Этические проблемы императива объяснимости систем искусственного интеллекта

В. Ю. Перов

Санкт-Петербургский государственный университет

vadimperov@gmail.com

Аннотация

Развитие этики в сфере систем искусственного интеллекта (СИИ) находит своё воплощение в разнообразных этических документах (декларациях, кодексах, рекомендациях и т. д.), которые различаются по источникам создания, статусу, содержанию, используемой терминологии. Такое положение дел, с одной стороны, не только проясняет этические вопросы, но и запутывает их, с другой стороны, может стимулировать пренебрежительное отношение к этическим императивам. В связи с этим, стремление кардинально сократить этические принципы в сфере СИИ выглядит оправданным. В статье рассматривается идея Л. Флориди, который предлагает ограничить этику в сфере СИИ принципами биомедицинской этики, добавив к ним в качестве пятого принцип объяснимости СИИ.

В ходе исследования было обнаружено, что в настоящее время принцип объяснимости (или близкие к нему по содержанию требования) присутствует в большинстве этических документов в сфере СИИ и даже рассматривается как необходимая предпосылка и основа для других этических принципов. Кроме того, объяснимость декларируется в качестве одного из основных моральных императивов этических политик компаний, являющихся лидерами в разработке СИИ (например, IBM, Google).

Проведённый этический анализ показал, что несмотря на привлекательность принципа объяснимости, он не является морально однозначным. Во-первых, для него существует ряд непреодолимых пока технических ограничений. Во-вторых, существуют обоснованные сомнения, что принцип объяснимости СИИ что-то содержательно дополняет к имеющимся четырём принципам биомедицинской этики. В-третьих, следование императиву объяснимости СИИ является парадоксальным и ведет к возможным этическим рискам. Наконец, принцип объяснимости является ассиметричным в плане его агентности и адресности, особенно в отношении пользователей. В совокупности, эти аргументы свидетельствуют о том, что принцип объяснимости, при всей его важности, не может рассматриваться как морально безупречный и обязывающий.

Ключевые слова: этика систем искусственного интеллекта (СИИ), этические кодексы, этические принципы, моральный императив, объяснимый СИИ, парадоксы, этические риски, ответственность, моральные парадоксы, асимметрия информированности

Библиографическая ссылка: Перов В. Ю. Этические проблемы морального императива систем искусственного интеллекта // Информационное общество: образование, наука, культура и технологии будущего. Выпуск 9 (Труды XXVIII Международной объединенной научной конференции «Интернет и современное общество», IMS-2025, Санкт-Петербург, 23 – 25 июня 2025 г. Сборник научных статей). — СПб: Университет ИТМО, 2025. С. 93-102. DOI: 10.17586/3033-5574-2025-9-93-102.

Развитие и широкое внедрение разнообразных технологий на основе систем искусственного интеллекта (СИИ) существенно изменяют взаимоотношения и практики, существующие в различных сферах жизнедеятельности людей. Такие трансформации требуют их нормативного, в том числе этического регулирования. За последние годы в России и мире появилось огромное количество этических документов (деклараций, рекомендаций, кодексов и т. д.), которые призваны стать основой для решения этических проблем, возникающих в связи с разработкой, производством и использованием СИИ. В настоящее время число этих документов, созданных инициативными группами, сообществами инженеров и учёных, международными и общественными организациями, корпорациями и т. д. уже перевалило за несколько сотен и с трудом поддаётся учёту.

С одной стороны, можно порадоваться, поскольку это свидетельствует, что, во-первых, этические вопросы в сфере СИИ вызывают интерес и рассматриваются как важные и существенные, во-вторых, этот интерес не ограничивается теоретическими изысканиями и есть стремление их практического воплощения. Кроме того, многочисленность этических документов может быть обусловлена как множеством продуктов и технологий на основе СИИ, так и разнообразием сфер их использования и применения. Если обратиться в качестве примера к этической деятельности российского Альянса в сфере ИИ, то на их сайте, помимо общего и базового документов «Кодекс этики в сфере искусственного интеллекта» [1], опубликованы: «Этические рекомендации по применению рекомендательных технологий и алгоритмов, основанных на искусственном интеллекте, в цифровых сервисах» [2]; «Подходы к проблеме вагонетки в разрезе беспилотного наземного транспорта» [3]; «Декларация об ответственной разработке и использовании сервисов в сфере генеративного искусственного интеллекта» [4]; «Рекомендации Комиссии по реализации Кодекса этики в сфере ИИ по теме: «Прозрачность алгоритмов искусственного интеллекта и информационных систем на их основе» (2024) [5]. Руководящие принципы в сфере роботов общего назначения [6]; Белая книга этики в сфере искусственного интеллекта [7] и др. В дополнение еще ряд этических кодексов и рекомендаций в связи с использованием СИИ в медицине, праве, образовании и т. д. находятся на стадии разработки.

С другой стороны, обилие этических документов с разными источниками создания, статусами, с многочисленностью этических принципов, с различиями в используемой терминологии и т. п. порождает как ситуацию неопределенности и растерянности, так и возможности пренебрежительного отношения к этическим императивам. Известный специалист в сфере этики СИИ Л. Флориди справедливо отмечает, что чрезмерное количество нормативных предписаний приводит к «рискам создания супермаркета принципов и ценностей, где частные и государственные субъекты могут выбирать тот вид этики, который лучше всего подходит для оправдания их поведения, вместо того, чтобы пересматривать свое поведение и приводить его в соответствие с общепринятыми этическими рамками» [8, с. 262]. Действительно, если обратиться к устоявшимся сферам прикладных и профессиональных этик, то можно обнаружить, что количество базовых этических императивов в них довольно ограничено. Так, сформулированный Р. Мертоном «этос науки» включает четыре этических предписания (коммунизм/коммунализм, универсализм, незаинтересованность, организованный скептицизм) [9], этика исследований с использованием животных базируется на принципах «трёх R»: замена (Replacement), сокращение (Reduction), усовершенствование (Refinement) [10], в биомедицинской этике существует всего четыре общепризнанных фундаментальных принципа: «не навреди», «делай благо», «уважение автономии пациента» и «справедливость» [11]. На этом фоне многообразие этических принципов в сфере СИИ выглядит избыточным. Так, в упомянутом ранее отечественном «Кодексе этики в сфере ИИ» сформулировано более 20 принципов (точное число сложно подсчитать, поскольку

они изложены в «Разделе 1. Принципы этики и правила поведения», и некоторые положения не могут быть однозначно идентифицированы именно как этические принципы [1]. В этом контексте вполне оправдано стремление некоторых исследователей существенно сократить перечень базовых этических императивов. Любопытна в этом отношении широко обсуждаемая идея Л. Флориди о том, что для этики в сфере СИИ достаточно взять четыре принципа биомедицинской этики и дополнить ее пятым принципом объяснимости (англ. — explicability): «Исходя из сравнительного анализа, необходимо добавить новый принцип: *объяснимость*. Объяснимость понимается как в эпистемологическом смысле *понятности* — в качестве ответа на вопрос «как это работает?» — так и в *этическом* смысле подотчётности — как ответ на вопрос «кто несёт ответственность за то, как это работает?» [12, с. 57-58]. На первый взгляд такое предложение выглядит разумным, учитывая, что биомедицинская этика является наиболее теоретически и практически развитой областью прикладных и профессиональных этик, опыт которой может быть использован и в других областях, в том числе и в этике в сфере СИИ. В этом плане следует констатировать, что в настоящее время почти во всех этических документах в сфере СИИ присутствует или чётко сформулированный принцип объяснимости, или же близкие по содержанию и смыслу упомянутые Л. Флориди требования понятности и подотчетности, или другие, например, объяснимость, но другим словом (англ. — explainable), а также прозрачность (англ. — transparency), интерпретируемость (англ. — interpretability) или поднадзорность [1]. При этом объяснимость и близкие к нему принципы иногда рассматриваются как основа для многих других моральных императивов, которые используются в этики сфере СИИ. Например, в «Рекомендации об этических аспектах искусственного интеллекта» ЮНЕСКО написано: «п. 37. Прозрачность и объяснимость работы ИИ-систем нередко является существенным предварительным условием обеспечения уважения, защиты и поощрения прав человека, основных свобод и нравственных принципов…» и «п. 41. Прозрачность и объяснимость тесно связаны с адекватными мерами ответственности и подотчётности, а также с надёжностью систем ИИ» [13].

Считается, что впервые принцип объяснимости СИИ в развернутом и обоснованном виде был предложен в 2016 г. в проекте Управления перспективных исследовательских проектов Министерства обороны США (англ. — Defense Advanced Research Projects Agency, DARPA): «Целью проекта Объяснимый Искусственный Интеллект (англ. — Explainable Artificial Intelligence (XAI)) является создание набора новых или модифицированных методов машинного обучения, создающих объяснимые модели, которые в сочетании с эффективными методами объяснения позволят конечным пользователям понимать, правильно доверять и эффективно управлять новым поколением систем искусственного интеллекта (ИИ)» [14, с. 5]. В дальнейшем термин в виде аббревиатуры XAI закрепился, и идея стала настолько популярной, что ведущие в сфере создания СИИ компании позиционируют принцип объяснимого СИИ как ключевой момент своей деятельности. В качестве примера — информация на сайте IBM, где есть специальный раздел «Что такое объяснимый ИИ?», в котором XAI даже противопоставляется «обычным» AI (ИИ) [15]. Кроме того, именно объяснимость стоит на первом месте в перечне «столпов политики IBM», вместе со справедливостью (англ. — fairnes), надежностью (англ. — robustness), прозрачностью (англ. — transparency) и конфиденциальностью (англ. — privacy). Любопытно, что эти столпы (следовательно, и объяснимость) позиционируется в качестве фундамента для остальных этических принципов в сфере СИИ [16]. Одновременно объяснимость присутствует в перечне «Принципов IBM для доверия и прозрачности» [17], и в качестве столпа для самих СИИ [18].

Приведённые примеры склоняют к тому, что введение принципа объяснимости в качестве морального императива для этики в сфере СИИ является правильным

и морально обоснованным. К преимуществам создания XAI относят то, что это позволяет понять логику работы технологий на основе СИИ, даёт возможность проверить обоснованность решений с использованием СИИ, упрощает процесс поиска ошибок, повышает доверие пользователей к СИИ, обеспечивает формирование критериев для оценки контроля качества функционирования СИИ и т. д. Иными словами, моральный императив объяснимости во многом призван минимизировать возможные, в т. ч. этические риски в сфере СИИ.

Но при внимательном этическом рассмотрении требование объяснимости вызывает достаточно существенные сомнения, которые могут быть сформулированы следующим образом.

1. Технические и программные трудности. Когда компании-разработчики СИИ пропагандируют XAI, то они сами обращают внимание на принципиальные ограничения. Чтобы не зацикливаться на примере с IBM (хотя и на их сайте представлено нечто похожее), можно обратиться к странице сайта Google «Ограничения Vertex Объяснимый ИИ!» [19], где на первое место выдвигаются вопросы качества исходных данных и проблемы с процедурами их предварительной обработки для обучения СИИ (классическая проблема «мусор на входе — мусор на выходе»). Не отрицая важности этих проблем, следует отметить, что это скорее техническая, а не этическая проблема. Для краткого пояснения (по аналогии) можно привести следующий пример. Если рассматривать исходные данные как своеобразные «кирпичи» для строительства и функционирования СИИ, то как и в любом строительстве они должны быть (а) хорошо/ качественно сделаны, в т. ч. отсортированы и отбракованы; (б) правильно уложены, что обеспечивает надёжность и устойчивость строения. Это есть этическое требование к архитекторам и строителям, но не является специфическим моральным требованием для самого строительства («этичности» постройки), а есть то, что сродни общему критерию добросовестности в профессиональной деятельности. Если эти «кирпичи/данные» будут плохими или неправильно уложенными, то строение развалится, а это потребует соответствующего разбирательства и поиска виновных. Аналогичные аргументы применимы к этическим вопросам в сфере СИИ, что свидетельствует о том, что подобные проблемы и способы их решения не являются какими-то исключительными и особенными для этики в сфере СИИ и не требуют выдвижения и обоснования дополнительных моральных принципов и императивов, в т. ч. и императива объяснимости. По большому счёту, мы не очень интересуемся тем, из какого именно кирпича и по какой конкретной технологии построен дом, в котором мы живём. Кроме этого, к проблемам разработки XAI часто относят удорожание разработки объяснимых СИИ, неопределённость и неоднозначность понятия «объяснимости», быстрое развитие технологий на основе СИИ, что усложняет их объяснение и может препятствовать их внедрению и т. д., Но эти трудности не могут считаться особыми проблемами этики в сфере СИИ, подобно тому, как не является моральной проблемой трудности логистики в доставке помидоров на ближайший к нам рынок, хотя это влияет на их цену и может создать нам неудобства. Точно также, возможные технические ограничения с созданием и функционированием объяснимых СИИ, не являясь специфическими моральными проблемами, затрудняют их использование в этически значимых целях.

2. Проблема «пятого принципа». Появление морального императива объяснимости СИИ в качестве дополнения к принципам биомедицинской этики вызывает сомнения у некоторых учёных и практиков. Так, группа исследователей в статье с провокационным названием «Должна ли объяснимость быть пятым этическим принципом этики ИИ?» показали, что требование объяснимости СИИ является важным, но мало что добавляет к имеющимся в биомедицинской этике принципам, поэтому «…использование алгоритмов ИИ *действительно ставит новые аспекты вопросов* объяснимости с этическими последствиями. Тем не менее, мы думаем, что даже если

это так, объяснимость не может рассматриваться в качестве «пятого принципа» в этике ИИ…» [20, p. 132-133]. По их мнению, это связано с тем, что принцип объяснимости и то, что требуется в соответствии с ним, уже содержится в имеющихся принципах биомедицинской этики или следует из них. Справедливости ради нужно отметить, что эта и подобные дискуссии сосредоточены не на самой этике в сфере СИИ, а на этических вопросах использования СИИ в медицине, то есть на особенностях именно биомедицинской этики. Тем не менее, высказанные аргументы порождают разумные сомнения в позиционировании морального императива объяснимости в качестве столь важного и обязательного.

3. Парадоксальность. Анализ принципа объяснимости обнаруживает, что следование ему сопровождается многочисленным парадоксам, игнорирование которых ведет к возможным этическим рискам. Не претендуя на полноту перечня, можно выделить несколько наиболее показательных разнонаправленных тенденций прямолинейного и настойчивого следования требованию объяснимости СИИ:

- парадокс «чёрного ящика»: возникновение морального императива объяснимости СИИ во многом мотивировано так называемым эффектом «чёрного ящика», под которым понимается невозможность или затруднительность отслеживания и контролирования того, как технологии СИИ выдают результаты. Поэтому формулируется идея, что нужно создавать такие СИИ, которые должны сами объяснять процессы получения ими результатов. Но это не гарантирует того, что выдаваемые СИИ объяснения в свою очередь не будут включать элементы «черного ящика». В таких случаях возможна ситуация возникновения «дурной бесконечности» самообъяснений деятельности СИИ, каждая из которых будет требовать объяснений следующего уровня, но при этом все они будут основаны на «черном ящике». Получается, что одна необъяснимость может порождать другую, но при этом возникает иллюзия объяснимости;
- парадокс прозрачности: чем более объяснимы СИИ, тем проще их взломать и манипулировать ими, в т. ч. для достижения этически неприемлемых целей. Требование безусловной объяснимости подразумевает доступ ко всей, в т. ч. критически важной для функционирования СИИ информации. Другими словами, есть риски снижения безопасности при увеличении открытости;
- парадокс простоты: существует риск снижения эффективности функционирования системы ради её понятности и объяснимости, и может возникнуть соблазн упростить технологии СИИ до ущербного уровня;
- парадокс честности: при объяснении решений технологии СИИ могут раскрывать конфиденциальные данные, и возникают риски утечки этически важной и морально чувствительной информации.

Перечень подобных парадоксов и этических рисков может быть продолжен. Их объединяет то, что прямолинейное следование принципу объяснимости СИИ сродни тому, что в юриспруденции называется «злоупотребление правом», когда соблюдение формальных правил и норм ведёт к негативным юридическим и этическим последствиям. Но если в обществе существует соответствующая практика правовой и судебной деятельности, которые призваны бороться и пресекать такие явления, то в отношении объяснимого СИИ ничего подобного пока нет и вряд ли появится в обозримом будущем. Это означает, что все решения, связанные с объяснимостью технологий на основе или с использованием СИИ возлагаются на акторов СИИ (разработчиков, производителей, поставщиков, операторов и т. д.) как моральных агентов. Именно с этим связан следующий комплекс проблем морального императива объяснимости СИИ.

4. Проблемы агентности и адресности этических требований. Во-первых, стоит напомнить, что особенностью прикладных и профессиональных этических кодексов и сформулированных в них требованиях является их субъектная определённость — они

предполагают понятно очерченный круг тех лиц, которые выступают моральными агентами (акторами) их соблюдения. Так, нормы судейской этики распространяются на судей и сотрудников суда, требования журналисткой этики на журналистов, императивы биомедицинской этики на медицинских работников и т. д. Ни в одном из подобных этических документов нет моральных норм, предназначенных для подсудимых, читателей или зрителей журналистских материалов, пациентов врачей. Для этого есть вполне понятные аргументы, обсуждение которых выходит за рамки данной статьи. Поэтому стоит ограничиться простой констатацией фактического положения дел. Именно поэтому среди акторов российского «Кодекса этики в сфере искусственного интеллекта» [1] потребители не указаны. Хотя к ним иногда можно отнести эксплуатантов и операторов, которые есть в этом кодексе, но нужно учитывать, что они (особенно, операторы) имеют специальную профессиональную подготовку, поэтому на них должны распространяться моральные требования прикладных и профессиональных этических документов в сфере СИИ. Тем не менее, в остальных упомянутых ранее этических разработках под эгидой Альянса с сфере этики ИИ, развивающих и конкретизирующих положения этого кодекса, «непрофессиональные» пользователи в качестве акторов почему-то появляются. Для этого возможны разные причины, требующие внимательного изучения, но даже при беглом взгляде возникает устойчивое подозрение, что разработчики и производители стремятся переложить ответственность ненадёжного функционирования СИИ на пользователей. Косвенным аргументом в пользу такого предположения может служить общая ситуация с беспилотными автомобилями (в отечественной терминологии — высокоавтоматизированные транспортные средства). В настоящее время нет примеров, когда в авариях были бы признаны разработчики или производители, даже в тех случаях, когда их причиной стали сбои программного обеспечения. Стремление переложить всю полноту ответственности на водителей даже воплотилось следующую сомнительную практику: «Эксперты Национального управления по безопасности дорожного движения США (National Highway Traffic Safety Administration, NHTSA) выяснили, что в большинстве зафиксированных случаев система автопилота Tesla отключается за несколько секунд до аварии. Это значит, что теперь невозможно привлечь компанию в суде по пункту обвинения в причинении умышленного ущерба из-за работы автопилота». [7, с. 22]. В качестве дополнительного подтверждающего аргумента можно сослаться на многочисленные дисклеймеры (англ. — disclaimer, т. е. отказ от ответственности), которыми разработчики и производители сопровождают почти все технологии на основе СИИ. В контексте морального императива объяснимости это означает, что производители вменяют пользователям в качестве моральной обязанности понимать особенности работы СИИ не хуже, а может быть даже лучше, чем их разработчики и производители.

Во-вторых, одной из этически значимых причин появления законодательства в области защиты прав потребителей является принципиальная и практически неустранимая асимметрия в доступе к информации или асимметрия информированности. За редчайшими исключениями производители любого продукта или услуги, в т. ч. и с использованием СИИ, знают и понимают о них в разы больше, чем даже самые продвинутые пользователи. Справедливости ради надо отметить, что ситуация в сфере СИИ является довольно специфической, поскольку иногда производители сами являются потребителями своих продуктов, но тогда это история про «сапожника без сапог». В большинстве случаев пользователи вроде бы имеют законное и моральное право требовать объяснений хода и результатов работы СИИ (неважно, кто будет давать эти объяснения: люди или СИИ), но на практике это выглядит совсем иначе. Кратко рассмотрим возможные ситуации. Вы приходите в банк за кредитом, Вам отказывают или предлагают какие-то несуразные с Вашей точки зрения проценты, говоря, что это было сделано при помощи технологии на основе СИИ.

А на Ваши требования объяснить, как и почему у СИИ получились такие результаты, Вы скорее всего получите категорический отказ, поскольку, с одной стороны, Вы являетесь клиентом банка, а не пользователем банковской СИИ. Иными словами, Вы пришли в банк за услугой в виде кредита, а не за «общением» с СИИ, хотя по факту именно СИИ выдаёт результат по кредиту, даже когда это озвучивают люди. С другой стороны, СИИ является собственностью банка и особенности работы системы являются коммерческой тайной. Иными словами, доступ с объяснимой СИИ имеют только работники банка. Похожая ситуация может возникнуть, если Вас задержали полицейские на основании результатов распознавания лиц при помощи СИИ. Очень сомневаюсь, что Вам раскроют и объяснят, как именно было произведено это распознавание. А когда и если ошибка будет обнаружена, то Вас отпустят с извинениями и словами вроде: «Это не мы виноваты, а СИИ ошибся». Даже когда люди сами пользуются соответствующими технологиями СИИ, то далеко не всегда им дают доступ к исходным программным кодам и прочей существенно важной информации. Эти и подобные случаи свидетельствуют, что связанные с принципом объяснимости практики, с одной стороны, поддерживают и усиливают имеющуюся асимметрию информированности, с другой стороны, пытаются включить пользователей, не очень разбирающихся в особенностях функционирования СИИ и морально не обязанных это делать, в число лиц, кому вменяется обязанность быть ответственными за результаты деятельности СИИ. Кроме того, в большинстве случаев они намеренно лишены доступа к такой информации, которая бы позволила им объяснить работу СИИ. Даже если не считать такие практики полностью неэтичными, то следует признать их как минимум этически спорными и сомнительными.

Подводя итоги проведенного анализа этических проблем морального императива объяснимого СИИ, можно сформулировать следующие выводы. Было обнаружено, что в настоящее время принцип объяснимости СИИ или близкие к нему по смыслу и содержанию требования присутствуют в подавляющем большинстве этических деклараций, кодексов и рекомендаций в сфере СИИ, в т. ч. в политиках компаний, являющихся признанными лидерами в разработке и производстве технологий с использованием СИИ. Проведённый этический анализ показал, что несмотря на привлекательность принципа объяснимости, он не является морально бесспорным. Во-первых, для него существует ряд непреодолимых технических ограничений, что затрудняет его использование в этически значимых целях. Во-вторых, существуют обоснованные сомнения, что принцип объяснимости СИИ что-то содержательно дополняет к имеющимся четырём принципам биомедицинской этики. В-третьих, следование императиву объяснимости СИИ ведет к парадоксальным результатам (парадокс «чёрного ящика», парадокс прозрачности, парадокс простоты, парадокс честности и т. д.) и возможным этическим рискам. Наконец, принцип объяснимости является ассиметричным в плане его агентности и адресности, особенно в отношении пользователей. Существующие практики объяснимости СИИ поддерживают и усиливают асимметричность информированности и способствуют этически неоправданному перекладыванию бремени ответственности на пользователей. В совокупности, эти аргументы свидетельствуют о том, что принцип объяснимости, при всей его важности, не может рассматриваться как морально безупречный и обязывающий.

Исследование проведено в рамках проекта РНФ № 24-28-00562 «Философские основания этических рисков в сфере систем искусственного интеллекта»).

Литература

[1] Кодекс этики в сфере искусственного интеллекта [Электронный текст] // Альянс в сфере искусственного интеллекта. URL: https://ethics.a-ai.ru/assets/ethics_files/2023/05/12/Кодекс_этики_20_10_1.pdf (дата обращения 18.03.2025).

[2] Этические рекомендации по применению рекомендательных технологий и алгоритмов, основанных на искусственном интеллекте, в цифровых сервисах [Электронный текст] // Альянс в сфере искусственного интеллекта. URL: https://ethics.a-ai.ru/assets/ethics_documents/2023/09/19/Recommendation_Services_Ethics_01_0RYxN8h.pdf (дата обращения 18.03.2025).

[3] Подходы к проблеме вагонетки в разрезе беспилотного наземного транспорта [Электронный текст] // Альянс в сфере искусственного интеллекта. URL: https://ethics.a-ai.ru/assets/ethics_documents/2023/12/20/Подходы_к_проблеме_вагонетки_в_разрезе_беспилотного_наземног_kRVGrJc.pdf (дата обращения 18.03.2025).

[4] Декларация об ответственной разработке и использовании сервисов в сфере генеративного искусственного интеллекта [Электронный текст] // Альянс в сфере искусственного интеллекта. URL: https://ethics.a-ai.ru/assets/ethics_files/2024/03/13/GenAi_Declaration_Ai_Alliance_Russia_FpNJ2Lc_82yB8pD.pdf (дата обращения 18.03.2025).

[5] Рекомендации Комиссии по реализации Кодекса этики в сфере ИИ по теме: «Прозрачность алгоритмов искусственного интеллекта и информационных систем на их основе» [Электронный текст] // Альянс в сфере искусственного интеллекта. URL: https://a-ai.ru/?page_id=2375 (дата обращения 18.03.2025).

[6] Руководящие принципы в сфере роботов общего назначения [Электронный текст] // Альянс в сфере искусственного интеллекта. URL: https://ethics.a-ai.ru/assets/ethics_files/2024/12/12/2_Руководящие_принципы_в_сфере_роботов_общего_назначения.pdf (дата обращения 18.03.2025).

[7] Белая книга этики в сфере искусственного интеллекта / под ред. А. В. Незнамова. М.: Nova Creative Group, 2024.

[8] Floridy L. Establishing the rules for building trustworthy AI // Nature Machine Intelligence. 2019. No. 1(6). P. 261-262. DOI: 10.1038/s42256-019-0055-y.

[9] Merton R. K. The Normative Structure of Science // Merton R. K. The Sociology of Science: Theoretical and Empirical Investigations. Chicago: University of Chicago Press, 1973.

[10] Russell W. M. S., Burch R. L. The Principles of Humane Experimental Technique. Methuen, London, 1959.

[11] Beauchamp T. L., James F., Childress J. F. Principles of Biomedical Ethics. Oxford University Press, 2001.

[12] Floridi L. The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities. Oxford University Press, 2023. 243 p.

[13] Рекомендация об этических аспектах искусственного интеллекта (ЮНЕСКО 2021) [Электронный текст] // UNESCO. URL: https://unesdoc.unesco.org/ark:/48223/pf0000380455_rus (дата обращения 18.03.2025).

[14] DARPA: Explainable Artificial Intelligence (XAI). Broad Agency Announcement DARPA-BAA-16-53. 2016.

[15] What is explainable AI? [Электронный текст] // IBM. URL: https://www.ibm.com/think/topics/explainable-ai (дата обращения 18.03.2025).

[16] AI ethics [Электронный текст] // IBM. URL: https://www.ibm.com/impact/ai-ethics (дата обращения 18.03.2025).

[17] IBM's Principles for Trust and Transparency [Электронный текст] // IBM. URL: https://www.ibm.com/policy/trust-transparency (дата обращения 18.03.2025).

[18]IBM Artificial Intelligence Pillars [Электронный текст] // IBM. URL: https://www.ibm.com/policy/ibm-artificial-intelligence-pillars/ (дата обращения 18.03.2025).
[19]Limitations of Vertex Explainable AI! [Электронный текст] // Google. URL: https://cloud.google.com/vertex-ai/docs/explainable-ai/limitations (дата обращения 18.03.2025).
[20]Cortese J. F. N. B., Cozman F. G., Lucca-Silveira M. P. et al. Should explainability be a fifth ethical principle in AI ethics? // AI Ethics. 2023. Vol. 3. P. 123-134. DOI: 10.1007/s43681-022-00152-w.

Ethical Challenges to the Imperative of the Explainability of Artificial Intelligence Systems (AIS)

V. Iu. Perov

Saint-Petersburg State University

The development of ethics in the field of artificial intelligence systems (AIS) is embodied in a variety of ethical documents (declarations, codes, recommendations, etc.), which differ in the sources of creation, status, content, terminology used, etc. In this regard, the desire to radically reduce the number of ethical principles in the field of artificial intelligence seems correct. The article discusses the idea of L. Floridi, who proposes to limit ethics in the field of AIS to the principles of biomedical ethics, adding to them as the fifth principle of the explainability of AIS. In the course of the study, it was found that at present the principle of explainability (or requirements close to it in content) is present in most ethical documents in the field of AIS and is even considered as a necessary prerequisite and basis for other ethical principles. In addition, explainability is declared as one of the main moral imperatives of the ethical policies of companies that are leaders in the development of AIS (for example, IBM, Google). The ethical analysis has shown that despite the attractiveness of the principle of explainability, it is not morally unambiguous. Firstly, there are a number of insurmountable technical limitations for it. Secondly, there are reasonable doubts that the principle of explainability of AIS is something meaningfully complementary to the existing four principles of biomedical ethics. Thirdly, following the imperative of explainability of AIS is paradoxical and leads to possible ethical risks. Finally, the principle of explainability is asymmetrical in terms of its agency and targeting, especially in relation to users. Taken together, these arguments show that the principle of explainability, important as it is, cannot be regarded as morally irreproachable and binding..

Keywords: artificial intelligence systems (AIS) ethics, codes of ethics, ethical principles, moral imperative, explainable AIS, paradoxes, ethical risks, responsibility, moral paradoxes, awareness asymmetry

Reference for citation: V. Iu. Perov Ethical Challenges to the Imperative of the Explainability of Artificial Intelligence Systems (AIS) // Information Society: Education, Science, Culture and Technology of Future. Vol. 9 (Proceedings of the XXVIII International Joint Scientific Conference «Internet and Modern Society», IMS-2025, St. Petersburg, June 23–25, 2025). – St. Petersburg: ITMO University, 2025. P. 93-102. DOI: 10.17586/3033-5574-2025-9-93-102.

Reference

[1] Code of Ethics in the Field of Artificial Intelligence [Electronic resource] // Artificial Intelligence Alliance. URL: https://ethics.a-ai.ru/assets/ethics_files/2023/05/12/Кодекс_этики_20_10_1.pdf (accessed: 18.03.2025). (in Russian)

[2] Ethical Guidelines for the Use of Recommender Technologies and Algorithms Based on Artificial Intelligence in Digital Services [Electronic resource] // Artificial Intelligence Alliance. URL: https://ethics.a-ai.ru/assets/ethics_documents/2023/09/19/Recommendation_Services_Ethics_01_0RYxN8h.pdf (accessed: 18.03.2025). (in Russian)
[3] Approaches to the Trolley Problem in the Context of Autonomous Ground Transport [Electronic resource] // Artificial Intelligence Alliance. URL: https://ethics.a-ai.ru/assets/ethics_documents/2023/12/20/Подходы_к_проблеме_вагонетки_в_разрезе_беспилотного_наземног_kRVGrJc.pdf (accessed: 18.03.2025). (in Russian)
[4] Declaration on the Responsible Development and Use of Services in the Field of Generative Artificial Intelligence [Electronic resource] // Artificial Intelligence Alliance. URL: https://ethics.a-ai.ru/assets/ethics_files/2024/03/13/GenAi_Declaration_Ai_Alliance_Russia_FpNJ2Lc_82yB8pD.pdf (accessed: 18.03.2025). (in Russian)
[5] Recommendations of the Commission for the Implementation of the AI Ethics Code on the Topic: «Transparency of Artificial Intelligence Algorithms and Information Systems Based on Them» [Electronic resource] // Artificial Intelligence Alliance. URL: https://a-ai.ru/?page_id=2375 (accessed: 18.03.2025). (in Russian)
[6] Guidelines in the Field of General-Purpose Robots [Electronic resource] // Artificial Intelligence Alliance. URL: https://ethics.a-ai.ru/assets/ethics_files/2024/12/12/2_Руководящие_принципы_в_сфере_роботов_общего_назначения.pdf (accessed: 18.03.2025). (in Russian)
[7] Neznamov A.V. (ed.). White Book of Ethics in the Field of Artificial Intelligence. Moscow: Nova Creative Group, 2024. (in Russian)
[8] Floridy L. Establishing the rules for building trustworthy AI // Nature Machine Intelligence. 2019. No. 1(6). P. 261-262. DOI: 10.1038/s42256-019-0055-y.
[9] Merton R. K. The Normative Structure of Science // Merton R. K. The Sociology of Science: Theoretical and Empirical Investigations. Chicago: University of Chicago Press, 1973.
[10] Russell W. M. S., Burch R. L. The Principles of Humane Experimental Technique. Methuen, London, 1959.
[11] Beauchamp T. L., James F., Childress J. F. Principles of Biomedical Ethics. Oxford University Press, 2001.
[12] Floridi L. The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities. Oxford University Press, 2023.
[13] Rekomendaciya ob eticheskih aspektah iskusstvennogo intellekta (2021) // UNESCO. URL: https://unesdoc.unesco.org/ark:/48223/pf0000380455_rus (accessed date:18.03.2025). (In Russian)
[14] DARPA: Explainable Artificial Intelligence (XAI). Broad Agency Announcement DARPA-BAA-16-53. 2016.
[15] What is explainable AI? // IBM. URL: https://www.ibm.com/think/topics/explainable-ai (accessed date: 18.03.2025).
[16] AI ethics // IBM. URL: https://www.ibm.com/impact/ai-ethics (accessed date: 18.03.2025).
[17] IBM's Principles for Trust and Transparency // IBM. URL: https://www.ibm.com/policy/trust-transparency (accessed date: 18.03.2025).
[18] IBM Artificial Intelligence Pillars // IBM. URL: https://www.ibm.com/policy/ibm-artificial-intelligence-pillars/ (accessed date:18.03.2025).
[19] Limitations of Vertex Explainable AI! // Google. URL: https://cloud.google.com/vertex-ai/docs/explainable-ai/limitations (accessed date: 18.03.2025).
[20] Cortese J. F. N. B., Cozman F. G., Lucca-Silveira, M.P. et al. Should explainability be a fifth ethical principle in AI ethics? // AI Ethics. 2023. Vol. 3. P. 123–134. DOI: 10.1007/s43681-022-00152-w.