DOI: 10.31857/S023620070026109-6

©2023 Г.А. КАРПОВА

ЭТИЧЕСКАЯ ЭКСПЕРТИЗА В СФЕРЕ ЗДРАВООХРАНЕНИЯ С ПРИМЕНЕНИЕМ ИИ: МЕТОДОЛОГИЯ ДИЗАЙНА ЦЕННОСТЕЙ

Карпова Елизавета Александровна — магистр Школы философии, стажер-исследователь проекта «Этическая экспертиза в сфере ИИ».
Национальный исследовательский университет «Высшая школа экономики».
Российская Федерация, 105066 Москва, ул. Старая Басманная, д. 21/4.
ORCID: 0009-0005-0499-7930
ea.karpova@hse.ru

Аннотация. По мере того как наш мир становится все более зависим от данных, алгоритмы чаще и чаще используются для принятия обоснованных решений в различных областях, начиная от финансов и заканчивая управлением персоналом. Сфера здравоохранения не является исключением, и искусственные интеллектуальные системы получают все большее распространение в этой области. В то время как искусственный интеллект (ИИ) может помочь нам принимать более обоснованные и эффективные решения, он также служит источником множества

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.

УДК 17

моральных и этических проблем. Одним из самых глобальных является вопрос доверия. Когда машина заменяет человека в сфере здравоохранения, пациентам и медицинским работникам может стать трудно доверять тому процессу и результатам лечения, которые предлагает ИИ. Кроме того, механизм «черного ящика» в системах искусственного интеллекта ставит вопросы: кто несет ответственность за принятые решения и кто должен за них отвечать в случае ошибки? К этим вопросам добавляется проблема эмоциональной фрустрации — разрыв контакта «человек—человек» рискует ввести пациентов и медицинских работников в состояние эмоционального разочарования от неспособности вести живую коммуникацию с роботизированным механизмом и, как следствие, привести к падению эффективности лечения и комплаентности.

Несмотря на увеличение внимания к этим проблемам в последние годы, технические решения этих сложных моральных и этических вопросов часто разрабатываются без учета социального контекста и мнения агентов, которых затрагивает технология. Кроме того, призывы к более этичному и социально ответственному ИИ часто фокусируются на основных правовых принципах, таких как «прозрачность» и «ответственность», и не учитывают намного более проблематичную область человеческих ценностей. Для решения этой проблемы в статье предлагается рассмотреть "ценностно-чувствительный" подход к разработке ИИ, который может помочь перевести основные права и ценности человека в контекстно-чувствительные требования к алгоритмам искусственного интеллекта. Этот подход может помочь создать маршрут от человеческих ценностей к ясным и понятным требованиям для дизайна ИИ. Помимо представления ключевых требований к алгоритмам этико-чувствительного ИИ, предлагается посмотреть, какие философские истоки имеют те ценности, о которых мы говорим сегодня в рамках обсуждения этичного искусственного интеллекта: честность, добросовестность, ответственность, эмпатия, долг перед жизнью. Автор предлагает рассмотреть их сквозь призму этики добродетели и деонтологии.

Ключевые слова: биоэтика, искусственный интеллект, дизайн ценностей, шесть принципов ИИ, проблема «черного ящика», ценностные сценарии, ценностно-чувствительный подход, совместное проектирование, дорожная карта, этичный ИИ.

Ссылка для цитирования: Карпова Е.А. Этическая экспертиза в сфере здравоохранения с применением ИИ: методология дизайна ценностей // Человек. 2023. Т. 34, № 3. С. 129–145. DOI: 10.31857/S023620070026109-6

Алгоритмы искусственного интеллекта (ИИ) различной сложности все быстрее становятся частью повседневности, проникая в жизнь каждого отдельного человека и общества в целом. Веря в обещания эффективности, надежности и автоматизации принятия безоговорочно верных решений, частные

компании и правительства включают ИИ во все более широкий круг общественных процессов и институтов, таких как аналитика потребительских предпочтений, управление персоналом, оценка кредитоспособности, формирование цен на страхование жизни и, конечно, здравоохранение. Прогнозируемые успехи и удачные кейсы применения искусственного интеллекта в различных сферах жизни не могут не распространять энтузиазм и веру в скорое установление симбиоза человека и машины. Однако вместе с тем все большее развитие систем искусственного интеллекта и попытки создания сильного ИИ выносят на свет все большее количество трудноразрешимых морально-этических проблем, и сфера медицины отнюдь не стала исключением. Дискриминация, нарушение конфиденциальности, распространение дезинформации, необоснованные решения, вопросы безопасности — все это только вершина айсберга, одной стороной находящаяся в правовой сфере и оттого артикулируемая чуть яснее. Куда глубже залегают проблемы морально-этического спектра — проблема «черного ящика», разрыв контакта «человек–человек» и эмоциональная фрустрация, проблема индивидуальной свободы выбора, ограничение автономии, а также проблема решения конфликтов предсказаний «человек–машина» и другие возникающие вопросы. Ответы на эти морально-этические вопросы сложно найти в областях юриспруденции или социологии, точно так же, как и получить их от разработчиков ИИ и создателей алгоритмов. Отсюда вытекает необходимость междисциплинарного диалога, привлекающего к обсуждению области знания, непосредственно занимающиеся изучением моральных ценностей и социальных норм, — философии и этики.

Создание этичного искусственного интеллекта — сложная задача, требующая глубокого понимания принципов работы ИИ. Эти принципы, в том числе прозрачность, подотчетность, инклюзивность и гибкость необходимы для обеспечения этичного и ответственного внедрения ИИ в рабочие процессы и социальную сферу. Однако простого перечисления этих принципов недостаточно. Чтобы действительно создать этичный ИИ, необходимо вовлечь в процесс разработки все заинтересованные стороны. В случае здравоохранения к ним относятся разработчики, компании и государственные учреждения, врачи, пациенты и другие стороны, такие как члены семей пациентов и поставщики медицинских услуг. Одной из методологий, которую можно использовать для разработки этичного ИИ, является ценностно-чувствительный дизайн [Friedman, 2008]. Этот подход включает в себя изучение и включение этических ценностей прямых и косвенных заинтересованных сторон в процесс проектирования. Другой подход — совместное

проектирование [Van der Velden, 2015], которое предполагает активное вовлечение всех ключевых заинтересованных сторон в процесс разработки систем ИИ, отвечающих их этическим требованиям и ценностям. Одним из ключевых аспектов этого подхода является необходимость создания алгоритмов, учитывающих социальный контекст, в который будет встроена технология. Это поможет сместить акцент с технологических артефактов и технических требований на социальные нужды будущих пользователей технологий ИИ. Для создания такого социотехнического подхода необходимо учитывать как взаимодействие между людьми и техническими артефактами, так и между людьми и организационными практиками. Примечательно, что авторы ценностно-чувствительного подхода к дизайну используют Хартию Европейского союза по правам человека [Charter, 2012] как основной источник общечеловеческих ценностей. Однако они сознательно уходят от рассмотрения этих ценностей через призму правового поля, концентрируясь вместо этого на морально-этических аспектах. Это различие существенно, поскольку вовлечение заинтересованных сторон в процесс развития на основе ценностей (а не установленных законов и стандартов) может привести к требованиям развития, выходящим за пределы существующих правовых рамок. Кроме того, как мы увидим во второй части статьи, эти ценности, на которых должен быть основан этико-ориентированный ИИ, прямо вытекают из двух классических философских теорий — этики добродетели и деонтологии.

Дизайн этических ценностей

В общем смысле дизайн ценностей представляет собой перевод морально-этических требований заинтересованных сторон в контекстно-зависимые требования к дизайну разрабатываемого алгоритма или системы искусственного интеллекта. Постепенный перевод абстрактных значений в требования к дизайну называется спецификацией и может быть визуально отображен в виде иерархии значений (рис. 1а).

В первую очередь ценности расширяются до норм, которые представляют собой свойства или возможности, которые должна демонстрировать разработанная технология, чтобы поддерживать желаемые ценности. Например, ценность «уважение частной жизни» (рис. 1б) может в определенном контексте интерпретироваться как информированное согласие на определенные действия, конфиденциальность или право на удаление нежелательной информации. Таким образом, эти три нормы рассматриваются

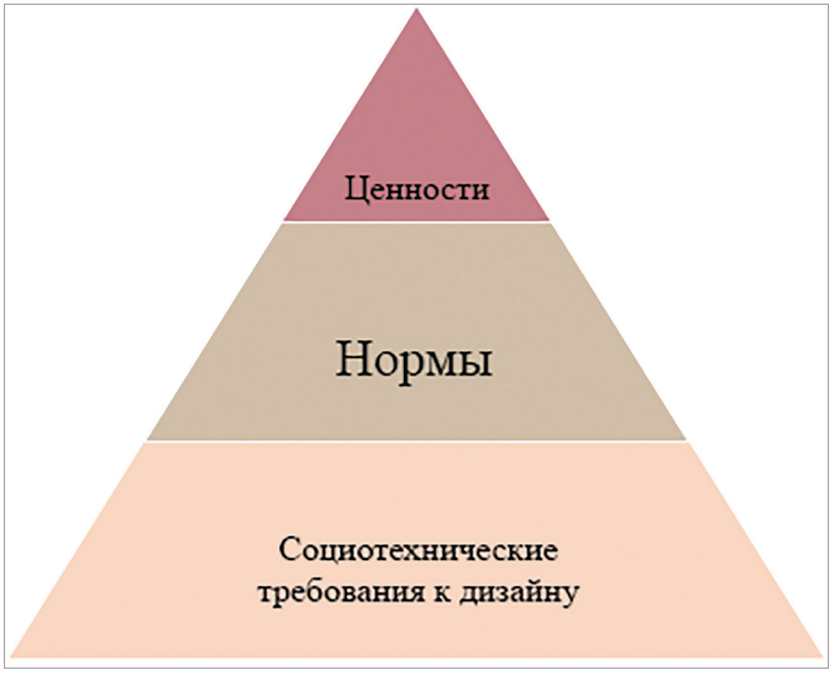

Рис. 1а [Aizenberg, 2020: 5].

заинтересованными сторонами как спецификация того, что означает ценность «уважение частной жизни» в данном контексте использования.

Каждая из этих норм, в свою очередь, уточняется в требованиях социально-технического проектирования. На рис. 1б информированное согласие на обработку персональных данных должно

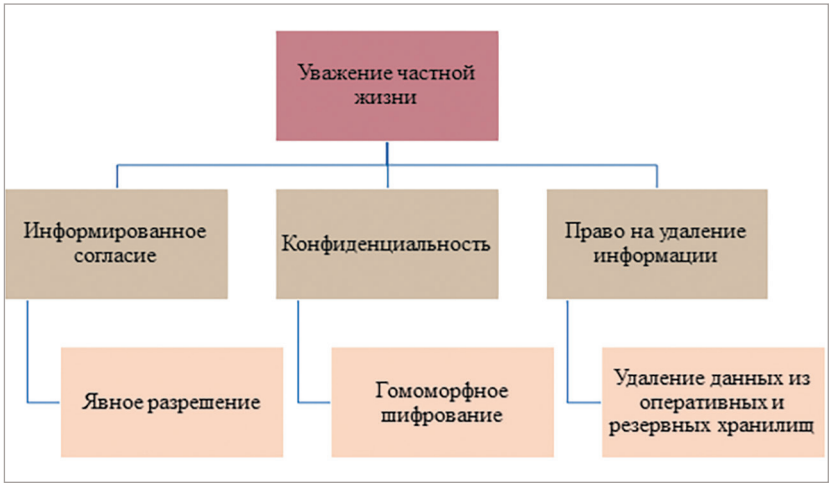

Рис. 1б [Aizenberg, 2020: 5].

быть реализовано как положительное согласие (требующее явного разрешения), а конфиденциальность — с точки зрения методов гомоморфного шифрования.

Важно понимать, что связь между более высокими и более низкими уровнями в иерархии не является дедуктивной. То есть на более низком уровне не просто выводятся следствия из более высоких уровней. Например, нормы, приведенные в примере, могут быть недостаточными или совершенно неуместными в другом контексте использования. Стороны дебатов о частной жизни могут не согласиться с предложенным разделением. Однако иерархия ценностей предлагает структуру для целенаправленного обсуждения этих разногласий и позволяет оценивать отношение между более высокими и более низкими уровнями в контексте, в котором они используются. Кроме того, спецификация обеспечивает объяснение абстрактного значения и всегда зависит от контекста. Это означает, что для каждого конкретного случая нужно учитывать нюансы и особенности контекста, чтобы более точно определить, каким образом элемент низшего уровня связан с элементом более высокого уровня.

В обратном направлении связь между элементом более низкого уровня (*L*) в иерархии и элементом более высокого уровня (*H*) характеризуется отношением «*L* ради *H*». Это означает, что элемент *L* выполняется ради достижения элемента *H*, однако из этого не вытекает, что это единственное возможное отношение между ними. Фактически элемент *L* может воплощать ряд более конкретных отношений [ibid.]:

1. *L* является средством для достижения *H*;
2. *L* — цель, достижение которой способствует достижению *H*;
3. *L* позволяет достичь *H*, не внося прямого вклада в это достижение;
4. *L* устраняет препятствие для достижения *H*.

Изучение иерархии ценностей и отношений между элементами на разных уровнях может помочь в осознании того, как различные элементы взаимосвязаны между собой, что может быть полезно для принятия решений и определения приоритетов.

Каждое из этих отношений может быть дополнительно проработано заинтересованными сторонами в конкретном контексте, чтобы определить социальные, этические и моральные соображения стейкхолдеров и их мотивы. Например, дополнительная проработка в конкретном случае позволяет заинтересованным сторонам прийти к выводу о том, что право на удаление нежелательной информации является средством сохранения уважения к частной жизни. В этой связи отношение «*L* ради *H*» отражает контекстуальные рассуждения и мотивы заинтересованных сторон, которые

требуют, чтобы элемент *H* более высокого уровня, такой как уважение частной жизни, поддерживался элементом *L* более низкого уровня, например правом на удаление нежелательной информации. Таким образом, тонкая, контекстно-зависимая, сформулированная заинтересованными сторонами аргументация требований к дизайну является важным механизмом для преодоления социотехнического разрыва. Она способствует инклюзивному, социально осознанному, междисциплинарному разговору, который необходим для учета широты и глубины сложных социально-этических проблем и их значения.

Иерархия ценностей показывает в структурированной и прозрачной форме, какие моральные и социальные ценности, то есть нормативные требования, должна поддерживать рассматриваемая технология, каковы интерпретации этих ценностей заинтересованными сторонами в конкретном контексте использования и какие соответствующие требования к дизайну должны быть реализованы. Это позволяет всем лицам, затронутым работой технологии (непосредственным пользователям, инженерам, полевым экспертам, практикующим юристам и т.д.), обсуждать варианты разработки, чтобы проследить причины, преимущества и недостатки каждого выбора в соответствии с общественными нормами и ценностями.

Одним из способов включения в процесс разработки этико-ориентированного ИИ с учетом человеческих ценностей является метод ценностных сценариев. Ценностные сценарии — это воображаемые ситуации с нюансами, в которых участвуют предлагаемая технология ИИ и различные прямые и косвенные заинтересованные стороны. Создание таких сценариев позволяет исследователям совместно со стейкхолдерами проработать различные варианты взаимоотношений между человеком и ИИ, прислушиваясь к мнению каждого участника исследования. Это способствует созданию более этичных, социально ориентированных продуктов. Важно отметить, что изучение альтернативных взглядов предполагает переосмысление того, что может показаться очевидным техническим решением сложной социально-этической проблемы. Это позволяет заинтересованным сторонам, «которые в противном случае могли бы быть невидимыми или маргинализированными, иметь право голоса в процессе проектирования» [Van der Velden, 2015: 53]. Такой подход к разработке продукта не только учитывает интересы всех заинтересованных сторон, но и позволяет создать продукты, которые активно удовлетворяют потребности пользователей, в то время как организации получают прибыль.

Несколько раз подчеркивалась важность спецификации ценностей, связанной социальным контекстом, в котором эта

ценность будет реализована через программные требования к ИИ. Различные авторы акцентируют внимание на моральной важности приспособления к местным значениям социальных практик и языка [см., напр.: MacIntyre, 1981]. Действия и язык имеют значение в конкретной социальной практике или институциональной среде, которые, в свою очередь, устанавливают собственные цели и задачи, а также нормы, регулирующие поведение в этой среде. Наши представления и дизайн должны уважать эти социальные и нормативные контексты. Социальный контекст играет важную роль в развитии и использовании искусственного интеллекта. Для того чтобы ИИ был эффективным и приносил пользу людям, он должен учитывать социальные и культурные особенности тех мест, где он используется. Это означает, что разработка систем и технологий ИИ должна быть основана на учете уникальных социально-культурных контекстов. Например, в разных культурах могут существовать разные представления о том, что считается этически правильным или неправильным. Поэтому при создании систем ИИ необходимо учитывать эти различия и разрабатывать индивидуальные подходы для каждой конкретной культуры.

В связи с этим ключевой вопрос, который часто возникает при проектировании ценностей в ИИ, заключается в том, в какой степени система ИИ должна заменять действия человека при выполнении определенной деятельности с учетом социальной, нормативной и институциональной цели этой деятельности. Концепция, которая может помочь ответить на этот вопрос, называется «The Nature of Activities» («Природа действий») [Santoni de Sio, 2016]. Автор этого подхода, Филиппо Сантони, предлагает разделение деятельности на целенаправленную и практически ориентированную. Чтобы проиллюстрировать это деление, рассмотрим пример использования роботов-помощников для подъема пациента с больничной койки, изученный Ф. Сантони. Эта деятельность является целенаправленной, поскольку ориентирована на безопасный подъем пациента с кровати под определенным углом и скоростью для его дальнейшего размещения в инвалидном кресле. В то же время выполнение этой деятельности человеком, осуществляющим уход, имеет практико-ориентированный аспект развития связи и отношений между врачом и пациентом. Эти отношения имеют социальную ценность для пациента, а также важны для долгосрочного ухода за ним и установления доверительных отношений с целью проведения наиболее эффективного лечения.

Использование этой концепции двух видов действия при изучении контекста помогает преодолеть социально-технические пробелы, раскрывая ценные взаимодействия между людьми, которые в противном случае могли бы остаться незамеченными. Это

может помочь определить, какие задачи (или подзадачи) лучше делегировать системе ИИ, а какие — человеку. Фактически может возникнуть инновационное взаимодействие между людьми и ИИ, когда разделение задач между ними в большей степени удовлетворяет потребности и ценности заинтересованных сторон по сравнению с тем, когда человек или ИИ выполняют эту деятельность в одиночку. В итоге использование систем и технологий ИИ должно быть основано на учете уникальных социально-культурных контекстов. Разработчики ИИ должны учитывать этот факт и принимать во внимание местные ценности и культурные особенности, чтобы создать системы, которые могут быть эффективными и полезными для всех людей.

«Дорожная карта»: от ценностей к требованиям

В концептуальном смысле «дорожная карта» перехода ценностей в нормы, а затем в конкретные требования к искусственному интеллекту представляет собой сложную иерархию. Эта иерархия включает в себя несколько уровней, которые последовательно отображают ценности, требования и контекст, в котором действует искусственный интеллект. На верхнем уровне находятся человеческие права, которые являются главным критерием оценки того, насколько этичен и надежен тот или иной продукт, работающий на базе искусственного интеллекта. На следующем уровне расположены моральные ценности, которые часто определяются на основе мнения общества и проводимых исследований, в которых участвуют заинтересованные стороны. Как мы увидим дальше, эти ценности являются прямым отражением добродетелей и норм, которые мы встречаем в теориях моральной философии — аретической этике и деонтологии. Далее, на уровне контекстных спецификаций, уточняются конкретные ситуации, в которых будет использоваться искусственный интеллект. Это позволяет разработчикам точнее определить требования к продукту, учитывая особенности применения и контекст. Наконец на последнем уровне расположены социально-технические требования к программам и алгоритмам. Они определяют, как продукт должен работать на практике и каким образом должны реализовываться ранее определенные ценности и требования.

Необходимо отметить, что авторы концепции дизайна ценностей предлагают рассматривать ограниченный список человеческих прав, принятых в Хартии, из которых впоследствии выводятся и морально-этические ценности. К таким правам авторы относят свободу, равенство, солидарность, право на жизнь и

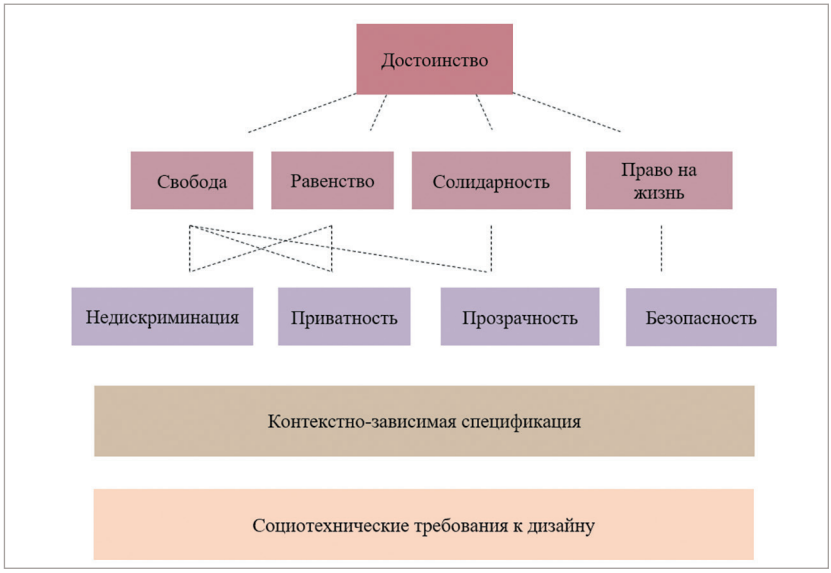

Рис. 2 [Aizenberg, 2020: 18].

достоинство, которое они считают фундаментальной ценностью и ставят выше всех других в иерархии. Сложности, возникающие при попытках дать точное и единственное определение для «достоинства», подтолкнули авторов концепции к использованию негативного подхода через описание трех базовых нарушений этой ценности:

1. Унижение: постановка человека в состояние беспомощности, внушение ощущения ничтожества, потеря автономии над собственной репрезентацией.

2. Инструментализация: отношение к человеку как к предмету обмена и просто средству для достижения цели.

3. Умаление вклада: создание ощущения ненужности, непризнание человека, его стремлений и потенциала.

Влияние ИИ на «достоинство» можно более четко увидеть в этих трех измерениях. Например, дискриминация, непрозрачность действий и проблема «черного ящика», которые могут возникнуть при использовании ИИ в медицине, влекут за собой эмоциональную фрустрацию как пациентов, так и врачей, которые не чувствуют себя в полной мере автономными в принятии решений, касающихся собственного здоровья и здоровья пациентов. Последствия автоматизации на рынке труда могут сделать целые сообщества работников излишними, с перспективой их обмена на более прибыльную альтернативу ИИ. Это, несомненно, является еще одним фактором, подрывающим базовый принцип «достоинства» и влекущим за собой формирование новых этических челленджей: проблемы

решения конфликта предсказаний «человек–машина», проблемы индивидуальной свободы выбора, эмоциональное беспокойство от разрыва контакта «человек–человек» и др.

Как показано на рис. 2, ценности второго порядка в дорожной карте дизайна могут быть конкретизированы в ценности третьего порядка, которые нам знакомы по дискуссиям, связанным с этикой ИИ: недискриминация, конфиденциальность, прозрачность и безопасность. Этот список не является исчерпывающим и открыт для расширения и уточнения. Обратите внимание, что некоторые из этих значений третьего порядка выполняют связь «*L* ради *H*» в отношении более чем одного значения второго порядка. Например, ценность конфиденциальности поддерживает как свободу, так и равенство, а прозрачность — как свободу, так и солидарность. С этого момента дальнейшая спецификация этих значений и их преобразование в системные свойства становятся сильно зависимыми от контекста. Именно здесь методы контекстуальной дифференциации при решающем участии заинтересованных сторон в обществе вступают в действие, чтобы выявить нормативные и социально-технические последствия ценностей более высокого порядка в конкретном контексте. Как отмечалось ранее, этот процесс проектирования требует повторения нескольких раундов концептуальных, эмпирических и технических исследований. С каждой итерацией будет появляться более тонкое и широкое понимание нормативных требований, ценностных противоречий и социально-технических требований к дизайну.

Следует ожидать, что в некоторых контекстах и сценариях заинтересованные стороны могут прийти к выводу, что ни одна из предложенных технических реализаций не удовлетворяет в достаточной мере важным ценностным требованиям. С одной стороны, это может быть возможностью для технических инноваций, мотивирующих нестандартное мышление, которое может привести к успешному решению. С другой стороны, мы не должны попасть в ловушку решимости, предполагая, что та или иная форма ИИ всегда является панацеей от проблемы или поставленной цели. На самом деле в ситуациях, когда морально-этические требования не могут быть согласованы с подходящей реализацией, иерархический характер «дорожной карты» ценностей предоставляет средства для точного отслеживания того, какие нормы, ценности и, в конечном счете, права человека будут недостаточно поддерживаться или нарушаться. В таких сценариях ответственное проектирование в отношении прав человека потребует признания того, что внедрение ИИ в этом контексте будет вредным. Идея отказа от решений ИИ, которые не соответствуют в достаточной мере нормативным ценностным требованиям, не будет

отличаться от того, как другие технологии, такие как автомобили, самолеты, бытовая техника и мосты, должны соответствовать различным стандартам сертификации, целью которых является поддержка таких ценностей, как безопасность, доступность и защита окружающей среды.

«Дорожная карта», описанная в данной статье, открывает ряд практических вопросов, касающихся ее реализации. Например, как и кто определяет соответствующие заинтересованные стороны в определенном контексте? Как оценить консенсус между заинтересованными сторонами? Как и кто определяет, было ли выполнено достаточное количество итераций в рамках концептуальных, эмпирических и технических исследований? Есть ли потребность во внешнем органе или учреждении, которое удостоверяло бы честность процесса проектирования и обеспечивало бы его прозрачность по отношению к обществу? На некоторые из этих вопросов можно ответить с помощью знаний и опыта, накопленных в ходе прошлых применений дизайна для ценностей в различных областях [Azenkot, 2011]. Однако многие из этих вопросов уводят нас в неизведанные воды, как и сам ИИ. Одним из наиболее волнующих становится вопрос о том, какие именно ценности мы должны транслировать и принимать как основополагающие для создания этичного ИИ? Основываясь на Хартии Европейского союза по правам человека, авторы статьи предлагают нам ограниченное число основополагающих прав, на базе которых с помощью ценностных сценариев разработчики должны выводить контекстно-ориентированные ценности. При этом, однако, остается не затронутым вопрос о существовании таких ключевых ценностей, которые с необходимостью должны присутствовать в каждом алгоритме искусственного интеллекта, дабы избежать тотального морального релятивизма. Для установления такого «списка» ценностей необходимо обратиться к классическим концепциям этики добродетели и деонтологии, которые уже нашли свое широкое применение в сфере медицины и могут быть экстраполированы дальше на область искусственного интеллекта в здравоохранении.

Истоки этичного ИИ — этика добродетели и деонтология

При попытке определить, откуда же нам черпать те ценности, о которых шла речь в предыдущих разделах, мы упираемся в давно знакомые нам философские концепции, ставшие хрестоматийными для этической теории. Речь идет об этике добродетели и деонтологии, которые нашли свое применение не только в моральной

философии, но и, в частности, используются в медицинской практике. Этика добродетели — философский подход, который делает упор на развитии нравственных черт характера и добродетелей, таких как сострадание, мудрость и честность, а не фокусируется исключительно на правилах или последствиях. При применении в контексте искусственного интеллекта в здравоохранении этика добродетели может обеспечить ценную основу для руководства разработкой и внедрением технологий ИИ таким образом, чтобы способствовать достижению этических и гуманных результатов. В конце концов именно на добродетелях честности, сострадания и эмпатии и базируются те самые ценности, которыми должен быть наделен этико-ориентированный ИИ.

Одним из ключевых достоинств, которое особенно важно для ИИ в здравоохранении, является эмпатия. Поставщики медицинских услуг и исследователи давно осознали важность эмпатии в оказании эффективной помощи и построении прочных отношений с пациентами. Однако эмпатию может быть сложно воспроизвести на машинах, которые не способны испытывать эмоции или реагировать на пациентов по-человечески. Несмотря на эти ограничения, по-прежнему возможно разрабатывать системы ИИ, которые отражают и продвигают эмпатические ценности. Например, чат-боты на базе искусственного интеллекта можно запрограммировать так, чтобы они отвечали на запросы и проблемы пациентов с сочувствием и пониманием, используя обработку естественного языка для участия в человеческом разговоре. Точно так же алгоритмы ИИ можно научить распознавать и реагировать на невербальные сигналы, такие как выражение лица и язык тела, таким образом, чтобы это отражало чуткое понимание потребностей и эмоций пациентов.

Еще одним ключевым достоинством, имеющим отношение к ИИ в здравоохранении, является мудрость. Разработка и внедрение технологий искусственного интеллекта требуют тщательного рассмотрения широкого круга этических и практических соображений, включая вопросы, связанные с конфиденциальностью, согласием, предвзятостью и потенциальными непредвиденными последствиями. Этика добродетели в данном случае рассматривается как гарант того, что системы ИИ разрабатываются и используются с учетом мудрой и вдумчивой совместной работы проектировщиков и этиков. Например, алгоритмы ИИ могут быть разработаны таким образом, чтобы свести к минимуму предвзятость и дискриминацию, используя разнообразные и репрезентативные наборы данных, а также постоянно отслеживая и корректируя их результаты для обеспечения справедливости и равноправия. Точно так же системы ИИ могут разрабатываться с упором на прозрачность и объяснимость, позволяя пациентам и поставщикам услуг

понимать, как принимаются решения, и участвовать в процессе их принятия.

В целом применение этики добродетели при разработке и внедрении ИИ в здравоохранении может помочь обеспечить использование этих технологий таким образом, который отражает наши самые высокие моральные ценности и устремления. Сосредоточив внимание на таких достоинствах, как эмпатия, мудрость и сострадание, мы можем создавать системы ИИ, которые способствуют этическим и гуманным результатам и поддерживают здоровье и благополучие пациентов и общества.

Помимо сочувствия и мудрости, другие достоинства, которые имеют отношение к алгоритмам ИИ в здравоохранении, включают честность, добросовестность и ответственность. Эти добродетели могут помочь обеспечить прозрачность, безопасность и подотчетность систем ИИ, которые впоследствии оформляются в виде ценностных требований к разработке алгоритмов. Например, механизмы искусственного интеллекта могут быть разработаны таким образом, чтобы предоставлять честную и точную информацию пациентам и поставщикам медицинских услуг без искажения или манипулирования данными таким образом, который может ввести в заблуждение или нанести вред. Точно так же системы ИИ могут разрабатываться с упором на конфиденциальность и безопасность данных, обеспечивая защиту информации о пациентах и постоянное соблюдение этических стандартов. В каждом из этих случаев истоком тех ценностей и норм, которые закладываются в ИИ на этапе разработки, являются этические добродетели, релевантные для решения проблем низкого доверия и эмоциональной фрустрации.

Еще одним ключевым аспектом этики добродетели в контексте ИИ в здравоохранении является признание ограничений этих технологий. Несмотря на то что искусственный интеллект может во многих отношениях улучшить результаты лечения, он не является панацеей от всех медицинских проблем и не может заменить человеческий фактор помощи, необходимый для построения доверия и отношений с пациентами. Чтобы решить эту проблему, системы ИИ могут быть спроектированы для работы в тесном сотрудничестве с поставщиками медицинских услуг, поддерживая их в принятии решений и предоставляя им инструменты и информацию, необходимые для оказания высококачественной помощи. Признавая взаимодополняющие роли ИИ и человеческого опыта в здравоохранении, мы можем создавать системы, отражающие действенный и целостный подход к уходу за пациентами.

Не менее важной базовой концепцией, на которой должен быть основан тот дизайн этических ценностей ИИ, о котором идет речь в этой статье, является деонтология как принцип,

сосредоточивающийся на моральных обязательствах и этике долга. Ее основным постулатом является идея уважения к личности. Этот принцип требует, чтобы люди рассматривались как цели сами по себе, а не как средства для достижения цели. В контексте ИИ в здравоохранении это означает, что к пациентам и поставщикам медицинских услуг следует относиться с уважением и достоинством, а их права и автономию необходимо постоянно защищать. Чтобы обеспечить уважение к людям при разработке и внедрении ИИ в здравоохранении, необходимо установить этические нормы и правила, регулирующие использование этих технологий. В этих рекомендациях могут быть указаны условия, при которых ИИ может использоваться для принятия медицинских решений, тип информации, которую можно собирать и хранить, а также способы, которыми пациенты могут дать информированное согласие на использование ИИ в своей работе.

Нельзя забыть и о деонтологическом принципе избегания вреда. В контексте ИИ в здравоохранении это означает, что разработчики и пользователи этих технологий должны предпринять шаги, чтобы гарантировать их безопасность, точность и эффективность, а также то, что они не причиняют вреда пациентам или поставщикам. Например, команда разработки этико-ориентированного ИИ может проводить тщательное тестирование и проверку алгоритмов ИИ, чтобы гарантировать их точность и надежность. Они также могут реализовать меры безопасности и отказоустойчивости, чтобы предотвратить ошибки или сбои, которые могут привести к причинению вреда. Кроме того, они могут установить протоколы для мониторинга и отчетности о нежелательных явлениях, связанных с использованием ИИ в здравоохранении, для выявления и решения любых возникающих проблем.

Как мы видим, концепция этико-ориентированного ИИ, продуманного достаточно, чтобы быть использованной в одной из самых сакральных сфер — сфере здоровья и жизни, должна быть основана на базовых человеческих ценностях — честности, открытости, добросовестности, долга перед жизнью, эмпатии. Все эти ценности, переходящие в нормы и требования по «дорожной карте», имплицитно вшиты и имеют свой исток в классических философских концепциях добродетелей и должного. Требования к этичному ИИ, которому мы можем доверять, которого мы как пациенты, можем не бояться, — это прямое наследие аристотелевской и кантовской традиции, оформленное в практические условия. Говоря об искусственном интеллекте в медицине, мы говорим о машине, наделенной пониманием и знаниями тех базовых ценностей, которые объединяют людей по всему миру. Вот почему дизайн ценностей настаивает на своей универсальности, а

приближение к этико-ориентированному ИИ возможно только с учетом ключевых человеческих ценностей. И все же впереди еще долгий путь. Технологии могут сыграть полезную роль в эволюции прав человека, но, как предупреждает нас обсуждаемая ранее «ловушка доверия», мы не должны вводить себя в заблуждение, полагая, что сложные этические проблемы могут быть решены только с помощью «умных» технологий, которые часто ошибочно считают этически нейтральными и объективными. ИИ, созданный человеком, — это ИИ, наделенный все теми же человеческими предрассудками, ложными суждениями, вероятностями ошибок и неточностей, и, чтобы сделать его, да и себя чуть ближе к объективности, мы всегда должны держать в поле своего зрения те базовые этические концепции, которые дают нам набор ключевых добродетелей и норм. Создавая этичный ИИ, мы одновременно с этим учимся сами быть чуть более чуткими к тем ценностям, которые транслируем машине. В этом симбиозе машины и человека определенно есть положительный потенциал, реализация которого зависит только от нас самих. В конце концов, следуя аристотелевской традиции, мы должны постоянно находиться в поиске золотой середины, балансируя между двумя полюсами ценностного экстремума — релятивизмом и догматизмом.

AI in Health Ethical Review: A Value Design Methodology

Elizaveta A. Karpova
Master of the School of Philosophy, Intern-researcher of the project “Ethical expertise in the field of AI”.
National Research University ″Higher School of Economics″.
21/4 Staraya Basmannaya Str., Moscow 105066, Russian Federation.
ORCID: 0009-0005-0499-7930
ea.karpova@hse.ru

Abstract. As our world becomes more dependent on data, algorithms are increasingly being used to make informed decisions in areas ranging from finance to HR. The healthcare sector is no exception, and artificial intelligence systems are becoming more and more widespread in this area. While AI can help us make more informed and efficient decisions, it also presents many moral and ethical challenges. One of the biggest issues is the issue of trust. When “machine” replaces “human” decision making, it can be difficult for patients and healthcare professionals to trust the outcome. In addition, the “black box” mechanisms in artificial intelligence systems make it unclear who is responsible for the decisions made, which can lead to ethical dilemmas. In addition, there is a risk of emotional frustration for patients and healthcare professionals, as AI may not be able to provide the

kind of human touch that is often needed in healthcare. Despite increased attention to these issues in recent years, technical solutions to these complex moral and ethical issues are often developed without regard to the social context and opinions of the advocates affected by the technology. In addition, calls for more ethical and socially responsible AI often focus on basic legal principles such as “transparency” and “responsibility” and leave out the much more problematic area of human values. To solve this problem, the article proposes a “value-sensitive” approach to the development of AI, which can help translate basic human rights and values into context-sensitive requirements for AI algorithms. This approach can help create a route from human values to clear and understandable requirements for AI design. It can also help overcome ethical issues that hinder the responsible implementation of AI in healthcare and everyday life.

Keywords: bioethics, artificial intelligence, value design, 6 principles of AI, black box problem, value scenarios, value-sensitive approach, collaborative design, roadmap, ethical AI.

For citation: Karpova E.A. AI in Health Ethical Review: A Value Design Methodology // Chelovek. 2023. Vol. 34, N 3. P. 129–145. DOI: 10.31857/S023620070026109-6

Литература/References

Aizenberg E., Van den Hoven J. Designing for human rights in AI. *Big Data & Society*. 2020. Vol. 7, Iss. 2. P.1–30. DOI: 10.1177/2053951720949566

Azenkot S., Prasain S., Borning A. et al. Enhancing independence and safety for blind and deaf-blind public transit riders. *Proceedings of the 2011 annual conference on Human factors in computing systems CHI ’11*. New York, NY, 2011. P. 3247–3256.

Davis J., Nathan L.P. Value Sensitive Design: Applications, Adaptations, and Critiques. *Handbook of Ethics, Values, and Technological Design.* Van den Hoven J., Vermaas P.E., Van de Poel I. (eds). Dordrecht: Springer, 2015. P. 11–40.

Friedman B. Human Values and the Design of Computer Technology. Cambridge: Cambridge Univ. Press, 1997.

MacIntyre A. After Virtue: A Study in Moral Theory. Notre Dame, IN: Univ. of Notre Dame Press, 1981.

Charter of Fundamental Rights of the European Union. *Official Journal of the European Union*. 2012. Vol. 55. P. 391–407.

Santoni de Sio F., Van Wynsberghe A. When Should We Use Care Robots? The Nature-of-Activities Approach. *Science and Engineering Ethics*. 2016. Vol. 22, N 6. P. 1745–1760. DOI: 10.1007/s11948-015-9715-4

Van de Poel I. Translating Values into Design Requirements. *Philosophy and Engineering: Reflections on Practice, 28 Principles and Process.* Michelfelder D.P., McCarthy N., Goldberg D.E. (eds). Dordrecht: Springer, 2013. P. 253–266.

Van der Velden M., Mörtberg C. Participatory Design and Design for Values. *Handbook of Ethics, Values, and Technological Design*, Van den Hoven J., Vermaas P.E., Van de Poel I. (eds). Dordrecht: Springer, 2015. P. 41–66.