



ПРЕДВАРИТЕЛЬНЫЙ
НАЦИОНАЛЬНЫЙ
СТАНДАРТ
РОССИЙСКОЙ
ФЕДЕРАЦИИ

ПНСТ
840—
2023
(ISO/IEC TR
24368:2022)

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

Обзор этических и общественных аспектов

(ISO/IEC TR 24368:2022, Information technology — Artificial intelligence —
Overview of ethical and societal concerns, MOD)

Издание официальное

Москва
Российский институт стандартизации
2023

Предисловие

1 ПОДГОТОВЛЕН Федеральным государственным автономным образовательным учреждением высшего образования «Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ) на основе собственного перевода на русский язык англоязычной версии документа, указанного в пункте 4

2 ВНЕСЕН Техническим комитетом по стандартизации ТК 164 «Искусственный интеллект»

3 УТВЕРЖДЕН И ВВЕДЕН В ДЕЙСТВИЕ Приказом Федерального агентства по техническому регулированию и метрологии от 16 ноября 2023 г. № 60-пнст

4 Настоящий стандарт модифицирован по отношению к международному документу ISO/IEC TR 24368:2022 «Информационные технологии. Искусственный интеллект. Обзор этических и общественных аспектов» (ISO/IEC TR 24368:2022 «Information technology — Artificial intelligence — Overview of ethical and societal concerns», MOD) путем изменения отдельных фраз и ссылок, которые выделены в тексте курсивом.

Внесение указанных технических отклонений направлено на учет национальных интересов Российской Федерации.

Наименование настоящего стандарта изменено относительно наименования указанного международного документа для приведения в соответствие с ГОСТ Р 1.5—2012 (пункт 3.5)

Правила применения настоящего стандарта и проведения его мониторинга установлены в ГОСТ Р 1.16—2011 (разделы 5 и 6).

Федеральное агентство по техническому регулированию и метрологии собирает сведения о практическом применении настоящего стандарта. Данные сведения, а также замечания и предложения по содержанию стандарта можно направить не позднее чем за 4 мес до истечения срока его действия разработчику настоящего стандарта по адресу: info@tc164.ru и/или в Федеральное агентство по техническому регулированию и метрологии по адресу: 123112 Москва, Пресненская набережная, д. 10, стр. 2.

В случае отмены настоящего стандарта соответствующая информация будет опубликована в ежемесячном информационном указателе «Национальные стандарты» и также будет размещена на официальном сайте Федерального агентства по техническому регулированию и метрологии в сети Интернет (www.rst.gov.ru)

© ISO, 2022

© IEC, 2022

© Оформление. ФГБУ «Институт стандартизации», 2023

Настоящий стандарт не может быть полностью или частично воспроизведен, тиражирован и распространен в качестве официального издания без разрешения Федерального агентства по техническому регулированию и метрологии

II

Содержание

1 Область применения	1
2 Нормативные ссылки	1
3 Термины и определения	2
4 Описание	3
4.1 Общие положения	3
4.2 Основополагающие источники	4
4.3 Этические рамочные концепции	5
5 Правозащитная практика	6
5.1 Общие положения	6
6 Темы и принципы	7
6.1 Общие положения	7
6.2 Описание ключевых тем и связанных с ними принципов	8
7 Примеры практик для создания и использования этичного и общественно приемлемого ИИ	14
7.1 Согласование внутреннего процесса с принципами ИИ	14
7.2 Рассмотрение рамок этического анализа	16
8 Рассмотрение вопросов создания и использования этичного и общественно приемлемого ИИ	16
8.1 Общие положения	16
8.2 Неполный перечень этических и общественных аспектов для рассмотрения	17
Приложение А (справочное) Документы, касающиеся принципов в области ИИ	22
Приложение В (справочное) Обзор примеров использования	30
Библиография	39

Введение

Искусственный интеллект (ИИ) обладает потенциалом для инновационных изменений в мире и несет в себе множество преимуществ для общества, организаций и частных лиц. Однако ИИ может нести в себе значительные риски и неопределенности. Специалисты, исследователи, регулирующие органы и физические лица должны знать об этических и общественных аспектах, связанных с системами и приложениями ИИ.

Потенциальные этические проблемы, связанные с ИИ, имеют широкий спектр. Примеры этических и общественных аспектов, связанных с ИИ, различны: от нарушений неприкосновенности частной жизни и безопасности до различных видов дискриминации и воздействия на автономию человека. Источники этических и общественных аспектов включают, но не ограничиваются следующими позициями:

- несанкционированные средства или способы сбора, обработки или раскрытия персональных данных;
- получение и использование необъективных, неточных или нерепрезентативных данных для обучения;
- непрозрачное принятие решений в области машинного обучения (ML) или недостаточное документирование, обычно называемое отсутствием объяснимости;
- отсутствие возможности отслеживания;
- недостаточное понимание общественного воздействия технологии после ее внедрения.

Системы ИИ могут демонстрировать несправедливые результаты работы, особенно когда они обучаются на предвзятых или несоответствующих (в т. ч. искаженных, нерепрезентативных, ошибочных и т. п.) данных или когда модель либо алгоритм не соответствует своему назначению. Ценности, заложенные в алгоритмы, а также выбор задач, для решения которых используют системы и приложения ИИ, могут быть намеренно или ненамеренно сформированы под влиянием мировоззрения и когнитивных предубеждений разработчиков и заинтересованных сторон.

Дальнейшее развитие ИИ может расширить возможности существующих систем и приложений, распространив их на новые области и повысив уровень автоматизации, которым обладают эти системы. Решение этических и общественных аспектов не успевает за быстрым развитием ИИ. Следовательно, проектировщики, разработчики, специалисты по внедрению и пользователи ИИ могут извлечь пользу из гибкого и сбалансированного вклада в этические рамки, принципы ИИ, инструменты и методы снижения рисков, оценку этических факторов, передовую практику тестирования, оценку воздействия и этические обзоры. Эти задачи можно решить с помощью инклюзивного, междисциплинарного, разнообразного и межсекторального подходов, включающих все заинтересованные стороны ИИ, при содействии международной стандартизации, в рамках которой рассмотрены различные вопросы, возникающие в связи с этическими и общественными аспектами ИИ, включая и деятельность Объединенного технического комитета ISO/IEC JTC 1, SC 42.

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

Обзор этических и общественных аспектов

Artificial intelligence. Overview of ethical and societal concerns

Срок действия — с 2024—01—01
до 2027—01—01

1 Область применения

В настоящем стандарте представлен высокоуровневый обзор этических и социальных (общественных) проблем искусственного интеллекта (ИИ). Кроме того, настоящий стандарт:

- содержит информацию о принципах, процессах и методах в этой области;
- предназначен для технологов, регулирующих органов, заинтересованных групп и общества в целом;
- не предназначен для продвижения какого-либо конкретного набора ценностей (системы ценностей).

Настоящий стандарт включает обзор международных стандартов, в которых рассмотрены вопросы, возникающие в связи с этическими и общественными аспектами ИИ.

2 Нормативные ссылки

В настоящем стандарте использованы нормативные ссылки на следующие стандарты:

ГОСТ Р ИСО 9241-210 *Эргономика взаимодействия человек—система. Часть 210. Человеко-ориентированное проектирование интерактивных систем*

ГОСТ Р ИСО 26000 *Руководство по социальной ответственности*

ГОСТ Р ИСО 31000 *Менеджмент риска. Принципы и руководство*

ГОСТ Р ИСО/МЭК 27002 *Информационные технологии. Методы и средства обеспечения безопасности. Свод норм и правил применения мер обеспечения информационной безопасности*

ГОСТ Р ИСО/МЭК 38500 *Информационные технологии. Стратегическое управление ИТ в организации*

Примечание — При пользовании настоящим стандартом целесообразно проверить действие ссылочных стандартов в информационной системе общего пользования — на официальном сайте Федерального агентства по техническому регулированию и метрологии в сети Интернет или по ежегодному информационному указателю «Национальные стандарты», который опубликован по состоянию на 1 января текущего года, и по выпускам ежемесячного информационного указателя «Национальные стандарты» за текущий год. Если заменен ссылочный стандарт, на который дана недатированная ссылка, то рекомендуется использовать действующую версию этого стандарта с учетом всех внесенных в данную версию изменений. Если заменен ссылочный стандарт, на который дана датированная ссылка, то рекомендуется использовать версию этого стандарта с указанным выше годом утверждения (принятия). Если после утверждения настоящего стандарта в ссылочный стандарт, на который дана датированная ссылка, внесено изменение, затрагивающее положение, на которое дана ссылка, то это положение рекомендуется применять без учета данного изменения. Если ссылочный стандарт отменен без замены, то положение, в котором дана ссылка на него, рекомендуется применять в части, не затрагивающей эту ссылку.

3 Термины и определения

В настоящем стандарте применены термины по [1], а также следующие термины с соответствующими определениями:

3.1 агентство (agency): Способность определять свои цели и действовать в соответствии с ними.

Примечание — См. [2], 3.6.2.

3.2 смещенность (bias): Систематическое различие в обработке определенных объектов, людей или групп по сравнению с другими.

Примечание — См. [3], 3.2.2.

3.3 управление данными (data management): Процесс отслеживания всех данных и/или информации, связанной с созданием, производством, распространением, хранением, утилизацией и использованием электронных медиа (e-media), а также сопряженных с этим процессов.

Примечание — См. [4], 3.5.4.

3.4 защита данных (data protection): Правовые, административные, технические или физические меры, принятые для предотвращения несанкционированного доступа к данным и их использования.

Примечание — См. [5], 3.13.5.01.

3.5 равенство (equality): Состояние равенства, особенно в статусе, правах или возможностях.

Примечание — См. [6], 3.9.

3.6 равноправие (equity): Практика устранения преодолемых или исправимых различий между группами людей независимо от того, определяются ли эти группы социально, экономически, демографически или географически.

3.7 справедливость (fairness): Обработка, поведение или результаты, которые основаны на установленных фактах, общественных нормах и убеждениях и не определяются или не зависят от превосходства или несправедливой дискриминации.

Примечания

1 Соображения справедливости в значительной степени контекстуальны и варьируются в зависимости от культуры, поколения, географии и политических взглядов.

2 Справедливость — это не то же самое, что отсутствие предвзятости. Предвзятость не всегда приводит к несправедливости, а несправедливость может быть вызвана факторами, отличными от предвзятости.

3.8 когнитивная предвзятость (cognitive bias): Когнитивная предвзятость человека (human cognitive bias); предубеждение, возникающее при обработке и интерпретации информации человеком.

Примечание — Когнитивная предвзятость человека влияет на суждения и принятие решений (см. [3], 3.2.4).

3.9 жизненный цикл (life cycle): Эволюция системы, продукта, услуги, проекта или другого созданного человеком объекта от замысла до вывода из эксплуатации.

Примечание — См. [7], 3.1.26.

3.10

организация (organization): Компания, корпорация, фирма, предприятие, орган власти или учреждение, лицо или лица или их часть или комбинация частей, объединенная или нет, государственная или частная, которые имеют свои собственные функции и административный аппарат.

[ГОСТ Р ИСО 30000—2013, пункт 3.10]

3.11 приватность (privacy): Права субъекта (обычно физического лица или организации), действующего от своего имени, определять степень сохранения конфиденциальности своей информации.

Примечание — См. [8], 3.1.46.

3.12

ответственность (responsibility): Обязанность действовать и принимать решения для достижения требуемых результатов.

[ГОСТ Р ИСО/МЭК 38500—2017, пункт 2.22]

3.13 **обработка** (treatment): Вид действия, такой как восприятие, наблюдение, представление, предсказание или решение.

Примечание — См. [3], 3.2.2.

3.14 **защищенность; защита; сохранность** (safety): Ожидание того, что при определенных условиях система не приведет к состоянию, при котором жизнь, здоровье, имущество или окружающая среда человека окажутся под угрозой.

Примечание — См. [7], 3.1.48.

3.15 **безопасность** (security): Аспекты, связанные с определением, достижением и поддержанием конфиденциальности, целостности, доступности, подотчетности, аутентичности и достоверности.

Примечания

1 Продукт, система или услуга считаются безопасными в той степени, в которой их пользователи могут быть уверены, что они функционируют (или будут функционировать) так, как задумано, т. е. предусмотренным образом. Обычно это рассматривается в контексте оценки фактических или предполагаемых угроз.

2 См. [9], 3.25.

3.16 **устойчивость** (sustainability): Состояние международной системы, включая экологические, общественные и экономические аспекты, при котором потребности настоящего удовлетворяются без ущерба для способности будущих поколений удовлетворять свои собственные потребности.

Примечание — См. [10], 3.1.

3.17 **прослеживаемость** (traceability): Способность идентифицировать или восстановить историю, происхождение, применение, использование и местонахождение объекта/элемента или его характеристик.

3.18 **цепочка создания стоимости** (value chain): Совокупность видов деятельности или сторон, которые создают или получают ценность в виде продукции или услуг.

Примечание — См. [11], 3.2.11.

4 Описание

4.1 Общие положения

Этические и общественные аспекты являются факторами, которые необходимо учитывать при разработке и использовании систем и приложений ИИ. Учет контекста, масштаба и рисков может смягчить нежелательные этические и социальные (общественные) последствия и вред. Примеры областей, где возрастает риск нежелательных этических и общественных последствий и вреда, включают следующие (см. [12]):

- финансовый ущерб;
- психологический вред;
- вред физическому здоровью или безопасности;
- нематериальная собственность [например, кража интеллектуальной собственности (ИС), ущерб репутации компании];
- социальные или политические системы (например, вмешательство в выборы, потеря доверия к властям);
- гражданские свободы (например, необоснованное тюремное заключение или другое наказание, цензура, нарушение неприкосновенности частной жизни).

В случае пренебрежения анализом таких факторов существует риск того, что технология может повлечь за собой значительные общественные или иные последствия с возможными непреднамеренными или непрогнозируемыми затратами, даже если она безупречна с технической точки зрения.

4.2 Основопологающие источники

В различных источниках этические и общественные аспекты рассмотрены конкретно или в общем виде. Некоторые из этих источников приведены ниже.

Прежде всего, в [10] приведены рекомендации для разработчиков стандартов по учету устойчивости в их деятельности с конкретной ссылкой на руководство по социальной ответственности в *ГОСТ Р ИСО 26000*. Таким образом, настоящий стандарт описывает социальную ответственность в форме, которая может послужить основой для деятельности, связанной со стандартизацией доверенного ИИ.

В *ГОСТ Р ИСО 26000* представлено руководство по социальной ответственности. Оно основано на фундаментальных практиках:

- признание социальной ответственности в организации;
- выявление и привлечение заинтересованных сторон.

Без *данных* разработка и использование ИИ невозможны. Поэтому значимость данных и их качества делает прослеживаемость и управление данными ключевыми аспектами при использовании и развитии ИИ.

В основе создания этического и устойчивого ИИ лежат следующие элементы, ориентированные на данные:

- сбор данных (включая средства или способы такого сбора данных);
- подготовка данных;
- контроль прослеживаемости;
- контроль доступа и обмена данными (аутентификация);
- защита данных;
- контроль хранения (добавление, изменение, удаление);
- качество данных.

Эти элементы влияют на объяснимость, прозрачность, безопасность и конфиденциальность, особенно в случаях создания, контроля или обработки персональной идентифицируемой информации. Прослеживаемость и управление данными являются существенными аспектами для той организации, которая использует или разрабатывает системы и приложения ИИ.

В [13] изложена ценность данных, риски и ограничения при регулировании вопросов, касающихся того, как данные собирают, хранят, распространяют, утилизируют, предоставляют (отражают) и используют в процессе принятия решений и организационных процедур. Результаты деятельности по извлечению данных или машинному обучению при составлении отчетов и принятии решений рассмотрены как еще одна форма данных, на которые, следовательно, распространены те же руководящие принципы управления данными.

Кроме того, описание этических и общественных аспектов, связанных с системами и приложениями ИИ, может быть основано на международной практике использования ИИ.

В [1] также приведен набор заинтересованных сторон, участвующих в разработке и использовании системы ИИ, а также описано участие различных заинтересованных сторон ИИ в цепочке создания стоимости системы ИИ, которые включают поставщика ИИ, производителя ИИ, заказчика ИИ, партнера ИИ и субъекта ИИ, в том числе различные второстепенные роли этих заинтересованных сторон. В настоящем стандарте все эти разные типы заинтересованных сторон обобщенно именуются заинтересованными сторонами.

Органы регулирования и выработки политики в качестве вспомогательной роли субъекта ИИ перечислены в [1]. Регулирующие роли для ИИ в настоящее время широко не определены, но сделан ряд предложений, включая: организации, назначенные отдельными заинтересованными сторонами; отраслевые представительные органы; самостоятельно назначенные субъекты гражданского общества или учреждения, созданные на основе национального законодательства или международного договора.

Перечисленные выше аспекты включены в описание этических и общественных аспектов, связанных с ИИ (см. [1]).

Поскольку ИИ способен оказать воздействие на широкий круг заинтересованных сторон общества, включая будущие поколения, на которые повлияют изменения в окружающей среде (косвенно затронутые заинтересованные стороны). Например, изображения пешеходов на тротуаре могут быть получены с помощью технологии автономных транспортных средств, или рядовые люди могут стать объектом полицейского наблюдения с использованием оборудования, предназначенного для изучения лиц, подозреваемых в преступлении.

Руководящие принципы по управлению рисками, связанными с ИИ, с которыми сталкиваются организации при разработке и применении технических методов и систем ИИ, приведены в [14]. В ГОСТ Р ИСО 31000 представлена структура, а также содержится руководство, необходимое при разработке и использовании систем ИИ. Система управления рисками, описанная в [14], помогает в описании этических и общественных аспектов в настоящем стандарте.

Типы и формы смещенности, способы их измерения и смягчения, а также концепция справедливости в системах ИИ описаны в [3]. Предвзятость и справедливость важны для описания специфических для ИИ этических и общественных аспектов.

Введение в прозрачность и объяснимость систем ИИ, которые являются существенными аспектами надежности и могут влиять на этические и общественные аспекты, приведено в [15].

Набор из 124 примеров использования приложений ИИ в 24 различных областях применения описан в [16]. В данных примерах определены заинтересованные стороны, активы и ценности заинтересованных сторон, а также перечислены угрозы и уязвимости описанной системы и приложения ИИ. Некоторые из примеров описывают общественные и этические проблемы.

Руководство по последствиям управления для организаций, участвующих в разработке и использовании систем ИИ, рассмотрено в [17]. Это руководство дополняет меры, определенные в ГОСТ Р ИСО/МЭК 38500 (см. также [13], [18]).

Управление является ключевым механизмом, с помощью которого организация может решать этические и социальные последствия своего участия в системах и приложениях ИИ.

Определены также требования к созданию, внедрению, поддержанию и постоянному усовершенствованию системы управления информационной безопасностью в контексте организации (см. [19]). Он также включает требования к оценке и обработке рисков информационной безопасности с учетом потребностей организации. Требования, изложенные в [19], носят общий характер и могут служить основой для системного управления информационной безопасностью в контексте ИИ. Это, в свою очередь, может оказать последующее влияние на этические и общественные вопросы в системах и приложениях ИИ. ГОСТ Р ИСО/МЭК 27002 содержит руководящие принципы для стандартов информационной безопасности организации и практики управления информационной безопасностью, включая выбор, внедрение и управление средствами контроля, а также для управления конфиденциальностью данных в организации.

Примечание — См. [19] для управления конфиденциальной информацией в контексте ИИ, а также для руководства по созданию, внедрению, поддержанию и постоянному совершенствованию системы управления конфиденциальной информацией в форме расширения.

4.3 Этические рамочные концепции

4.3.1 Общие положения

Этика ИИ — это область прикладной этики, т. е. принципы и практика редко являются результатом применения этических теорий. Тем не менее многие проблемы тесно связаны с традиционными этическими концепциями и проблемами, например: такие понятия, как конфиденциальность, объективность и автономия, могут быть рассмотрены в существующих этических рамках. Дополнительные возможные этические рамки (этические концепции и подходы) приведены в [20]. Этот список этических рамок не является ни исчерпывающим, ни взаимоисключающим. Следовательно, могут быть рассмотрены этические рамки, выходящие за границы перечисленных в [21].

В 4.3.2—4.3.4 приведены основные этические системы, которые условно обобщают философские теории и концепции целого ряда мыслителей, представителей различных философских школ. В рамках каждой из представленных систем существуют различные подходы, которые исторически отражают философские поиски в сфере этических проблем. Такие подходы и этические системы не должны быть определены и истолкованы как правильные или неправильные, верные или ошибочные, полезные или вредные, они приведены исключительно с целью показать многообразие теоретических изысканий философов разных эпох. Исходя из этого представленное общее описание основных этических систем не имеет целью продвигать определенные философские теории и не носит рекомендательный характер¹⁾.

¹⁾ Внесено на основании замечаний и предложений экспертов для учета в настоящем стандарте национальных интересов и защиты пользователей и населения.

4.3.2 Этика добродетели

Этика добродетели — это этическая система, определяющая совокупность добродетелей, к которым следует стремиться (например, уважение, честность, мужество, сострадание, щедрость), и совокупность пороков или недостатков (например, нечестность, ненависть), которых следует избегать. Достоинством этики добродетели является ее гибкость и устремленность. Ее основной недостаток заключается в том, что она не предлагает конкретных рекомендаций для реализации. Систему ИИ можно считать объективной при наличии механизма, с помощью которого эта добродетель реализуется на практике. Однако с учетом технических ограничений этика добродетели может служить полезным инструментом для определения того, является ли применение ИИ отражением человеческих добродетелей.

4.3.3 Утилитаризм

Утилитаризм — это этическая система, которая стремится к максимальному благу и минимальному вреду. Утилитарный выбор — это выбор, который приносит наибольшее благо и наносит наименьший вред всем заинтересованным сторонам. После логического объяснения этических аспектов проблемы утилитарные подходы универсальны и интуитивно понятны. Основным недостатком утилитаризма заключается в том, что утилитарные подходы допускают причинение вреда некоторым ради блага общего. В качестве примера можно привести проблему тележки [22], где утилитаризм поддерживает убийство, или пример с трансплантационными пациентами в больнице, где утилитаризм поддерживает вскрытие здорового донора для пересадки его органов нескольким пациентам (см. [23]). Кроме того, многие моральные соображения трудно поддаются количественной оценке (например, достоинство) или являются субъективными — то, что хорошо для одного человека, может не быть хорошо для другого. Моральные соображения настолько различны, что их трудно сопоставить друг с другом, например загрязнение окружающей среды против общественной правды. Кроме того, утилитаризм как система является разновидностью консеквенциализма — доктрины, согласно которой цель оправдывает средства. Консеквенциализм поддерживает создание решений, обеспечивающих чистую выгоду, но не требует, чтобы эти решения функционировали этично, например беспристрастно (непредвзято) (см. [24]).

4.3.4 Деонтология

Деонтология — это этическая система, которая оценивает мораль по набору заранее определенных обязанностей или правил. Конкретным механизмом для такого определения является набор правил или кодифицированных норм, которые можно проанализировать в данный момент, не просчитывая, какими могут быть последствия этих действий. Примером такого правила является равенство возможностей в рамках обеспечения справедливости. Равенство возможностей диктует, что люди, которые претендуют на возможность, имеют равные шансы сделать это, независимо от принадлежности к социальной группе. Основным недостатком деонтологии заключается в том, что такие универсальные правила трудно вывести на практике и они могут быть неустойчивыми при применении в перекрестных контекстах или в сильно изменчивой среде.

5 Правозащитная практика

5.1 Общие положения

Международные права человека, изложенные во Всеобщей декларации прав человека (см. [25]), цели в области устойчивого развития ООН (см. [26]) и руководящие принципы ООН в области предпринимательской деятельности и прав человека (см. [27]) являются фундаментальными моральными принципами, на которые человек имеет право по определению. Они могут служить руководящей основой для управления корпоративной ответственностью в отношении систем и приложений ИИ с учетом международного одобрения в качестве более зрелой основы для оценки политики и технологий. Международные права человека могут также обеспечить установленный процесс для проведения должного соблюдения предосторожности и оценки воздействия. Вопросы прав человека для управления ИИ в организациях рассмотрены в [17].

Рамочные подходы, такие как этика в отношении медицинского обслуживания или социальная справедливость, способствуют многим темам, представленным в 6.2, включая конфиденциальность, объективность и недискриминацию, продвижение человеческих ценностей, безопасность и надежность, а также уважение международных норм поведения. Кроме того, многие источники международного права и правовые принципы могут индивидуально дополнять несколько тем. Они включают, но не ограничиваются следующим:

- всеобщая декларация прав человека (см. [25]);
- руководящие принципы ООН в области предпринимательской деятельности и прав человека (см. [27]);
- Международная конвенция о ликвидации всех форм расовой дискриминации (см. [28]);
- декларация о правах коренных народов (см. [29]);
- конвенция о ликвидации дискриминации в отношении женщин (см. [30]);
- конвенция о правах инвалидов (см. [31]).

Эти источники можно понять с точки зрения их цели — повышения стандартов и практики в отношении бизнеса и прав человека, а также достижения значительных результатов для подверженных воздействию лиц и сообществ. Соответствующие вопросы включают должную предосторожность в деятельности организации для выявления и смягчения воздействия на права человека. В тех случаях, когда права человека затрагиваются деятельностью организации в области ИИ, четкие, доступные, прогнозируемые и справедливые механизмы могут быть применены для разрешения жалобы.

Некоторые примеры потенциального воздействия ИИ на гражданские и политические права человека включают:

- право на жизнь, свободу и личную безопасность (например, использование автономного оружия или мотивированных ИИ практик сбора данных);
- право на свободу мнения, выражения и доступ к информации (например, использование фильтрации или синтеза цифрового контента с помощью ИИ);
- свобода от дискриминации и право на равенство перед законом (например, на это влияет использование алгоритмов оценки судебных рисков с помощью ИИ, предиктивного полицейского инструмента для профилактики преступлений или финансовых технологий);
- свобода от неоправданного вмешательства в частную жизнь, семью, дом или корреспонденцию (например, несанкционированные, основанные на ИИ средства и меры по сбору чувствительных биометрических и физиологических данных);
- право на образование и желаемую работу (например, использование ИИ при найме людей на работу или предоставлении доступа к образованию и обучению).

6 Темы и принципы

6.1 Общие положения

В дополнение к правозащитной практике, упомянутой в разделе 5, принципы ИИ могут помочь организациям разрабатывать и использовать ИИ ответственным образом. Цель этих принципов заключается в том, чтобы помочь организациям выйти за рамки непричинения вреда и сосредоточиться на обеспечении пользы технологии, например: разработка ИИ, призванного содействовать общественному благу и выполняющего эту конкретную функцию, а не просто направленного на предотвращение вреда.

Эти принципы распространяются не только на поставщиков и производителей ИИ, но и на предполагаемое использование систем ИИ. Когда системы ИИ становятся доступными для потребителей ИИ и других заинтересованных сторон, следует также изучить возможности их ненадлежащего (непредусмотренного) использования, злонамеренного применения или неиспользования (см. [15], 9.9.1):

- чрезмерную зависимость от систем ИИ, ведущую к негативным результатам (злоупотребление);
- недостаточное использование систем ИИ, приводящее к негативным результатам (неиспользование);
- негативные результаты, возникающие в результате использования или перепрофилирования систем ИИ в области, для которой они не были разработаны и испытаны (например, неправомерное или непредусмотренное использование).

Системы ИИ особенно подвержены злоупотреблениям и нецелевому использованию из-за того, что они имитируют человеческие возможности. Когда система кажется похожей на человеческую, но при этом не имеет контекста, который бы учитывали люди, пользователи могут использовать ее неправильно или не по назначению. Такое неправильное использование или неиспользование может возникнуть из-за того, что системе доверяют больше или меньше, чем это оправдано, например: при автономном вождении, медицинской диагностике или одобрении кредитов.

В ответ на эти опасения, в ожидании государственного регулирования или в попытке саморегулирования отрасли, в международном сообществе появилось несколько наборов принципов для ИИ. Они были задокументированы в различных публикациях (см. приложение А).

Настоящий раздел соответствует структуре, изложенной в докладе Центра Беркмана Кляйна (см. А.1), в котором принципы ИИ сгруппированы по темам. Темы возникли из вопросов этического характера, для решения которых принципы созданы. Принципы в рамках этих тематических групп могут значительно различаться и даже противоречить друг другу. Темы, специфичные для ИИ, дополняют темы, представленные в *ГОСТ Р ИСО 26000*, в нем изложены принципы, которые организация должна учитывать с целью социально ответственного поведения.

6.2 Описание ключевых тем и связанных с ними принципов

6.2.1 Подотчетность

Подотчетность (см. [32]) обеспечивается, когда организация принимает на себя ответственность за воздействие на заинтересованные стороны, общество, экономику и окружающую среду. Подотчетность означает, что организация проводит соответствующую проверку (контроль, отчетность) с последующими обязательствами по реагированию на результаты этой проверки (контроля). Следовательно, подотчетность в отношении решений ИИ означает, что организация способна принять на себя ответственность за решения, принятые от ее имени, и не освобождается от ответственности за ошибочные решения, основанные, например, на результатах машинного обучения ИИ.

Подотчетность определяет, что организация и ее представители несут ответственность за негативные последствия, возникшие в результате проектирования, разработки, использования или неправильного использования систем и приложений ИИ любым лицом, применяющим технологии ИИ. Подотчетность также обеспечивает сосредоточенность и внимание на рассмотрении непреднамеренных последствий, которые могут возникнуть из-за эволюционного характера систем и приложений ИИ и сложности прогнозирования того, как системы и приложения ИИ могут быть использованы и перепрофилированы после развертывания. Без четких требований к подотчетности ограничения и границы не установлены, и потенциальный вред может остаться незамеченным.

Ответственность за решения организации в конечном итоге лежит на группе людей, которые руководят и контролируют организацию. Однако зачастую ответственность делегируется соответствующим ответственным лицам, поэтому сотрудников можно обучить разбираться в последствиях их работы по разработке, внедрению или использованию инструментов ИИ и нести ответственность за свою зону деятельности. Они также могут понять, какие действия необходимо предпринять для принятия соответствующих решений как в организационном, так и в инженерном контексте. Например, организация несет ответственность за разработку недискриминационной и прозрачной политики. В обязанности инженера входит разработка систем и приложений ИИ, которые следуют этой политике, обеспечивая разработку и использование недискриминационных, прозрачных и объяснимых алгоритмов.

Подотчетность обеспечивает необходимые ограничения, помогающие ограничить потенциальные негативные последствия и установить реалистичное и действенное управление рисками для организации. В совокупности они помогают определить, как расставить приоритеты в зоне ответственности. Некоторые аспекты, охватываемые этой темой, следующие:

- работа с заинтересованными сторонами для оценки потенциального влияния системы на ранних этапах проектирования;
- подтверждение того, что потребности заинтересованных сторон действительно удовлетворены;
- проверка того, что система ИИ работает так, как предполагалось;
- обеспечение прослеживаемости данных и алгоритмов на протяжении всей цепочки создания стоимости ИИ;
- обеспечение возможности проведения аудита третьей стороной и принятие мер в соответствии с его результатами;
- предоставление возможностей для оспаривания решений ИИ;
- исправление ошибочных или негативных решений ИИ, когда оспаривание или обжалование невозможно.

6.2.2 Справедливость и недискриминация

Тема справедливости и недискриминации (см. [33], [34]) направлена на обеспечение того, чтобы ИИ качественно работал для людей из различных социальных групп, особенно для тех, кто лишен социальной, политической или экономической власти в местном, национальном и международном контекстах. Эти социальные группы различаются в разных контекстах и включают, но не ограничиваются теми, которые требуют защиты от дискриминации по признаку пола, расы, цвета кожи, этнического или социального происхождения, генетических особенностей, языка, религии или убеждений, политических

или любых других взглядов, принадлежности к национальному меньшинству, имущественного положения, рождения, инвалидности, возраста или сексуальной ориентации. Некоторые аспекты, охватываемые этой темой, следующие:

- смягчение предвзятого отношения к представителям различных групп;
- обеспечение сбора и применения учебных данных и данных о пользователях таким образом, чтобы они отражали интересы представителей различных групп;
- объективное, равноправное и равнозначное отношение к представителям различных групп;
- рассмотрение того, как ИИ может по-разному воздействовать на представителей различных групп;
- обеспечение равных возможностей для человеческого развития и обучения для всех членов различных групп;
- обеспечение смягчения влияния когнитивных или общественных предубеждений человека в процессе сбора и обработки данных, проектирования системы, обучения модели и принятия других решений по развитию, которые принимают отдельные лица;
- *обеспечение и принятие мер, препятствующих разработке и использованию систем, способных манипулировать человеком, в прямой или опосредованной форме навязывать ему определенные цели, решения или предпочтения, оказывать другое воздействие, снижающее или негативно влияющее на автономию и самостоятельность человека¹⁾*;
- предоставление заинтересованным сторонам возможности обжаловать решение, если они считают его несправедливым.

6.2.3 Прозрачность и объяснимость

Принцип прозрачности в *ГОСТ Р ИСО 26000* расширен и включает в себя объяснимость систем ИИ, чтобы гарантировать, что, когда заинтересованные стороны взаимодействуют с системой ИИ и приложением, ее решения являются одновременно прозрачными и объяснимыми, тем самым обеспечивая понимание ее работы.

Тема прозрачности и объяснимости направлена на то, чтобы люди понимали, когда они взаимодействуют с системой ИИ, как она принимает решения и как она разработана и протестирована, чтобы гарантировать, что система работает так, как задумано. Принцип в *ГОСТ Р ИСО 26000* направлен на обеспечение прозрачности целей и процессов организации, в то время как специфические принципы ИИ направлены на обеспечение того, чтобы система ИИ была понятна в том, как она работает. Некоторые аспекты, которые охватываются этой темой, следующие:

- раскрытие прослеживаемости, информации об алгоритмах, обучающих данных и пользовательских данных, включая то, как они собраны;
- раскрытие методов оценки и метрик, используемых для подтверждения того, как работает система;
- объяснение заинтересованным сторонам исходных данных, которые использованы для принятия решения;
- объяснение заинтересованным сторонам, насколько это возможно, того, как система ИИ пришла к тому или иному решению;
- уведомление заинтересованных сторон, когда система ИИ принимает решение в отношении данных сторон;
- уведомление заинтересованных сторон, когда они взаимодействуют с системой ИИ;
- *уведомление заинтересованных сторон, когда они взаимодействуют с системой ИИ, о конкретных целях и назначении системы, а также недопустимости ее использования для других целей, в т. ч. связанных со сбором данных или информации, не относящейся к заявленным целям и назначению системы¹⁾*;
- *рассмотрение возможности понятной и легко доступной для считывания маркировки контента, созданного с использованием ИИ, для уведомления заинтересованных сторон о его происхождении¹⁾*;
- рассмотрение возможности предоставления заинтересованным сторонам тестовых примеров, чтобы увидеть, как система ИИ и приложение реагируют на различные ситуации.

¹⁾ Внесено на основании предложений экспертов для учета в настоящем стандарте национальных интересов и защиты пользователей и населения.

Аналогично подотчетности прозрачность и объяснимость применяются к решениям, принимаемым как на организационном, так и на алгоритмическом уровнях. Более подробная информация о прозрачности и объяснимости приведена в [34]—[39].

6.2.4 Профессиональная ответственность

Подотчетность и этическое поведение тесно связаны с профессиональной ответственностью. Тема профессиональной ответственности направлена на то, чтобы специалисты, которые проектируют, разрабатывают или внедряют системы и приложения ИИ или продукты и системы на основе ИИ, осознавали свое уникальное положение, позволяющее им оказывать влияние на людей, общество и будущее ИИ — тем более что политика, нормы и принципы часто отстают от новых и развивающихся технологий. ИИ стимулирует их профессионализм и честность обеспечить ответственное развитие систем и приложений ИИ и направить их опыт на общественное обсуждение, образование и управление. Это также относится к экспертам в предметной области и к другим специалистам, которые используют технологии ИИ для выводов или принятия решений для осуществления действий в будущем (например, представители по найму, сотрудники приемных комиссий, ученые, судьи и сотрудники правоохранительных органов). Эта тема тесно связана с этикой добродетели (см. 4.3.2), которая фокусируется на чертах, характеристиках и достоинствах специалиста по ИИ. Некоторые аспекты, охватываемые этой темой, следующие:

- строгие методы оценки качества системы ИИ, понимание вреда, который может быть причинен проблемами в области качества, а также меры по снижению вреда;
- уделение целенаправленного внимания вероятным последствиям применения системы ИИ на этапе разработки, включая долгосрочные, глобальные последствия, а также те, которые могут возникнуть в результате неиспользования, неправильного использования и злоупотребления;
- консультации с соответствующими группами заинтересованных сторон при разработке и управлении использованием систем и приложений ИИ, включая политиков, профильных экспертов и заинтересованные организации, особенно при определении соответствия целям;
- этические принципы при разработке, включая инклюзивное проектирование, разнообразие, прозрачность и защиту частной жизни;
- профессиональные ценности и практика научной добросовестности, включая строгие научные методы, приверженность открытому исследованию, междисциплинарное сотрудничество. Это касается оценки научной обоснованности систем и приложений ИИ, делающих прогнозы, предоставляющих рекомендации или решения на основе коррелирующих данных, в конкретном контексте, в котором система может быть развернута.

6.2.5 Продвижение человеческих ценностей

Тема продвижения человеческих ценностей направлена на обеспечение того, чтобы ИИ внедрялся и использовался таким образом, чтобы приносить максимальную пользу обществу, способствовать благополучию и процветанию человечества. Некоторые человеческие ценности потенциально универсальны (см. [40]). Некоторые уникальные человеческие ценности включают (но не ограничиваются ими) целостность, свободу и уважение. ИИ можно внедрять таким образом, чтобы уважались общественные нормы, наилучшие интересы человечества, основные культурные убеждения и ценности, а также ценности заинтересованных сторон (см. [41]). Эта тема выходит за рамки непричинения вреда и фокусируется на пользе технологии, например: разработка ИИ, предназначенного для содействия общественному благу и выполняющего эту конкретную функцию в дополнение к стремлению избежать вреда. Продвижение человеческих ценностей в процессе разработки, внедрения и развертывания ИИ имеет решающее значение.

Конкретные области применения ИИ, направленные на продвижение человеческих ценностей, включают (но не ограничиваются ими):

- улучшение здоровья и здравоохранения;
- улучшение условий жизни;
- улучшение условий труда;
- усилия по охране окружающей среды и обеспечению устойчивости.

6.2.6 Конфиденциальность и защита частной жизни, приватность

Приватность направлена на то, чтобы системы и приложения ИИ разрабатывались и внедрялись с учетом права физических лиц на приватность, а также это касается и умерших лиц (через управляющего их наследством или номинального представителя, если это применимо). Право на неприкосновенность частной жизни стало одной из наиболее заметных тем в разработке ИИ, во многом благодаря Общему положению о защите данных в Европейском союзе (см. [42]). Организация может обеспечить

использование механизмов для реализации и демонстрации соответствия законодательным требованиям, касающимся неприкосновенности частной жизни и правил защиты данных (см. [42]). Процедуры, методы и критерии необходимы для обеспечения того, чтобы алгоритмы ИИ подбирались, разрабатывались и использовались для репрезентативного и прозрачного применения человеком, обеспечивая соблюдение неприкосновенности частной жизни. Общие аспекты неприкосновенности частной жизни включают:

- ограничение источника, сбора, использования или раскрытия данных только тем объемом, который необходим для выполнения намеченных целей и задач;
- информирование о целях обработки личной идентифицируемой информации и о любом ее совместном использовании;
- согласие: прозрачность данных о физическом или умершем лице, данные физических или умерших лиц не должны собираться или использоваться без их ведома (или ведома представителя) или разрешения;
- контроль над использованием данных: контроль физического или умершего лица (здесь и далее — через представителя) над использованием его личной идентифицируемой информации;
- степень влияния физических или умерших лиц на то, как и почему использована их личная информация;
- возможность ограничения обработки данных: право физических или умерших лиц на ограничение сбора или использования данных в связи с технологией ИИ;
- исправление: предоставление физическим или умершим лицам возможности изменить информацию, если она неверна;
- удаление: предоставление физическим или умершим лицам возможности удалить личные данные из системы и приложения ИИ;
- предоставление физическим или умершим лицам возможности просматривать персональные данные, используемые системой и приложением ИИ;
- конфиденциальность при проектировании: включение соображений конфиденциальности данных в разработку систем и приложений ИИ и на протяжении всего жизненного цикла использования данных (см. [43]);
- разрешение споров: предложение механизмов для разрешения споров в связи с этими аспектами.

6.2.7 Защищенность и безопасность

Защищенность и безопасность — это разные, но тесно связанные понятия (см. [44]). Этика применительно к теме защиты и безопасности, которую можно назвать предотвращением потерь, фокусируется на предполагаемых возможностях систем и приложений ИИ и их устойчивости к компрометации. Защищенность (защита) направлена на то, чтобы избежать уменьшения травмирования и ущерба здоровью людей, имуществу и окружающей среде, и часто является существенным аспектом регулирования конкретных продуктов. Безопасность направлена на снижение вероятности несанкционированного и ненадлежащего доступа к данным, активам или имуществу, поскольку ИИ может создавать новые угрозы безопасности, такие как состязательные атаки, отравление данных и кража моделей (см. [45]). Особое значение имеют системы, взаимодействующие с физическим миром, которые могут влиять на людей, сообщества или и то и другое. Эта тема связана с другими темами, которые касаются человеческого контроля над технологиями и подотчетностью, поскольку механизмы реализации этих тем также могут помочь повысить безопасность и защищенность систем и приложений ИИ. Некоторые аспекты, которые охватываются этой темой и часто являются частью подхода к ИИ, основанного на обеспечении безопасности при проектировании, следующие:

- надежность: системы и приложения ИИ, которые постоянно работают в соответствии с поставленными задачами;
- тестирование и мониторинг могут помочь минимизировать последующее неправильное использование и злоупотребление системой и приложением ИИ. Тестирование и мониторинг также помогают определить, являются ли системы надежными и спроектированными таким образом, чтобы противостоять непредвиденным событиям и состязательным атакам, которые могут нанести ущерб или способны манипулировать;
- предсказуемость: повышение предсказуемости систем и приложений ИИ обычно представляется как ключевой механизм, помогающий снизить риск компрометации систем ИИ;
- оценка и снижение рисков, связанных с безопасностью, с учетом сложности среды, в которой могут быть использованы системы и приложения ИИ (например, автономное вождение);

- обеспечение работы систем и приложений ИИ в соответствии с их назначением (робастность) во всех сценариях. Это относится, в частности, к системам непрерывного обучения.

6.2.8 Контроль технологии со стороны человека

Тема человеческого контроля технологий направлена на обеспечение того, чтобы значимые решения оставались предметом рассмотрения человеком (см. ссылки [33, с. 5], [40], [46] и [47, пункт 1]). В рамках этой темы рассмотрена важность уважения автономии пользователей, на которых влияют автоматизированные решения. Человеческий контроль может быть достигнут путем проектирования систем и приложений ИИ таким образом, чтобы те, на кого влияют автоматизированные решения, могли запросить и получить человеческий анализ этих решений. Для систем с высокой степенью риска это может также принять форму требования о наличии квалифицированного специалиста для авторизации автоматизированных решений. В этой теме рассматриваются следующие аспекты:

- разработка систем и приложений ИИ, позволяющих операторам (людям) проверять или санкционировать автоматизированные решения;
- обеспечение возможности отказаться от принятия автоматизированных решений;
- критическая оценка того, как и когда делегировать решения системам и приложениям ИИ, оценка того, как такие системы и приложения могут передать управление человеку таким образом, чтобы оно было осмысленным и понятным.

Системы и приложения ИИ могут поддерживать автономию человека и принятие им решений, как предписывает принцип уважения автономии человека. Для этого необходимо, чтобы системы и приложения ИИ способствовали созданию демократического, процветающего и справедливого общества, поддерживая самостоятельность пользователя, поощряя права и позволяя человеку осуществлять соответствующий надзор.

6.2.9 Вовлечение и развитие сообщества

Социальная ответственность, связанная с вовлечением и развитием сообществ, подразумевает выявление заинтересованных сторон, на которые влияют система и приложение ИИ, а также выяснение и удовлетворение их проблем прозрачным образом. Особое внимание уделяется уязвимым и менее развитым сообществам, проживающим в тех географических зонах, в которых работают системы и приложения ИИ. Поскольку многие продукты и услуги на основе ИИ являются виртуальными и транснациональными, выявление таких сообществ может представлять собой сложную задачу. Системы ИИ могут оказывать потенциальное воздействие на жизнь всех людей во всем мире независимо от того, являются ли они прямыми потребителями услуг на основе ИИ или косвенно затронутыми заинтересованными сторонами. Поэтому организация может рассмотреть возможность не только выявления и взаимодействия с заинтересованными сторонами, но и признания (и построения) отношений с сообществом, места организации в этом сообществе и ее ответственности за социальное, политическое, экономическое и культурное развитие этого сообщества.

Еще одно соображение для организаций, стремящихся привлечь затронутые сообщества, заключается в том, чтобы включить вклад сообщества в инновационный процесс при создании и эксплуатации товаров и услуг на основе ИИ. Это особенно актуально, когда внедрение ИИ нарушает существующие модели жизни сообщества, его сплоченность и возможности трудоустройства. В частности, воздействие на занятость посредством автоматизации на основе ИИ поднимает вопросы образования и культуры сообщества, создания рабочих мест, развития навыков и инвестиций в социальные и медицинские меры сообщества для смягчения негативных последствий этого воздействия. Организации могут привлекать сообщество различными способами, например: в рамках интервью, фокус-групп, семинаров, конференций, анкетирования, опросов, исследований и консультативных советов заинтересованных сторон. Более подробная информация приведена в *ГОСТ Р ИСО 26000* (см. также [34], [41]). Значимость консультаций с заинтересованными сторонами, что для некоторых проектов может потребовать участия сообщества и, возможно, общественного мнения, отражена в [35].

6.2.10 Проектирование, ориентированное на человека

Тема проектирования, ориентированного на человека, в *ГОСТ Р ИСО 9241-210* направлена на обеспечение учета интересов людей, использующих систему и приложение ИИ, а также других заинтересованных групп, включая тех, на кого может повлиять (прямо или косвенно) их применение. Обеспечение того, что система ИИ и приложение ориентированы на человека [34], включает в себя следующее:

- а) проектирование основано на четком понимании пользователей, задач и среды;
- б) пользователи вовлекаются в процесс проектирования и разработки;
- с) проектирование осуществляется и дорабатывается на основе оценки, ориентированной на пользователя;

- d) процесс является итеративным (т. е. до получения нужного результата);
- e) проектирование охватывает и учитывает весь пользовательский опыт;
- f) команда разработчиков включает в себя представителей различных дисциплин.

Одно из предложений о том, как сделать проектирование, удовлетворяющее этим целям, приведено в [41].

6.2.11 Уважение верховенства закона

Верховенство закона требует, в частности, чтобы даже влиятельные организации и системы соблюдали закон. Соблюдение правовых требований, включая обращение в суд в случае необходимости в отношении решений, вынесенных системами и приложениями ИИ, является общепризнанным аспектом информационно-коммуникационных технологий (ИКТ), управления данными и управления рисками. *При этом важно соблюдение верховенства закона на протяжении всего жизненного цикла системы ИИ (см. [47]), это включает в себя отказ от незаконного использования данных человека (см. [46]). Ответственному отношению к верховенству закона может способствовать просветительская и разъяснительная работа о возможностях неправомерного использования ИИ (см. [20])¹⁾.*

Согласно ГОСТ Р ИСО 26000 следование верховенству закона в каждой юрисдикции, в пределах которой компания ведет свою деятельность, может включать:

- соблюдение законов и нормативных актов, даже если они не применяются в данной юрисдикции;
- соблюдение всех юридических обязательств на протяжении всей цепочки создания стоимости ИИ и периодический анализ соответствия деятельности и взаимоотношений заинтересованных сторон;
- обеспечение того, чтобы цели, для которых разрабатывают и используют ИИ, были законными и конкретно обозначенными.

6.2.12 Соблюдение международных норм поведения

ГОСТ Р ИСО 26000 содержит рекомендацию о том, что в ситуациях или в юрисдикциях, где верховенство закона не обеспечивает надежных гарантий социального и экологического воздействия, организации могут стремиться соблюдать международные нормы поведения и избегать сотрудничества с организациями, которые не соблюдают такие нормы. Кроме того, в ГОСТ Р ИСО 26000 рекомендовано организациям пересмотреть характер своей деятельности в юрисдикциях, где законодательство или его применение отклоняется от международных норм до такой степени, что это может иметь значительные последствия. IEEE EAD предлагает заинтересованным сторонам принять участие в создании новых норм, связанных с системами ИИ (см. [20]).

6.2.13 Экологическая устойчивость

Организация, занимающаяся производством продукции или оказанием услуг, может взять на себя обязательство минимизировать воздействие на окружающую среду, здоровье и благополучие заинтересованных сторон и влияние на общество в целом (см. [48]). Организация может взять на себя обязательство существенно сократить выбросы парниковых газов, ограничить повышение средней мировой температуры (см. [49]), соблюдать цели в области устойчивого развития Организации Объединенных Наций, такие как борьба с изменением климата и его последствиями, обязательство обеспечить чистую воду, санитарные нормы, доступную и чистую энергию, ответственное потребление и производство — как для жизни в водной среде, так и для жизни на суше (см. [26]). Одной из дилемм устойчивости, стоящих перед технологиями, требующими больших объемов данных и вычислений, такими как ИИ, является постоянно растущая потребность в энергоресурсах, поскольку большие массивы данных и алгоритмы требуют потребления еще большего количества вычислительных мощностей. Эта потребность возрастает даже в тех условиях, когда глобальные цели в области устойчивого развития призывают к повышению энергоэффективности и снижению потребления невозобновляемых ресурсов. Поэтому следует предоставлять заинтересованным сторонам прозрачную информацию о потреблении энергии, изменении климата и смягчении негативных последствий во всей цепочке создания стоимости услуг на основе ИИ, чтобы заинтересованные стороны могли принимать рациональные решения. Организация также может использовать системы и приложения ИИ для повышения устойчивости и управления воздействием на окружающую среду и изменением климата, используя подход, основанный на жизненном цикле, направленный на сокращение отходов, повторное использование продуктов и компонентов, а также на переработку материалов. В качестве примеров можно привести оптимизацию энергетических сетей, точное сельское хозяйство, устойчивые цепочки поставок, мониторинг климата и прогнозирование экологических катастроф.

¹⁾ Внесено на основании предложений экспертов для учета в настоящем стандарте национальных интересов и защиты пользователей и населения.

6.2.14 Практика в сфере труда

Практика в сфере труда рассмотрена в *ГОСТ Р ИСО 26000*. Кроме того, Международная организация труда и трехстороннее агентство ООН объединяют правительства, работодателей и работников для установления трудовых стандартов, разработки политики и программ, которые принимаются на основе консенсуса, для продвижения достойного труда. Особый интерес представляют основополагающие конвенции, охватывающие то, что считается основными правами в сфере труда (см. [50]):

- свобода объединений и ассоциаций и эффективное признание права на ведение коллективных переговоров;
- ликвидация любого принудительного или обязательного труда;
- эффективное упразднение детского труда и соблюдение законодательства о минимальном рабочем возрасте;
- равное вознаграждение и ликвидация дискриминации.

Потенциальные аспекты относительно того значения, которое имеет ИИ в трудовых отношениях, включают:

- обеспечение понимания правил, касающихся занятости, трудовых отношений и вовлечения людей в принятие решений, требующих эффективного человеческого контроля и сочувствия [например, использование ИИ в управлении работниками, (включая работников на временных должностях и с частичной занятостью, вакансии заполняются независимыми подрядчиками и фрилансерами), для предотвращения дискриминации между работниками, непропорционального и неоправданного контроля и наблюдения на рабочем месте, особенно при удаленной работе, защиты частной жизни работников, устранения всех форм принудительного или обязательного труда и эффективного упразднения детского труда];
- обеспечение справедливого вознаграждения, условий труда, охраны здоровья и безопасности, защиты работников и решения других проблем (например, работники, привлекаемые на основе краудсорсинга и аутсорсинга, готовые данные для обучения, или модераторы контента, подвергающиеся воздействию контента социальных сетей с использованием ИИ);
- вопросы развития и подготовки кадров, особенно в условиях, когда внедрение ИИ устраняет роль работника, рабочие места или существенно меняет их характер (например, переподготовка);
- прогнозирование последствий внедрения ИИ и переквалификация рабочей силы;
- гарантии того, что уважение к человеческой жизни и человеческому достоинству сохраняется и что ИИ и системы обработки больших данных не оказывают негативного влияния на деятельность, свободу и достоинство человека;
- установление правил ответственности предприятий и разработчиков;
- задачу сделать ИИ предметом общественного диалога и коллективных переговоров в соответствии с правилами и практикой, действующими в каждой организации.

7 Примеры практик для создания и использования этического и общественно приемлемого ИИ

7.1 Согласование внутреннего процесса с принципами ИИ

7.1.1 Общие положения

Данный раздел содержит исчерпывающий перечень практик для создания и использования этического и общественно приемлемого ИИ. Такие аспекты и подходы, как правило, являются частью более широкого процесса управления организацией.

7.1.2 Определение этических принципов ИИ, которые может принять и использовать организация

Помимо создания первоначального набора принципов, соответствующих ценностям организации, интерес может представлять также создание процесса для оценки и обновления этих принципов на регулярной основе. Ключевые темы, связанные с ИИ, и соответствующие принципы описаны в 6.2.

Принципы могут быть более действенными, если они транслируются на внешний уровень и изложены в виде четких, однозначных предложений о том, что организация может или не может делать. Данные предложения могут помочь организации в оценке соблюдения/несоблюдения своих принципов.

7.1.3 Определение областей применения, которыми организация не может заниматься

Помимо установления принципов ИИ организация может также перечислить области применения, в которых невозможны разработка или внедрение ИИ. Принципы ИИ, как правило, утилитарны, что

означает, что если выгоды достаточно существенны, то они могут перевесить потенциальный вред. Следовательно, решение о сферах применения, которыми организация не может заниматься, принимается на основе деонтологии, определяющей области применения, которыми не допускается. Такой список может быть основан на широко признанных правах человека и может отражать и проецировать приверженность организации определенным базовым ценностям.

7.1.4 Процесс рассмотрения новых проектов

Ключевые темы, связанные с ИИ, и соответствующие принципы, описанные в 6.2, носят высокоуровневый характер и требуют тщательного и детального рассмотрения при их применении к конкретным проектам. Надежный, повторяемый и заслуживающий доверия процесс рецензирования может включать несколько элементов:

- орган экспертизы, охватывающий диапазон перспектив, дисциплин и опыта, необходимый для проведения тщательной экспертизы. В его состав может входить одна группа, которая занимается повседневными операциями и первоначальными оценками, и вторая, которая занимается наиболее сложными и трудными вопросами, включая решения, затрагивающие множество продуктов и технологий, или вопросы, в которых существует конфликт между несколькими этическими темами и принципами. Наличие третьей, специальной, группы по анализу может помочь обеспечить такие условия, при которых опыт организации в области этики ИИ со временем расширился и люди во всей организации могли использовать такой опыт;
- процедура передачи проектов на экспертизу, а также осведомленность об этой процедуре в организации;
- руководящие принципы, рамки и протоколы, позволяющие органу экспертизы последовательно и надежно оценивать проект и выносить решение о его соответствии принципам ИИ, другим этическим соображениям или и тому и другому;
- группы или отдельные лица, которые могут заниматься разработкой технических стратегий смягчения этических проблем, выявленных органом экспертизы, например: связанных с несправедливой предвзятостью, объяснимостью и человеческим контролем;
- процедуры регистрации прецедентного права и обеспечения того, чтобы созданные прецеденты могли быть перенесены на соответствующие дела;
- последующие рабочие процессы для обеспечения адекватного выполнения рекомендованных или требуемых мер по смягчению последствий;
- регулярный мониторинг для обеспечения того, чтобы система и приложение ИИ работали так, как задумано, на протяжении всего жизненного цикла. Это может включать мониторинг дрейфа концепции и распределения, периодическое переобучение для обеспечения справедливости модели, обновление языковых моделей для отражения изменений в лингвистической практике, а также оценку неправильного использования или неиспользования, которое может произойти со временем, например из-за предвзятости автоматизации.

Эти элементы, в зависимости от размера и характера организации, могут формироваться при участии как внутренних, так и внешних заинтересованных сторон, причем последние включают внешние агентства, консультативные советы и приглашенных экспертов в данной области. Рассмотрение проекта может также включать использование формального метода оценки рисков, такого как метод оценки рисков в ГОСТ Р ИСО 31000 (см. также [14]).

7.1.5 Обучение вопросам этики в области ИИ

Обучение в области этики по вопросам, связанным с принципами и организационными ценностями для всех членов организации, является ключом к созданию культуры этической разработки, внедрения и использования ИИ. В ходе обучения можно акцентировать внимание на том, почему важна этическая разработка ИИ, как привести в соответствие принципы и ценности, как внедрить этический процесс принятия решений в жизненном цикле продукта, как применять этические рамки, методологии и оценки воздействия, а также как процессы и структуры управления связаны с работой над принципами, которыми руководствуется организация. Обучение также может раскрыть этические проблемы, возникающие при разработке и внедрении ИИ, углубляясь в вопросы обоснования для разработки систем и приложений ИИ с точки зрения бизнеса и подробно разбирая, какими ценностями следует руководствоваться при внедрении новой системы. Обучение также может быть адаптировано с учетом конкретных ролей с акцентом на то, как разные заинтересованные стороны могут выявить проблемы в зависимости от их опыта, сферы ответственности и положения в жизненном цикле. Наконец, обучение может включать схемы цифровой грамотности в области ИИ для конкретных заинтересованных сторон.

7.2 Рассмотрение рамок этического анализа

7.2.1 Определение этической проблемы

Первым шагом при принятии решения является признание наличия этических проблем в конкретной ситуации. Это требует от лиц, принимающих решения, выйти за рамки рассмотрения бизнес-обоснований (что законно, что эффективно и т. д.) для разработки, внедрения и реализации ИИ и углубиться в этические вопросы (что приемлемо для человечества, что соответствует ожиданиям и уважает права пользователя и т. д.). Обучение этике ИИ, упомянутое в 7.1.5, может помочь лицам, принимающим решения, развить этические навыки идентификации (выявления) проблемы. Выявление этических проблем может происходить на всех этапах цепочки создания стоимости и жизненного цикла ИИ. Это может происходить на стадии идеи («можем ли мы создать эту систему ИИ и приложение?»), на стадии разработки («какие данные мы можем использовать?», «как мы можем снизить потенциальные риски?»), на стадии внедрения («как система ИИ и приложение могут быть использованы не по назначению?») и после запуска («как используется система ИИ и приложение?»).

7.2.2 Сбор фактической информации

Лицам, принимающим решения, необходимо иметь доступ к максимально возможному количеству фактов и информации. Сбор дополнительной информации для принятия решений, связанных с этикой ИИ, может включать в себя обращение к заинтересованным сторонам (менеджерам продукта, техническим руководителям, исследователям и т. д.) за дополнительной документацией и информацией. Кроме того, заинтересованными сторонами могут быть также те, кто не входит в непосредственную команду, разрабатывающую технологию, а дополнительные внутренние и внешние заинтересованные стороны (представители демографической группы, на которую может повлиять технология, неправительственные организации, обладающие опытом в этой области, и т. д.).

7.2.3 Рассмотрение и оценка альтернативных действий

Заинтересованные стороны, скорее всего, предложат различные действия и сосредоточатся на различных проблемах, например деловых, юридических и этических. Лица, принимающие решения, могут перечислить и рассмотреть как предлагаемые действия, так и список проблем, которые они используют для оценки каждого действия. Для целей этики в области ИИ одним из решаемых вопросов должно быть соблюдение принципов ИИ, которые организация решила принять.

7.2.4 Принятие решения и действия по его реализации

После принятия решения лицо или орган, принимающее(ий) решение (те, которые несут ответственность за принятие решения), может четко сформировать и изложить все рекомендации и меры по смягчению последствий, которые являются частью решения. Решение может определять ответственность за его реализацию, устанавливать четкие роли подотчетности, включая постоянный мониторинг решения (кто отслеживает результаты решения, насколько эффективно используется обратная связь с внешними заинтересованными сторонами или как информация доводится до тех, кого это касается, и т. д.). Решения могут быть задокументированы с надлежащим объяснением и обоснованием. Соответствующая документация помогает зафиксировать и каталогизировать созданные прецеденты и описать меры по смягчению последствий, чтобы повысить эффективность принятия решений в будущем и установить четкие исходные данные для аналогичных случаев в будущем.

7.2.5 Реализация и отражение результатов

В конце процесса может быть полезным проанализировать проделанное, чтобы понять, как обнаружить этические проблемы на более ранней стадии, и можно ли повторно использовать какие-либо из смягчающих действий в более широком контексте. Такой анализ может помочь организации извлечь уроки и усовершенствовать свои процессы, особенно если организация реагирует на этическую проблему, которая возникла внезапно и не была замечена в рамках действующих процессов выявления и анализа.

8 Рассмотрение вопросов создания и использования этического и общественно приемлемого ИИ

8.1 Общие положения

Настоящий стандарт содержит неполный перечень предложений и мер по созданию и использованию этического и общественно приемлемого ИИ, так как их полный перечень обычно является частью более широкого процесса управления организацией.

Этические вопросы, касающиеся систем и приложений ИИ, часто требуют осознания контекста (см. 6.2). Иными словами, наличие, характер, степень и серьезность этических проблем, связанных с системой и приложением ИИ, часто зависят от конкретного социально-политического, экономического, физического контекста их разработки, внедрения, аудитории или использования. Поэтому этическое использование систем и приложений ИИ связано с достаточно глубоким осознанием и пониманием соответствующего контекста, поскольку то, что может быть этически и общественно приемлемым в одном контексте (сектор, географический регион, предполагаемая цель, демографическая аудитория, политическая или физическая среда и т. д.), неприемлемо в другом контексте.

Предложения, изложенные в настоящем стандарте, могут быть дополнены с учетом типа системы ИИ и приложения, дизайна (конструкции) и предполагаемых функций системы или приложения ИИ. Кроме того, неполный перечень примеров использования приведен в приложении В.

8.2 Неполный перечень этических и общественных аспектов для рассмотрения

8.2.1 Общие положения

Данный раздел содержит неполный перечень предложений и мер по созданию и использованию этических и общественно приемлемых систем и приложений ИИ, так как их полный перечень обычно является частью более широкого процесса управления организацией.

8.2.2 Международные права человека

Международные права человека включают следующие вопросы:

- полностью ли соблюдаются международные права человека применительно к вопросам, связанным с системами и приложениями ИИ?
- существуют ли какие-либо процессы, позволяющие конечным пользователям или другим лицам, прямо или косвенно затронутым системой ИИ, сообщать о потенциальных проблемах с правами человека, а организации — реагировать и принимать во внимание такие отзывы для усовершенствования системы ИИ?
- существуют ли процессы для оценки случаев краткосрочного или долгосрочного воздействия на автономию и самостоятельность субъекта данных и информирования об этом заинтересованных сторон?
- существуют ли процессы, обеспечивающие заинтересованным сторонам право на возмещение, помощь, средство правовой защиты или способ получения помощи или средства правовой защиты?

8.2.3 Подотчетность

Подотчетность включает в себя рассмотрение соответствующих процессов в отношении следующего:

- существует ли сквозное отслеживание решений, принятых в отношении системы ИИ?
- можно ли отключить систему ИИ и приложение в случае отклонения от запланированных результатов?
- можно ли обнаружить или предсказать злоупотребление, неправильное использование или неиспользование системы ИИ и приложения?
- существуют ли стратегии и процессы тестирования и мониторинга, чтобы избежать создания или усиления смещенности (предвзятости) в данных и алгоритмах?
- проверяются ли внедренные алгоритмы на предмет их надежности, корректности и воспроизводимости и контролируются ли меры надежности и корректности, а также условия воспроизводимости?
- как и где могут быть задокументированы методология тестирования, результаты и внесенные на основе результатов изменения?
- существуют ли четко определенные обязанности и ответственность для заинтересованных сторон, участвующих в разработке или внедрении систем и приложений ИИ?
- является ли результат работы систем и приложений ИИ прозрачным и объяснимым?
- соответствует ли применение систем и приложений ИИ действующим нормам и стандартам?

8.2.4 Справедливость и недискриминация

Соображения в отношении объективности и недискриминации включают следующее:

- какие действия предприняты для выявления и смягчения предвзятости в наборе данных, используемом для обучения модели?
- *какие действия предприняты, чтобы в процессе работы системы предотвратить выявление и сбор данных о потребительских или других предпочтениях людей, а также других данных, ко-*

которые не относятся к назначению системы и могут быть использованы для оказания воздействия, в т. ч. манипулятивного характера, на пользователей или население¹⁾?

- может ли источник данных содержать исторические предубеждения, которые рискуют закрепиться?
- являются ли данные для обучения репрезентативными для населения, которое подвергается воздействию?
- существуют ли четко прописанные механизмы отчетности о том, как поднимать вопросы, касающиеся различных исключений, особенно со стороны пользователей системы и приложения ИИ или других лиц, на которых они повлияли?

8.2.5 Прозрачность и объяснимость

Соображения прозрачности и объяснимости включают следующее:

- ясно ли, кто или что может получить пользу или вред от предполагаемого функционального назначения системы ИИ?
- доведены ли доступным и понятным образом до сведения предполагаемых пользователей и людей, на которых потенциально может повлиять система, потенциальные или предполагаемые риски, а также проинформированы ли они о характере системы ИИ?
- имеются ли функции отслеживания (например, протоколирование действий) системы ИИ, чтобы сделать ее проверяемой, особенно в критических ситуациях?
- существует ли стратегия отчетности и коммуникации с заинтересованными сторонами в отношении использования, сбора и поддержания учебных наборов данных?
- существует ли полноценная, разработанная от начала и до конца стратегия отчетности и коммуникации в отношении сбора, обслуживания и использования данных?
- существует ли стратегия отчетности и коммуникации заинтересованных сторон в отношении того, как собираются, используются и поддерживаются любые данные, накопленные внедренной системой?
- предусматривает ли стратегия отчетности и коммуникации связь с заинтересованными сторонами, включая руководство, государственные регулирующие органы, гражданское общество, журналистов и другие наблюдательные органы?
- проводились ли консультации с работниками и представителями работников о характере и масштабах системы ИИ?
- сообщает ли организация о назначении системы ИИ и приложений о процессах, связанных с потоком данных, заинтересованным сторонам и другим лицам, информацию о которых собирают, выводят, используют, хранят или обрабатывают в определенной комбинации (например, делается ли это с помощью уведомления, размещенного рядом с местом внедрения, или QR-кода со ссылкой на политику)?
- доводятся ли до всех заинтересованных сторон выгоды, риски, связанные с системой ИИ, и то, что выгоды перевешивают потенциальные риски?
- определены ли потенциальные источники изменений (вариативности) в принятии решений, которые происходят во время прогнозирования и можно ли измерить эту вариативность?
- при внедрении системы ИИ, созданной третьей стороной, какого рода документация, обучение и поддержка требуются третьей стороне для предотвращения неправильного использования, неиспользования или злоупотребления?
- существуют ли процессы оценки воздействия разработки и внедрения системы ИИ на окружающую среду (например, электроэнергия, вода, землепользование).

8.2.6 Профессиональная ответственность

Вопросы профессиональной ответственности включают следующее:

- подкреплены ли корреляции, на основе которых система ИИ делает прогнозы, решения и рекомендации, соответствующими научными исследованиями?
- существуют ли меры по обеспечению безопасности, целостности, достоверности и точности данных, используемых системой ИИ и приложением?
- помимо точности данных какое тестирование точности системы необходимо провести до ее внедрения в среде, чтобы контролировать установленные и обоснованно прогнозируемые риски?

¹⁾ Внесено на основании предложений экспертов для учета в настоящем стандарте национальных интересов и защиты пользователей.

- действуют ли процессы, учитывающие известные научные недостатки и факторы, связанные с внедрением в проблемных условиях, которые могут способствовать неблагоприятным (вредным) результатам?

- соответствует ли технология цели, особенно для внедрения в конкретном контексте?
- являются ли модели, предназначенные для внедрения, достаточно надежными и адекватного качества, были ли они построены на основе всеобъемлющих (комплексных) сценариев?

8.2.7 Продвижение и развитие человеческих ценностей

Продвижение человеческих ценностей включает следующие вопросы:

- предоставляется ли информация в случае возможного вреда для человека в результате прогнозов, рекомендаций или решений систем и приложений ИИ?
- проводились ли консультации с экспертами, представителями затрагиваемых групп (включая работников) или другими заинтересованными сторонами при определении того, является ли внедренная система и приложение ИИ законной (легитимной) в глазах лиц, которые в наибольшей степени будут подвержены воздействию?
- если это применимо, имеется ли описание заинтересованных сторон в приложении ИИ?
- существуют ли процессы для раскрытия установленных уязвимостей, рисков или предвзятости и существует ли процесс для информирования о них?
- существуют ли процессы оценки влияния систем и приложений ИИ на занятость и дальнейшую трудовую деятельность и сообщается ли о таких случаях заинтересованным сторонам?

8.2.8 Конфиденциальность и защита частной жизни, приватность

Соображения конфиденциальности включают следующее:

- имеются ли механизмы коммуникации, указывающие, где могут быть подняты вопросы, связанные с конфиденциальностью?
- существуют ли процессы, позволяющие конечным пользователям или другим заинтересованным сторонам, прямо или косвенно затронутым системой ИИ, сообщать о проблемах, связанных с конфиденциальностью?
- существуют ли процессы, позволяющие заинтересованным сторонам реагировать на проблемы, связанные с конфиденциальностью, и принимать во внимание такие отзывы от пользователей (когда это оправдано) для усовершенствования системы ИИ?
- если пользователь не предоставляет все данные, необходимые для покупки, аренды или иного доступа к услуге или продукту, отказывают ли ему полностью в предоставлении услуг или продуктов или предлагают альтернативные варианты?
- насколько тщательно оценены риски, связанные со сбором и хранением персональных данных?
- существуют ли внутренние механизмы управления данными, включающие отслеживание того, как получены персональные данные?
- соблюдается ли внутреннее управление данными и обработка персональных данных, можно ли их отследить или повторно идентифицировать?
- *проинформирован ли субъект данных о том, что полученные данные будут обработаны с использованием технологий ИИ¹⁾?*
- было ли надлежащим образом получено информированное согласие?
- может ли пользователь или другая заинтересованная сторона полностью реализовать права на собственные персональные идентифицируемые данные?
- проинформирован ли субъект данных о том, как дается действительное согласие и в случае необходимости или по требованию субъекта данных как такое согласие может быть отозвано?
- проведена ли оценка, рассмотрение и реализация мер, которые касаются рассмотрения и определения конфиденциальности в контексте (сектор, географическая область и т. д.), в котором разработана и внедрена система ИИ?
- какие потоки данных и возможности проверки предоставлены для обеспечения того, чтобы собранная личная информация использовалась по назначению?

¹⁾ Внесено на основе предложений экспертов для учета в настоящем стандарте национальных интересов и защиты пользователей.

- имеются ли возможности и функции для гарантированного сохранения подлинности (недопущение искажений, изменений, модификации) архивных или иных фактических данных (оригинальные тексты, фото, аудио- или видеозаписи и т. п.) при их обработке с использованием технологий ИИ¹⁾?

- имеются ли функции для поддержки возможности субъектов данных проверять наличие, актуальность и точность их личной информации?

8.2.9 Защищенность и безопасность

Вопросы защищенности и безопасности включают следующее:

- имеются ли механизмы коммуникации для обсуждения вопросов, связанных с защитой и безопасностью?

- существуют ли какие-либо известные последствия для защиты и безопасности при использовании системы и приложения ИИ по назначению или в случае невыполнения предусмотренных функций (т. е. предоставление недостоверных результатов, недоступность или использование не по назначению либо для выполнения задачи, для которой она не протестирована)?

- какие технические меры могут быть приняты для защиты собранных данных от потери и несанкционированного доступа, уничтожения, использования, модификации и раскрытия?

8.2.10 Контроль технологии ИИ со стороны человека

Вопросы, связанные с контролем технологии ИИ человеком, включают следующее:

- требует ли система ИИ присутствия квалифицированного человека для минимизации риска?

- существуют ли меры по предотвращению злоупотреблений, неправильного использования или неиспользования системы ИИ?

- существует ли документация, определяющая контексты и тех лиц, для которых технология должна (и не должна) работать оптимально, учитывая состав набора данных и другие ограничения?

- существуют ли соответствующие меры для обеспечения того, чтобы контроль над системой ИИ мог быть безопасно передан человеку-оператору и осуществляться им в случае необходимости?

8.2.11 Вовлечение и развитие сообщества

Вопросы вовлечения, участия и развития сообщества включают следующее:

- когда системы и приложения ИИ используют для управления государственными пособиями и здравоохранением, существует ли достаточный надзор при определении соответствия требованиям и надзор при распределении государственных пособий?

- возможна ли и доступна ли помощь человека конечным пользователям, взаимодействующим с системами и приложениями ИИ?

- является ли вся соответствующая информация о системах и приложениях ИИ, которые используют в рамках оказания государственных услуг, открытой и доступной?

8.2.12 Проектирование, ориентированное на человека

Вопросы проектирования, ориентированного на человека, включают следующее:

- признает ли организация, что люди различаются по своим способностям и потребностям, использует ли организация эргономические и социальные данные о природе (сущности) и степени этих различий?

- является ли удобство использования и доступность стратегическими целями бизнеса в организации?

- применяет ли организация комплексный системный подход?

- являются ли здоровье, безопасность и благополучие приоритетами бизнеса?

- ценит ли организация своих сотрудников и предпринимает ли шаги, чтобы сделать работу сотрудников осмысленной и содержательной?

- является ли организация открытой и заслуживающей доверия?

- действует ли организация социально ответственно?

8.2.13 Уважение верховенства закона

Уважение верховенства закона (см. [51]) включает следующие вопросы:

- обладают ли судебные органы достаточным уровнем понимания систем и приложений ИИ для обеспечения уважения верховенства закона?

- обеспечивается ли в отношении систем и приложений ИИ надлежащее и тщательное судебное разбирательство, а также справедливое и равное обращение в соответствии с верховенством закона?

¹⁾ Внесено на основе предложений экспертов для учета в настоящем стандарте национальных интересов и защиты пользователей.

- выступает ли судебная система в качестве гаранта обращения в случае нарушения международных прав человека или свобод, которые вызваны решениями систем и приложений ИИ?
- может ли человек, в отношении которого принято алгоритмическое решение, юридически изучить и оспорить его обоснованность?
- гарантирует ли верховенство закона точность обработки данных во всей цепочке создания стоимости ИИ?

8.2.14 Уважение международных норм поведения

Уважение международных норм поведения включает следующие вопросы (см. [51]):

- соблюдаются ли международные нормы поведения при использовании или поощрении (продвижении) применения систем и приложений ИИ?
- защищены ли законные интересы (например, безопасность), права (в частности, права интеллектуальной собственности) или соответствующие ценности заинтересованных сторон (см. [41])?

8.2.15 Экологическая устойчивость

Вопросы экологической устойчивости в отношении систем и приложений ИИ включают в себя следующее:

- проведена ли оценка углеродного следа с учетом преимуществ, снижения воздействия и дополнительного потребления?
- существует ли политика использования возобновляемых источников энергии и избежание или использование теплоотдачи, особенно в центрах обработки данных?
- проводились ли исследования для оценки и максимального сокращения выбросов загрязняющих веществ в воздух, воду и почву (где это уместно)?
- существует ли практика «зеленых» закупок для оценки поставщиков товаров и услуг с точки зрения их воздействия на окружающую среду?
- существует ли подход к жизненному циклу (включая утилизацию) для сокращения отходов, повторного использования продуктов или компонентов и вторичной переработки материалов?

8.2.16 Практика в сфере труда

Вопросы практик, применяемых в сфере труда, включают следующее:

- проинформированы ли органы, отвечающие за общественный диалог в организации (независимо от их формы; профсоюзы, представители персонала, комитеты по охране труда и технике безопасности и т. д.) и проконсультированы ли они, когда это уместно, о внедрении системы и приложения ИИ?
- проинформирован ли персонал компании о внедрении системы и приложения ИИ?
- когда ИИ включает в себя наблюдение за работниками, предпринимаются ли шаги для проверки того, что наблюдение ограничено его единственной целью и не имеет несоразмерных и неоправданных последствий? Существует ли процедура сообщения о соответствующих проблемах?
- существует ли возможность дискриминационного отношения к работникам по причине применения предвзятых алгоритмов?
- приняты ли меры по предотвращению злоупотреблений в области защиты данных и конфиденциальности на рабочем месте?
- выявлено ли влияние ИИ на рабочие места, когда при необходимости требуется переквалификация работников?
- существует ли план по обеспечению грамотности в области ИИ и цифровой грамотности для пострадавших работников?
- четко ли понимаются и доводятся до сведения работников правила, касающиеся обязанностей и ответственности организации или разработчиков?

Приложение А
(справочное)

Документы, касающиеся принципов в области ИИ

А.1 Доклад Центра Беркмана Кляйна

Карта консенсуса в этических подходах и подходах, основанных на правах человека в отношении принципов для ИИ (см. [52])

В международном сообществе появились и продолжают разрабатываться новые наборы принципов или тем для ИИ. В исследовании Центра Беркмана Кляйна (см. 6.1) проанализированы 36 международных документов по принципам ИИ, опубликованных в период с 2016-го по 2019-й год, на основе материалов из разных отраслей и разных географических регионов. Более поздние международные публикации сопоставимого характера, в которых также перечислены несколько подборок принципов или тем ИИ, приведены в А.2—А.9.

А.2 Документы международных организаций

Десять принципов для этического ИИ (см. [53])

После изучения 90 организаций PWC¹⁾ выделил десять основных принципов, которые помогут определить этичность применения ИИ. На основе работы компании как внутри организации, так и с ее клиентами в статье представлены несколько предложений о том, как реализовать эти принципы на практике. Принципами являются интерпретируемость, надежность и устойчивость, безопасность, подотчетность, выгодность, конфиденциальность, человеческое участие, законность, справедливость и безопасность.

Незавершенность морального выбора и эволюция к полностью автономному ИИ (см. [54])

Принятие этических решений, рассматриваемое с точки зрения компьютерных, технических и естественных наук, связано исключительно со сложностью этой темы. Ученые и разработчики ИИ в основном исходят из модели машины Тьюринга, предполагая, что для решения любых задач (включая вопросы принятия этических решений) можно построить машину, способную механически вычислить определенную функцию, если эту функцию можно записать в алгоритм. Таким образом, принятие этических решений воспринимается как абстрактное понятие, влияние которого не зависит ни от конкретного физического организма, в котором происходит алгоритм, ни от того, из чего он сделан. Если на практике будет построен достаточно сложный алгоритм, то он будет демонстрировать и достаточно сложное поведение, которое можно охарактеризовать как этическое в полном смысле этого слова. Данная статья (см. [54]) отражает основной аргумент: если задача требует морального авторитета в той или иной форме, когда она выполняется человеком, то ее полная автоматизация, передача той же задачи автономным машинам, платформам и алгоритмам ИИ обязательно подразумевает передачу моральной компетенции. Вопрос о том, что включает в себя эта компетенция, предполагает проведение эмпирических исследований и переоценку чисто нормативных подходов в этике ИИ.

Этика в области ИИ и принципы (см. [55])

Хьюлетт-Паккард²⁾ считает ИИ мощной, преобразующей технологией, которая может усилить человеческие возможности, но также несет в себе риски. Организация инвестирует в этические принципы ИИ и соблюдает следующие этические и ответственные принципы ИИ: безопасность ИИ с поддержкой конфиденциальности, принцип ориентированности ИИ на человека, принцип инклюзивности ИИ, принцип надежности ИИ и принцип ответственного ИИ.

Руководство по стратегиям ИИ ЮНЕСКО (см. [56])

ИИ стремительно развивается в условиях, когда устойчивое развитие является главной целью международного сообщества. ООН призвала правительства разработать национальные стратегии устойчивого развития, включающие политические меры для достижения Повестки дня в области устойчивого развития на период до 2030 г. Хотя технологии ИИ могут способствовать прорыву в достижении целей устойчивого развития, они также могут иметь непредвиденные последствия, которые усугубят неравенство и негативно повлияют на отдельных людей, общества, экономику и окружающую среду. Реализация ИИ должна быть подкреплена междисциплинарным обзором, чтобы направить ИИ в русло, обеспечивающее уважение прав человека и человеческого достоинства. Данное справочное руководство представляет собой сборник ключевых ссылок, которые могут обеспечить глобальный обзор дискуссий по этике ИИ, технических стандартов и примеров национальных стратегий. Оно включает в себя три основные главы по ИИ: этические принципы и последствия; технические стандарты и международные стратегии; национальные стратегии. В этом документе отсутствуют конкретные определения принципов этики ИИ, вместо

¹⁾ PricewaterhouseCoopers (PwC) является торговой маркой PwC Business Trust. Эта информация приведена для удобства пользователей настоящего стандарта и не подразумевает одобрения или согласия со стороны ИСО или МЭК.

²⁾ Hewlett Packard Enterprise является зарегистрированной торговой маркой данной компании и/или ее филиалов. Эта информация приведена для удобства пользователей настоящего стандарта и не подразумевает одобрения или согласия со стороны ИСО или МЭК.

этого предлагается ознакомиться с работой международных и национальных правительств, частных организаций, гражданского общества и ресурсов из Европы, Америки и Азии.

Рамочная структура для этического ИИ в ООН (см. [57])

В данном документе не только сделана попытка дать определение «этический ИИ», но и рассмотрено, как такая организация, как ООН, могла бы приступить к внедрению этического ИИ. Предлагается рассматривать этические ценности «как состоящие из различных уровней, начиная с глобального и заканчивая групповым и индивидуальным: основные этические ценности, основанные на неотъемлемых правах человека (человеческое достоинство, автономия); конституционные ценности (верховенство закона, объективность, неприкосновенность частной жизни); групповые ценности (убеждения или культурные нормы) и индивидуальные этические ценности (личные убеждения)». В нем рекомендуется, чтобы ООН разработала структуру, состоящую из этических принципов, стандартов архитектуры, методов оценки, инструментов и методологий, а также политику, регулирующую внедрение и соблюдение этой структуры, сопровождаемую образовательной программой для персонала.

*Рекомендация по этике искусственного интеллекта*¹⁾ ЮНЕСКО (см. [58])

В документе проведено разграничение между ценностями и принципами этики. К ценностям относят: уважение, защита и продвижение человеческого достоинства, прав человека и основных свобод; окружающая среда и развитие экосистемы; обеспечение разнообразия и инклюзивность; жизнь в гармонии и мире. В рекомендации определены следующие этические принципы ИИ: соразмерность и непричинение вреда; безопасность и надежность; справедливость и недискриминация; устойчивость; конфиденциальность; человеческий надзор; прозрачность и объяснимость; ответственность и подотчетность; осведомленность и грамотность; управление и сотрудничество с участием многих заинтересованных сторон, адаптивное управление и сотрудничество.

Этика в искусственном интеллекте (см. [59])

Статья содержит тщательно подготовленный материал и указывает на актуальные ресурсы по данной теме. Фонд также предоставляет оценки готовности к ИИ для стран, не входящих в так называемый первый и второй мир. Последние отчеты включают Филиппины, Сальвадор, Нигерию и Мьянму. В своем анализе фонд приводит общую пятерку этических принципов: прозрачность, справедливость и честность, непричинение вреда, ответственность и конфиденциальность. Кроме того, в нем указано на то, что «ни в одном из рассмотренных документов не обсуждался ИИ в рамках опеки, воспитания, в терминах помощи, благосостояния, социальной ответственности, экологических сетей, а также достоинства или солидарности». Более того, документы в основном подготовлены в экономически развитых странах, в то время как Центральная и Южная Америка, Африка и Центральная Азия представлены незначительно и, таким образом, исключены из этического дискурса.

А.3 Северная Америка

Директива об автоматизированном принятии решений (см. [60])

Правительство Канады все больше стремится использовать ИИ для принятия решений или помощи в принятии административных решений в целях повышения качества обслуживания. Правительство стремится делать это таким образом, чтобы это соответствовало основным принципам административного права, таким как прозрачность, подотчетность, законность и процессуальная справедливость. Технология быстро меняется, и данная директива будет продолжать развиваться, чтобы обеспечить ее актуальность. Приложение А содержит определения, в том числе определение процедурной справедливости; приложение В — четыре уровня оценки воздействия, которые варьируются от воздействия, которое является обратимым и кратковременным (уровень I), до решений, которые являются необратимыми и постоянными (уровень IV); при этом они не связаны с риском. В приложении С перечислены процедурные требования для каждого уровня воздействия: от экспертных оценок до уведомлений, размещаемых по каналам предоставления услуг, от принятия решений с участием человека до требований к разъяснениям, обучению, планированию действий в чрезвычайных ситуациях и до принятия решения о вводе системы в эксплуатацию.

Принципы политики в области ИИ, краткое содержание (см. [61])

В этом документе рассмотрен широкий круг вопросов, охватывающих ИИ, промышленную и государственную политику. Что касается ценностей и этики, документ обращает внимание на ответственность промышленного сектора в продвижении продуманной разработки и использования ИИ, что включает: ответственное проектирование и внедрение, безопасность и управляемость, надежные и репрезентативные данные, интерпретируемость и ответственность систем ИИ в связи с автономностью. Последняя тема, которая рассмотрена в нескольких других публикациях, представляет интерес.

От препятствий к масштабированию. Глобальная гонка исследований в области ИИ (см. [62])

ИИ внедряется в повседневную жизнь, бизнес, государственное управление, медицину и другие сферы. Внедрение этических принципов в приложения и процессы ИИ позволяет создавать системы, основанные на доверии. Среди обсуждаемых тем: принципы доверия и прозрачности, повседневная этика для ИИ, обоснованность ИИ, объяснимость ИИ, расширение человеческого интеллекта и владение данными.

¹⁾ Внесена на основе предложений экспертов (актуализировано наименование принятой в ЮНЕСКО Рекомендации, в исходном тексте указан проект документа).

А.4 Южная Америка

Принципы и оси политики Чили в области ИИ (см. [63])

27 октября 2021 г. в Чили опубликована национальная политика в области ИИ, в которой рассмотрены преимущества технологии ИИ и цель позиционирования страны как лидера в области инноваций, исследований и разработок, а также демократизации технологических решений. Данная политика в области ИИ определяет четыре сквозных принципа: ИИ с упором на благосостояние людей, соблюдение прав человека и безопасность; ИИ для устойчивого развития; инклюзивный ИИ, а также глобальный и эволюционный ИИ.

Искусственный интеллект в Бразилии: бразильская стратегия ИИ (BSAI/EBIA) и билль № 21/2020 (см. [64])

В этой статье проанализирован билль № 21/2020 в палате депутатов и более подробно раскрыты некоторые существенные моменты бразильской стратегии ИИ (BSAI/EBIA). Уникальным среди статей и документов этого билля является то, что в данном документе рассмотрены ИКТ как следствие пандемии COVID19. Статья увязывает цель BSAI/EBIA «внести вклад в разработку этических принципов для развития и использования ответственного ИИ» и критику закона как обобщенного и не имеющего технической основы.

Правовая база для развития ИИ в Бразилии (см. [65])

Конгресс Бразилии принял билль, который создает правовую базу, определяющую правила использования ИИ в стране. Документ включает права пользователей систем ИИ, такие как получение информации об организации, ответственной за разработку системы ИИ, с которой они взаимодействуют, а также право на доступ к четкой и адекватной информации о критериях и процедурах, используемых системой. В нем содержатся правила, касающиеся уважения прав человека и демократических ценностей, и перечислены часто поднимаемые этические проблемы. Бразилия придерживается принципов Организации экономического сотрудничества и развития (ОЭСР) в области ИИ, ориентированных на человека, в которых приведены рекомендации в таких областях, как прозрачность и объяснимость.

Аргентина и Уругвай (см. [66])

Статья представляет собой вводный материал к стратегии ИИ в Аргентине и Уругвае и служит указателем на различные ресурсы, которые в других случаях недоступны на английском языке. Аргентинская стратегия (опубликована в 2019 г.) направлена на продвижение ИИ в частном секторе, минимизацию этических рисков и развитие талантов среди прочих целей. В аргентинском документе признается необходимость решения этических и общественных аспектов, которые возникают в связи с технологической трансформацией, одновременно с этим документ способствует ее развитию и реализации, что благоприятствует экономическому росту и социальному развитию. Предлагается применять методы ИИ, в которых антропологические, культурные и психологические аспекты могут иметь решающее значение, подчеркивается необходимость включения экспертов в этих областях в состав исследовательской группы, что позволит не только обучить специалистов различных дисциплин методам ИИ, но и разработать решения с потенциалом внедрения и применения в различных областях.

История и уроки первой стратегии ИИ в Латинской Америке. Мексика (см. [67])

В 2017 г. правительство Мексики совместно с авторами, представляющими гражданское общество, частный сектор, научные круги и сферу международного сотрудничества, решили разработать первую стратегию ИИ для страны в частности и Латинской Америки в целом. Она направлена на поиск ответов на основные вопросы, связанные с ИИ, включая вопрос относительно того, как юрисдикции будут решать этические вопросы, связанные с подотчетностью, открытостью и прозрачностью алгоритмов. В краткой аналитической справке содержится отчет руководителей инициативы, предложенной правительственными органами и представителями гражданского общества, а также рекомендации для правительств других стран, стремящихся разработать стратегии ИИ в будущем.

А.5 Африка

Определение этических норм в ИИ требует участия африканцев (см. [68])

В статье «Применение принципа объяснимости к исследованиям ИИ в Африке: должны ли мы это делать?» (см. [68]) отмечается, что «учитывая, что ценности в разных культурах различны, дополнительной этической проблемой является обеспечение того, чтобы эти системы ИИ не разрабатывались в соответствии с неким неоспоримым, но тем не менее спорным предположением об универсальных нормах, а были бы на самом деле совместимы с теми обществами, в которых они действуют. Это особенно актуально для исследований и внедрения ИИ в Африке — на континенте, где системы ИИ используются и будут использоваться, но также и на континенте с историей навязанных внешних ценностей». Объяснимость ИИ может помочь «внести вклад в ответственное и продуманное развитие ИИ, которое учитывает интересы и ценности Африки».

Главные проблемы ИИ в Африке: вычисления (см. [69])

Статья посвящена основным проблемам доступа к вычислительным ресурсам для ИИ в Африке. Разные представители сообщества лидеров ИИ из Ганы, Нигерии, Кении, Танзании, Марокко и других государств Африки рассказали о нюансах в своих странах и предложили собственные решения. Ключевыми задачами, которые необходимо решить, прежде чем станет возможным практическое обсуждение этики ИИ, являются: низкий доступ к графическим процессорам, проблемы с объемом памяти, отсутствие понимания того, какой графический процессор выбрать для достижения желаемой цели.

Африка, искусственный интеллект и этика (см. [70])

Ресурс представляет собой стенограмму подкаста Совета Карнеги. Основные барьеры на пути внедрения ИИ и обсуждения вопросов этики касаются следующих вопросов: неспособность большинства ИИ работать с африканскими языками (как письменный, так и разговорный); подотчетность организаций с иностранным гражданством, работающих в Африке; защита неопытных потребителей платформ социальных сетей, работающих под управлением алгоритмов с использованием токсичного контента; культурные нюансы африканских стран; доступ к инструментам крупных компаний, которые могут быть использованы для местного развития; доверие к технологиям; риск административной универсальности концепции ИИ для всех африканских стран. В статье обсуждается, как разговоры на общинных собраниях расшифровываются с помощью естественного языка и включаются в юридический и законодательный процессы Ботсваны, обеспечивая тем самым значительную социальную интеграцию.

На пути к этике ИИ в Африке: норма образования (см. [71])

Широкомасштабный обзор ресурсов авторами африканского континента является отправной точкой для всестороннего исследования ИИ и этики ИИ в Африке. В частности, в нем обсуждаются уроки Ubuntu в Африке и далее указывается, что «развитие ИИ должно соответствовать африканским ценностям о человеке. Таким ценностям, которые ставят во главу угла человеческие отношения, а не индивидуализм. К сожалению, рациональный взгляд на людей всегда был отмечен противоречиями, исключениями и неравенством. Развитие ИИ на основе данных, полученных в результате такого взгляда, будет способствовать усилению уже существующего неравенства».

Главные вызовы ИИ в Африке: этика (см. [72])

Вопросы, рассматриваемые как имеющие существенное значение для африканских сообществ: необъективность из-за недостаточной представленности; потеря рабочих мест; применение ИИ в широком смысле; как и чьи жизни должен улучшить ИИ; какие действия должен разрешить ИИ; открытый доступ к данным, особенно к африканским данным. Помимо представления этики ИИ в африканской среде статья привлекает внимание к работе Лакуна Фанд, который предоставляет финансирование для сбора данных со стороны малообеспеченного населения в области сельского хозяйства, здравоохранения и языка, чтобы преодолеть фундаментальные барьеры для достижения объективных результатов.

А.6 Европа

Предложение по регулированию европейского подхода к искусственному интеллекту. Европейская комиссия (см. [73])

Данное предложение по регулированию направлено на предоставление разработчикам, специалистам по внедрению и пользователям ИИ четких требований и обязательств в отношении конкретных видов использования ИИ и в то же время на снижение административного и финансового бремени для бизнеса, в частности для малых и средних предприятий.

Группа экспертов высокого уровня по искусственному интеллекту. Оценочный список для заслуживающего доверия искусственного интеллекта (ALTAI) для самооценки (см. [46])

Документ содержит окончательный список оценок для надежного ИИ (ALTAI), представленный экспертной группой высокого уровня по ИИ (AL HELG), созданной Европейской комиссией, чтобы помочь оценить, соответствует ли система ИИ, которая разрабатывается, развертывается, приобретается или используется, следующим семи требованиям: человеческое управление и надзор, техническая надежность и защищенность, конфиденциальность и управление данными, прозрачность, разнообразие, недискриминация и объективность, социальное и экологическое благополучие и подотчетность.

Решение этических проблем в области ИИ путем кодификации (см. [74])

В публикации изложено несколько подходов, основанных на рассмотренных материалах, которые могут быть использованы для кодификации этических задач при практическом применении систем ИИ. В первой части публикации основное внимание уделено двум группам вопросов, которые считают значимыми для практического применения систем ИИ, в то время как во второй части рассмотрены некоторые современные дискуссии по этике ИИ, общепринятые этические принципы, международное регулирование для достижения общих интересов в развитии ИИ, а также смежные вопросы, например вопросы этики ИИ, имеющие важное значение. В третьей и четвертой частях изложены практические подходы, используемые профессиональными сообществами для регулирования этических вопросов.

Резолюция с рекомендациями Комиссии по режиму гражданской ответственности для искусственного интеллекта. Европейский парламент (см. [75])

Предлагается создать ориентированный на будущее и единый подход, устанавливающий стандарты для граждан и бизнеса с целью обеспечения последовательности прав и правовой определенности, тем самым стимулируя цифровые инновации и обеспечивая высокий уровень защиты прав граждан и потребителей путем гармонизации режимов ответственности для систем ИИ.

Субъекты данных, цифровое наблюдение, ИИ и будущее в сфере труда. Европейский парламент (см. [76])

В докладе представлен глубокий обзор социальных, политических и экономических вопросов, связанных с выявлением так называемого «нового наблюдения за рабочими местами». В докладе оценивается ряд технологий, которые внедряются для мониторинга, отслеживания и в конечном счете для наблюдения за работниками, и рассматриваются значимые изменения, которые они несут в нескольких сферах. В отчете изложен ряд принципов, ка-

сающихся права на частную жизнь и защиты от сбора и использования данных о работниках. В докладе изложены некоторые противоречия между принципами.

Отчет фокус-группы. Дорожная карта по ИИ (см. [77])

Дорожная карта содержит обзор деятельности по стандартизации в IEEE, ETSI, ИСО/МЭК, ITU-T и CEN-CENELEC (подраздел 1.3 и приложение А). Фокус-группа определила 13 тем, среди которых следующие семь рассмотрены для европейской стандартизации (раздел 3): подотчетность, качество, данные для ИИ, безопасность и конфиденциальность, этика, проектирование систем ИИ и защищенность систем ИИ.

A.7 Азия

Новые инициативы Китая по управлению ИИ (см. [78])

Мета-обзор китайского управления ИИ содержит введение и ссылки на политику и документы, рассматривающие этику и управление со стороны администрации киберпространства Китая (САС), китайской академии информационно-коммуникационных технологий и Министерства науки и технологий. В статье отмечается, что САС является наиболее компетентной организацией, а ее проект 30 правил регулирования рекомендательных алгоритмов в Интернете открывает новые возможности на международном уровне в таких областях, как объяснимость и возмещение вреда. В статье утверждается, что эти правила могут открыть новые возможности на международном уровне.

Что нужно знать о китайских правилах в области этики ИИ (см. [79])

В статье приводится вводная информация о предлагаемых Китаем правилах управления и регулирования «Спецификации в области этики ИИ нового поколения» с западной точки зрения. Целью является интеграция этики и морали во весь жизненный цикл ИИ: продвижение ответственности, объективности, гармонии и безопасности; предотвращение таких проблем, как предрассудки, дискриминация и утечка конфиденциальности/информации. Детализированные положения распространены на физических и юридических лиц, а также на другие соответствующие организации, занимающиеся деятельностью, связанной с управлением, исследованиями и разработками, поставками и использованием ИИ. Указаны шесть основных этических норм: повышение благосостояния человека, продвижение ответственности и объективности, защита конфиденциальности и безопасности, обеспечение подконтрольности и достоверности и повышение этической грамотности. Кроме того, определены цели управления проектами, связанными с ИИ, в том числе: содействие гибкому управлению и устойчивости развития, активная интеграция этических стандартов в процессы управления, обеспечение упорядоченного развития и устойчивости ИИ, уточнение ответственности заинтересованных сторон и стандартизация условий и процедур управления, усиление предотвращения рисков путем усовершенствования оценки рисков при разработке ИИ, поощрение использования и диверсификации технологий ИИ для решения экономических и социальных проблем и поощрение междисциплинарных, межрегиональных и трансграничных обменов и сотрудничества.

Принципы и практика этического использования ИИ в бизнесе социальных инноваций Hitachi (см. [80])

Управление и этика ИИ являются ключевыми вопросами, значительно воздействующими на общество в связи с использованием ИИ. Выделяется множество различных аспектов этики ИИ, среди которых — конфиденциальность, справедливость, прозрачность и подотчетность, а также безопасность. На странице представлены документы: этические принципы ИИ компании Hitachi¹⁾; стандарты поведения на этапе планирования, на этапе социальной реализации и на этапе обслуживания и управления, а также конфиденциальность, справедливость, равенство и предотвращение дискриминации, прозрачность, объяснимость и подотчетность, безопасность и соответствие требованиям.

Отчет о зрелости этики в области ИИ в Малайзии, 2021 г. Доктрина ИИ (см. [81])

В документе кратко изложена Национальная рамочная программа Малайзии по этике в области ИИ, в которой отмечаются два подхода: с одной стороны, стимулирование инновационного потенциала ИИ; с другой — разработка и внедрение этического и надежного ИИ. Эти процессы стали центральным фактором развития и внедрения ИИ. Развитие ИИ порождает ряд вопросов и задач, требующих тщательного рассмотрения, особенно когда разработка и внедрение ИИ создают этические дилеммы. В документе основное внимание уделено вопросам недоверия, которое вызывает ИИ. В нем изложены принципы, которые могут противостоять этим проблемам, такие как объяснимость, раскрытие информации, конфиденциальность и управление данными, качество и целостность данных, доступ к данным и практическая применимость. Документ включает индекс зрелости этики ИИ, который отражает степень осознания полезности системы ИИ, степень прозрачности и объяснимости, а также степень соблюдения конфиденциальности и принятия мер по управлению данными.

Готовность к рассмотрению ИИ на Филиппинах (см. [82])

Обзор готовности к рассмотрению ИИ на Филиппинах. Это один из примеров того, как Агентство по развитию ИИ ежемесячно освещает внедрение и готовность ИИ не только в западных странах. В число других стран входят: Мексика, Аргентина, Мьянма, Южная Корея (с упоминанием культуры «быстро, быстро»), Нигерия, Египет, Колумбия, Узбекистан, Зимбабве и Сингапур.

¹⁾ Hitachi является торговой маркой компании Kabushiki Kaisha Hitachi Seisakusho. Эта информация приведена для удобства пользователей настоящего стандарта и не означает одобрения или поддержки со стороны ИСО или МЭК.

Международный исследовательский центр по этике и управлению ИИ. Этические нормы для ИИ нового поколения. Китай (см. [83])

25 сентября 2021 г. Национальный комитет по управлению искусственным интеллектом нового поколения опубликовал «Этические нормы для ИИ нового поколения». Его цель — интегрировать этические нормы на всех этапах жизненного цикла ИИ, обеспечить этические рекомендации для физических и юридических лиц, а также для других организаций, занимающихся деятельностью, связанной с ИИ. Документ содержит полный набор норм. В своде этических норм рассмотрены актуальные этические задачи неприкосновенности частной жизни, предубеждений, дискриминации и объективности со стороны всех слоев общества. Свод этических норм включает в себя общие принципы, этические нормы для конкретных видов деятельности, руководство по организации и внедрению и выдвигает шесть фундаментальных этических требований, таких как повышение благосостояния человечества, продвижение ответственности и объективности, защита конфиденциальности и безопасности, обеспечение контролируемости и достоверности, усиление подотчетности и повышение этической грамотности. В то же время выдвинуто 18 конкретных этических требований для конкретных видов деятельности, таких как управление, исследование и разработки, поставки и использование ИИ.

Экспертная группа по вопросам реализации принципов ИИ. Рабочая группа (см. [84])

Япония опубликовала документ «Социальные принципы человеко-ориентированного ИИ», принятый советом по продвижению комплексной инновационной стратегии, который вносит вклад в формулирование рекомендаций ОЭСР по ИИ. Социальные принципы ИИ состоят из семи принципов: (1) ориентированность на человека, (2) образование/грамотность, (3) защита конфиденциальности, (4) обеспечение безопасности, (5) честная конкуренция, (6) справедливость, подотчетность и прозрачность и (7) инновации. В документе представлены цели, которые должны быть реализованы компанией, занимающейся ИИ, для оказания содействия при реализации принципов ИИ, что облегчит внедрение ИИ.

Что буддизм может сделать для этики ИИ (см. [85])

В публикации рассматривается, что буддизм может предложить для этики ИИ. В материале отмечается, что «любое этическое использование ИИ должно быть направлено на уменьшение боли и страданий», а также принцип «Не навреди: бремя доказательства лежит на тех, кто стремится доказать, что конкретное применение ИИ не причиняет вреда».

Переосмысление алгоритмической справедливости в Индии и за ее пределами (см. [86])

Исследование подчеркивает, что вопросы, связанные с ИИ в Индии, отличаются от рассматриваемых в западных странах и могут потребовать различных подходов для достижения объективности.

Угроза ИИ и наш ответ. Этическая хартия ИИ в Южной Корее (см. [87])

В начале хартии отмечается, что изменения, вызванные внедрением ИИ, продолжаются и что эффективность ИИ практически не опровергается. Южнокорейцы активно участвуют в дискуссиях по минимизации побочных эффектов использования ИИ и его ответственному применению, и публикация Корейской этической хартии ИИ является одним из таких результатов. В исследовании рассматривается, как южнокорейское общество реагирует на угрозы при использовании ИИ, которые могут возникнуть в будущем, путем изучения различных этических хартий ИИ в Республике Корея. Кроме того, семь угроз ИИ классифицируют по трем категориям: ценностное суждение ИИ (человеческая дискриминация в ИИ, оценка ИИ человеческой ценности); злонамеренное использование ИИ (смертельное оружие ИИ, кибератаки на основе ИИ, чрезмерное вторжение в частную жизнь) и отчуждение человека (поглощение ИИ человеческих профессий, углубление отчуждения уязвимых в цифровом отношении людей). Авторы выделяют 14 тем, основанных на трех категориях: защита социальных ценностей (включая предотвращение социальной дискриминации, социальную интеграцию в целом, уважение человеческого достоинства, стремление к человеческому благу и счастью); контроль ИИ (объяснимый алгоритм, использование данных на основе социальной этики, подготовка к сбоям и опасным ситуациям, обязательный контроль со стороны человека, ограничение целей использования ИИ, применение системы самоуправления, четкое разделение ответственности и возможность проверки применения ИИ) и укрепление цифрового гражданства (культура постоянного многостороннего общения, расширение возможностей использования ИИ). Анализ показывает, что Корея еще не смогла должным образом отреагировать на угрозу поглощения ИИ человеческих профессий (рабочих мест). Кроме того, хотя корейские этические хартии ИИ представляют адекватные ответы на смертоносное оружие ИИ, их положения будет трудно реализовать, поскольку конкуренция за оружие ИИ между военными державами усиливается. Статья содержит полезный обзор семи этических хартий ИИ в Корее, первая из которых (Проект этической хартии роботов) опубликована еще в 2007 г.

А.8 Евразия

Первый этический кодекс ИИ подписан в России (см. [88])

Кодекс станет частью федерального проекта «Искусственный интеллект» и Стратегии развития информационного общества на 2017—2030 гг. Он устанавливает общие этические принципы и нормы поведения, которыми должны руководствоваться лица, участвующие в деятельности с использованием ИИ. Кодекс распространяется на отношения, связанные с этическими аспектами создания (проектирование, конструирование, опытная эксплуатация), внедрения и использования технологий ИИ на всех этапах жизненного цикла, которые в настоящее время не регулируются российским законодательством или другими нормативными актами.

Сбер в числе пионеров, принявших этические принципы ИИ в России (см. [89])

В пресс-релизе сообщается, что правление Сбербанка¹⁾ приняло этические принципы разработки и использования ИИ в Группе Сбер. Приняты следующие пять принципов: безопасный ИИ, объяснимый ИИ, надежный ИИ, ответственный ИИ и справедливый ИИ. Каждый принцип задокументирован с поясняющими примерами. Принципы разработаны с учетом требований Национальной стратегии развития искусственного интеллекта до 2030 года.

Альянс в сфере ИИ. Кодекс этики в области ИИ (см. [90])

Кодекс распространяется на отношения, связанные с этическими аспектами создания (проектирование, конструирование, опытная эксплуатация), внедрения и использования технологий ИИ на всех этапах жизненного цикла, которые в настоящее время не регулируются законодательством Российской Федерации и/или положениями технического регулирования. В разделе 1 «Принципы этики и правила поведения» изложены основные принципы и сопутствующие каждому из них характеристики: защита интересов и прав человека как коллективных, так и индивидуальных (включая человекоориентированный и гуманистический подход, уважение автономии и свободы воли человека, соблюдение закона, отсутствие дискриминации, оценку рисков и гуманитарных аспектов); необходимость осознанной ответственности при создании и использовании ИИ (включая риск-ориентированный подход, ответственное отношение, рассмотрение и принятие мер предосторожности, принцип «не навреди», идентификацию ИИ при общении с человеком, безопасность данных, информационную безопасность, добровольную сертификацию и соблюдение кодексов, контроль рекурсивного самосовершенствования систем ИИ); человек всегда несет ответственность за последствия применения системы ИИ (контроль, ответственность); технологии ИИ должны применяться и внедряться там, где это может принести пользу людям (в том числе применение системы ИИ в соответствии с ее назначением, стимулирование развития ИИ); интересы развития технологий ИИ выше интересов конкуренции (корректность сравнений систем ИИ, развитие компетенций, сотрудничество разработчиков); значимость максимальной прозрачности и правдивости информации в отношении уровня развития, о возможностях и рисках технологий ИИ (в том числе достоверность информации об ИИ и повышение осведомленности об этике применения ИИ).

А.9 Азиатско-Тихоокеанский регион

Департамент промышленности, науки, энергетики и ресурсов Правительства Австралии. Австралийские рамочные положения в области этики искусственного интеллекта (см. [91])

Добровольные рамочные положения по этике ИИ служат руководством для предприятий и правительств по ответственному проектированию, разработке и внедрению ИИ и состоят из четырех базовых компонентов: принципы этики ИИ, применение принципов, проверка принципов и разработка концептуальных положений. Такой подход, напоминающий жизненный цикл, хотя и не является уникальным для Австралии, нечасто встречается в других ресурсах. Восемь принципов этики ИИ в Австралии разработаны для обеспечения защищенности, безопасности и надежности ИИ. Они учитывают следующее: благополучие человека, общества и окружающей среды, ценности, ориентированные на человека, справедливость, защиту частной жизни и конфиденциальность, надежность и безопасность, прозрачность и объяснимость, возможность конкуренции и подотчетность.

ИИ Сингапур. Обзор этики и управления ИИ (см. [92])

В статье рассмотрено управление через призму практического применения принципов ИИ и определены пять общих тем или всеобъемлющих принципов этичного ИИ: польза для людей, непричинение вреда, автономия, объективность и объяснимость. В статье отмечается, что первые четыре принципа соответствуют четырем традиционным принципам биоэтики, и к ним добавляется новый принцип объяснимости для ИИ. Документ содержит ссылки на технологические инструменты с открытым исходным кодом от IBM²⁾ и Microsoft³⁾, которые могут помочь разработчикам Python⁴⁾ при внедрении этих принципов в работу.

Форум по ИИ в Новой Зеландии. Надежный ИИ в Aotearoa (Новая Зеландия на языке коренного народа маори). Принципы ИИ (см. [93])

Чтобы помочь поддержать общественное доверие к разработке и использованию ИИ в Новой Зеландии, рабочая группа «Закон, общество и этика» форума ИИ опубликовала набор из пяти руководящих принципов, призванных обеспечить руководство высокого уровня для всех, кто занимается проектированием, разработкой и использованием ИИ в Новой Зеландии (заинтересованные стороны ИИ), с целью обеспечить новозеландцам доступ

¹⁾ Сбер является товарным знаком Сбербанка России. Данная информация приведена для удобства пользователей настоящего стандарта и не является одобрением или поддержкой со стороны ИСО или МЭК.

²⁾ IBM является зарегистрированной торговой маркой корпорации International Business Machines Corporation (IBM). Эта информация приведена для удобства пользователей настоящего стандарта и не является одобрением или поддержкой со стороны ИСО или МЭК.

³⁾ Microsoft является зарегистрированной торговой маркой корпорации Microsoft. Эта информация приведена для удобства пользователей настоящего стандарта и не является одобрением или поддержкой со стороны ИСО или МЭК.

⁴⁾ Python является торговой маркой Python Software Foundation. Эта информация приведена для удобства пользователей настоящего стандарта и не является одобрением или поддержкой со стороны ИСО или МЭК.

к надежному ИИ. Пять руководящих принципов: справедливость и правосудие (включая уважение применимых законов, прав человека, прав маори, демократических ценностей и принципов равенства и справедливости), надежность, безопасность и конфиденциальность, прозрачность, надзор за людьми и подотчетность и благополучие. Эта публикация примечательна тем, что она одна из немногих, в которой рассмотрены и соблюдены права коренных народов.

Сингапурский подход к управлению ИИ (см. [94])

В статье рассмотрена эволюция в период с 2019-го по 2020-й год сингапурской модели рамочного механизма управления ИИ (от первого ко второму изданию). Второе издание основано на двух руководящих принципах первого издания: решения, принимаемые ИИ, должны быть объяснимыми, прозрачными и справедливыми, а системы ИИ должны быть ориентированы на человека. Второе издание включает дополнительные соображения (такие как надежность и воспроизводимость) и уточняет оригинальную модельную основу для большей актуальности и удобства использования. Например, раздел об управлении взаимоотношениями с клиентами расширен за счет включения соображений о взаимодействии и коммуникации с более широкой сетью заинтересованных сторон. Во втором издании модельной основы по-прежнему использован подход, не зависящий от выбора сектора и технологии, который может дополнить отраслевые требования и руководства. Практическое руководство охватывает внутренние структуры управления и соответствующие меры, определение уровня участия человека в принятии решений с использованием ИИ, управление операциями, взаимодействие и общение с заинтересованными сторонами. Работа комиссии по защите персональных данных поддерживается консультативным советом по этическому использованию ИИ и данных.

Сборник примеров использования, практические иллюстрации модельной основы управления ИИ (см. [95])

Дополнением к сингапурской модельной основе и подходу к управлению ИИ является сборник примеров использования, который демонстрирует, как национальные и международные организации различных секторов и разного масштаба внедрили или привели свои практики управления ИИ в соответствие со всеми разделами модельной основы. В сборнике также показано, как представленные организации эффективно внедрили ответственную практику управления ИИ и получили выгоду от использования ИИ в своей сфере деятельности. Эти реальные примеры использования ИИ предназначены для того, чтобы стимулировать другие компании делать то же самое и способствовать ответственному развитию и внедрению ИИ.

Консультативный совет по этичному использованию ИИ и данных, Управление по развитию информационно-коммуникационных средств и комиссия по защите персональных данных. Руководство по реорганизации рабочих мест в эпоху ИИ (см. [96])

Первое в Сингапуре руководство, которое помогает организациям и сотрудникам понять, как можно пересмотреть существующие должностные функции, чтобы использовать потенциал ИИ и повысить ценность их работы. В руководстве представлен практический подход, не зависящий от отрасли, который поможет компаниям управлять воздействием ИИ на сотрудников, а организациям, внедряющим ИИ, подготовиться к цифровому будущему. В руководстве представлены практические шаги в четырех областях реорганизации рабочих мест: трансформация рабочих мест, определение четких взаимосвязей между рабочими местами, устранение барьеров на пути цифровой трансформации и обеспечение эффективной коммуникации между работодателями и сотрудниками. Руководство является практическим примером решения этических рисков и воздействия на устойчивое развитие общества, часто упоминаемых в других документах настоящего приложения.

Компьютерное общество Сингапура. Свод знаний по этике и управлению ИИ (см. [97])

Свод знаний по этике и управлению ИИ, разработанный Сингапурским компьютерным обществом и одобренный управлением по развитию информационно-коммуникационных средств массовой информации, поддерживает предприятия, нанимающие обученных специалистов по этике ИИ для ответственного внедрения ИИ. VoK — это документ, основанный на модельной основе управления ИИ, который со временем будет совершенствоваться. Он является основой для подготовки и сертификации специалистов по этике и управлению ИИ в будущем как в сфере ИКТ, так и в других сферах, а также призван способствовать разработке учебных программ по управлению ИИ. Данное пособие было предшественником VoK и может быть востребовано у тех специалистов, кто менее продвинулся на пути к этике ИИ, оно дает ответы на такие фундаментальные вопросы, как: «зачем нужна этика ИИ?», «какие преимущества понимания этики ИИ?» и «кто несет ответственность за этику ИИ?».

Приложение В
(справочное)

Обзор примеров использования

В.1 Общие положения

Специалисты, исследователи, регулирующие органы и физические лица должны иметь представление об этических и общественных проблемах, связанных с системами и приложениями ИИ. Общедоступная информация содержит примеры положительного воздействия систем и приложений ИИ (см. [16]). Примеры использования были выбраны для иллюстрации этических и общественных проблем или вопросов, связанных с системами и приложениями, использующими функциональность ИИ (системы и приложения на основе технологии ИИ).

В.2 Примеры использования

В.2.1 Пример использования 1

Таблица В.1 иллюстрирует проблемы подотчетности, профессиональной ответственности, конфиденциальности, защиты и безопасности, участия и развития сообщества, соблюдения верховенства закона и практики в трудовой сфере.

Т а б л и ц а В.1 — Компоненты ИИ для движения транспортных средств на дорогах общего пользования (беспилотные транспортные средства)

Элемент	Комментарий	
Описание	Общая концепция автоматизированного управления взводом заключается в том, что ведущее транспортное средство управляется обученным (профессиональным) водителем, а следующие за ведущим транспортные средства — системой полностью автоматически, что позволяет водителям выполнять другие задачи, кроме управления автомобилем. Дорожная карта ЕС по взводу грузовиков предполагает внедрение на рынок мультибрендового взвода к 2025 г. (пример использования 31, см. [16]). Примерно с 2000 г. было несколько пилотных проектов (примеры использования 27, 28, 29, 32 и 33, см. [16]). Хотя несколько компонентов ИИ уже применяются в пилотных проектах (например, поддержание полосы движения), в будущем продукты, вероятно, будут включать решения ИИ на нескольких функциональных уровнях	
Этические и общественные аспекты	Стресс или незадействованность для водителей, постоянный контроль, безопасность, защищенность и надежность системы, риск взлома и угона дальнобойного грузовика представляет большую опасность, доверие к надежности системы при вождении рядом с взводом, управляемым компьютером. Высокая непредсказуемость дорожной обстановки, законодательная ситуация, стандартизация. Пример использования иллюстрирует проблемы подотчетности (см. 6.2.1), профессиональной ответственности (см. 6.2.4), конфиденциальности (см. 6.2.6), защиты и безопасности (см. 6.2.7), участия и развития сообщества (см. 6.2.9), соблюдения верховенства закона (см. 6.2.10), трудовой практики (см. 6.2.13)	
Ссылки	См. [16]	Выдержка из примера использования 9

В.2.2 Пример использования 2

Таблица В.2 иллюстрирует проблемы прав человека, справедливости и недискриминации, конфиденциальности, безопасности и защиты, контроля человеком технологий и соблюдения верховенства закона.

Т а б л и ц а В.2 — Аналитика поведения и настроений (безопасность и правоохранительные органы)

Элемент	Комментарий	
Описание	Использование систем внутреннего видеонаблюдения и анализа для определения эмоционального состояния и цели человека по его жестам, выражению лица и действиям. Использование полученной информации для предотвращения нежелательного поведения, адаптации описываемой ситуации в соответствии с состоянием человека, предоставления динамического контента в соответствии с эмоциональными реакциями человека или обнаружения кражи предметов и других преступных действий. Однако камеры наблюдения часто имеют низкое разрешение и могут находиться в плохо освещенном помещении с плохим углом обзора сверху. Многие подозрительные действия могут быть неразличимы за прохожими или большими толпами людей. Нежелательное поведение встречается гораздо реже, чем обычное, и может выражаться в различных формах	
Этические и общественные аспекты	Предвзятость, угрозы безопасности, угрозы конфиденциальности. Пример использования иллюстрирует проблемы прав человека (см. 5.1), справедливости и недискриминации (см. 6.2.2), конфиденциальности (см. 6.2.6), безопасности и защиты (см. 6.2.7), контроля человеком технологий (см. 6.2.8), соблюдения верховенства закона (см. 6.2.10)	
Ссылки	См. [16]	Выдержка из примера использования 14

В.2.3 Пример использования 3

Таблица В.3 иллюстрирует проблемы конфиденциальности.

Т а б л и ц а В.3 — Повышение эффективности управления дорожным движением и точности выявления нарушений с помощью ИИ технологий

Элемент	Комментарий	
Описание	Технологии ИИ, использующие большие данные, применяют для мониторинга и управления дорожным движением в крупном муниципалитете Китая. Осуществляют мониторинг данных из разных источников (транспортный поток, данные о транспортных средствах, движение пешеходов и т. д.), на основе которых с помощью методов машинного обучения (ML) выявляются и анализируются незаконная эксплуатация транспортных средств, непредвиденные случаи, резкое увеличение трафика и т. д. Задачи ML (включая обучение и развертывание) выполняют на платформе, поддерживающей интеграцию различных ML-фреймворков, моделей и алгоритмов. Платформа основана на гетерогенных вычислительных ресурсах. Эффективность и точность обнаружения нарушений, а также эффективность управления дорожным движением значительно повышаются, при этом значительно снижаются человеческие усилия и общая стоимость решения	
Этические и общественные аспекты	Угрозы конфиденциальности. Пример использования иллюстрирует проблемы конфиденциальности (см. 6.2.6). Проблемы безопасности (из-за риска взлома, опасных нарушений в работе систем управления движением, например одновременного включения зеленого света на всех светофорах на перекрестке) (см. 6.2.7)	
Ссылки	См. [16]	Выдержка из примера использования 29

В.2.4 Пример использования 4

Таблица В.4 иллюстрирует проблемы подотчетности, справедливости и недискриминации, прозрачности, конфиденциальности, безопасности и защиты, соблюдения верховенства закона и соблюдения международных прав человека.

Т а б л и ц а В.4 — Правоохранительные органы, отправление правосудия и демократические процессы

Элемент	Комментарий	
Описание	ИИ и робототехника могут значительно расширить возможности правоохранительных органов по осуществлению наблюдения. Международная организация уголовной полиции — это межправительственная организация, членами которой в 2020 г. являлись 194 страны. В ходе глобального совещания Интерпола-ЮНИКРИ по ИИ для правоохранительных органов, состоявшегося в ноябре 2020 г., обсуждены следующие темы (см. [77]): - потенциальные злоупотребления в области ИИ; - использование ИИ правоохранительными органами, включая специальные коллегии по применению ИИ для борьбы с сексуальным насилием над детьми в Интернете и террористическим использованием Интернета и социальных сетей; - последние разработки в области ИИ для правоохранительных органов в частном секторе; - развитие смежных областей, таких как использование ИИ в системе уголовного правосудия, ИИ и уголовная ответственность, а также взаимодействие между ИИ и беспилотными аппаратами	
Этические и общественные аспекты	Этические проблемы, возникающие в связи с внедрением технологий ИИ, которые необходимо решать, например проблемы конфиденциальности, связанные с этими технологиями, включая такие вопросы, как, когда и где допустимо использовать датчики, потенциальные злоупотребления в области ИИ и различных систем ИИ. Сохранение демократических процессов. Пример использования иллюстрирует проблемы подотчетности (см. 6.2.1), справедливости и недискриминации (см. 6.2.2), прозрачности (см. 6.2.3), конфиденциальности (см. 6.2.6), безопасности и защиты (см. 6.2.7), соблюдения верховенства закона (см. 6.2.10), а также соблюдения международных прав человека (см. 5.1)	
Ссылки	См. [98]	Подчеркивание необходимости сотрудничества и разработки пригодных для использования рамочных подходов
	См. [99]	Возможности и риски ИИ и робототехники для правоохранительных органов

В.2.5 Пример использования 5

Таблица В.5 иллюстрирует проблемы подотчетности, справедливости и недискриминации, прозрачности, конфиденциальности, защиты и безопасности.

Т а б л и ц а В.5 — Разговорный ИИ и чат-боты

Элемент	Комментарий
Описание	В [100] приведены высказывания Роба Хая на тему этических соображений в отношении чат-ботов. Установление доверия между машинами и людьми работает аналогично установлению доверия между людьми. Бренд может укрепить доверие, приводя свои ожидания в соответствие с реальностью, обучаясь на своих ошибках и исправляя их, прислушиваясь к отзывам клиентов и являясь прозрачным. Следовательно, прозрачность может стать решающим фактором при разработке чат-бота для обслуживания клиентов. Все сводится к простому вопросу: очевидно ли, что пользователи разговаривают с человеком или машиной? Клиенты обычно могут отличить одно от другого, и они ожидают, что бренды (организации) будут честны в этом вопросе. Клиенты вряд ли ожидают, что чат-бот будет совершенным, но они хотели бы знать, что он может делать, а что не может

Окончание таблицы В.5

Элемент	Комментарий	
Этические и общественные аспекты	Непонимание человека, что он разговаривает с системой ИИ, может привести его к замешательству и разочарованию, а также к ощущению, что его обманули или ввели в заблуждение, это не способствует ни психическому здоровью, ни уверенности, ни межличностному или общественному доверию. Одним из ключевых моментов в обеспечении этичности ИИ является его прозрачность. Автор утверждает, что, когда человек взаимодействует с чат-ботом, он должен знать, что разговаривает с чат-ботом, а не с живым человеком. Пример использования иллюстрирует проблемы подотчетности (см. 6.2.1), прозрачности (см. 6.2.3), конфиденциальности (см. 6.2.6), защиты и безопасности (см. 6.2.7)	
Ссылки	См. [16]	Пример использования 106
	См. [100]	Этические соображения в отношении чат-ботов
	См. [101]	В публикации говорится о том, что использование языка не является «нейтральным» или «объективным» средством. Скорее всего, оно отражает существующие в обществе ценности и суждения

В.2.6 Пример использования 6

Таблица В.6 иллюстрирует проблемы подотчетности, справедливости и недискриминации, прозрачности, конфиденциальности, защиты и безопасности.

Т а б л и ц а В.6 — Разговорный ИИ и чат-боты

Элемент	Комментарий	
Описание	Использование ИИ в психиатрических учреждениях для обеспечения ухода за пациентами или дополнительных социальных взаимодействий	
Этические и общественные аспекты	Психиатрические службы встречают людей в наиболее уязвимом состоянии, поэтому использование разговорного ИИ в психиатрических службах требует тщательного рассмотрения. В частности, любое разговорное взаимодействие с такими пользователями требует пристального внимания с этической и критической точек зрения. Например, разговорный ИИ может нанести вред пользователям из-за его ограниченной способности воссоздать человеческое взаимодействие и обеспечить индивидуальное лечение. Это, однако, может быть сбалансировано преимуществами более вербального взаимодействия, чем это возможно в учреждениях по уходу, например из-за ограниченного количества персонала. Также необходим контроль, чтобы убедиться, что система ИИ не усвоила вредные понятия или язык. Пример использования иллюстрирует проблемы подотчетности (см. 6.2.1), справедливости и недискриминации (см. 6.2.2), прозрачности (см. 6.2.3), конфиденциальности (см. 6.2.6), защиты и безопасности (см. 6.2.7)	
Ссылки	См. [16]	Пример использования 106

В.2.7 Пример использования 7

Таблица В.7 иллюстрирует проблемы справедливости и недискриминации, профессиональной ответственности, продвижения человеческих ценностей, участия и развития сообщества и соблюдения верховенства закона.

Таблица В.7 — ИИ для понимания фальсификации в широко используемых продуктах питания

Элемент	Комментарий	
Описание	Фальсификация продуктов питания становится все более опасной, особенно с использованием фальсификатов, которые являются канцерогенными или вредными для таких частей тела, как почки. Например, молоко фальсифицируется содой, мочевиной и моющими средствами, а манго и бананы преждевременно созревают с помощью карбида кальция. Не существует низкочастотного (экономного) способа выявить эти виды фальсификации. Проведен эксперимент по контролируемой фальсификации и сняты показания гиперспектрального отражения. ИИ помог найти закономерности в гиперспектральной сигнатуре и смог надежно классифицировать (90 % ++) образцы, которые были нефальсифицированными либо фальсифицированными	
Этические и общественные аспекты	Справедливое отношение. Различные источники предвзятости, неправильное использование системы ИИ, неправильно обученная модель, неправильная классификация (≥ ложные обвинения). Неспособность применять такие системы одинаково для всего общества; риск пропуска правильных обвинений из-за неспособности применения технологии. Пример использования иллюстрирует проблемы справедливости и недискриминации (см. 6.2.2), профессиональной ответственности (см. 6.2.4), продвижения человеческих ценностей (см. 6.2.5), участия и развития сообщества (см. 6.2.9), соблюдения верховенства закона (см. 6.2.10)	
Ссылки	См. [16]	Выдержка из примера использования 19

В.2.8 Пример использования 8

Таблица В.8 иллюстрирует проблемы справедливости и недискриминации, прозрачности и объяснимости, а также профессиональной ответственности.

Таблица В.8 — Кредитный скоринг с использованием данных КЧС (Банковские и финансовые услуги)

Элемент	Комментарий	
Описание	Как правило, трудно построить систему оценки рисков, используя только данные КЧС («знай своего клиента»), которые часто содержат шум и неполноту. Однако если это реализовать, то можно использовать их для получения объективной оценки всех заявителей на получение кредита, даже тех, кто только начинает кредитоваться. Для этой цели подходят алгоритмы нелинейной классификации. В процессе КЧС от клиента собирается несколько переменных, таких как возраст клиента, заявленный им доход, род деятельности, цель кредита и т. д. Все эти характеристики можно добавить в нелинейную модель риска и обеспечить их сложное взаимодействие	
Этические и общественные аспекты	Данные КЧС, полученные из наиболее удаленных, сельских районов, могут быть с шумом и иметь несколько пропущенных значений. Перед передачей в алгоритм модели необходима соответствующая предварительная обработка. Пример использования иллюстрирует проблемы справедливости и недискриминации (см. 6.2.2), прозрачности и объяснимости (см. 6.2.3), профессиональной ответственности (см. 6.2.4)	
Ссылки	См. [16]	Выдержка из примера использования 27

В.2.9 Пример использования 9

Таблица В.9 иллюстрирует проблемы подотчетности, профессиональной ответственности, продвижения человеческих ценностей, конфиденциальности, безопасности и защиты, контроля человеком технологий, а также участия и развития сообщества.

Таблица В.9 — Сервис для разъяснения ситуации с помощью ИИ для людей с нарушениями зрения

Элемент	Комментарий	
Описание	Этот пример использования поддерживает повседневную жизнь людей с нарушениями зрения с помощью технологий машинного зрения (AI vision technologies). Этот сервис помогает таким людям распознавать и избегать опасных объектов на пути, идентифицировать людей, текст, предметы и знакомых людей, принимая во внимание различную окружающую ситуацию. Также поддерживается сервис субтитров, помогающий людям с нарушениями зрения понять текущую ситуацию или фотографии	
Этические и общественные аспекты	Развертывание подобных сервисов потенциально подвергает повседневную жизнь пользователя автоматической записи и мониторингу, если отсутствуют гарантии конфиденциальности или меры безопасности недостаточны. Непредоставление таких услуг на достаточном уровне там, где они доступны, чревато усугублением материального и социального неравенства. Пример использования иллюстрирует проблемы подотчетности (см. 6.2.1), профессиональной ответственности (см. 6.2.4), продвижения человеческих ценностей (см. 6.2.5), конфиденциальности (см. 6.2.6), безопасности и защиты (см. 6.2.7), контроля человеком технологий (см. 6.2.8), участия и развития сообщества (см. 6.2.9)	
Ссылки	См. [16]	Выдержка из примера использования 64

В.2.10 Пример использования 10

Таблица В.10 иллюстрирует проблемы продвижения человеческих ценностей, конфиденциальности, участия и развития сообщества.

Таблица В.10 — Решение ИИ для умного кампуса

Элемент	Комментарий	
Описание	Основанная на большом объеме данных и технологии ИИ, схема объединяет преподавание, экзамены, обучение и управление в интегрированную систему взаимного сотрудничества, основанную на сопутствующем сборе данных и динамическом анализе больших данных в сочетании с оценкой процесса. Это помогает преподавателям и студентам реализовать индивидуальное обучение и обучение в соответствии со способностями, помочь менеджерам в контроле и принятии решений и значительно способствует интеллектуализации образования, обучения и управления	
Этические и общественные аспекты	Раскрытие конфиденциальной информации преподавателям и студентам. Пример использования иллюстрирует проблемы продвижения человеческих ценностей (см. 6.2.5), конфиденциальности (см. 6.2.6), участия и развития сообщества (см. 6.2.9)	
Ссылки	См. [16]	Выдержка из примера использования 85

В.2.11 Пример использования 11

Таблица В.11 иллюстрирует проблемы подотчетности, профессиональной ответственности и контроля человеком технологий.

Т а б л и ц а В.11 — Автономность транспортных средств

Элемент	Комментарий
Описание	Автомобили, управляемые человеком, имеют значительный уровень аварийности. Функции помощника водителя продемонстрировали снижение уровня травматизма и смертности. Например, исследование показывает, что автомобили с системой ADAS (усовершенствованная система помощи водителю) приводят к снижению частоты заявлений на 27 % о телесных повреждениях и на 19 % о повреждениях имущества. Это исследование показывает, что если каждый автомобиль, реализуемый в США, оснастить системой контроля слепых зон, то можно предотвратить огромное количество аварий и травм. Однако конкретные случаи дорожно-транспортных происшествий с участием функций помощи водителю, а не полностью автономных автомобилей, привлекли большое внимание средств массовой информации и запятнали репутацию автономных транспортных средств в глазах некоторых людей (см. [102], [103])
Этические и общественные аспекты	Пример использования поднимает проблемы, связанные с подотчетностью (см. 6.2.1) и профессиональной ответственностью (см. 6.2.4) в отношении точного информирования о происшествиях и с контролем человеком технологий (см. 6.2.8)
Ссылки	См. [102]
	См. [103]

В.2.12 Пример использования 12

Таблица В.12 иллюстрирует проблемы конфиденциальности.

Т а б л и ц а В.12 — Распознавание лиц

Элемент	Комментарий
Описание	Технология распознавания лиц имеет широкий спектр потенциальных применений — от разблокировки мобильных телефонов до массового наблюдения за протестами со стороны государственных органов
Этические и общественные аспекты	Распознавание лиц — это технология, которая по-разному воспринимается в зависимости от контекста. Отдельные проявления могут быть запрещены из-за отсутствия согласия людей, чьи изображения использованы для обучения системы. Распознавание лиц регулируется по крайней мере в одном штате США (см. [104]), а планы по регулированию в Европейском союзе (см. [73]) материализованы в 2021 г. Пример использования потенциально иллюстрирует нарушение конфиденциальности (см. 6.2.6)
Ссылки	См. [104]
	См. [73]

В.2.13 Пример использования 13

Таблица В.13 иллюстрирует проблемы нарушения международных прав человека, конфиденциальности, защиты и безопасности, а также соблюдения верховенства закона.

Т а б л и ц а В.13 — ИИ в сочетании с подключенными медицинскими устройствами или устройствами интернета тела (IoB)

Элемент	Комментарий	
Описание	Исследователи Массачусетского технологического института (MIT) совместно с учеными из больницы Бригам энд Уоменс разработали новый способ питания и связи с устройствами, имплантированными глубоко внутри человеческого тела. Такие устройства могут быть использованы для доставки лекарств, мониторинга состояния организма или лечения заболеваний путем стимулирования мозга электричеством или светом. Имплантаты питаются от радиочастотных волн, которые могут безопасно проходить через ткани человека. Несмотря на то что эти крошечные имплантируемые устройства не имеют батарей, исследователи могут взаимодействовать с ними на расстоянии. Это открывает совершенно новые возможности для применения в медицине. Более того, другие исследователи пришли к выводу, что собственные проводящие свойства тела также могут быть использованы в качестве средства связи для устройств в так называемой телесной сети человека вместо использования беспроводных сигналов, которые легко перехватить (считать)	
Этические и общественные аспекты	Имплантируемое устройство, такое как искусственная нервная система или подключенное медицинское устройство, а также телесная сеть человека, как и IoB-устройство, может быть удалено, атаковано, использовано для несанкционированного сбора данных или взломано по различным причинам, вызванным ИИ. Также возможны ситуации, когда человеку без согласия или без его осознанного решения устанавливают имплантаты с ИИ (например, в сочетании с нанотехнологиями) или когда взламывают систему по различным причинам, связанным с ИИ, что свидетельствует о нарушении международных прав человека: права на жизнь, свободу и личную безопасность (см. 5.1), конфиденциальности (см. 6.2.6), защиты и безопасности (см. 6.2.7), соблюдения верховенства закона (см. 6.2.10)	
Ссылки	См. [105]	Разработка нового способа питания и связи с устройствами, имплантированными глубоко внутри человеческого тела. Такие устройства могут быть использованы для доставки лекарств, мониторинга состояния организма или лечения заболеваний путем стимулирования мозга электричеством или светом. Имплантаты питаются от радиочастотных волн, которые могут безопасно проходить через ткани человека. Несмотря на то что эти крошечные имплантируемые устройства не имеют батарей, исследователи теперь могут взаимодействовать с ними на расстоянии за пределами тела
	См. [106]	В статье описан следующий пример: «чтобы парализованный человек мог поднять предмет, в моторную кору (головного мозга) вживляются электродные решетки», которые «улавливают нейронные сигналы, возникающие, когда человек представляет, как он двигает рукой и кистью. Эти шумные сигналы мозга затем декодируются процессором с ИИ»
	См. [107]	В статье рассмотрено использование человеческого тела в качестве канала связи для IoB-устройств: «Вместо того чтобы выбирать между беспроводными сигналами, которые легко можно перехватить, и проводными сигналами, которые несут риск для организма, можно использовать человеческое тело в качестве средства связи для устройств в телесной сети человека, используя собственные проводящие свойства тела»

В.2.14 Пример использования 14

Таблица В.14 иллюстрирует проблемы нежелательной предвзятости и дискриминации.

Т а б л и ц а В.14 — Автоматизированная фильтрация кандидатов на соискание

Элемент	Комментарий
Описание	Если система и приложение ИИ использованы для фильтрации резюме, поданных соискателями, и находят соответствия с помощью модели, которая обучена распознавать людей, похожих на работающих сотрудников, это может закрепить существующие закономерности в рабочей силе, в том числе закономерности, возникшие в результате предыдущего или текущего необъективного найма, например за счет неосознанной предвзятости людей (см. [108], [109])
Этические и общественные аспекты	Сохранение нежелательной предвзятости и дискриминации. Пример использования иллюстрирует проблемы нежелательной предвзятости и дискриминации (см. 6.2.2)
Ссылки	См. [108]
	См. [109]

Библиография

- [1] ИСО/МЭК 22989 Информационная технология. Искусственный интеллект. Концепции и терминология искусственного интеллекта (Information technology — Artificial intelligence — Artificial intelligence concepts and terminology)
- [2] ISO/TR 21276:2018 Плиты кухонные и экологически чистые технологии для приготовления пищи. Словарь (Clean cookstoves and clean cooking solutions — Vocabulary)
- [3] ISO/IEC TR 24027:2021 Информационные технологии. Искусственный интеллект (ИИ). Смещенность в системах ИИ и при принятии решений с помощью ИИ [Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making]
- [4] ИСО 20294:2018 Технология полиграфии. Количественная оценка и обмен информацией для расчета углеродного следа электронных средств массовой информации (Graphic technology — Quantification and communication for calculating the carbon footprint of e-media)
- [5] ИСО 5127:2017 Информация и документация. Основные положения и словарь (Information and documentation — Foundation and vocabulary)
- [6] ИСО 30415:2021 Менеджмент человеческих ресурсов. Личностное многообразие и учет индивидуальных особенностей персонала (Human resource management Diversity and inclusion)
- [7] ISO/IEC/IEEE 12207:2017 Системная и программная инженерия. Процессы жизненного цикла программных средств (Systems and software engineering — Software life cycle processes)
- [8] ИСО/МЭК 24775-2:2021 Информационная технология. Управление памятью. Часть 2. Общая архитектура (Information technology — Storage management — Part 2: Common Architecture)
- [9] ИСО/МЭК 15444-8:2007 Информационные технологии. Система кодирования изображения JPEG 2000. Часть 8. Безопасность JPEG 2000 (Information technology — JPEG 2000 image coding system: Secure JPEG 2000 — Part 8)
- [10] ISO Guide 82:2019 Руководящие указания по рассмотрению уровня устойчивого развития в стандартах (Guidelines for addressing sustainability in standards)
- [11] ИСО 22948:2020 Углеродный след морепродуктов. Правила определения товарной категории костных рыб (Carbon footprint for seafood — Product category rules (CFP—PCR) for finfish)
- [12] Incident Database AI <https://incidentdatabase.ai/research/1-criteria>
- [13] ИСО/МЭК 38505-1 Информационная технология. Стратегическое управление ИТ. Стратегическое управление данными. Часть 1. Применение ISO/IEC 38500 для менеджмента данных (Information technology — Governance of IT — Governance of data — Part 1: Application of ISO/IEC 38500 to the governance of data)
- [14] ИСО/МЭК 23894 Информационная технология. Искусственный интеллект. Руководство по менеджменту риска (Information technology — Artificial Intelligence — Risk Management)
- [15] ISO/IEC TR 24028:2020 Информационная технология. Искусственный интеллект. Обзор достоверности систем искусственного интеллекта (Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence)
- [16] ISO/IEC TR 24030 Информационная технология. Искусственный интеллект (AI). Примеры использования [Information technology — Artificial intelligence (AI) — Use cases]
- [17] ИСО/МЭК 38507 Информационные технологии. Стратегическое управление ИТ. Последствия влияния стратегического управления при использовании искусственного интеллекта организациями (Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations)
- [18] ИСО 37000 Управление организациями. Руководство (Governance of organizations — Guidance)

- [19] ИСО/МЭК 27001 Информационная технология. Методы обеспечения безопасности. Системы менеджмента информационной безопасности. Требования (Information technology — Security techniques — Information security management systems — Requirements)
- [20] The IEEE Global Initiative on Ethics of Autonomous and Intelligent System, “Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems”, First ed. 2019
- [21] Wong P.-H. Responsible Innovation for Decent Nonliberal Peoples: A Dilemma? *Journal of Responsible Innovation* 3, no. 2 (2016): 154—168. <https://www.un.org/sites/un2.un.org/files/udhr.pdf>
- [22] Foot P. *The Problem of Abortion and the Doctrine of Double Effect*. *Oxford Review* 5, 1967. <https://philpapers.org/archive/FOOTPO-2.pdf>
- [23] Gensler H. *Ethics: A Contemporary Introduction*, 3rd ed. Routledge, 2018
- [24] Poff D. *Business Ethics in Canada*, 4th ed. Pearson, 2019
- [25] United Nations General Assembly. *Universal Declaration of Human Rights (Article 12)*. Paris, France, 10 Dec 1948. <https://www.un.org/en/universal-declaration-human-rights/>
- [26] United Nations Sustainable Development Goals *The 2030 Agenda for Sustainable Development*. 21 Oct 2015. https://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&Lang=E
- [27] United Nations Human Rights, Office of the High Commissioner. *Guiding Principles on Business and Human Rights*. United Nations, 2011. https://www.ohchr.org/documents/publications/GuidingprinciplesBusinesshr_eN.pdf
- [28] United Nations Human Rights, General Assembly. *International Convention on the Elimination of All Forms of Racial Discrimination*. 4 January 1969. <https://www.ohchr.org/en/professionalinterest/pages/cerd.aspx>
- [29] United Nations Department of economic and social affairs. *United Nations Declaration on the Rights of Indigenous Peoples (UNDRIP)*, New York, USA, adopted 13 September 2007. <https://www.un.org/development/desa/indigenouspeoples/declaration-on-the-rights-of-indigenous-peoples.html>
- [30] United Nations Human Rights, General Assembly. *Convention on the Elimination of All Forms of Discrimination against Women*. (CEDAW). New York, USA, adopted 18 Dec 1979. <https://www.ohchr.org/en/professionalinterest/pages/cedaw.aspx>
- [31] United Nations, Department of economic and social affairs. *Convention on the Rights of Persons with Disabilities (CRPD)*. Adopted Dec 13 2006. Entered into force 3 May 2008. <https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities.html>
- [32] Council of Europe Committee of experts on human rights dimensions of automated data processing and different forms of artificial intelligence, *A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework*. 22 May 2019. <https://rm.coe.int/a-study-of-the-implications-of-advanced-digital-technologies-including/168094ad40>
- [33] Council of Europe Study on human rights dimensions of automated data processing techniques, *Study on the human rights dimensions of automated data processing techniques (in particular algorithms) and possible regulatory implications*. March 2018. <https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5>
- [34] Commission to the European Parliament the Council, the European Economic and Social Committee and the Committee of the Regions. *Communication — Building Trust in Human Centric Artificial Intelligence*. European Commission. Brussels, Belgium, 4 April 2019. <https://digital-strategy.ec.europa.eu/en/library/communication-building-trust-human-centric-artificial-intelligence>
- [35] CWA 17145-2:2017, *Ethics assessment for research and innovation — Part 2: Ethical impact assessment framework*., <https://satoriproject.eu/media/CWA17145-23d2017.pdf>
- [36] AI Now Institute *AI Now 2017 Report*. New York, USA. 2017 https://assets.ctfassets.net/8wprhhvnpfc0/1A9c3ZTCZa2KEYM64Wsc2a/8636557c5fb14f2b74b2be64c3ce0c78/_AI_Now_Institute_2017_Report_.pdf
- [37] Wright D. 2011. A framework for the ethical impact assessment of information technology. *Ethics and information technology*, 13(3), pp.199—226. September 2011. Published online: 8 July 2010 Springer Science+Business Media B.V. 2010. <https://link.springer.com/article/10.1007/s10676-010-9242-6>

- [38] Adadi A., Berrada M. *Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI)*. IEEE. 17 Sep 2018.
- [39] Aether Committee Microsoft. *Microsoft AI principles*. 2017. <https://www.microsoft.com/en-us/ai/our-approach-to-ai>
- [40] Schwartz S.H., An overview of the Schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1), 2307-0919.) Dec 2012. <https://scholarworks.gvsu.edu/orpc/vol2/iss1/11/>
- [41] Noriega P., Verhagen H., Padget J., d'Inverno M. *Ethical Online AI Systems Through Conscientious Design*. IEEE Internet Computing no. 6, vol. 25 (2021) pp. 58—64. Dec 2021. <https://doi.org/10.1109/MIC.2021.3098324>
- [42] European Parliament and Council *Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)*. 27 April 2016. <https://eur-lex.europa.eu/eli/reg/2016/679/oj>
- [43] Article 29 Data Protection Working Party. *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679*. 3 October 2017 as last revised and adopted on 22 February 2018. European Commission. Brussels, Belgium, 2018. https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053
- [44] Cavoukian A. *Privacy by Design The 7 Foundational Principles*. November 2011. <https://www.privacysecurityacademy.com/wp-content/uploads/2020/08/PbD-Principles-and-Mapping.pdf>
- [45] Young & Leveson, 2014; Young W., & Leveson N. G. *An integrated approach to safety and security based on systems theory*. *Communications of the ACM*, Vol. 57(2), 31-35. Feb 2014. <https://cacm.acm.org/magazines/2014/2/171683-an-integrated-approach-to-safety-and-security-based-on-systems-theory/fulltext>
- [46] European Commission *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment*. July 17 2020. <https://futurium.ec.europa.eu/en/european-ai-alliance/pages/altai-assessment-list-trustworthy-artificial-intelligence>
- [47] OECD *Artificial Intelligence in Society*. 2019. <https://oecd.ai/en/ai-principles>
- [48] European Commission *White Paper on Artificial Intelligence — A European Approach to Excellence and trust*. Brussels, Belgium. 19 February 2020. <https://digital-strategy.ec.europa.eu/en/white-paper-artificial-intelligence-public-consultation-towards-european-approach-excellence-and-trust>
- [49] United Nations *Paris Agreement*. Paris, France. 12 Dec 2015. https://unfccc.int/sites/default/files/english_paris_agreement.pdf
- [50] International Labour Organization, 1. *Freedom of Association and Protection of the Right to Organise Convention*, 1948 (No. 87), 2. *Right to Organise and Collective Bargaining Convention*, 1949 (No. 98), 3. *Forced Labour Convention*, 1930 (No. 29) (and its 2014 Protocol), 4. *Abolition of Forced Labour Convention*, 1957 (No. 105), 5. *Minimum Age Convention*, 1973 (No. 138), 6. *Worst Forms of Child Labour Convention*, 1999 (No. 182), 7. *Equal Remuneration Convention*, 1951 (No. 100), 8. *Discrimination (Employment and Occupation) Convention*, 1958 (No. 111). <https://www.ilo.org/global/standards/introduction-to-international-labour-standards/conventions-and-recommendations/lang--en/index.htm>
- [51] Leslie D., Burr C., Aitken M., Cows J., Katell M. and Briggs. *Artificial intelligence, human rights, democracy, and the rule of law: a primer*. The Council of Europe. The Alan Turing Institute. The Council of Europe's ad hoc committee on artificial intelligence. 26 March 2021. <https://doi.org/10.5281/zenodo.4639743>
- [52] Fjeld J., Achten N., Hilligoss H., Nagy A., Srikumar M. 2020. *Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI*. Berkman Klein Center Research Publication No. 2020-1. 2020. <https://cyber.harvard.edu/publication/2020/principled-ai>
- [53] Maria Luciana Axente and Ilana Golbin *Ten principles for ethical AI*. April 13, 2022. <https://www.pwc.com.au/digitalpulse/ten-principles-ethical-ai.html>
- [54] Hauer T. *Incompleteness of moral choice and evolution towards fully autonomous AI*. 1 February 2022. <https://www.pwc.com.au/digitalpulse/ten-principles-ethical-ai.html>
- [55] Hewlett Packard Enterprise (Belgium) *AI Ethics and principles*. 2022. <https://www.hpe.com/be/en/ai-ethics.html>
- [56] UNESCO *Resource Guide on Artificial Intelligence (AI) strategies*. April 2021. https://sdgs.un.org/sites/default/files/2021-04/Resource%20Guide%20on%20AI%20Strategies_April%202021_rev_0.pdf

- [57] Hogenhout L. *Unite Paper. A Framework for Ethical AI at the United Nations*. March 2021. https://unite.un.org/sites/unite.un.org/files/unite_paper_-_ethical_ai_at_the_un.pdf
- [58] UNESCO *Recommendation on the Ethics of Artificial Intelligence, SHS/BIO/REC-AIETHICS/2021*. <https://unesdoc.unesco.org/ark:/48223/pf0000380455>
- [59] AI4DA. *Ethics in Artificial Intelligence*. 18 December 2020. <https://ai4da.com/ethics-in-artificial-intelligence/>
- [60] Government of Canada *Directive on Automated Decision-Making*. Updated 1 April 2021. <https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592>
- [61] Information Technology Industry Council *AI Policy principles, Executive summary*. March 2021. <https://www.itic.org/resources/AI-Policy-Principles-FullReport2.pdf>
- [62] IBM *From Roadblock to Scale: The Global Sprint Towards AI Study*. 2020. <https://www.ibm.com/au-en/artificial-intelligence/ethics>
- [63] Macarena Gatica L., M. Ignacia Ormeño Sarraide, and Jaime Urzúa — Chile. *Principles and axes of Chile's AI Policy*. December 2021. <https://www.dataguidance.com/opinion/chile-principles-and-axes-chiles-ai-policy>
- [64] Roman J. *Artificial Intelligence in Brazil: The Brazilian Strategy for Artificial Intelligence (BSAI/EBIA) and Bill No. 21/2020*. October 2021. <https://irisbh.com.br/en/artificial-intelligence-in-brazil-the-brazilian-strategy-for-artificial-intelligence-bsai-ebia-and-bill-no-21-2020/> (Article)
- [65] Mari A. *Legal framework for artificial intelligence advances in Brazil*. News article. September 2021 <https://www.zdnet.com/article/legal-framework-for-artificial-intelligence-advances-in-brazil/>
- [66] Kendall J. *Oxford Insights' AI Strategy Series: Argentina and Uruguay*. January 2021. <https://www.oxfordinsights.com/insights/2021/1/8/oxford-insights-ai-strategy-series-argentina-and-uruguay>
- [67] Enrique Zapata — Mexico *The story and lessons behind Latin America's first AI strategy*. June 2020. https://scioteca.caf.com/bitstream/handle/123456789/1587/Mexico_the_story_and_lessons_behind_Latin_America's_first_AI_strategy.pdf?sequence=1&isAllowed=y
- [68] Carman M., Rosman B. *Defining what's ethical in artificial intelligence needs input from Africans*. November 2021. <https://link.springer.com/article/10.1007/s10676-020-09534-2>
- [69] Akorfa J. *Top AI challenges in Africa: Compute*. November 2021 <https://www.alliance4ai.org/top-ai-challenges-in-africa-compute/>
- [70] James Cottrell et al *Carnegie Council Podcasts: Africa, Artificial Intelligence, & Ethics*. April 2021. <https://www.carnegiecouncil.org/studio/multimedia/20210405-africa-artificial-intelligence-ethics>
- [71] Sountongnoma Martial Anicet Kiemde & Ahmed Dooguy Kora *Towards an ethics of AI in Africa: rule of education*. 2021. <https://link.springer.com/article/10.1007/s43681-021-00106-8>
- [72] Teve M. *Top AI challenges in Africa: Ethics*. 2020. <https://www.alliance4ai.org/top-ai-challenges-in-africa-ethics/>
- [73] European Commission *Proposal for regulation on a European approach on artificial intelligence*. April 2021. <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52021PC0206&from=EN>
- [74] Kuleshov A., Ignatiev A., Abramova A., Marshalko G. *Addressing AI ethics through codification*, DOI: 10.1109/EnT48576.2020.00011. IEEE publisher. 15 July 2020. <https://ieeexplore.ieee.org/document/9140627/>
- [75] European Parliament *European Parliament Resolution with recommendations to the Commission on a civil liability regime for artificial intelligence*, (2020/2014(INL)). 20 Oct 2020. https://www.europarl.europa.eu/doceo/document/TA-9-2020-0275_EN.html
- [76] European Parliament *Data subjects, digital surveillance, AI and the future of work*. 23 December 2020. [https://www.europarl.europa.eu/RegData/etudes/STUD/2020/656305/EPRS_STU\(2020\)656305_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2020/656305/EPRS_STU(2020)656305_EN.pdf)
- [77] CEN-CENELEC *Focus Group report, Road Map on Artificial Intelligence (AI)*. 7 Sep 2020. https://ftp.cencenelec.eu/EN/EuropeanStandardization/Sectors/AI/CEN-CLC_FGR_RoadMapAI.pdf
- [78] Sheehan M. *China's New AI Governance Initiatives Shouldn't Be Ignored*. January 2022. <https://carnegieendowment.org/2022/01/04/china-s-new-ai-governance-initiatives-shouldn-t-be-ignored-pub-86127>

- [79] Chan S. *What you need to know about China's AI ethics rules*. 2022. <https://techbeacon.com/enterprise-it/what-you-need-know-about-chinas-ai-ethics-rules>
- [80] Yoshida J., Yanai K. Ph.D. *Principles and Practice for the Ethics Use of AI in Hitachi's Social Innovation Business*. 2022. <https://www.hitachi.com/rev/archive/2022/r2022-sp/01/index.html>
- [81] Malaysia AI Ethics Maturity Report 2021. *AI Doctrina*. November 2021. <https://www.ai-doctrina.info/malaysian-ai-ethics-maturity-report-2021>
- [82] AIDA *AI Readiness in the Philippines*. October 2021. <https://ai4da.com/ai-readiness-in-the-philippines/>
- [83] International Research Center for AI Ethics and Governance, The Ethical Norms for the New Generation Artificial Intelligence, China. September 2021. <https://ai-ethics-and-governance.institute/2021/09/27/the-ethical-norms-for-the-new-generation-artificial-intelligence-china>
- [84] *Expert Group on How AI Principles Should be Implemented — AI Governance Guidelines working group* (Japan). July 2021. https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20210709_9.pdf
- [85] Hongladarom S. *What Buddhism can do for AI ethics*. 2021. <https://www.technologyreview.com/2021/01/06/1015779/what-buddhism-can-do-ai-ethics/>
- [86] Sambasivan et al. *Re-imagining Algorithmic Fairness in India and Beyond*. 2021. <https://research.google/pubs/pub50002>
- [87] Hwang H. Min-Hye Park. *The Threat of AI and Our Response: The AI Charter of Ethics in South Korea*. 2020. <https://www.koreascience.or.kr/article/JAKO202013965596961.pdf>
- [88] Tass. First code of ethics of artificial intelligence signed in Russia. October 2021. <https://tass.com/economy/1354187>, <https://ethics.a-ai.ru/>¹⁾
- [89] Sberbank. *Sber among pioneers adopting AI ethics principles in Russia*. March 2021. <https://www.sberbank.com/news-and-media/press-releases/>
- [90] AI Alliance Russia *Artificial Intelligence Code of Ethics*. 2021. <https://a-ai.ru/wp-content/uploads/2021/10/Code-of-Ethics.pdf>
- [91] Australian Government Department of Industry Science, Energy and Resources. *Australia's Artificial Intelligence Ethics Framework*. 2022. <https://www.industry.gov.au/data-and-publications/australias-artificial-intelligence-ethics-framework>
- [92] Singapore A.I. *An Overview of AI Ethics and Governance*. June 2020. <https://aisingapore.org/2020/06/an-overview-of-ai-ethics-and-governance/> (Overview) <https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.ashx> (Model Framework, 2nd Edition)
- [93] Forum New Zealand A.I. *Trustworthy AI in Aotearoa: The AI Principles*. March 2020. <https://aiforum.org.nz/wp-content/uploads/2020/03/Trustworthy-AI-in-Aotearoa-March-2020.pdf>
- [94] Personal Data Protection Commission Singapore (PDPC) *Singapore's Approach to AI Governance (ISAGO)*. 2020. <https://www.pdpc.gov.sg/Help-and-Resources/2020/01/Model-AI-Governance-Framework>
- [95] Digital S. (SG:D). *Compendium of Use Cases, Practical illustrations of the Model AI governance framework*. 2020. <https://file.go.gov.sg/ai-gov-use-cases-2.pdf>
- [96] Advisory Council of the Ethical Use of AI and Data the Infocomm Media Development Authority (IMDA) and the PDPC. *A Guide to Job Redesign in the Age of AI*. 2020. <https://file.go.gov.sg/ai-guide-to-jobredesign.pdf>
- [97] Singapore Computer Society — AI Ethics and governance body of knowledge. 2020. <https://files-scs-prod.s3-ap-southeast-1.amazonaws.com/public/uploads/200721016+SCS+Booklet+6+Nov.pdf> (Toolkit) <https://ai-ethics-bok.scs.org.sg> (Book of Knowledge)
- [98] Interpol UNICSRI Global Meeting. *Artificial Intelligence and law enforcement challenges and opportunities*. 1 Dec 2020. <https://www.interpol.int/News-and-Events/News/2020/Artificial-Intelligence-and-law-enforcement-challenges-and-opportunities>

¹⁾ В ссылку [88] добавлена ссылка на сайт, где представлен полный текст российского «Кодекса этики в сфере ИИ» на русском и английском языках (внесено для удобства поиска документа пользователями).

- [99] The International Criminal Police Organization United Nations Interregional Crime and Justice Research Institute. *Artificial Intelligence and Robotics For Law Enforcement*. 2019. [https:// www.europarl.europa.eu/cmsdata/196207/ UNICRI%20-%20Artificial%20intelligence%20and%20robotics%20for%20law%20enforcement.pdf](https://www.europarl.europa.eu/cmsdata/196207/UNICRI%20-%20Artificial%20intelligence%20and%20robotics%20for%20law%20enforcement.pdf)
- [100] Morgan B. *Ethics and Artificial Intelligence with IBM Watson's Rob High*. 13 June 2017. [https:// www.forbes.com/sites/blakemorgan/2017/06/13/ethics-and-artificial-intelligence-with-ibm-watsons-rob-high/](https://www.forbes.com/sites/blakemorgan/2017/06/13/ethics-and-artificial-intelligence-with-ibm-watsons-rob-high/)
- [101] Ruane E., Birhane A., Ventresque A. *Conversational AI. Social and Ethical Considerations*. 27th AI Irish Conference on Artificial Intelligence and Cognitive Science. Dec 2019, pp. 104-115. https://ceur-ws.org/Vol-2563/aics_12.pdf
- [102] Elliott C. *Are Vehicle Safety Features Actually Reducing Car Accidents?* 2 Sep. 2020 <https://risk.lexisnexis.com/about-us/press-room/press-release/20200618-vehicle-build>
- [103] Davenport J., French T. *Device; dedicated attack vectors in existing Automotive Cyber & Digital resilience*. 2020
- [104] Hill K. New York Times. *How One State Managed to Actually Write Rules on Facial Recognition*. New York Times. 2021. <https://www.nytimes.com/2021/02/27/technology/Massachusetts-facial-recognition-rules.html>
- [105] Trafton A. *Wireless systems can power devices inside the body*. Science Daily. Massachusetts Institute of Technology. 4 June 2018. <https://news.mit.edu/2018/wireless-system-power-devices-inside-body-0604>
- [106] Bouton C. *Brain Implants and Wearables Let Paralyzed People Move Again*. IEEE. 26 Jan 2021. [https:// spinalcordinjuryzone.com/news/54839/brain-implants-and-wearables-let-paralyzed -people-move-again](https://spinalcordinjuryzone.com/news/54839/brain-implants-and-wearables-let-paralyzed -people-move-again)
- [107] Shreyas Sen Shovan Maity Debayan Das. *The Body is the Network*. IEEE. Dec 2020
- [108] Imana B., Korolova A., Heidemann J. *Auditing for Discrimination in Algorithms Delivering Job Ads*. In Proceedings of The Web Conference 2021. 2021. [https://www.isi.edu/~johnh/PAPERS/ Imana21b.pdf](https://www.isi.edu/~johnh/PAPERS/Imana21b.pdf)
- [109] Powell A. The Harvard Gazette. *An awakening of data privacy*. The Harvard Gazette. 27 February 2020. [https:// news.harvard.edu/gazette/story/2020/02/surveillance-capitalism-author-sees-data-privacy-awakening/](https://news.harvard.edu/gazette/story/2020/02/surveillance-capitalism-author-sees-data-privacy-awakening/)

УДК 004.8:006.034:006.034

ОКС 35.020

Ключевые слова: искусственный интеллект, этические проблемы, общественные аспекты

Редактор *Л.С. Зимилова*
Технический редактор *И.Е. Черепкова*
Корректор *Л.С. Лысенко*
Компьютерная верстка *Е.А. Кондрашовой*

Сдано в набор 17.11.2023. Подписано в печать 23.11.2023. Формат 60×84%. Гарнитура Ариал.
Усл. печ. л. 5,58. Уч.-изд. л. 5,02.

Подготовлено на основе электронной версии, предоставленной разработчиком стандарта

Создано в единичном исполнении в ФГБУ «Институт стандартизации»
для комплектования Федерального информационного фонда стандартов,
117418 Москва, Нахимовский пр-т, д. 31, к. 2.
www.gostinfo.ru info@gostinfo.ru