DOI: 10.31857/S0236200724040067

©2024 С.В. ГАРБУК, А.В. УГЛЕВА

АВТОМАТИЗИРОВАННЫЕ ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ: ЭТИЧЕСКИЙ И НОРМАТИВНО-ТЕХНИЧЕСКИЙ ПОДХОДЫ К РЕГУЛИРОВАНИЮ

Гарбук Сергей Владимирович — кандидат технических наук, директор по научным проектам.
Национальный исследовательский университет «Высшая школа экономики».
Российская Федерация, 109028 Москва, Покровский бульвар, д. 11.
ORCID 0000-0001-5385-3961
sgarbuk@hse.ru

Углева Анастасия Валерьевна — кандидат философских наук, PhD, профессор Школы философии и культурологии, заместитель директора Центра трансфера и управления социоэкономической информацией НИУ ВШЭ.
Национальный исследовательский университет «Высшая школа экономики».
Российская Федерация, 109028 Москва, Покровский бульвар, д. 11.
ORCID 0000-0002-9146-1026
augleva@hse.ru

Аннотация. Искусственный интеллект настолько прочно вошел в нашу жизнь, что его непосредственное влияние на формирование картины мира будущего не вызывает сомнений. Однако потребовалось время для того, чтобы наряду с теоретическими спекуляциями на тему «восстания машин» и прочих угроз человечеству с его стороны постепенно начал вырабатываться конструктивный подход к предупреждению

УДК 62-5

СОЦИАЛЬНЫЕ ПРАКТИКИ

рисков и нормативному регулированию технологий на всех этапах их жизненного цикла. Предметом особого внимания являются так называемые автоматизированные искусственные системы, регулирование которых ограничено до сих пор нормативно-техническими требованиями. Особенностью такого подхода является убеждение его сторонников в истинности технологического детерминизма, для которого «техника» является ценностно нейтральной. Предупреждение рисков этического характера с точки зрения такого подхода оказывается практически нерешаемой задачей в силу того, что вопросы нормативного регулирования касаются лишь функциональных характеристик и нарушения требований эксплуатации конкретной системы. В данной статье технологическому детерминизму противопоставляется технологический субстантивизм, для которого «техника» приобретает самостоятельную этическую ценность независимо от ее инструментального использования. Основанная на нем этическая оценка состоит в процедуре регулярного соотнесения социального «блага» и «надежности» системы. Разработка методологии такой процедуры соотнесения требует специальных компетенций, отличающих новую профессиональную сферу — этику в сфере ИИ.

Ключевые слова: философия техники, искусственный интеллект, технологический детерминизм, субстантивизм, этика искусственного интеллекта, автономные интеллектуальные системы, благо, надежность.

Ссылка для цитирования: Гарбук С.В., Углева А.В. Автоматизированные интеллектуальные системы: этический и нормативно-технический подходы к регулированию // Человек. 2024. Т. 35, № 4. С. 97–116. DOI: 10.31857/S0236200724040067

Бурное развитие цифровых технологий в последние 10-15 лет способствовало широким дискуссиям по поводу антропологических и этических особенностей технологического прогресса, связанных с радикальными, часто непредсказуемыми изменениями во всех сферах жизни общества под влиянием внедряемых в них технологий искусственного интеллекта (далее — ИИ). Одной из главных дилемм философии техники[1], как известно, является технодицея[2] — противостояние рационального дискурса и активного внедрения в общественное сознание представлений о неизбежном крахе человеческой цивилизации в связи с рисками, возникающими в силу автономизации интеллектуальных систем,

[1] В данном случае мы используем понятие техники в его расширительном значении, имея в виду в том числе технологии ИИ.

[2] Термин был введен в научный оборот Г.В. Лейбницем по аналогии с «теодицеей» в трактате «Опыты теодицеи о благости Бога, свободе человека и первопричине зла» (1710).

действия которых для человека не прозрачны и которые ему все менее подконтрольны. В этом контексте ИИ нередко рассматривается как технологическое зло, которое может мыслиться двояко: как зло физическое, выражающееся в потенциальных угрозах для жизни и свободы человека, и как зло моральное, заключающееся в целенаправленности нанесения вреда человеку. При этом рациональные дискуссии об этике, предметом которой исконно и являются вопросы «добра» и «зла», остаются вне поля внимания участников полемики. Однако даже если оставить в стороне вопросы собственно самой морали, переосмысления ее сущности, а обратить свой взор на прикладные задачи этики в сфере ИИ, то очевидно, что человечество заинтересовано в определении некой положительной перспективы для себя с опорой на человеческую рациональность и творчество тех, кто непосредственно участвует в архитектуре и разработке новых технологий.

Возможности и риски автоматизированных интеллектуальных систем

Прогресс в области ИИ происходит, прежде всего, благодаря использованию алгоритмов машинного обучения (далее — МО), для которых характерны [Garbuk, 2022]: 1. Отсутствие полной интерпретируемости. Для существенной части операций над данными, предусмотренных алгоритмом МО, человеком сколь угодно высокой квалификации не может быть принято решение о правильности или неправильности этих операций на основании критериев, истинность которых подтверждена внешними соображениями; 2. Обязательное использование специальным образом подготовленных наборов данных (далее — НД), содержащих примеры решения конкретных прикладных интеллектуальных задач. Такие обучающие НД могут формироваться как с участием человека, так и без него (автоматически, в режиме самообучения); 3. Возможность дообучения алгоритмов МО в процессе эксплуатации систем ИИ (далее — СИИ) путем расширения обучающих НД на дополнительных примерах решения прикладных задач. Уровень конфиденциальности данных в процессе такого обобщения на стадии эксплуатации систем может возрастать; 4. Универсальность алгоритмов МО, позволяющая использовать СИИ «для автоматизации сложных задач обработки данных, не поддающихся решению с помощью одних лишь полностью интерпретируемых алгоритмов. Это приводит к существенному повышению уровня автоматизации процессов обработки информации и, как следствие, к актуализации вопроса социальной приемлемости применения

алгоритмов МО [Gendron, 2014]; 5. Необходимость сравнения характеристик качества СИИ и функциональных возможностей человека-оператора (например, водителя транспортного средства).

Широкие возможности по применению алгоритмов МО обуславливаются, в свою очередь, развитием средств получения информации (сенсоров), которые становятся все более дешевыми, надежными и информативными, повышением пропускной способности информационно-телекоммуникационной инфраструктуры, а также совершенствованием средств вычислительной техники, обеспечивающих эффективную реализацию алгоритмов обработки данных и хранение больших массивов информации.

Новые возможности, возникающие благодаря широкому внедрению технологий ИИ, с неизбежностью приводят к новым вызовам и рискам, в том числе — рискам социальной приемлемости (то есть к этическим рискам). При этом следует учитывать, что этика — характеристика социальных отношений. Эксплицитно СИИ, как и любые другие технические системы, не обладают свойствами «этичности», вместе с тем они способны оказывать влияние на отношения между отдельными людьми и социальными группами. По большому счету, на уровень этичности социальных отношений влияют не только интеллектуальные, но и многие другие технологии — технические средства охраны, включая средства акустического контроля помещений, pinhole-видеокамеры, биометрические средства контроля и управления доступом и др.; носимые средства видео- и аудиорегистрации, к которым в полной мере относятся современные смартфоны, позволяющие вести негласную запись переговоров и действий людей, и др.

В контексте СИИ проблематизация этического аспекта внедрения новых технологий приобрела особую актуальность в связи с тем, что средства автоматизации, основанные на алгоритмах МО, являются беспрецедентно универсальными и получают все большее распространение при решении задач обработки данных, традиционно являвшихся исключительной прерогативой человека. К таким задачам следует отнести, например, узнавание людей по лицу, походке, жестам, понимание человеческих эмоций; прогнозирование поведения людей, в том числе при управлении автотранспортными средствами и при пешем движении по дорогам; при совершении покупок, выборе и приобретении самых разных товаров и услуг — от квартиры до мелодии на телефоне; при выборе карьерной траектории, управлении личными финансами и принятии других решений, определяющих жизненный путь человека; синтез мультимодальных данных, включая художественные высокореалистичные изображения, тексты и музыкальные произведения; принятие решения в сложных непредвиденных

Рис. 1. Модель жизненного цикла системы ИИ[3].

ситуациях, основываясь на большом объеме разнородных, неполных и порой противоречивых данных и т.п.

Универсальность методов МО объясняется тем, что они могут быть использованы даже в условиях отсутствия основанных на знаниях моделей наблюдаемых объектов, явлений и процессов. Достаточно лишь накопленного эмпирического опыта. Именно этим, например, объясняется активное применение ИИ в медицине: адекватных моделей человека в целом и его отдельных органов до настоящего времени не создано (объект моделирования слишком сложен и вариативен), но накоплен и продолжает накапливаться огромный опыт профилактики и лечения заболеваний, чему способствует бурное развитие различных медицинских сенсоров, фиксирующих разнообразные параметры как состояния человека, так и его образа жизни. В результате накапливаются большие наборы данных, обеспечивающие построение моделей человека с использованием методов МО. Они в дальнейшем используются для диагностики и прогнозирования состояния здоровья человека, в том числе в условиях принятия тех или иных врачебных решений.

Модель жизненного цикла (далее — ЖЦ) для типовой СИИ, учитывающая приведенные выше особенности алгоритмов МО, представлена на рис. 1 [Гарбук, 2022]. Модель предусматривает выделение следующих этапов ЖЦ: внешнее проектирование и выбор типовых решений (R), обучение (L1), тестирование при вводе в эксплуатацию (T1), эксплуатация (U), пробное дообучение на вновь поступающих данных (L2), повторное тестирование

[3] Буквами на рисунке обозначены точки возникновения несоответствия требованиям, установленным к различным информационным компонентам.

после пробного дообучения (T2) и модификация СИИ при успешном дообучении (M).

Нормативно-техническое регулирование этических аспектов создания и применения ИИ

Влияние ИИ на этику социальных отношений удобно рассматривать в разрезе реализуемых на протяжении всего его ЖЦ процессов, на которых проявляется специфика СИИ: сбор и накопление обучающих данных (этапы R, L1); тестирование СИИ (T1); принятие решения о возможности применения и эксплуатация СИИ (U); дообучение СИИ на стадии эксплуатации (L2, T2, M); утилизация (вывод из эксплуатации) системы (U).

Сбор обучающих наборов данных (далее — НД) может сопровождаться нарушениями в сфере этики при накоплении, например, персональных данных. В данном случае речь не идет о нарушениях в сфере защиты персональных данных, подпадающих под действие соответствующего федерального закона (Федеральный закон…, 2006) и относящихся, таким образом, к области с вполне жестким государственным регулированием – области защиты информации. Этические нарушения могут заключаться во введении в заблуждение субъектов персональных данных относительно целей сбора данных, их коммерческого использования, особенностей результатов их обработки и т.п.

Особые риски на стадии формирования обучающих НД могут возникать при реализации процедур «обезличивания» или «деклассификации» информации ограниченного распространения. Необходимость в таких процедурах обусловлена стремлением государственных или частных заказчиков расширить круг разработчиков СИИ за счет предоставления доступа к данным как можно большему числу профильных компаний, исключив при этом возможность компрометации ограниченного распространения. При этом возникают специфические риски «обратного восстановления» сведений о субъектах персональных данных, при которых ответственность в соответствии с законом не возникает, однако затрагивается этическая сторона вопроса. В качестве примера можно привести случай, когда обратное восстановление неких дискредитирующих сведений не позволяет персонально идентифицировать человека, однако позволяет локализовать группу людей по этническому, конфессиональному или иному признаку.

Отметим, что процедуры сбора обучающих НД могут сопровождаться повышением уровня конфиденциальности данных в процессе их агрегирования, что также может затронуть интересы

граждан в этической сфере. Некорректные действия при тестировании систем могут, как и на стадии проектирования и создания СИИ, заключаться в нарушениях при сборе тестовых НД, а также в неправильном оценивании функциональных характеристик (далее — ФХ) систем. Под ФХ в данном случае понимается совокупность показателей качества алгоритмов МО в составе СИИ, определяющих возможность применения СИИ по назначению. Неприемлемые социальные последствия их применения могут быть вызваны неточным и/или неполным пониманием эксплуатирующей стороной ФХ системы. Это, однако, не означает «неэтичность» системы как таковой, а связано исключительно с неквалифицированным ее использованием. Для любых видов угроз и любых заинтересованных сторон степень соответствия требованиям СИИ рассматривается, как правило, в контексте определенной прикладной задачи и в предусмотренных условиях эксплуатации (далее — ПУЭ). Так, в п.3.22 стандарта американской Ассоциации автомобильных инженеров (SAE J3016, 2018) «предусмотренная область применения» понимается как условие, при котором использование данного автотранспортного средства является безопасным, исходя из параметров окружающей среды, географических и временных условий, обязательного наличия или отсутствия определенных характеристик транспортного потока и проезжей части [Гарбук, 2023].

Таким образом, проблемы в функционировании ИИ объясняются двумя основными причинами. Во-первых, это низкое качество СИИ, которое проявляется в несоответствии ее функциональных характеристик установленным требованиям при решении конкретных задач из-за недостаточной информативности используемых обучающих данных, недостатков методов машинного обучения и других факторов, что может привести к потере доверия к системе. Во-вторых, это нарушение условий эксплуатации из-за недостаточной подготовки оператора ИИ, что не всегда является объективным основанием для потери доверия к самой системе. В контексте широкого спектра автоматизированных задач, включая те, которые ранее выполнялись только людьми, нарушение условий эксплуатации может вызвать у пользователей потерю доверия к системе, что негативно скажется на развитии технологий ИИ в целом.

При этом важно подчеркнуть, что негативные социальные последствия, вызванные нарушениями в функционировании ИИ на этапах тестирования и эксплуатации, не ограничиваются только этическими аспектами. Они также включают угрозы для жизни, здоровья людей и экологической безопасности (физические угрозы), угрозы в области информационной безопасности,

когда некорректная работа поисковых систем может иметь следствием распространение неправомерной информации и оказание дестабилизирующего психологического воздействия на общественное сознание (угрозы информационной безопасности). Функциональные нарушения могут также приводить к угрозам экономической безопасности, уменьшая эффективность использования ИИ в соответствии с его назначением.

Наконец, этические нарушения на этапе утилизации (вывода из эксплуатации) СИИ могут заключаться в игнорировании эксплуатантом морального устаревания системы, введении разработчиком или поставщиком системы в заблуждение эксплуатанта системы относительно ее конкурентоспособности, а также в нарушениях в области защиты информации, накопленной за время практического применения системы. Последнее обстоятельство для СИИ является особо существенным с учетом вышеупомянутой возможности неконтролируемого повышения уровня конфиденциальности данных, накапливаемых при функционировании системы.

Как на международном, так и на национальном уровнях разрабатываются нормально-технические документы, в которых находят отражение некоторые этические риски, возникающие при создании и применении СИИ. Например, в национальном стандарте ГОСТ Р 70889 (2023) указывается на необходимость учета этических и социальных аспектов при создании и принятии решений о возможности применения СИИ. Уделено им внимание и в гармонизированном предварительном национальном стандарте ПНСТ 840 (2022), в котором представлен обзор этических и социальных проблем ИИ. Разработанный на основе международного документа ПНСТ 841-2023, развивающий подходы программной и системной инженерии SQuaRE, стандарт вводит дополнительную характеристику ее качества — свободу от этического и социального риска. В стандарте особо отмечается, что организации, разрабатывающие и применяющие СИИ, должны выявлять и решать общественные и этические проблемы на протяжении всего ее жизненного цикла.

Ограниченность технологического детерминизма

Надо признать, что нормативно-техническая проработка этических аспектов создания и применения ИИ находится еще в начальной стадии. Не определены четкие правила разделения этических рисков и случаев несоблюдения требований физической и

информационной безопасности. Так, в ПНСТ 840 (2022, пп. 4.1) в качестве примеров областей, где возрастает риск нежелательных этических и общественных последствий, приводятся «физический ущерб, вред физическому здоровью и безопасности, вмешательство в выборы и т.п.)». Кроме того, сторонники этого подхода нередко указывают на излишнюю «гуманитаризацию» дискурса, связанного с обсуждением технических свойств СИИ (см. ПНСТ 840, 2022, пп.4.3.2), и необоснованное приписывание им таких качеств, как уважение достоинства человека, честность, мужество, сострадание и т.п. В целом, сбалансированный учет различных рисков, возникающих при обучении и внедрении СИИ с учетом особенностей их реализации, обеспечивается применением риск-ориентированного подхода, основные принципы которого изложены в ПНСТ 776-2022.

Вместе с тем у такого подхода есть ряд недостатков, обусловленных во многом логикой технологического детерминизма, в основе которого лежит идея автономной и ценностно нейтральной природы самой технологии. Этот подход изначально провоцирует социальное напряжение и рост недоверия к разработчикам технологий и технологическим компаниям, поскольку буквально понятый технологический детерминизм не исключает принятия решения о применении технологий исключительно на основе их технических возможностей, не учитывая этические, социальные и культурные последствия.

В самом широком смысле слова доверие связано с благонадежностью системы, а значит вопрос о ее «этичности» должен решаться не столько на стадии обучения и внедрения ИИ, а уже на стадии проектирования. С технической точки зрения надежной является система, способная решать задачи с определенным качеством, то есть в ней должны быть реализованы релевантные методы, применена математическая модель, которая позволяет считать, что задача решается верно, должны быть проведены соответствующие испытания, получен сертификат качества и т.п. Этот аспект доверия относится к функционалу разработчика или ответственного пользователя, оператора системы. С моральной точки зрения надежной является система, отвечающая требованию социальной приемлемости, а также действующая в соответствии с контекстуально определенным понятием общественного блага. А это означает, что разработчик и оператор системы не могут быть отчуждены от принятия ответственности за последствия использования технологий. Только в этом случае СИИ может быть «доверенной».

Вопрос о доверии возникает в силу того, что пока оператор СИИ может задавать только условия решения при последующей принципиальной необъяснимости ответа в каких-то разумных

предметных терминах и неделимости алгоритма, в отличие, например, от базы данных или операциональной системы, вроде Windows, где, напротив, оператор контролирует метод решения при принципиальной прослеживаемости ответа и понятности алгоритма, то есть в принципе знание и понимание того, как было принято то или иное решение, ему доступно. Ведь, по сути, модель машинного обучения возникает либо как результат большого разнонаправленного взаимодействия со средой (обучение с подкреплением), либо как результат обобщения большого объема данных. Причем максимум, знание чего нам доступно, — это знание алгоритма, который обучал эту систему, но каждый раз решение может быть принято разное. То есть фактически принимающий решение алгоритм нам неизвестен.

Доверие к системе основано на некотором наборе норм, правил и принципов, которые определяются теми, кто разрабатывает систему, и теми, кто пользуется ею. Речь идет о трансфере имеющейся у оператора устоявшейся системы ценностей на СИИ, которая должна обладать условной доброй волей, то есть быть нацеленной на благо. Именно так — не рационально, но эмоционально пользователь относится к системе: если он ей доверяет решение значимой для него задачи, то имеет право ожидать, что доверенная СИИ в ситуации выбора предпочтет решение, которое либо несет пользователю непосредственное благо, либо выберет вариант наименьшего вреда. Иначе говоря, доверие к ИИ может быть определено как через технические требования надежности и безопасности, так и через процесс рационального выбора определенных ценностных ориентиров и стремление к благой цели (или, как минимум, отсутствие злого умысла у системы).

Однако следует иметь в виду, что средства естественного языка не приспособлены для того, чтобы приписать и описать наличие доброй воли и вообще какой-либо «этичности» у СИИ, то есть очевидна проблема перевода терминов естественного языка на язык таких систем. Поэтому до сих пор не решена задача выработки инструментария для оценки на основании приемлемых для всего человечества критериев уровня доверия к СИИ. Такая оценка всегда будет включать в себя элемент субъективности, предрасположенности или непредрасположенности к доверию того, кто обращается к такой системе. Кто-то с радостью доверится СИИ, кто-то будет относиться к ней с известной долей подозрительности. Поэтому для утверждения о социальной приемлемости внедрения конкретной СИИ как следствия общественного доверия к ней требуется формализация и инструментализация подхода к оценке доверия к ИИ.

Непонимание и, как следствие, недоверие к ИИ связано и с тем, что общество практически не участвует в принятии решений, связанных с научно-технологическим развитием. Отсутствие гуманитарной оценки и обоснованного предвидения социокультурных последствий внедрения передовых цифровых технологий приводит к необходимости привлечения экспертов соответствующего профиля для такой оценки [Шевченко и др., 2021]. Это проявляется в осознании социальной потребности в инструментах для анализа антинаучных мифов, дезинформации, технофобии и неолуддизма. Развитие трансдисциплинарного диалога между обществом, бизнесом и государством необходимо для эффективного решения этих проблем и реализации партисипативной модели «общество — бизнес — государство». Социо-гуманитарная экспертиза подразумевает переход от узко технической оценки рисков к диалогу между этими тремя элементами системы и от узкоутилитарных интересов рынка к культуре совместной ответственности. Это означает, что принятие решений о технологическом развитии должно основываться не только на коммерческих интересах, но и проспективно — на оценке их социальной приемлемости. В противном случае убеждение в технологическом детерминизме может препятствовать развитию альтернативных видений будущего и инновационных подходов к решению стоящих сегодня перед обществом задач.

В качестве альтернативы технологическому детерминизму можно рассматривать технологический субстантивизм. Он согласуется с технологическим детерминизмом в том, что технология обладает определенной автономией по отношению к человеку и обществу в целом. Однако если с точки зрения технологического детерминизма, технология играет определяющую роль в развитии общества, причем общество изменяется в соответствии с возможностями, которые предоставляют ему технологии, и в этом смысле технология имеет собственную динамику, которая напрямую влияет на социальные структуры, культуру и даже ценности общества, то с точки зрения технологического субстантивизма технология зависит от общества и его целей. Он признает, что не только технологии формируют общество, но и общество влияет на направление развития технологий. Субстантивизм подчеркивает важность человеческих ценностей, норм и стремлений в проектировании технологий, а также необходимость учитывать социокультурный контекст и последствия их использования.

Не случайно технопессимисты утверждают, что приписывание технологии самостоятельной этической ценности означает власть и господство тех, кто имеет к ней доступ. И тогда торжество технологии над любыми другими этическими ценностями видится

неизбежным злом (см. О. Хаксли «О дивный новый мир», 1932). В итоге, как пишет Хайдеггер, власть и господство технологии приводят к тому, что мы воспринимаем человека и природу как сырье, которое должно быть использовано в технологических разработках, а общество — как отражение технических систем [Heidegger, 1954]. В том же духе рассуждает Н. Бостром в своей теории репликатора скрепок, описывая гипотетический сценарий возникновения суперинтеллекта, постоянно оптимизирующего процесс решения стоящей перед ним задачи в том числе за счет разрушения привычной структуры общественных связей. ИИ есть не что иное, как максимизатор функции-полезности, который при игнорировании этической составляющей может нанести человечеству существенный вред, даже если его действия запрограммированы на решение самых безобидных или полезных задач [Bostrom, 2016].

Хайдеггер тоже не отрицал важность технологии и признавал ее значимость для решения отдельных, в том числе социальных, задач. Однако высказывал справедливые опасения относительно технологического детерминизма и его влияния на человеческую сущность и смысл бытия. Технология не является просто инструментом, но имеет свою «сущность». Она представляет собой определенный способ мышления и взаимодействия с миром, который оказывает влияние на наше понимание себя, других людей и окружающей среды. Необходимо избегать опасности превращения технологии в безразличный способ обращения с миром, где все становится объектом манипуляции и в конце концов потеряет свою уникальность. Это может привести к утрате глубокого понимания нашего собственного бытия, поскольку технологический прогресс и автоматизация могут привести к потере значимости человеческого опыта, ведь человеческое понимание и мудрость, интуиция и эмпатия основаны на неформализованных, контекстуальных и ситуативных знаниях, которые ИИ традиционно не способен достоверно воспроизвести [Дрейфус, 1990].

Нормативно-этическое регулирование в сфере ИИ

В дополнение нормативно-техническому регулированию технологий ИИ может быть предложен нормативно-этический подход, проявляющийся: (1) в регуляции технологий через «принципы», то есть выделение и описание основных требований, которыми должны руководствоваться все акторы, вовлеченные в разработку или использование СИИ (открытость, безопасность для человека, защита личных данных и пр.), (2) в регуляции через «процессы»,

то есть включение в алгоритмы ограничений, которые не позволят неэтическое или противоправное использование СИИ, (3) в регуляция через «этическое сознание», то есть через предписание акторам этической осознанности и необходимости предвидеть последствия и риски использования интеллектуальных систем [Kazim, Koshiyama, 2021]. Уже в проектировании ИИ невозможно полностью избежать нормативных вопросов, потому что этика требует от разработчика и оператора системы пропорциональности и соответствия принципу непричинения вреда. А это предполагает, что «выбранный метод ИИ должен быть соразмерным достижению поставленной законной цели; выбранный метод ИИ не должен нарушать фундаментальные ценности; его использование не должно нарушать права человека; метод ИИ должен соответствовать контексту и опираться на строгие научные основания» [Шталь, 2024: 82]. Мы не должны стремиться согласовывать ценности, программируемые в ИИ, только с инструкциями или требованиями к эксплуатации, выраженными намерениями или выявленными предпочтениями [Gabriel, 2020]. Правильно согласованный ИИ должен учитывать возможные проявления некорректного с точки зрения этики поведения и предусматривать объективные ограничения такого поведения у искусственных агентов.

Конечно, можно задаться целью выявления общеприемлемых принципов поведения, несмотря на существование различных систем убеждений между людьми, что потребует также технического уточнения понятий, фигурирующих в этих принципах. Основная задача ценностного согласования состоит не в том, чтобы определить одну единственную универсальную моральную доктрину, а в том, чтобы определить принципы, которые признаются обществом в качестве справедливых. Проблема согласования в этом смысле политическая, а не метафизическая. Ее решение требует глобального ценностного консенсуса, и процесс для его достижения должен быть инклюзивным и процедурно справедливым в том смысле, что он не должен отдавать произвольного преимущества какой-либо из сторон.

Еще одним важным качеством процесса отбора является его способность справляться с возможностью широко распространенной моральной ошибки. Учитывая, что мы можем совершать ошибки такого рода, было бы неправильным привязывать ИИ к морали текущей эпохи [ibid.]. Поэтому сегодня разработано несколько возможных рамочных стратегий в области этики в сфере ИИ, которые, как ни странно, плохо согласуются друг с другом [Forbes, 2021], хотя все они ориентированы на достижение большей прозрачности и объяснимости ИИ. При отсутствии единого языка описания для обсуждения этических аспектов на основе принципалистского

подхода предлагается выработать путем консенсуса набор общих этических принципов, релевантных для всех областей, в которых возникают сложные этические задачи. И эти принципы предлагается транслировать через различные образовательные и просветительские практики на сообщество социально ответственных архитекторов и пользователей ИИ. Посредством образования предлагается способствовать развитию и более ответственной инвестиционной системы, которая будет способствовать росту социально ответственных, или «этичных» компаний, отражающих общие идеалы этики в сфере ИИ. Наконец, без систематической подотчетности будет трудно перейти от теории к практике.

В настоящий момент можно выделить 14 ключевых сфер применения цифровых технологий в области этики ИИ [Ashok et al., 2022] с предварительной разметкой на четыре онтологические области (физическую, когнитивную, информационную и управленческую). В когнитивном домене сферы применения ИИ следующие: понятность, подотчетность, процветание, справедливость, солидарность и автономия. В физическом домене — достоинство и благополучие, безопасность и устойчивое развитие. В информационной — конфиденциальность и безопасность. В управленческой — регулирующее воздействие, финансово-экономическое воздействие и индивидуальное и общественное воздействие ИИ. Такое разнообразие сфер применения ИИ вызывает, с одной стороны, озабоченность, а с другой стороны, порождает растущее количество откликов со стороны специалистов. За последние годы количество публикаций в области этики ИИ растет экспоненциально, так же как увеличивается количество практических руководств и предписаний, хотя не всегда этот рост публикаций приводит к реальным трансформациям в этой области. Зачастую этические размышления в работах по этике в сфере ИИ остаются ограниченными узкими областями применения без учета глобальных этических проблем, в которые они встроены [Bruneault, Laflamme, 2021].

Наиболее многообещающий подход основан на информационной этике, предложенной Лучано Флориди [Floridi et al., 2020; Hager et al., 2017] в его концепции «ИИ для общественного блага». Он приводит примеры новых технологий, обладающих потенциалом для решения в том числе социальных проблем за счет разработки решений на основе ИИ. Тем не менее пока есть лишь ограниченное теоретическое понимание того, что такое «социально ориентированный ИИ», что считать таковым на практике и как воспроизвести его первоначальные успехи при повторных его применениях. Флориди выделил семь этических факторов таких систем: фальсифицируемость, защита от манипулирования,

объяснимость и прозрачность целей, защита конфиденциальности и согласие субъекта данных на их использование, ситуативная справедливость и удобная для человека семантизация. Каждый из этих факторов может быть положен в основу отдельной рекомендации для разработчика и пользователя ИИ в целях обеспечения общественного блага.

Почему этика в сфере ИИ терпит неудачу?

Однако на практике фундаментальные теории «ответственного ИИ» далеко не всегда находят отражение в документах и практических руководствах. Так, Хагендорфф [Hagendorff, 2020] проанализировал и сравнил 22 сборника руководящих принципов (guidelines) по этике ИИ, подчеркнув, что изученные им руководства ничего не говорят про риски общего ИИ, несмотря на то, что эта тема является одной из самых обсуждаемых в философии. Хагендорфф также рассмотрел, в какой степени этические принципы и ценности реализуются в практике и как можно было бы повысить эффективность этических требований к ИИ. Вывод неутешительный: в настоящее время этика ИИ во многих отношениях терпит неудачу, поскольку у нее нет механизма поощрения и мотивации. Отклонение от различных этических кодексов не носит никаких последствий ни для разработчика, ни для пользователя. А в тех случаях, когда этика интегрирована в различные учреждения, она в основном работает в качестве маркетинговой стратегии. Кроме того, эмпирические эксперименты показывают, что знание руководящих принципов этики не оказывает значительного влияния на принятие решений разработчиками программного обеспечения. Напротив, на практике этика часто рассматривается как постороннее влияние, как излишек или своего рода «надстройка» к техническим проблемам, как не имеющая обязательной силы рамка, навязываемая учреждениями, находящимися за пределами технического сообщества. Распределение ответственности в сочетании с отсутствием знаний о долгосрочных социальных последствиях технологий приводит к тому, что разработчики программного обеспечения не чувствуют ответственности или не понимают морального значения своей работы. А экономические стимулы легко перевешивают приверженность этическим принципам и ценностям. Это означает, что цели, для которых разрабатываются и применяются СИИ, не соответствуют общественным ценностям или фундаментальным принципам социально ответственного проектирования, таким как ориентация на общественное благо,

непричинение вреда, справедливость и объяснимость [Taddeo, Floridi, 2018; Pekka et al., 2018].

Внедрение автономных интеллектуальных систем рискует создать «разрыв ответственности» за недостатки продуктов, которые создаются при помощи этих систем, потому что проектировщики могут думать, что они не несут полной ответственности за реализованные системами продукты. На основе разграничения каузальной ответственности и капацитарной ответственности Дуглас [Douglas et al., 2021] сопоставил представления о моральной ответственности в трех базовых теориях — инструментализме, машинной этике и концепции гибридной ответственности — и сделал вывод о том, что разработчики и пользователи СВП несут моральную ответственность за любые недостатки или неисправности в продуктах, разработанных этими системами.

Антропоцентричный или беспристрастный ИИ?

Аналогичным образом Боддингтон [Boddington, 2020] считает, что люди никогда не должны уступать свою моральную агентность машинам, но машины должны быть приведены в соответствие с человеческими ценностями. Нам необходимо учитывать, как общие допущения о наших моральных способностях и возможностях ИИ влияют на наше собственное мышление об ИИ. Так, идея о том, что мы можем закодировать человеческую этику в машины, может основываться на невыраженном предположении о существовании морального прогресса, сопровождающего прогресс технический. Однако во избежание такой иллюзии стоит обратить внимание на процесс принятия моральных решений человеком и остерегаться любой эрозии нашей собственной моральной агентности и ответственности. Внимание к различиям между людьми и машинами, наряду с вниманием к тому, как люди терпят моральные неудачи, может быть полезно для выявления конкретных способов, которыми ИИ помогает нам совершенствовать нашу собственную моральную агентность, но не узурпировать ее.

Размышления о переносе моральной агентности на искусственные алгоритмы открывает перспективу размышления о возможном моральном статусе общего ИИ. В текущих исследованиях экзистенциальных рисков ИИ общим местом является убеждение в необходимости разработки человеко-ориентированного искусственного морального агента. Вместе с тем в альтернативу антропоцентричному «дружественному людям» ИИ все чаще звучат призывы к созданию беспристрастного ИИ [Daley, 2021] на двух основаниях: во-первых, люди — не единственные существа, достойные моральной заботы; во-вторых, люди, вероятно, уничтожили и еще уничтожат виды, которые вполне могли бы эволюционировать в виды, равные нам или даже превосходящие нас по своим моральным качествам.

Понимая все риски беспристрастного ИИ, обсуждение этических и антропологических его аспектов на всех этапах ЖЦ технологии становится важнейшей социальной задачей [Borenstein, Howard, 2021] профессиональных экспертов. Хотя все члены сообщества, имеющие дело с ИИ, должны отдавать себе отчет в том, какие этические вопросы порождает развитие цифровых технологий, особую ответственность за разработку и внедрение конкретных этических норм и практических предписаний должны взять на себя специалисты по этике в сфере ИИ и государственные органы власти, регулирующие данную сферу. Руководящим органам необходимо следить за оперативным принятием нормативных актов, касающихся вопросов собственности, управления и контроля продуктов в области робототехники и ИИ [Iphofen, Kritikos, 2019]. Но сделать это можно только с опорой на компетенции нового типа профессионала, который способен осуществить комплаенс между технологиями и этическими нормами: специалиста по этике ИИ, имеющего глубокие знания в истории моральных понятий, разбирающегося в функциональных характеристиках СИИ, а также способного применять и передавать этические принципы в рамках корпоративной структуры. Но самое главное качество, которое должно отличать этого специалиста, — это отважность (bravery). В условиях быстро меняющихся технологий, неопределенности языка и концептуального каркаса, множественных противоречий между ожиданиями и интересами стейкхолдеров, новый формирующийся профессионал должен уметь создавать точки соприкосновения между прогрессом технологий, потребностями общества и непреложными человеческими ценностями — и делать это смело и социально ответственно [Gambelin, 2021]. А это значит, что в основе его деятельности должно лежать постоянное соотнесение общественного блага и нормативно-технически толкуемой «надежности» системы, при котором внедрение в широкую общественную практику автоматизированной интеллектуальной системы всегда должно быть ограничено условиями ее социальной приемлемости и технической целесообразности.

Automating Intelligent Systems: Technological Determinism and Substantivism

Sergey V. Garbuk
PhD in Technical Sciences, Director on Scientific Projects.

HSE University.
11, Pokrovsky Bld., 109028 Moscow, Russian Federation.
ORCID 0000-0001-5385-3961
sgarbuk@hse.ru

Anastasia V. Ugleva
PhD, Professor, School of Philosophy and Cultural Studies.
HSE University, Deputy Director of the Center for Transfer and Management of Socioeconomic Information at the National Research University Higher School of Economics.
11, Pokrovsky Bld., 109028 Moscow, Russian Federation.
ORCID 0000-0002-9146-1026
augleva@hse.ru

Abstract. Artificial Intelligence has become so firmly embedded in our lives that its direct influence on shaping the world of the future is inevitable. However, it has taken time for a constructive approach to risk prevention and regulation of technologies at all stages of their life cycle to gradually emerge alongside theoretical speculation about «machine uprising» and other threats to humanity. The subject of special attention is the so-called automated artificial systems, the regulation of which is still limited by normative and technical requirements. The peculiarity of this approach is the conviction of its proponents in the truth of technological determinism, for which “technology” is value neutral. The prevention of ethical risks from the perspective of this approach is practically impossible because regulatory issues are only concerned with the functional characteristics and operational violations of a particular system. This article contrasts technological determinism with technological substantivism, for which “technology” has an independent ethical value, regardless of its instrumental use. The ethical evaluation based on it consists in the procedure of regular correlation of social “good” and “reliability” of the system. The development of a methodology for such a correlation procedure requires special competences that distinguish a new professional field — ethics in the field of AI.

Keywords: philosophy of technology, artificial intelligence, technological determinism, substantivism, ethics AI, highly automated systems.

For citation: Garbuk S.V., Ugleva A.V. Automating Intelligent Systems: Technological Determinism and Substantivism // Chelovek. 2024. Vol. 34, N 4. P. 97–116. DOI: 10.31857/S0236200724040067

Литература / References

Гарбук С.В. Метод оценки влияния параметров стандартизации на эффективность создания и применения систем искусственного интеллекта // Информационно-экономические аспекты стандартизации и технического регулирования. 2022. № 3. С. 4-14.

Garbuk S. V. *Metod ocenki vliyaniya parametrov standartizacii na effektivnost’ sozdaniya i primeneniya sistem iskusstvennogo intellekta* [Method for assessing the influence of standardization parameters on the effectiveness of the creation and implementation of systems of experimental intelligence]. *Information and economic aspects of standardisation and technical regulation*. 2022. N 3. P. 4–14.

Гарбук С.В. Особенности обоснования представительного набора требования к интеллектуальным автотранспортным средствам // Труды НАМИ. 2023. № 4. С. 69–86.

Garbuk S.V. *Osobennosti obosnovaniya predstavitel’nogo nabora trebovaniya k intellektual’nym avtotransportnym sredstvam* [Features of justification of a representative set of requirements for intelligent vehicles]. *Proceedings of NAMI*. 2023. N 4. P. 69–86.

Грунвальд А., Железняк В.Н., Середкина Е.В. Беспилотный автомобиль в свете социальной оценки техники // Технологос. 2019. № 2. С. 41–51.

Grunwald A., Zheleznyak V.N., Seredkina E.V. *Bespilotnyj avtomobil’ v svete social’noj ocenki tekhniki* [The unmanned car in the light of social valuation of technology]. *Technologos*. 2019. N 2. P. 41-51.

Шевченко С.Ю., Шкомова Е.М., Лаврентьева С.В. Гуманитарная экспертиза полного цикла // Горизонты гуманитарного знания. 2021. №2 2. С. 3–17.

Shevchenko S.Y., Shkomova E.M., Lavrentieva S.V. *Gumanitarnaya ekspertiza polnogo cikla* [Full-cycle human health expertise]. *Horizons of Humanitarian Knowledge*. 2021. N 2. P. 3–17.

Шталь Б.К. Этика искусственного интеллекта: Кейсы и варианты решения этических проблем. М.: Издательство ВШЭ, 2024.

Shtal B.K. *Etika iskusstvennogo intellekta: Kejsy i varianty resheniya eticheskih problem* [Ethics of Artificial Intelligence: Cases and Options for Addressing Ethical Challenges]. Moscow: HSE Publ., 2024.

Ashok M., Rohit M., Anton J., Uthayasankar S. Ethical framework for Artificial Intelligence and Digital technologies. *International Journal of Information Management*, 2022. Vol. 62, N 2. P. 102433.

Boddington P. AI and moral thinking: how can we live well with machines to enhance our moral agency? *AI and Ethics*. 2021. N 1. P. 109–111.

Borenstein J., Howard A. Emerging challenges in AI and the need for AI ethics education. *AI and Ethics*. 2020. N 1. P. 61–65.

Bostrom N. Superintelligence: Paths, Dangers, Strategies. Oxford Academ, 2016.

Brendel A.B., Mirbabaie M., Lembcke T.-B., Hofeditz L. Ethical Management of Artificial Intelligence. *Sustainability*. 2021. N 13(4).

Bruneault Fr., Sabourin Laflamme A. AI Ethics: how can information ethics provide a framework to avoid usual conceptual pitfalls? An Overview. *AI & SOCIETY.* 2021. N 36(3). P. 757–766.

Daley K. Two arguments against human-friendly AI. *AI and Ethics*. 2021. N 1. P. 435–444.

Douglas D.M., Howard D., Lacey J. Moral responsibility for computationally designed products. *AI and Ethics*. 2021. N 1. P. 273–281.

Dreyfus H. (1990). Being-in-the-World: A Commentary on Heidegger’s Being and Time, Division I. The MIT Press, 1990.

Floridi L., Cowls J., King T.C. et al. (2020). How to Design AI for Social Good: Seven Essential Factors. *Sci Eng Ethics*, 26: 1771–1796.

Человек. 2024. Т. 35, № 4

Forbes K. Opening the path to ethics in artificial intelligence. *AI Ethics*. 2021. N 1. P. 297–300.

Gabriel I. Artifcial Intelligence, Values, and Alignment. *Minds and Machines*. 2020. N 30. P. 411-437.

Gambelin O. Brave: what it means to be an AI Ethicist. *AI Ethics.* 2021. N 1. P. 87–91.

Garbuk S.V. Intellimetry as a way to ensure AI trustworthiness. *The Proceedings of the 2018 International Conference on Artificial Intelligence Applications and Innovations (IC-AIAI)*. Limassol, Cyprus, 6-10.10.2018. P. 27–30.

Gendron C. Penser l'acceptabilité sociale: au-delà de l'intérêt, les valeurs. *Éthique et relations publiques: pratiques, tensions et perspectives*, 2014. N 11. P. 117–129.

Hagendorff T. The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds & Machines.* 2020. N 30. P. 99–120.

Hager G.D., Drobnis A., Fang F., Ghani R., Greenwald A., Lyons T., Parkes D. C. et al. *Artificial intelligence for social good.* 2017.

Heidegger M. Die Frage nach der Technik. *Die Kunste im technischen Zeitalter*. München, 1954. S. 70–108.

Iphofen R., Kritikos M. Regulating artificial intelligence and robotics: ethics by design in a digital society. *Contemporary Social Science*. 2019. N 16(2). P. 1–15.

Kazim E., Soares Koshiyama A. A high-level overview of AI ethics. *Patterns.* 2021. N 2(9). P. 100314.

Pekka A.-P., Bauer W., Bergmann U., Bieliková M., Bonefeld-Dahl C., Bonnet Y., Bouarfa L. et al. The European Commission's high-level expert group on artificial intelligence: Ethics guidelines for trustworthy ai. *Working Document for stakeholders' consultation*. Brussels, 2018. P. 1–37.

Taddeo M., Floridi L. How AI Can be a Force for Good. *Science*. 2018. Vol. 361(6404). P. 751–752.

ГОСТ Р 70889-2023 Информационные технологии. Искусственный интеллект. Структура жизненного цикла данных (ISO/IEC 8183:2023, MOD).

ПНСТ 840 Искусственный интеллект. Обзор этических и общественных аспектов (ISO/IEC TR 24368:2022 MOD).

ПНСТ 841-2023 Системная и программная инженерия. Требования и оценка качества систем и программного обеспечения (SQuaRE). Руководство по оценке качества систем ИИ (ISO/IEC DTS 25058, MOD).

ПНСТ 776-2022 Информационные технологии. Интеллект искусственный. Управление рисками. (ISO/IEC FDIS 23894, NEQ).

Федеральный закон «О персональных данных» от 27 июля 2006 г. № 152-ФЗ.

SAE J3016 Surface vehicle recommended practice. Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. June, 2018.