные с ними намерения и мотивы, которые могут быть поставлены под угрозу) и этике добродетели (ориентируясь на те изменения моральных и профессиональных качеств, которые могут быть спровоцированы внедрением ИИ-системы, а также формулируя идеальное сочетание личных или групповых моральных установок ее пользователей) [Kazim, Koshiyama, 2021]. Принимая во внимание результативность каждой из них в конкретных обстоятельствах, при разработке данного «Индекса» предпочтение было отдано в пользу современной этики и эпистемологии добродетели, а также оценке принципиально операционализируемых с помощью опросника этических понятий – таких как доверие, ответственность, справедливость, прозрачность ИИ-моделей, автономия, недискриминация. Преимущество аретологического подхода видится в том, что он не выводит этические принципы из каких-либо априорных оснований, а акцентирует внимание на качествах субъекта действия, определяемых его внутренними убеждениями (в соответствии в том числе с его профессиональной принадлежностью) и одновременно – надежностью его когнитивного аппарата. При всем разнообразии возможных толкований «добродетели», в самом общем виде она понимается как, с одной стороны, способность субъекта обеспечить надлежащий (нормативный) характер действия благодаря добродетельности его характера [Zagzebski, 2010], а с другой стороны, как его эпистемическое свойство (процессуальная и инструментальная добродетель), контекстно зависимое и конститутивное качество

[8] Эта возможность обращения к квалифицированному мнению эксперта позволяет сохранить человеческий контроль за системой и оценкой ее социальной приемлемости на всем ее жизненном цикле, с учетом требования проектирования ИИ-систем и приложений. Для систем с высокой степенью риска – таких как ИИ в медицине – сохранение человеческого фактора в оценке их эффективности является критически важным.

ответственного индивида, пребывающего в перманентном состоянии поиска наилучшего решения с точки зрения его надежности и эффективности. Таким образом понимаемая «добродетель» требует от субъекта действия способности принимать автономное ответственное решение с учетом общего блага и социальной допустимости поступка.

Что касается операционализируемых этических понятий, то их многомерность и принципиальная функциональная сложность потребовали прибегнуть к двум взаимодополнительным методическим приемам. В первой части «Индекса» требуется оценить релевантность различных критериев в ситуации моральной оценки («Предполагает ли внедрение данной системы искусственного интеллекта возможность нарушения принципа автономии пациента?»). Во второй его части оценивается степень согласия испытуемых с различными утверждениями («Оцените состояние решенности Вами вопроса об управлении требованиями к данной системе на всех этапах ее жизненного цикла в целях недопущения ошибок алгоритмов машинного обучения при предъявлении исходных данных, отличающихся от однотипных условий эксплуатации. Оценку нужно произвести по шкале от 0 до 3 (где “0” – вопрос не обсуждался, и задача так не поставлена; “1” – задача поставлена, но не решена; “2” – задача решается, но пока не выработан механизм полного контроля за работой ИИ-системы; “3” – ИИ-система находится под полным контролем со стороны разработчика и пользователя)»). Оба подхода показали свою эффективность, в частности, в ходе разработки опросника моральных оснований (Moral Foundations Questionnaire, MFQ) [Graham et al., 2011], который в свою очередь основан на теории моральных оснований Дж. Хайдта и Кр. Джозефа. Эта теория предполагает, что набор врожденных интуиций[9] приводит людей к определенным эмоциональным реакциям на конкретные события [Haidt, Joseph, 2004]. Эти же интуиции легко прослеживаются в качестве базовых оснований медицинской этики: забота, справедливость (равенство и пропорциональность), приверженность ценностям профессионального сообщества (преданность) и ненанесение вреда пациенту [Atari et al., 2022].

Основные результаты тестирования

Исследование показало, что эксперты – врачи имеют представление о новейших разработках в сфере ИИ, преимущественно ограниченные сферой их профессиональной деятельности, многие принимают личное участие в качестве консультантов в проектировании ИИ-систем. Эксперты – разработчики ИИ-систем часто не имеют специальных знаний в сфере медицины, но обладают

[9] Существует довольно широкая критика теории моральных оснований, отрицающая идею существования некоторого рода интуитивной моральной грамматики [Mikhail, 2007]. Однако, соглашаясь с критикой этой концепции в случае ее претензии на универсальность, применительно к медицинской этике она кажется вполне удовлетворительной в силу того, что ключевой категорией в ней оказывается ненанесение вреда пациенту. Это есть своего рода «когнитивный моральный шаблон», позволяющий оценивать поступки как морально приемлемые в соответствии со степенью вреда, который они способны нанести [Schein, Gray, 2018].

общими квалификациями, позволяющими решать в том числе специальные медицинские задачи. При этом вопросы этики в большей степени важны для медицинского сообщества, чем для IT-специалистов, однако для обеих групп респондентов их решение представляется значимым на современном уровне развития ИИ-технологий.

При этом осуществленная классификация критериев, по которым формируется доверие к ИИ, показывает, что медицинские работники готовы доверять ИИ-системам в том случае, если алгоритмы их работы максимально прозрачны и понятны, прежде всего, должно быть ясно, на каких именно материалах (их объем и качество) обучалась система. Важнейшее значение для доверия к ИИ-системе имело сотрудничество разработчиков с тем или иным признанным в медицинской сфере научным учреждением, при этом то, где была разработана система – в России или за рубежом, – существенного устойчивого значения и однозначно определяющего влияния на доверие к ИИ-системе экспертов – медицинских работников не имеет. Большинство опрошенных врачей (77%) отмечали, что доверие к ИИ-системе сформируется у них лишь в том случае, если на начальном этапе работы с этой системой, когда врач имеет возможность перепроверить принимаемые ею решения, они безошибочны или совпадают с решением врача. В целом отношение медицинских работников к ИИ-системам варьируется от осторожного доверия до крайнего недоверия. В случае же с разработчиками, это отношение чаще можно охарактеризовать как доверительное.

Опросы также показали, что при попытке оценить конкретные кейсы допущения медицинских ошибок при использовании ИИ-систем, респонденты затруднялись определить степени ответственности, которая лежит на разработчиках, медицинских учреждениях, врачах, пациентах и контролирующих органах. Это говорит о том, что в сфере применения новых технологий в области медицины на данный момент не проработаны протоколы и нормативные акты, четко определяющие сферу ответственности каждой из сторон. Одновременно этика ответственности применительно к ИИ также не является распространенной практикой (стандартом) профессионального поведения в медицинской среде, т.е. пока не выработан консенсус по этому поводу. Так, одни респонденты настаивают на том, что врач всегда ответственен за любое принятое решение (52%); другие, напротив, полагают, что за ошибку ИИ-системы, при соблюдении врачом условий ее эксплуатации, должен отвечать разработчик и выдавший ему технический сертификат орган (48%).

Оценив, в какой степени артикулируются и как понимаются важные общечеловеческие ценности разработчиками ИИ при разработке, внедрении и использовании ИИ в медицинской практике, был сделан промежуточный вывод о том, что ценность непричинения вреда (как физического, так и психологического) пациенту по-прежнему является высшей ценностью для обеих категорий респондентов, тогда как ценности автономии и недискриминации пациента являются второстепенными. Более того, значительная часть респондентов (57%) допускают возможность с внедрением ИИ в медицину формирования новых видов неравенств.

Проанализировав явные и скрытые этические риски при внедрении ИИ в области медицины, можно сделать вывод, что в данной сфере в настоящий момент присутствуют отдельные интуитивно осознаваемые этические нормы и принципы, но они четко не сформулированы и не артикулированы ни внутри медицинского сообщества, ни в среде разработчиков ИИ. Это способствовало в процессе валидации «Индекса»[10] формированию более четкого запроса со стороны как врачей, так и IT-специалистов, на выработку формализованного языка описания моральных категорий (глоссария); генерацию нравственных дилемм, в которых имеется ощутимый конфликт моральных оснований с целью тестирования ИИ-систем на адекватность конкретной заранее заданной моральной рамке, а также предложения стандартных решений таких ситуаций на всем жизненном цикле системы; разработку системы тестирования ИИ-технологии, в которой объяснимость и возможность интерпретации конкретного медицинского ИИ была бы эксплицирована предельно ясно. Это должно быть необходимым элементом при прохождении этической самодиагностики в отношении прозрачности ИИ уже на этапе его проектирования, а также служить инструментарием для независимых этических комитетов (по аналогии с биоэтическими комитетами), позволяющим осуществлять наряду с предварительными процедурами этической оценки глубинные аналитические процедуры с обоснованным экспертным заключением.

Применимость «Индекса»

Применение данной методики призвано, во-первых, способствовать минимизации вреда от внедрения той или иной ИИ-технологии (системы) в медицине путем оценки характера ее внедрения, т.е. выявления степени ее социальной и моральной приемлемости. Под последней понимается соответствие внедряемой ИИ-системы, с одной стороны, принятым в обществе нормам и ценностям, а с другой стороны, – установленным в профессиональном медицинском сообществе стандартам поведения и общепринятым практикам решения сложных этических дилемм; во-вторых, сместить фокус внимания всех участников процесса с сугубо технических характеристик внедряемых ИИ-систем в сторону социальных потребностей их будущих пользователей; в-третьих, способствовать пересмотру принципа единоличной ответственности медицинского работника за все принимаемые клинические решения в сторону распределенной ответственности всех участников процессов жизненного цикла ИИ-системы; в-четвертых, отказаться от рассмотрения социальной приемлемости внедрения той или иной технологии исключительно в терминах позитивного права, концентрируясь на ее морально-этических аспектах. Это различение «морального» и «правового» имеет ключевое значение как при формулировке отдельных вопросов теста, так и для интерпретации полученных в его ходе результатов, поскольку позволяет выйти за рамки действующего

[10] Готовая методика была апробирована на 15 врачах и 5 разработчиках ИИ для медицины, что позволило скорректировать итоговую формулу расчета индекса этичности.

закона и сконцентрировать внимание разработчика и оператора ИИ-системы на контекстно-зависимых условиях ее внедрения и необходимости следования принципу сохранения достоинства человека; в-пятых, предоставить экспертному сообществу (этические комитеты, судебно-медицинское экспертное сообщество и проч.) инструмент, который может быть использован как для предварительной оценки социальной приемлемости конкретной ИИ-системы, предлагаемой к внедрению в медицинских целях, так и в оценке post factum случаев причинения вреда здоровью по причине врачебных ошибок или «нежелательного исхода».

В силу высокой доли непредсказуемости машинного и тем более глубинного обучения ИИ-технологий, данный «Индекс» может являться одним из базовых элементов социо-гуманитарного сопровождения технологии на всех этапах ее жизненного цикла, а также как дополнение к ее добровольной технической сертификации. Вместе с тем требуется предусмотреть этико-правовой и социальный мониторинг рутинного применения ИИ-системы для установления отдаленных последствий ее внедрения с учетом отраслевой специфики медицины.

Index of "Ethicality" of AI Systems in Medicine: From Theory to Practice

Anastasia V. Ugleva, Valentina A. Shilova, Elizaveta A. Karpova

A.V. Ugleva – HSE University. 11 Pokrovsky bd., Moscow, 109028, Russian Federation.

ORCID: 0000-0002-9146-1026
e-mail: augleva@hse.ru

V.A. Shilova – HSE University. 11 Pokrovsky bd., Moscow, 109028, Russian Federation; Institute of Sociology, Federal Center of Theoretical and Applied Sociology, RAS. Krzhizhanovsky str., bld. 5, 24/25, Moscow, 117218, Russian Federation

ORCID: 0000-0002-8899-2707
e-mail: vshilova@yandex.ru

E.A. Karpova – HSE University. 11 Pokrovsky bd., Moscow, 109028, Russian Federation.

ORCID: 0009-0005-0499-7930
e-mail: ea.karpova@hse.ru

The article presents the methodology developed in the HSE University – *Index of Ethics of Artificial Intelligence Systems*. The task of developing this *Index* was to assess real and possible ethical risks arising at all stages of the life cycle of AI systems. The system itself does not possess any "ethics", while socially acceptable, morally permissible, and necessary may be the actions of developers and data providers in the process of its design, as well as of operators and consumers in the process of piloting and implementation. Several issues related, for example, to the confidentiality of personal data, to liability, are partly regulated by the current legislation, while most of the threats are only predicted. The methodology is designed for the purposes of the medical community, which needs to confirm the moral soundness and compliance with professional standards for the introduction of new technologies

into clinical practice; and as a tool that can be used in the activities of ethics commissions, similar to bioethics committees, when approving relevant medical research involving AI; as a supplement to voluntary technical certification procedures, as well as in forensic medical examinations. The *Index* reflects the most pressing issues – such as issues of trust, distributed responsibility, data privacy, transparency and explainability of AI models, fairness, and non-discrimination. The uniqueness of the *Index* lies in its underlying interdisciplinary approach, which includes the methodology of "value design" and field sociological research, which made it possible to combine theoretical humanitarian approaches to understanding the content of key moral concepts, with their interpretation in professional medical ethics. This makes the "Index" not only interesting from a theoretical point of view to formalize ethical concepts, but also useful from a practical point of view – as an example of the applied meaning of ethics and the use of its tools to solve socially significant problems.

Keywords: medical artificial intelligence, ethics index, ethics in artificial intelligence, responsible artificial intelligence, "trusted" artificial intelligence, non-discrimination, data privacy, explainability, forensics, ethical expertise

Литература / References

Алексеев А.Ю., Винник Д.В., Гарбук С.В., Лекторский В.А., Черногор Н.Н. (ред.) Методика оценки доверия к «искусственному интеллекту». М.: Президиум РАН, 2022.

Alekseev, A.Yu., Vinnik, D.V., Garbuk, S.V., Lektorskij, V.A., Chernogor, N.N. (eds.) *Metodika ocenki doveriya k "iskusstvennomu intellektu"* [A Methodology for Assessing Trust in "Artificial Intelligence"]. Moscow: Prezidium RAN Publ., 2022. (In Russian)

Брызгалина Е.В., Гумарова А.Н., Шкомова Е.М. Искусственный интеллект в медицине: рекомендации по проведению социально-гуманитарной экспертизы // Сибирский философский журнал. 2023. Т. 21. № 1. С. 51–63.

Bryzgalina, E.V., Gumarova, A.N., Shkomova, E.M. "Iskusstvennyj intellekt v medicine: rekomendacii po provedeniyu social'no-gumanitarnoj ekspertizy" [Artificial Intelligence in Medicine: Recommendations for Social and Humanitarian Expertise], *Sibirskij filosofskij zhurnal*, 2023, Vol. 21, No. 1, pp. 51–63. (In Russian)

Всеобщая декларация о биоэтике и правах человека. Организация Объединенных Наций. URL: https://www.un.org/ru/documents/decl_conv/declarations/bioethics_and_hr.shtml (дата обращения: 28.02.2024)

Universal Declaration on Bioethics and Human Rights. *United Nations* [https://www.un.org/ru/documents/decl_conv/declarations/bioethics_and_hr.shtml, accessed on 28.02.2024]. (In Russian)

Кодекс этики в сфере искусственного интеллекта (принят 26.10.2021). Альянс в сфере ИИ. URL: https://a-ai.ru/ethics/index.htm (дата обращения: 28.02.2024).

Code of Ethics for Artificial Intelligence (26.10.2021), Alliance in AI [https://a-ai.ru/ethics/index.htm, accessed on 28.02.2024]. (In Russian)

ГОСТ Р 59276-2020. Системы искусственного интеллекта. Способы обеспечения доверия. Общие положения. М.: Стандартинформ, 2020.

GOST R 59276-2020 Sistemy iskusstvennogo intellekta. Sposoby obespecheniya doveriya. Obshchie polozheniya [National Standards of Russia 59276-2020. Artificial Intelligence Systems. Ways of Ensuring Trust. General Provisions]. Moscow: Standartinform Publ., 2020. (In Russian)

Гусев А.В., Шарова Д.Е. Этические проблемы развития технологий искусственного интеллекта в здравоохранении // Общественное здоровье. 2023. № 1. С. 42–50.

Gusev, A.V., Sharova, D.E. “Eticheskie problemy razvitiya tekhnologij iskusstvennogo intellekta v zdravoohranenii” [Ethical Issues in the Development of Artificial Intelligence Technologies in Health Care], *Obshchestvennoe zdorov’e*, 2023, No. 1, pp. 42–50. (In Russian)

Квале С. Исследовательское интервью. М.: Смысл, 2003.

Kvale, S. *Issledovatel’skoe interv’yu* [Research Interview]. Moscow: Smysl Publ., 2003. (In Russian)

Практики анализа качественных данных в социальных науках: учеб. пособие / Отв. ред. Е.В. Полухина. М.: Изд. дом ВШЭ, 2023.

Poluhina, E.V. (ed.) Praktiki analiza kachestvennyh dannyh v social’nyh naukah: ucheb. posobie [Practices for Analyzing Qualitative Data in the Social Sciences: A Training Manual]. Moscow: Izd. dom HSE Publ., 2023. (In Russian)

Этические принципы и использование искусственного интеллекта в здравоохранении: руководство ВОЗ. Резюме. Женева: Всемирная организация здравоохранения, 2021.

Eticheskie principy i ispol’zovanie iskusstvennogo intellekta v zdravoohranenii: rukovodstvo VOZ. Rezyume [Ethics and Governance of Artificial Intelligence for Health: WHO Guidance. Executive Summary]. Geneva: World Health Organisation Publ., 2021. (In Russian)

Aizenberg, E., Van Den Hoven, J. “Designing for Human Rights in AI”, *Big Data & Society*, 2020, No. 2, pp. 1–30.

Anderson, B., Sutherland, E. “Collective Action for Responsible AI in Health”, *OECD Artificial Intelligence papers*, 2024, No. 10, pp. 3–41.

Asilomar AI Principles [https://futureoflife.org/2017/08/11/ai-principles/, accessed on 28.02.2024].

Atari, M., Haidt, J., Graham, J., Koleva, S., Stevens, S.T. “Dehghani M. Morality Beyond the WEIRD: How the Nomological Network of Morality Varies Across Cultures” (Preprint), *PsyArXiv*, 2022, pp. 1–79.

Baber, N. “International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH)”, *British Journal of Clinical Pharmacology*, 1994, No. 37 (5), pp. 401–404.

Convention for the Protection of Human Rights and Dignity of the Human Being with regard to the Application of Biology and Medicine: Convention on Human Rights and Biomedicine, 1997. [https://rm.coe.int/168007cf98, accessed on 28.02.2024].

Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems [https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf, accessed on 28.02.2024].

European Group on Ethics in Science and New Technologies. Statement on Artificial Intelligence, Robotics and ‘Autonomous’ Systems. Luxemburg: Publication Office of European Union, 2018.

Floridi, L., Cowls, J. “A Unified Framework of Five Principles for AI in Society”, *Machine Learning and the City: Applications in Architecture and Urban Design*, ed. by S. Carta. London: John Wiley and Sons Ltd., 2022, pp. 535–545.

Gibson, K. *Ethics and Business: An Introduction*, Cambridge: Cambridge UP, 2007.

Graham, J., Nosek, B.A., Haidt, J., Iyer, R., Koleva, S., Ditto P.H. “Mapping the Moral Domain”, *Journal of Personality and Social Psychology*, 2011, Vol. 101, pp. 366–385.

Haidt, J., Joseph, C. “Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues”, *Daedalus*, 2004, No. 4, pp. 55–66.

Kazim, E., Koshiyama, A.S. “A High-Level Overview of AI Ethics”, *Patterns*, 2021, No. 9, pp. 1–12.

Mikhail, J. “Universal Moral Grammar: Theory, Evidence and the Future”, *Trends in Cognitive Sciences*, 2007, Vol. 11 (4), pp. 143–152.

Montreal Declaration for a Responsible Development of Artificial Intelligence [https://www.montrealdeclaration-responsibleai.com/the-declaration, accessed on 28.02.2024].

PAI's Guidance for Safe Foundation Model Deployment, *Partnership on AI (PAI)*. 2023. [https://partnershiponai.org/about/, accessed on 28.02.2024].

Schein, C., Gray, K. "The Theory of Dyadic Morality: Reinventing Moral Judgment by Redefining Harm", *Personality and Social Psychology Review*, 2018, Vol. 22 (1), pp. 32–70.

Select Committee on Artificial Intelligence. *AI in the UK: Ready, Willing, and Able?* London: House of Lords, 2018.

Smajdor, A., Herring, J., Wheeler, R. *Oxford Handbook of Medical Ethics and Law*. Oxford: Oxford UP, 2022.

Trials of War Criminals before the Nuremberg Military Tribunals under Control Council, Law. Washington, D.C.: U.S. Government Printing Office, 1949, No. 10, Vol. 2.

WMA International Code of Medical Ethics, 2022. [https://www.wma.net/policies-post/wma-international-code-of-medical-ethics/, accessed on 28.02.2024].

Wong, D.B. *Moral Relativism and Pluralism*. Cambridge: Cambridge UP, 2023.

World Medical Association Declaration of Geneva, 2006. [https://www.wma.net/policies-post/wma-declaration-of-geneva/, accessed on 28.02.2024].

"World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects", *JAMA*, 2013, No. 20, pp. 2191–2194.

Zagzebski, L.T. "Exemplarist Virtue Theory", *Metaphilosophy*, 2010, Vol. 41, No. 1–2, pp. 41–57.