

Министерство науки и высшего образования
Российской Федерации
ФГБОУ ВО «Бурятский государственный университет
имени Доржи Банзарова»

М. Х. Бадмаева, М. В. Золхоева

**ПРИМЕНЕНИЕ СИСТЕМ
ИСКУССТВЕННОГО ИНТЕЛЛЕКТА:
СОЦИАЛЬНО-ФИЛОСОФСКИЕ ПРОБЛЕМЫ
И ПРИНЦИПЫ**

Монография

Улан-Удэ
Издательство Бурятского госуниверситета имени Доржи Банзарова
2026

УДК 1:316 Утверждено к печати редакционно-издательским советом
ББК 60.0 Бурятского государственного университета
Б 153 Протокол № 5 от 16 сентября 2025 г.

Рецензенты

О. Б. Истомина, д-р филос. наук, доц., зав. кафедрой социально-экономических дисциплин, Иркутский государственный университет
И. И. Осинский, д-р филос. наук, проф. кафедры философии,
Бурятский государственный университет имени Доржи Банзарова
И. В. Цвык, д-р филос. наук, проф. кафедры философии, Московский авиационный институт (национальный исследовательский университет)

Издание размещено в системе РИНЦ на платформе научной
электронной библиотеки eLibrary.ru

Авторы

Бадмаева М. Х. (гл. 1, 3, закл.), Золхоева М. В. (предисл., гл. 2)

Бадмаева М. Х.

Б 153 **Применение систем искусственного интеллекта: социально-философские проблемы и принципы:** монография / М. Х. Бадмаева, М. В. Золхоева. — Улан-Удэ: Издательство Бурятского госуниверситета им. Д. Банзарова, 2026. 204 с. ISBN 978-5-9793-2162-2
DOI 10.18101/978-5-9793-2162-2-1-204

В монографии на социально-философском уровне рассмотрения научной проблемы исследована сущность искусственного интеллекта, выявлены основные сферы его использования, изучены специфические проблемы применения ИИ в современном обществе, а также выработаны философские принципы, призванные предупредить, исключить их появление и эскалацию в современном обществе.

Издание предназначено для философов, политологов, социологов, историков и всех интересующихся проблемами современного общества и применения систем искусственного интеллекта.

Badmaeva M. Kh.

Applications of Artificial Intelligence Systems: Social and Philosophical Problems and Principles: monograph / M. Kh. Badmaeva, M. V. Zolkhoeva. — Ulan-Ude: Dorzhi Banzarov Buryat State University Publishing Department, 2026. — 204 p. ISBN 978-5-9793-2162-2

In the monograph, the essence of artificial intelligence is studied at the socio-philosophical level of consideration of the scientific problem, the main areas of its use are identified, specific problems of the use of AI in modern society are studied, and philosophical principles are developed that are designed to prevent, exclude their emergence and escalation in modern society.

The publication is intended for philosophers, political scientists, sociologists, historians and all those interested in the problems of modern society and the use of artificial intelligence systems.

УДК 1:316
ББК 60.0

ISBN 978-5-9793-2162-2

© Бурятский госуниверситет
им. Д. Банзарова, 2026

ПРЕДИСЛОВИЕ

В последние годы искусственный интеллект (далее — ИИ) прочно входит в нашу жизнь, раскрывая перед человеком многообразные возможности, улучшая качество человеческой жизни, расширяя горизонты самореализации человека. Однако ИИ таит в себе и новые опасности, оказывая значительное влияние на человека, общество, окружающую среду.

ИИ обладает способностью воздействовать на культуру, существенно трансформируя ее базовые составляющие. Можно с полным правом уже сегодня заявить о том, что ИИ — это сложное и многоплановое явление, источник масштабных социокультурных изменений, как положительных, так и негативно воздействующих на человека и человеческую цивилизацию.

В этих условиях философия, будучи рациональным типом мировоззрения, определяет задачу выработать аргументированное и взвешенное отношение к внедрению систем искусственного интеллекта, изучить, какие риски сопровождают этот процесс, какое именно воздействие они оказывают на миропонимание современного индивида, на его отношение к окружающему миру.

Проблемы применения ИИ необходимо рассматривать сквозь призму подлинно философского толкования места и роли человека в мире, гуманистических целей и ценностей современного социума. Любые достижения в сфере ИИ имеют смысл только в том случае, если они коррелируют с идеалами процветания человека, человеческой цивилизации. Чрезмерное делегирование полномочий от человека системам ИИ, напротив, способно негативно повлиять на экзистенциальные основы бытия человека, с особой остротой поставив вопрос о смысле его существования и жизненном предназначении.

В социально-философском дискурсе не раз обращалось внимание на тему искусственного интеллекта, но теоретическое осмысление происходящих внутри его систем изменений, современной специализации ИИ и воздействия результатов новейших разработок в

этой сфере на человека и общество все еще существенным образом отстает от прогресса и внедрения этих технологий в жизнь современного социума. Теоретико-методологические и социально-философские основания планирования процесса развития и внедрения систем ИИ сегодня разработаны все еще недостаточно.

Все это может быть устранено при разработке философских принципов взаимодействия с ИИ, нацеленных на эффективное и безопасное использование искусственного разума человеком и обществом. Подобные исследования позволят сформулировать содержание основополагающих требований к разработчикам и пользователям систем ИИ на любых стадиях его развития и совершенствования, минимизировать негативные последствия, определить границы применения ИИ, разумно и гуманно реализовать его потенциал для решения глобальных проблем нашей цивилизации. В настоящее время на уровне государственных структур, научных организаций, бизнеса и гражданского общества многое сделано для выявления указанных принципов, но содержание их, как правило, не раскрыто, расплывчато, многозначно или вообще лишено каких-либо пояснений.

В настоящее время существует значительное количество отечественных и зарубежных исследований технического, юридического, социологического и философского характера, в которых рассмотрены отдельные аспекты взаимодействия человека с искусственным интеллектом. Среди них прежде всего выделяются работы, посвященные влиянию техники на человека. В самом широком смысле искусственный интеллект можно рассматривать как часть философии техники. К этому пулу относятся классические труды У. Бека (2000), В. Г. Горохова (2011), М. Кастельса (2000), Г. Маркузе (1994), М. Маклюэна (2003), Л. Мэмфорда (1986), Х. Ортеги-и-Гассета (1997), В. М. Розина (2006), Ф. Рело (1885), А. Ридлера (1900), М. А. Розова [Степин, Горохов, Розов, 1996], С. А. Смирнова (2006), В. С. Степина (1989), П. Флоренского (1993), Ф. Фукуямы (2004), Ю. Хабермаса (2007), М. Хайдеггера (1993), И. В. Черниковой и Д. В. Черниковой (2021) и др.

Так, К. Ясперс, в частности, видит в технизации угрозу утраты внутренней свободы и ценности личности. Он приходит к выводу, что именно за человеком должна оставаться целеполагающая функция, а техника является лишь средством достижения целей, поставленных человеком. Размышляя об экзистенциальных угрозах, вызванных тотальным распространением техники, он пишет, что техника — это «нечто такое, что подавляет, влияет на все их существование, противостоит им, не осознано ими, что словно бы происходит на заднем плане, не раскрыто» [Ясперс, 1991, с. 137]. А. Гринфилд называет современные технологии «радикальными», поскольку они оказывают беспрецедентное в истории развития техники влияние на социальный порядок и повседневный уклад жизни общества. По его мнению, искусственный интеллект — это нечто, ни на что не похожее и способное кардинально изменить будущее человечества. Ф. Фукуяма в своих трудах выражает обеспокоенность о допустимости преобразований человека и общества с помощью применения современных технологий. Он предостерегает, что изменениям подвергнутся те существенные черты, которые делают человека человеком.

Ж. Бодрийяр (1999), Л. Мэмфорд, Д. Нейсбит (2005), Ф. Фукуяма резко критиковали технологизацию и внедрение искусственного интеллекта без осмысления социальных и этических последствий. Отечественные исследователи В. А. Кутырев (2001), И. В. Черникова и Д. В. Черникова (2021) полагают, что человеческая идентичность подвержена опасности из-за необдуманного внедрения высоких технологий. А. Нордманн (Nordmann, 2015, р. 3–7), Э. Тоффлер (2002), Ф. Фукуяма, Ю. Хабермас исследуют использование новых технологий, задаваясь главным образом вопросом о вызванных им возможных изменениях человеческой природы. Д. В. Иванов (2002), М. Кастельс, И. С. Мелюхин (1999), А. И. Ракитов (1989) анализируют социокультурные последствия использования информационных технологий. В. Г. Горохов рассматривает возможные результаты внедрения новых технологий с позиций этических норм и вводит новое понятие «наноэтика» [Горохов, 2008, с. 84–98].

Не менее важны работы, целью которых стало определение, а также классификация ИИ. К этой группе относятся исследования таких авторов, как А. Ю. Алексеев (2013), В. В. Архипов, В. Б. Наумов (2017), Р. Брукс [Stone, Rodney, Erik, Ryan, Etzioni, 2016], Н. Бостром (2016), Р. Бродхэurst [Broadhurst, Brown, Maxim, Trivedi, Wang, 2019, p. 1–70], В. В. Васильев (2009), Н. Винер [Wiener, 1954], Б. Герцель [Goertzel, 2014, p. 1–46], П. М. Готовцев, Г. В. Ройзенсон (2020), Х. Де Гарис (2010), И. Дубровский (2005), А. Э. Евстратов, И. Ю. Гученков (2020), В. Э. Карпов, П. М. Готовцев, Г. В. Ройзенсон (2020), Р. Курцвейл [Kurzweil, 1990], Ш. Легг [Legg, Hutter, 2007, p. 17–24], Дж. Маккарти [McCarthy, 2007], К. Макниш [MacNish, Pearce, Pereira, 1994], Р. Мерфи [Murphy, 2019], М. Мински (1979), П. М. Морхат (2017), В. Наумов (2018), А. В. Незнамов (2020), Н. Нильсон (1973), П. Норvig [Norvig, 1991], Р. Пенроуз (2008), М. Райан [Ryan, 2020, p. 2749–2767], С. Рассел (2006), О. В. Ревинский (2017), А. В. Резаев, Г. Ф. Ручкина (2020), Дж. Серль (1998), В. Н. Синельникова (2017), Н. Д. Трегубова (2019), А. Тьюринг [Turing, 1950, p. 433–460], А. Турчин (2013), В. Фулин (2009), М. Хэнлэйн [Kaplan, Haenlein, 2019].

Впервые тема искусственного интеллекта в виде идеи «мыслящих машин» была введена в 1950 г. в работе британского математика А. Тьюринга «Вычислительные машины и разум» [Тьюринг, 2016]. В 1956 г. Дж. Маккарти использовал термин «искусственный интеллект» для обозначения научных исследований, связанных с математическими, лингвистическими и алгоритмическими проблемами, необходимыми для имитации интеллекта человека с помощью компьютера [McCarthy, 2007].

Понятие «сильный искусственный интеллект» было введено американским философом Дж. Серлем и контекстуально определено следующим образом: «Более того, такая программа будет не просто моделью разума; она в буквальном смысле слова сама и будет разумом, в том же смысле, в котором человеческий разум — это разум» [Серль, 2017].

Определение «слабого (или узкого) ИИ» предложили, в частности, российские ученые В. Н. Синельникова и О. В. Ревинский, по мнению которых слабый ИИ — это компьютерная программа, спроектированная людьми и обладающая способностью в соответствии с заложенной командной архитектурой создавать новую информацию [Синельникова, Ревинский, 2017, с. 17–27].

В дальнейшем представители аналитической философии (Н. Блок [Block, 1978, р. 261–325], Д. Деннет (2004), Т. Нагель (2003), Х. Патнэм [Putnam, 1992, р. 1–18], Дж. Сёрль (1998), Дж. Фодор [Fodor, 1998, р. 17], Д. Чалмерс (2019) и др.) в своих трудах различали сильный и слабый ИИ как два основных подхода к искусственному интеллекту, сложившихся в современной науке.

Разнообразие многочисленных подходов к пониманию ИИ представлено в исследовании Ш. Легг и М. Хаттер «Коллекция определений ИИ» [Legg, Hutter, 2007, р. 17–24].

Отдельного внимания заслуживает обсуждение проблем и принципов применения искусственного интеллекта в трудах Р. Г. Апресяна (2019), П. Готовцева (2020), А. А. Гусейнова (2004), В. Карпова (2020), А. В. Разина (2019), Г. В. Ройзензона (2018), В. А. Цвык и И. В. Цвык (2022), Б. Г. Юдина (2012) и др. Авторы пришли к выводу о том, что проблемы разработки, внедрения и использования ИИ обладают ярко выраженной спецификой, отличающей их от комплекса вопросов, обсуждаемых в рамках биоэтики, генной инженерии, информатики и других областей научного знания. В. А. Цвык и И. В. Цвык отмечают, что последствия от внедрения ИИ оказывают глубокое влияние на развитие общества, науки, культуры и коммуникации.

По их мнению, несмотря на то, что ИИ имеет потенциал изменения человечества в лучшую сторону, он порождает риски, угрожающие основным правам и свободам человека. П. Готовцев, В. Карпов, А. В. Разин, Г. В. Ройзензон в своих работах ставят вопросы о том, какие этические нормы должны быть заложены в ИИ на этапе его разработки.

Целый ряд отечественных и зарубежных исследователей занимались исследованием проблем применения ИИ в конкретных сферах жизнедеятельности общества: Дж. Боссманн [Bossmann, 2018], А. Джобин, М. Йенка и Е. Вайена [Jobin, Lenca, Vayena, 2019, p. 389–399], В. Э. Карпов (2018), М. Кэролан [Carolan, 2020, p. 184–207], В. А. Лаптев (2017), П. М. Морхат (2017), А. В. Попова (2020), М. Райан [Ryan, 2019], С. Рассел [Russell, 2021, p. 3–23], Таддео [Taddeo, 2019, p. 349–354], Э. Тополь (2016), Л. Флориди [Floridi, 2021, p. 612–622], Т. Хагендорф [Hagendorff, 2020, p. 99–120], Э. Юджовски (2008), Н. А. Ястреб (2015) и др.

Так, в трудах Л. Флориди обсуждались вопросы, связанные с соблюдением конфиденциальности личных данных [Mittelstadt, Allo, Taddeo et. al., 2016, p. 1–21]. С. Паркенсон и Э. Харпер пытались выявить препятствия и трудности, возникающие в процессе включения конфиденциальных данных пользователей в наборы больших данных [Harper, Parkerson, 2015]. Необходимость обеспечения прозрачности процессов, связанных с данными пользователей, была обоснована в работах Дж. Баррел [Burrell, 2016, p. 1–12].

Преодоление и недопущение социальной несправедливости и предвзятого отношения — предмет исследований Т. Панч [Panch, Mattie, Atun, 2019, p. 23–32]. Российско-французский философ А. Гринбаум в труде «Машина-доносчица» поднимает вопросы, касающиеся «нравственности» узкого искусственного интеллекта, его ответственности перед человеком [Гринбаум, 2017]. Автор называет узкий ИИ некой цифровой особой, которая сама по себе не может быть признана способной нести ответственность за совершенные действия и принятые решения. По-настоящему ответственной, нравственной ее может сделать только сам человек. Вообще проблемы нравственности, соответствия критериям добра при использовании систем ИИ затрагивались еще во времена А. Тьюринга. По сей день в трудах многих ученых, исследователей анализируется возможность причинения вреда человечеству со стороны сильного ИИ, выражается тревога, а также опасение за будущее, за безопасность че-

ловческой цивилизации в целом. Определенные результаты в данном контексте достигнуты в отечественном общественном сознании. Так, труды П. М. Морхата посвящены правовой регламентации процессов жизненного цикла систем ИИ [Морхат, 2017]. Л. В. Баева, С. А. Храпов исследуют риски и последствия цифровых технологий, в том числе ИИ, в области образования [Баева, Храпов, 2022, с. 16–24]. В. Э. Карпов сосредоточил внимание на этических аспектах применения ИИ в жизни современного общества, базовых принципах взаимодействия ИИ с человеком.

Несмотря на исследования ученых отдельных проблем, возникающих в отношениях человека с ИИ, они, как правило, не рассматриваются в комплексе, в целостности, на уровне социально-философского познания, что не позволяет разработать адекватную и безопасную стратегию взаимодействия общества с искусственным разумом, а также философские принципы применения систем ИИ в современном мире.

Осмысление взаимодействия человека и техники в современном обществе стало предметом исследования в работах А. Гринфилда, В. А. Кутырева, Х. Ортега-и-Гассета, Ф. Фукуямы, М. Хайдеггера, И. В. Черниковой, Д. В. Черниковой, Б. Г. Юдина, К. Ясперса и др. Социально-философскими проблемами, вызванными применением систем искусственного интеллекта, занимались А. Гринбаум [Grinbaum, Chatila, Devillers, Ganascia, 2017, p. 139–145], М. Райан [Ryan, 2019], М. Форд (2016), Л. Флориди и др. В трудах М. Форда, например, обсуждаются проблемы роста безработицы, профессиональной поляризации, необходимости перестройки структуры управления организации в связи с внедрением технологии ИИ.

Проведенное нами исследование имело своей целью стремление доказать необходимость рассмотрения проблем искусственного интеллекта с позиций социально-философского понимания места и роли человека в мире. Избранная теоретическая позиция, основанная на принципах гуманизма, по нашему мнению, позволяет выявить и раскрыть содержание существующих философских принци-

пов применения систем ИИ и тем самым содействовать развитию концептуальных положений о безопасном, надежном и эффективном ИИ.

Полученные в ходе исследования результаты основаны на широком круге источников, включающих помимо научной и философской литературы документы, регламентирующие этико-правовые аспекты использования ИИ, созданные и опубликованные государственными структурами, международными и неправительственными организациями, представителями бизнеса как в России, так и за рубежом. Тем самым в научный оборот вводится около 100 работ (в том числе на английском и китайском языках), отражающих международный опыт применения систем ИИ, а также актуальные проблемы ИИ, стратегии и пути их преодоления.

Не менее важными являются представленные в монографии итоги работы по выявлению, классификации и систематизации основных проблем, обусловленных ИИ, полученные на основе их системного понимания, а также в отношении центрального принципа непричинения вреда, составляющего сущность и ядро концепции безопасного и надежного искусственного интеллекта.

Выводы работы могут быть использованы для разработки документов, устанавливающих внутренние этические принципы компаний, занимающихся созданием и продвижением систем ИИ, а также послужить теоретической основой дальнейшего развития существующих общественных норм и правил в сфере применения ИИ, ориентированных на широкие социальные слои и группы. Основные положения исследования могут вызвать интерес у IT-специалистов, создающих новые программные продукты, людей, которые применяют ИИ в промышленности, здравоохранении, транспорте, образовании и других сферах общественной жизни, людей, реализующих применение ИИ в малом и среднем бизнесе, государственных структур и частных компаний, институтов гражданского общества, поскольку фундаментальные философские принципы применения ИИ являются универсальными по своему характеру, содержат в себе общее руководство по применению человеком любых систем ИИ.

Выводы исследования могут также найти применение при разработке широкого спектра научно-исследовательских программ социально-экономического развития, учебных курсов, посвященных актуальным проблемам социальной философии, спецкурсов, лекций и семинаров по машинной этике, философии техники, при подготовке учебных и методических пособий по соответствующим дисциплинам. В средней школе представленные материалы могут быть использованы для создания учебных курсов в качестве источника информации и средства повышения общей осведомленности учащихся о проблемах и возможностях применения ИИ.

Глава I

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ: СУЩНОСТЬ, ОБЛАСТИ ПРИМЕНЕНИЯ, ОСОБЕННОСТИ ИССЛЕДОВАНИЯ

1.1 Определение и классификация ИИ

Развитие науки и появление новых технологий, воспроизводивших те или иные возможности человеческого интеллекта и потому обозначавшихся термином «искусственный интеллект», привело к тому, что понятие «искусственный интеллект» стало чрезвычайно размытым и проблематичным. Но точное понимание смысла и значения термина «искусственный интеллект» является необходимым условием выявления и интерпретации социально-философских проблем, непосредственно связанных с внедрением систем искусственного интеллекта во все сферы жизни социума. Рефлексия над определением соответствующего ему понятия продолжается по сей день. Это понятие представляет собой спорную научную категорию, по-разному трактуемую представителями различных специальностей. Большая часть имеющихся определений не универсальна. Эти определения зачастую не охватывают тех значений понятия, которые уже существуют в современном обществе, трактуют ИИ упрощенно, искажают его сущность или делают акцент лишь на отдельных аспектах его функционирования (например, на системах машинного обучения или на способности к автоматическому принятию решений).

Несмотря на отсутствие общепринятого определения, многочисленные дефиниции ИИ в своем содержании зачастую пересекаются, совпадают. Потому не удивительно, что Ш. Легг и М. Хаттер посвятили им отдельную работу «Коллекция определений ИИ»

[Legg, Hutter, 2007, p. 17–24]. В ней авторы рассмотрели наиболее известные и упоминаемые в научных работах определения ИИ, выявив имеющиеся в них сходства и различия.

Пионером в области создания искусственного интеллекта по праву считается Н. Винер, автор работы «Кибернетика или управление и связь в животном и машине» [Wiener, 1948]. Тогда понятия «искусственный интеллект» еще не существовало, однако ученым была высказана основная идея конструирования систем ИИ — универсальность алгоритмов решения любых, в том числе практических, задач независимо от их воплощения. Н. Винер тем самым фактически начал поиск общего алгоритма решения любых задач.

В 1950 г. А. Тьюрингом в работе «Вычислительные машины и разум» отражен разработанный им специальный тест, который, по его замыслу, должен был отличить поведение объекта, наделенного ИИ, от поведения человека [Turing, 1950, p. 433–460]. Тест можно было считать успешно пройденным, если человек-экспериментатор, задающий ИИ вопросы в письменном виде, не сможет определить, кто на самом деле отвечает: человек или интеллектуальное устройство.

Впервые термин «искусственный интеллект» был введен в научный оборот информатиком Д. Маккарти, выступившим в 1956 г. на конференции в Дартмутском колледже, поэтому именно эту дату появления ИИ официально признали. Д. Маккарти понимал под ИИ возможность «обучить машины использовать различные навыки для решения задач, подвластных на данном этапе только людям» [McCarthy, 2007]. В 1958 г. он создал язык программирования LISP, разработанный для искусственного интеллекта.

В 1960 г. создана программа GPS (General Problem Solver), которая могла решать головоломки, заниматься вычислением неопределенных интегралов, решать такие интеллектуальные задачи, как доказательство теорем, игра в шашки или шахматы [Newell, Shaw, Simon, 1959, p. 256–264]. В 1964 г. С. Маслов сформулировал идею метода автоматического поиска доказательства теорем в исчислении предикатов [Маслов, 1964, с. 17–20]. В 1966 г. В. Турчин стал

разработчиком алгоритмического метаязыка рекурсивных функций РЕФАЛ для описания языков и разнообразных способов их обработки [Турчин, 1968]. В этом же году был создан первый анимационный робот «Shakey» в Стэнфордском университете [Shakey the Robot – SRI International. URL: <https://www.sri.com/hoi/shakey-the-robot/>].

Появление новых подобных программ и развитие существующих с течением времени обусловило проблему недостаточности, предела имеющихся у ИИ знаний. В результате при развитии ИИ произошли так называемые «зимы искусственного интеллекта», которые пришлось на начало 1970-х и конец 1980-х годов, когда надежды, возлагаемые на ИИ, потерпели крах и финансирование их дальнейшего существования заметно сократилось [Альманах «Искусственный интеллект», 2020. № 6]. В промежутке между периодами затишья запустили интернет (1974 г.), а Г. Саймон получил Нобелевскую премию за теорию ограниченной рациональности, которая повлияла своим появлением на дальнейшее развитие области ИИ. После продолжительной «зимы» появляется первый персональный компьютер от IBM, человекоподобный робот от Cog, Deep Blue побеждает Г. Каспарова, появляются роботы Asimo, София и т. д. Поворотным пунктом в совершенствовании систем ИИ стало появление чат-бота GPT 4, новой генеративной языковой модели, способной обрабатывать как текстовые, так и визуальные данные [GPT 4 – OPENAI. URL: <https://openai.com/product/gpt-4>].

Обобщая широкий круг научных работ, посвященных особенностям и эволюции ИИ, следует отметить, что упомянутая многозначность толкования термина в них сохранилась. ИИ понимается как область знаний, технология, наука, искусственная система, совокупность метапроцедур представления знаний и т. д.

В понимании В. Архипова и В. Наумова искусственный интеллект — это системное явление, обладающее рядом признаков: наличие встроенного технического устройства, которое способно воспринимать и передавать данные; наличие некоторой автономности; способность анализировать, обобщать, выводить интеллектуальные

решения; свойство обучаться, самостоятельно заниматься поиском информации и принимать на ее основе соответствующие решения [Архипов, Наумов, 2017, с. 46–62].

В преамбуле Монреальской декларации искусственный интеллект характеризуется как «автономная система, способная выполнять сложные задачи, которые ранее совершались только «природным» интеллектом: обработку большого количества информации, вычисление, прогнозирование, обучение и адаптацию ответов к изменяющимся условиям, а также распознавание и классификацию объектов» [Montréal Declaration: Responsible AI. URL: https://monoskop.org/images/d/d2/Montreal_Declaration_for_a_Responsible_Development_of_Artificial_Intelligence_2018.pdf].

Европейская комиссия по искусственному интеллекту определяет его как интеллектуальные системы, которые анализируют окружающую среду и действуют с некоторой степенью автономности [European Commission. Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions on Artificial Intelligence for Europe. URL: <https://ec.europa.eu/digital-singlemarket/en/news/communication-artificial-intelligence-europe>]. Эти интеллектуальные системы могут быть представлены в виде программных продуктов, действующих в виртуальном мире (голосовые помощники, распознавание изображений, поисковые системы распознавания речи и лиц). Кроме того, ИИ может быть встроен в аппаратные устройства (передовые роботы, автономные автомобили, дроны или приложения интернета вещей).

Достаточно обширное определение ИИ приводит Л. С. Болотова. По ее мнению, ИИ можно интерпретировать как «совокупность метапроцедур — представления знаний, рассуждений, поиска релевантной информации в среде имеющихся знаний, логического вывода, пополнения знаний, их корректировки и т. д., то есть процедур, имитирующих мыслительную деятельность человека» [Болотова, 2012, с. 38–39]. Системы искусственного интеллекта — это «ап-

паратный и информационно-программный комплекс, действие которого аналогично действию механизмов человека» [Болотова, 2019, с. 29]. В данном определении утверждается, что система ИИ имеет внешнюю часть в виде аппарата и внутреннюю часть в качестве программного и информационного обеспечения. Также в определении отражено то, что системы ИИ обладают способностью к действиям и поведению, которые аналогичны человеческому. Но достигло ли развитие систем ИИ того уровня, когда их можно приравнивать человеческому интеллекту? По нашему мнению, ошибочно представление тех разработчиков, которые пытаются при программировании имитировать человеческий разум, обходя вопрос о сущности, природе этого разума.

Отдельные авторы пытаются определить ИИ посредством перечисления его наиболее существенных признаков. В работе И. В. Понкина и А. И. Редькиной среди характерных черт ИИ указываются высокая степень субстативности, автономности, способность самореферентно адаптировать собственное поведение и самообучаться, самостоятельно моделировать алгоритмы для решения проблем [Понкин, 2018, с. 91–109].

П. М. Морхат определяет ИИ как независимый комплекс программных или программно-аппаратных средств, обладающий целым рядом возможностей: способность этих систем осуществлять и демонстрировать антропоморфно-разумные мыслительные и когнитивные действия, такие как распознавание, понимание, интерпретация и генерирование образов, символьных систем и языков, рефлексия, рассуждение, моделирование, образное (смысло-порождающее и смысло-воспринимающее) мышление, обобщение, анализ и оценка информации; способность самореферентности, саморегулирования, самоограничения, самоадаптирования под изменяющиеся условия, автономного самоподдержания себя в гомеостазе; способность самостоятельно накапливать информацию и опыт; способность автономно осуществлять генетический поиск и обработку информации; способность обучаться и самообучаться (также

на своих ошибках и своём опыте); самостоятельно разрабатывать и самостоятельно применять алгоритмы самоомологации и др. [Понкин, 2018, с. 92–93].

Ассоциация развития искусственного интеллекта описывает ИИ как «научное понимание механизмов, лежащих в основе мышления и разумного поведения, и их воплощение в машинах» [AI Overview: Broad Discussions of Artificial Intelligence, AI Topics. URL: <http://aitopics.org/topic/ai-overview>]. При этом уровень интеллекта в любой конкретной реализации ИИ может значительно варьироваться, этот термин не обязательно подразумевает соответствие интеллекту человеческого уровня. ИИ включает в себя множество функций: а) глубокое обучение, трансферное обучение, обучение с подкреплением и их комбинации; б) понимание или представление глубоких знаний, необходимых для задач, специфичных для предметной области, таких как кардиология, бухгалтерский учет и юриспруденция; в) рассуждение, которое бывает нескольких разновидностей, таких как дедуктивное, индуктивное, временное, вероятностное и количественное; г) взаимодействие с людьми или другими машинами для совместного выполнения задач и обучения в окружающей среде.

С. Рассел и П. Норвиг определяют ИИ как системы, созданные людьми, которые могут выполнять сложные задачи и обрабатывать информацию так же, как мы [Russell, Norvig, 2010, p. 78–79]. Этот подход был разработан в рамках функционализма, где познание представляет собой вычислительный процесс над символическими представлениями [Brooks, 1990, p. 3–15]. Основная цель исследований состояла в том, чтобы узнать больше о естественном интеллекте путем искусственного воспроизведения этих вычислений. Соответственно, системы ИИ были задуманы как запрограммированные для выполнения конкретных действий.

Однако начиная с 1990-х годов цель исследований ИИ изменилась на создание интеллектуальных агентов, т. е. «сущностей, которые ощущают свое окружение и действуют в нем» [Russell, 2016]. В этой новой структуре ИИ не обязательно связан с алгорит-

мами для принятия решений. ИИ в этом случае выступает адаптивным процессом вычислений, взаимодействующим с окружающей средой для более эффективного принятия решений [Там же]. Другими словами, ИИ рассматривается не просто как процесс вычисления символов по определенным инструкциям, а как адаптивное и гибкое взаимодействие с окружающей средой.

В российском законодательстве, а именно указе Президента РФ от 10.09.2019 г. № 490 «О развитии искусственного интеллекта в Российской Федерации», ИИ рассматривается как комплекс технологических решений, позволяющий имитировать когнитивные функции человека (в том числе самообучение и поиск решений без заранее заданного алгоритма) и получать при выполнении конкретных задач результаты, сопоставимые как минимум с результатами интеллектуальной деятельности человека [О развитии искусственного интеллекта в Российской Федерации: указ Президента РФ от 10.10.2019 № 490 (вместе с «Национальной стратегией развития искусственного интеллекта на период до 2030 г.»). URL: <https://base.garant.ru/72838946/>]. Этот комплекс охватывает информационно-коммуникационную инфраструктуру, программное обеспечение (в т. ч. использующее методы машинного обучения), процессы и сервисы по обработке данных, поиску решений).

Уже приведенный перечень дефиниций свидетельствует о чрезвычайно сложном, комплексном характере феномена, обозначаемого термином «искусственный интеллект», широком спектре мнений ученых о его сущности и предназначении, о принципиальных разногласиях в понимании основ и принципов взаимодействия искусственного интеллекта с сознанием, разумом, интеллектом человека. Не случайно специалисты в области ИИ, как правило, указывают на многоаспектность в толковании его сущности. Так, И. А. Филиппова пишет: «Можно говорить об искусственном интеллекте как о явлении, как о группе технологий и как о научно-техническом направлении» [Филиппова, 2022, с. 9].

Рассмотрев целый ряд теоретических подходов к выявлению сущности ИИ, мы пришли к выводу, что она не может быть познана, раскрыта без понимания разнообразия ИИ, без его классификации. Подобным образом рассуждали Д. Кастро и Д. Нью, которые, также рассмотрев различные попытки определения ИИ, пришли к мысли о том, что разница подходов к пониманию ИИ сводится к различию между «слабым» и «сильным» ИИ [Castro, New, 2016, p. 3].

Действительно, большинство исследователей в области ИИ выделяют два типа ИИ: слабая версия искусственного интеллекта и сильная версия искусственного интеллекта [Шевченко, 2016, с. 7–15].

Слабый ИИ (или узкий ИИ) всеми специалистами определяется как система для выполнения конкретных, узкоспециализированных интеллектуальных задач. Например, А. Каплан и М. Хэнлайн характеризуют данный ИИ как тип аналитического интеллекта, который используется для определенных функций и приложений [Kaplan, Haenlein, 2019, p. 15–25]. При этом слабый ИИ ограничен задачами, для выполнения которых обучен [Macnish, Ryan, Stahl, 2019, p. 1–34]. Но в решении узконаправленных задач он может превосходить присущие человеку интеллектуальные способности. Все современные системы ИИ и роботы с ИИ являются примерами того, что в современной литературе принято называть слабой версией ИИ.

Что касается сильного ИИ (или общего ИИ), то под этим термином, как правило, понимается программное обеспечение, обладающее такими же когнитивными способностями, как у человека, или даже превосходящими его в этом отношении. Он предназначен для решения огромного количества задач и проблем разными способами в автономном порядке. Так, по мнению П. Ванг, Б. Гёрцель и С. Франклин, общий ИИ (далее — ОИИ) — это настоящие мыслящие машины с интеллектом, подобным человеческому [Wang, Goertzel, Franklin, 2008]. Сайт Search Enterprise AI определяет ОИИ как систему ИИ с обобщенными когнитивными способностями человека. При возникновении незнакомой задачи общий ИИ способен найти решение без вмешательства человека» [Search Enterprise AI.

Artificial intelligence. URL: <https://searchenterpriseai.techtarget.com/definition/AI-Artificial-Intelligence>].

В настоящее время общего ИИ не существует. Эту мысль в 1980 г. сформулировал Дж. Сёрль в работе «Сознание, мозг и программы». В ней, опираясь на созданный им мысленный эксперимент «Китайская комната», автор пришел к заключению, что создание общего ИИ невозможно, поскольку компьютер оперирует данными, не придавая им никакого значения и смысла, как это делает человек [Серль, 1998].

В то же время такие мыслители, как Р. Курцвейл [Kurzweil, 1990], Б. Герцель [Goertzel, 2012, p. 96–111] и Х. Де Гарис [De Garis, 2008, p. 362–373] считают, что мы вступаем в мир чрезвычайно интеллектуальных машин, в котором ОИИ будет способен непосредственно оказывать влияние на благополучие людей и сможет равным образом как помочь, так и навредить человеку. Например, представим, что ОИИ выходит из строя или достается небольшим политически мотивированным террористическим группам или крупным военным организациям. С помощью ОИИ они, например, обретут возможность шпионить, собирать конфиденциальную информацию о любых организациях, политиках и частных лицах, используя его в дальнейшем при решении собственных задач, нанося тем самым значительный ущерб мирному населению и государственным институтам. Исследователь В. М. Маслов также делает вывод, что становится все больше аргументов за то, что сильный искусственный интеллект возможен [Маслов, 2014].

К сожалению, нельзя однозначно утверждать, что эти опасения являются только плодом воображения. Сегодня мы становимся свидетелями того, что узкий ИИ уже используется странами первого мира в военных целях. Речь идет, в частности, о беспилотных летательных аппаратах (дронах) Northrop Grumman X-47B, которые проходят испытания в ВМС США [Defense Tech. Navy's second stealthy X-47B drone flies. DefenseTech. org. URL: <http://defensetech.org/2011/11/28/second-x-47buav-flies/#more-15485>]. Управ-

ление этой системой, выбор способа выполнения задачи полностью поручены не человеку, а узкому ИИ — дрону. В настоящее время неизвестно, будут ли подобные системы развиваться до уровня общего интеллекта. Тем не менее, по свидетельству Рональда К. Аркина, военные США уже проявили интерес к производству интеллектуальных систем, предназначенных для убийства [Arkin, 2008, p. 51–62].

Х. Де Гарис в работе «Война за арtilекты» задается вопросом о том, должны ли люди создавать ОИИ и чем в итоге это может для них обернуться [De Garis, 2008, p. 362–373]. Данный вопрос будет доминировать в глобальной политике XXI в., и, по мнению автора, положительный ответ на него может привести к войне, в результате которой погибнут миллиарды людей [Там же].

Р. Курцвейл, в свою очередь, утверждает, что в связи с применением ИИ люди будут постепенно превращаться в киборгов и четкая грань между машинами и людьми исчезнет. ОИИ может заменить врачей и других специалистов, выгодно отличаясь тем, что подобные системы не будут уставать, требовать длительной подготовки и при этом будут совершать меньше ошибок, чем живые люди. Как результат, в будущем во всех уголках мира доступная медицинская помощь станет нормой.

Экспоненциальный прогресс в технологиях обеспечит своеобразный интеллектуальный взрыв, в результате чего появится отборный ИИ (Seed AI). Этот ИИ, как прогнозирует Р. Курцвейл, будет способен модифицировать свою программу, создавать более разумное собственное «я». Обновленная версия ИИ будет еще лучше программироваться, что, в свою очередь, позволит ей создавать еще более умные обновления [Singularity Institute for Artificial Intelligence, Seed AI. General Intelligence and Seed AI. URL: http://singinst.org/ourresearch/publications/GISAI/paradigms/seedAI.html#glossary_crystalline].

Безусловно, развитие ИИ будет продолжаться, и результаты уже сегодня являются одним из важнейших предметов обсуждения уче-

ных во всем мире. Но в своей работе мы рассматриваем различные подходы к пониманию ИИ, учитывая существующий уровень развития его систем. Поскольку в настоящее время дебаты об отборном, универсальном, общем и т.п. интеллекте ведутся исключительно в сфере прогнозов и предсказаний, мы приходим к выводу о необходимости использовать самую общую классификацию ИИ, которая подразделяет множество существующих видов ИИ на сильный ИИ и слабый (узкий) ИИ. При этом сильный ИИ все еще остается спекулятивной темой, а объектом научного познания, по нашему мнению, сегодня может выступать лишь ИИ, созданный человеком для решения определенных практических задач, т.е. узкий ИИ.

Таким образом, узкий ИИ может быть охарактеризован как группа технологий, как результат применения человеком научного знания для решения разнообразных практических задач. При этом большинство исследователей подчеркивают системный характер узкого ИИ. В частности, И. А. Филиппова пишет: «Сегодня искусственным интеллектом признается полностью или частично автономная самоорганизующаяся система, обладающая возможностями мыслить, обучаться, самостоятельно принимать решения. ...искусственная интеллектуальная система является программно-аппаратным комплексом, она включает аппаратное и программное обеспечение. Аппаратное обеспечение охватывает все физические части компьютера или машины — носителя искусственного интеллекта, то есть электронные и механические части, входящие в состав системы. Программное обеспечение включает программы, используемые для управления машиной. Компьютерщики нередко называют эти части «хард» и «софт». Проще говоря, искусственная интеллектуальная система — это компьютер, способный выполнять функции, ранее свойственные только человеку» [Филиппова, 2022, с. 9].

Другими словами, узкий ИИ — это системы, элементами которых являются аппаратный комплекс, программное обеспечение, набор данных. Аппаратный комплекс — это физические устройства,

которые обеспечивают работу программ. Программное обеспечение представляет собой набор из компьютерного кода и связей между его частями, который функционирует благодаря аппаратному комплексу. Набор данных включает в себя входные и выходные данные, при этом последние являются результатом обработки входных данных и существуют в виде решения и действия системы ИИ. Они могут быть представлены как в виде программ, работающих в виртуальном пространстве (например, программа для распознавания изображений или текста), так и в виде аппаратных устройств с программным обеспечением (например, беспилотный автомобиль, робот-уборщик). Подчеркнем еще раз, что все современные системы ИИ и роботы с ИИ являются примерами узкого ИИ.

На системный характер узкого ИИ указывает в своем определении экспертная группа высокого уровня европейской комиссии, которая сформулировала следующую дефиницию ИИ. «Системы искусственного интеллекта (ИИ) — это программные (и, возможно, также аппаратные) системы, разработанные людьми, которые, имея сложную цель, действуют в физическом или цифровом измерении, осуществляя сбор данных об окружающей их среде, интерпретируя собранные и структурированные или неструктурированные данные в виде рассуждений, основанных на ранее полученном знании или результатах обработки вновь поступившей информации, и способные к принятию решения о наилучшем способе действия с точки зрения достижения поставленной человеком цели. Подобные системы искусственного интеллекта способны следовать заданным разработчиком правилам, сконструировать цифровую модель того или иного процесса либо явления, а также могут изменить собственное поведение, адаптируя его к изменившимся вследствие ранее совершенных действий условиям окружающей среды» [High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI. URL: <https://digitalstrategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>]. Указанную дефиницию можно найти в отдельном документе, подготовленном экспертной группой

и опубликованном под названием «Определение ИИ: основные возможности и дисциплины» [A definition of Artificial Intelligence: main capabilities and scientific disciplines. URL: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=56341].

Именно это определение, по нашему мнению, можно назвать наиболее корректным, поскольку оно содержит указание на основные свойства и способности ИИ, отличающие его от любых других предметов и явлений социальной реальности, отражая тем самым глубинную сущность этого феномена.

Отметим, что в отчете специального межправительственного комитета экспертов по искусственному интеллекту Совета Европы (САНАИ) именно это определение ИИ признано наиболее точным [САНАИ. Draft analysis of the Multi-Stakeholder Consultation. URL: https://rm.coe.int/cahai-cog2021-01-draft-report-msc-2763-7501-0051-v-1/1680_a2e2d5]. Однако данное определение нельзя считать окончательным, поскольку технические науки и технология ИИ в современном мире находятся в постоянном развитии.

Часто возникает путаница в понятийном аппарате из-за отождествления ИИ и методов, способов, алгоритмов, математических моделей, создаваемых для обеспечения его рабочих функций. Речь, в частности, идет о машинном обучении и искусственной нейронной сети. Машинное обучение (далее — МО) — это группа методов (обучение с учителем, обучение без учителя, обучение с подкреплением), которые обеспечивают анализ данных программой для принятия решений. Посредством указанных методов осуществляется решение отдельных интеллектуальных задач, таких как сбор и анализ данных, построение выводов и др.

Искусственная нейронная сеть — это математическая модель, созданная для анализа обработки данных и принятия на их основе решений при использовании методов МО. Создатели искусственной нейронной сети вдохновлялись биологическими нейронными сетями, поэтому она построена из искусственных нейронов (процессоров). Они делают «интеллектуальным» программное обеспечение

ИИ и выступают в роли методов и инструментов достижения системами ИИ желаемого результата. Сущность систем ИИ выражена в полной мере в представленном ранее определении.

Таким образом, искусственный интеллект, получивший широкое распространение в современном обществе, существует в форме его слабой (или узкой) версии, предназначенной для выполнения строго определенных, поставленных человеком задач. Его отличает системный характер, обеспечивающий целостность взаимодействующих компонентов, представленных программным и информационным обеспечением, а также аппаратными устройствами. Он способен осуществлять сбор данных, их интерпретацию в виде рассуждений, принимать решение на основании имеющейся информационной базы о возможном и наилучшем способе решения поставленной перед ним задачи.

1.2 Основные сферы применения и особенности исследования ИИ

Изучение систем ИИ, уже нашедших применение в современном обществе, неизбежно ставит вопрос об актуальных сферах их внедрения и использования, результатах, проблемах, рисках и перспективах развития указанных систем. Ученые проводят самостоятельные исследования или опираются на известные им и описанные в литературе эмпирические данные. Несмотря на изобилие информации, в настоящее время по-прежнему не сложилось единого мнения, поскольку автоматизация производственных процессов, стремительный рост информационного оборота, инвестиций в эту сферу расширяют область применения ИИ практически ежедневно. Так, в работе, подготовленной Специальным межправительственным комитетом экспертов по искусственному интеллекту Совета Европы (САНАИ), обсуждаются следующие области социальной жизни, в которых применение систем ИИ получило широкое распространение: государственное управление, транспорт, сельское хозяйство,

контроль окружающей среды и климатических изменений, здравоохранение, образование, правосудие, банковская, финансовая и страховая деятельность, военная сфера и др. [Специальный межправительственный комитет экспертов по искусственному интеллекту (CAHAI). URL: https://sk.ru/media/documents/CAHAI_AI_Research.pdf]

В статье Б. К. Шталя, Дж. Антониу, М. Райана, посвященной исследованию этических проблем применения ИИ, представлен следующий перечень отраслей жизни социума, в которых активно используются системы ИИ: контроль и администрирование сотрудников, работа правительства, сельское хозяйство, обеспечение устойчивого развития окружающей среды, наука, страхование, энергетика и коммунальные услуги, коммуникации, СМИ и развлечения, розничная и оптовая торговля, производство и добыча природных ресурсов [Stahl, Antoniou, Ryan, Macnish, Jiya, 2022, p. 23–37].

В другом источнике М. Райан, Дж. Антониу, Л. Брукс выделяют банковское дело, ценные бумаги, здравоохранение, страхование, розничную и оптовую торговлю, науку, образование, энергетику и коммунальные услуги, производство и добычу природных ресурсов, сельское хозяйство, коммуникации, СМИ и развлечения, транспорт, контроль и администрирование сотрудников, обеспечение работы правительства, правоохранительных органов и правосудия, устойчивое развитие окружающей среды, оборону и национальную безопасность [Antoniou, Ryan, Brooks, Jiya, Macnish, Stahl, 2021, p. 16].

И. А. Асеева, также обсуждая этические проблемы применения ИИ, указывает на бизнес, здравоохранение, регулирование городской среды, государственные услуги, страхование и статистику, индустрию развлечений, маркетинг [Асеева, 2022, с. 89–98].

С. И. Конев, рассматривающий этико-правовые проблемы, выделяет такие области использования ИИ, как медицина, сельское хозяйство, транспорт, юриспруденция, повседневный быт [Конев, 2020].

Проведенное нами исследование позволило, в частности, сделать вывод о том, что одной из важнейших сфер социальной жизни, в которой успешно находят применение системы ИИ, является транспорт. Здесь тестируют свои беспилотные автомобили крупные западные компании Google, General Motors, Tesla, BMW, Ford. Умное освещение дорожного покрытия, обеспеченное системами ИИ, может помимо выполнения основной функции анализировать состояние дорог, оказывать помощь в регулировании дорожного движения, предотвращая возникновение «пробок» на сложных и востребованных участках трассы. Беспилотные роботы-такси и автобусы разрабатываются и тестируются для удовлетворения нужд людей, использующих общественный транспорт. Уже доступны или разрабатываются процессоры, способные с помощью соответствующих датчиков следить за тем, что происходит рядом с автомобилем в режиме реального времени, определять местоположение автомобиля на карте и планировать маршрут, динамически адаптируясь к трафику. Программа European Truck Platooning Challenge запустила парк автономных грузовиков, которые успешно преодолели более 2000 километров в режиме взвода. Они отслеживали скорость друг друга, поддерживали оптимальное расстояние между машинами с тем, чтобы максимально эффективно пройти по заданному маршруту движения [Limer, 2016].

В сфере образования большая часть технологий ИИ пока задействована для проверки посещаемости занятий и выполнения учениками заданий, оценки и анализа экзаменационных ответов, составления персональных планов обучения. IBM разработала Teacher Advisor, что помогает учителям математики третьего класса в США подготовить персонализированные планы уроков [Harris, 2016]. Эта система позволяет учителям адаптировать учебные материалы для учащихся одного класса, но с разным уровнем навыков. В будущем планируется увеличить количество предметных областей и классов, чтобы помочь в организации и проведении занятий.

Школьный округ Такома, штат Вашингтон, совместно с Microsoft разработал модель машинного обучения для анализа данных учащихся. Система, используя данные об успеваемости, демографические и биографические сведения, составляет прогнозы о том, кто из учащихся с большой долей вероятности может бросить учебу и предлагает профилактические меры воздействия на учеников и их окружение с целью недопущения негативного результата [ML Predicts School Dropout Risk & Boosts Graduation Rates. URL: <https://blogs.technet.microsoft.com/machinelearning/2015/06/04/ml-predicts-school-dropout-risk-boostsgraduation-rates/>]. После многолетнего использования пилотной системы школьный округ Такома смог увеличить количество выпускников с 55% в 2010 г. до 78% к 2014 г. [Там же].

Технологический институт Джорджии запустил автоматизированного помощника преподавателя Джилл, работающего на платформе когнитивных вычислений IBM Watson. Джилл помогает отвечать на различные запросы студентов в отношении онлайн-курсов (например, где найти материалы курса) и получает в среднем около 10 000 сообщений от студентов в течение семестра [Maderer, 2016].

Компания Duolingo, занимающаяся разработкой программного обеспечения для освоения иностранных языков, использует МО для анализа активности пользователей и оценки прогресса в их обучении. Она разрабатывает персонализированные планы уроков, а также регулярно тестирует новые стратегии повышения их эффективности [Goliani, 2016]. Duolingo структурирует планы уроков таким образом, чтобы каждый пользователь мог узнать, как ему улучшить собственные результаты и реализовать наиболее эффективные варианты обучения иностранным языкам.

В области энергетики IBM разработала систему машинного обучения SMT (или «Самообучающаяся модель погоды и технология прогнозирования возобновляемых источников энергии»). Этот инструмент предназначен для анализа данных с 1600 метеостанций, солнечных электростанций, ветряных электростанций и метеороло-

гических спутников. На основе этих данных он генерирует прогнозы погоды на 30% точнее, чем Национальная Weather Service, и прогнозирует доступность возобновляемой энергии на несколько недель вперед [Mearian, 2015].

Окриджская национальная лаборатория Министерства энергетики США создала инструмент Autotune для суперкомпьютера Titan. Система использует МО для создания высокодетализированных моделей энергоэффективного здания [ORNL Researchers Develop ‘Autotune’ Software to Make It Quicker, Easier, and Cheaper to Model Energy Use of Buildings. URL: <http://energy.gov/eere/buildings/articles/ornl-researchers-develop-autotune-software-make-it-quicker-easier-and>]. Инструмент контролирует около 150 параметров, влияющих на энергоэффективность, таких, например, как освещение и вентиляция, и анализирует, как оптимизировать эти факторы, чтобы использование электроэнергии в здании оставалось максимально эффективным.

Google внедрила подобное программное обеспечение искусственного интеллекта, разработанное Deep Mind, дочерней компанией Alphabet, для автоматической оптимизации энергоэффективности [Clark, 2015]. Система постоянно отслеживает 120 переменных и учится, как лучше настроить производительность оборудования и системы охлаждения, чтобы обеспечить высокий уровень эффективности работы центра обработки данных. В результате использования указанной системы энергопотребление центра обработки данных снизилось на 15%.

Еще один успешный пример использования систем ИИ в сфере экологии — проект Orbital Insight, предназначенный для анализа спутниковых изображений лесов. Сравнивая снимки «до» и «после», система способна обнаруживать признаки незаконных рубок, отмечая те изменения, которые могут остаться вне поля зрения людей, отвечающих за сохранность лесного массива. Например, появление новых дорог для осуществления лесозаготовительных операций [Peters, 2015].

В мегаполисах системы, нацеленные на решение конкретных задач, также находят успешное применение. Исследователи IBM, например, спроектировали в Пекине систему машинного обучения для анализа данных об уровне загрязнения воздуха в городе. Уже сегодня система может прогнозировать изменение качества воздуха в течение ближайших трех суток на 30% точнее по сравнению с результатами, полученными традиционными способами [Knight, 2015]. Проект продолжают совершенствовать для прогнозирования изменения количества водителей на дорогах столицы, причем таким образом, что результаты будут сохранять свое значение на срок до 10 дней.

В области государственной службы и охраны правопорядка была протестирована программа Series Finder, выявляющая на основании анализа шаблона преступлений личности потенциальных преступников и прогнозирующая тем самым возможность возникновения противоправных деяний в будущем [Rudin, Sloan, 2013]. Так, для раскрытия уже совершенных краж и прогноза новых программа использовала данные отдела анализа преступности Кембриджского полицейского управления (КПУ), позволившие выявить своего рода *modus operandi* преступника, т. е. тип поведения, набор привычек, которые отличают правонарушителя. На основании полученных девяти поведенческих моделей, начав буквально с пары преступлений, Series Finder смогла восстановить большинство совершенных преступлений, зарегистрированных КПУ. Программа также выявила еще девять преступлений, укладывающихся в рамки созданных ею моделей, о которых КПУ ранее ничего не было известно.

В сфере государственной службы широкое применение получили системы распознавания изображений. Однако, как только точность распознавания лиц преодолела рубеж 94%, в обществе возникли опасения в связи с опасностью нарушения конфиденциальности, вторжения в частную жизнь и возможной дискриминации прав граждан как неизбежных последствий использования данных технологий [Snow, 2017, p. 189].

В США в сфере обеспечения общественной безопасности используется система Shot Spotter для обнаружения выстрелов из огнестрельного оружия. Система использует сетевые аудиосенсоры, разбросанные по городским кварталам, и МО для автоматического определения звуковых сигнатур выстрелов. Shot Spotter с высокой степенью точности сообщает полиции о стрельбе и районе, в котором она регистрируется [Shot in the Dark: New Surveillance Tool Called ShotSpotter Traces and Records Incidents of Gunfire. URL: <https://www.sciencedaily.com/releases/2016/04/160416130850.html>]. Алгоритмы Shot Spotter могут различать выстрелы, определять координаты места, откуда произошел выстрел, а также устанавливать, было ли задействовано несколько видов огнестрельного оружия и в каком направлении двигался стрелок [Roze, 2015].

Hitachi тестирует программное обеспечение в нескольких городах США, которое использует обработку естественного языка и МО для анализа сотен точек данных, таких как активность в социальных сетях, местоположение звонков в службу экстренной помощи, близость постов с геотегами к находящимся поблизости школам, для создания тепловых карт районов с повышенным уровнем криминальной активности [Captain, 2015]. Данное программное обеспечение умеет, помимо прочего, анализировать общедоступную геотегированную активность в социальных сетях с тем, чтобы идентифицировать фразы, соответствующие им ключевые слова, которые с большой долей вероятности являются закодированными отсылками к наркотикам, процессам их распространения, к людям, занятым в этой противоправной деятельности.

Системы ИИ сегодня используются даже в творческих процессах. Например, при создании произведений живописи или художественной литературы. Так, проект Sketch-RNN на самообучающейся нейросети может рисовать эскизы различных предметов [David, Douglas, 2017]. Sketch-RNN распознает то, что нарисовано человеком, и пытается повторить этот рисунок, но не как точную копию. Опираясь на эскиз, приложение создает свой оригинальный рису-

нок. Есть даже случаи, когда картины, написанные ИИ, было невозможно отличить от тех, которые создал человек [Ефимова, 2017]. Японский робот-писатель, в свою очередь, написал роман, который стал финалистом литературной премии имени Хоси Синъити [Clark, 2016]. Работа оказалась действительно хорошо структурированной, интересной и заслужила высокую оценку читателей.

Перечень подобных примеров использования систем ИИ может быть продолжен, поскольку область их применения в современном мире постоянно расширяется.

Обширный перечень областей применения систем ИИ, свидетельствующий о фундаментальном встраивании их в жизнь современного социума, все возрастающем влиянии на общество и людей, их привычные формы бытия, сложившиеся мировоззренческие системы, ценностные установки, сам образ жизни человека в мире, обусловил невероятный интерес к исследованию систем ИИ со стороны самых разных специалистов, организаций, социальных и профессиональных групп.

Специфической особенностью исследования ИИ стало своеобразие источниковой базы его изучения. Наряду с научными статьями, монографиями, диссертационными исследованиями проблемы разработки, внедрения и применения систем ИИ рассматриваются в документах, содержащих этические и правовые требования и принципы регулирования систем ИИ в различных областях жизни социума.

Как объяснить появление таких источников, как многочисленные этические рекомендации, кодексы, стандарты, нормативно-правовые акты, регламентирующие сегодня применение систем ИИ в мире? По нашему мнению, ответ заключается в том, что, во-первых, системы ИИ строятся по примеру интеллекта человека. Они, по сути, представляют собой искусственный человеческий разум, т.е. своего рода аналог важнейшего, основного признака человека, его родовой сущности. С древних времен человек определялся как разумное существо, наделенное сознанием, способное к целенаправленной дея-

тельности, основанной на процессах абстрактного мышления. Именно разум позволил человеку выделиться из природного мира и стать единственным существом на Земле, которое не просто приспособливается к миру, но и приспосабливает мир к удовлетворению собственных потребностей, внося в мироздание порой принципиальные изменения.

В то же время разумность человека неотъемлема от его природно-биологической составляющей. Он подвержен многочисленным слабостям, страстям, болезням. Он устает, его дееспособность может снижаться, и он смертен. Кроме того, человек живет в социуме, также накладывающем на его жизнедеятельность многочисленные и разнообразные ограничения в виде этических, правовых, политических, социальных, ценностных требований, установок и законов, что также способно оказывать сдерживающее влияние на результаты его интеллектуальной и природопреобразующей деятельности.

Когда мы говорим об искусственном интеллекте, то все перечисленные препятствия отсутствуют, поскольку ИИ — это не естественный интеллект. Ему неведомы проблемы, с которыми сталкивается любой биологический организм. Он не болеет, не устает, не ошибается и не умирает. Для него не существует преград в виде призраков человеческого рода, пещеры, рыночной площади или театра, о которых писал Ф. Бэкон. Он быстро и эффективно обучается всему новому, уже сегодня решает сложнейшие для обычного человека задачи (например, играет в шахматы на уровне гроссмейстера). Не удивительно, что его столь быстрый прогресс вызывает обеспокоенность и тревогу со стороны человека. Не станем ли мы по истечении непродолжительного отрезка времени неэффективными для систем ИИ? Не коснется ли проблема «лишних» людей всего человечества, всех представителей рода людского?

Иначе говоря, внедрение и совершенствование систем ИИ уже сегодня представляется не просто результатом эволюции, новым достижением человека. Оно вызывает трепет, беспокойство, страх, грозит стать уже в ближайшем будущем настоящей экзистенциаль-

ной угрозой человечеству, которая может означать пересмотр привычного для человека места в мире, смещение «любимейшего дитя Бога, созданного по образу и подобию Творца», на самый край, на периферию современного мира, освобождение человеком своего центрального положения в мире для быстро прогрессирующих «умных машин».

Возможно, стремясь избежать подобного финала, человечество заранее стремится не допустить столь нежелательного результата посредством создания многочисленных этических кодексов, требований, норм, руководств для разработчиков систем ИИ, для тех, кто занимается их внедрением и эксплуатацией. Человек надеется, что это поможет ему сохраниться как виду и защитить привычное для него место в мире.

Ответом на обсуждение социально-философских проблем внедрения ИИ стали попытки установления неких правил, принципов, регулирующих его использование. С этой целью различные научно-исследовательские, политические, профессиональные организации, комитеты национального уровня, неправительственные и частные коммерческие компании разработали многочисленные законодательные акты, этические руководства и стандарты, решения этических комитетов и множество других аналогичных по смыслу документов.

Рассматривая эту группу источников, мы включили в первую очередь манифесты крупных политических образований (ЕС) и профессиональных организаций (IEEE, Google), национальные стратегии государств, отдельные законодательные акты. Указанные документы, разработанные на основе научных исследований, носят предписывающий характер, созданы для специалистов, практически работающих с системами ИИ, и лиц, принимающих решения и управляющих процессами создания и применения ИИ. Следовательно, их потенциальное влияние на дальнейшее развитие систем ИИ может быть чрезвычайно высоким. Источником рассмотренных документов являются правительственные структуры отдельных государств (например, Специальный комитет лордов Великобрита-

нии, Экспертная группа высокого уровня европейской комиссии), авторитетные в научном сообществе организации (например, Future of Life Institute, IEEE, AI4People), компании, представляющие крупный международный бизнес (например, Google, IBM, Microsoft, Intel).

Отдельным предметом изучения стали нормативно-правовые акты, этические стандарты и руководства по регулированию систем ИИ, резолюции, международные декларации по внедрению этических систем ИИ, периодические издания.

Нормативно-правовые акты представлены международно-правовыми актами, законодательными актами отдельных государств, среди которых, например, «Закон о дорожном движении Германии» [URL:<https://www.bundesregierung.de/breg-de/suche/automatisiertes-fahren-auf-dem-weg-326108>], проект Федерального закона «О внесении изменений в Гражданский кодекс Российской Федерации в части совершенствования правового регулирования отношений в области робототехники» [URL:<https://www.dentons.com/ru/insights/alerts/2017/january/27/dentons-develops-first-robotics-draft-law-in-russia>], «Закон об автоматизированных и электрических транспортных средствах» [Automated and Electric Vehicles Act. URL: <https://www.legislation.gov.uk/ukpga/2018/18/contents/enacted>].

Самостоятельную группу документов, имеющих регламентирующий характер, составляют Резолюция о запрете применения автономных смертельных систем вооружения [Résolution visant à interdire l'utilisation, par la Défense belge, de robots tueurs et de drones armés // Chambre des représentants de Belgique. URL: <https://www.lachambre.be/FLWB/pdf/54/3203/54K3203005.pdf>] (Бельгия, 2018), Рекомендация по беспилотным автомобилям от Комиссии по этике при Министерстве транспорта и цифровой инфраструктуры Германии [Ethics Commission Automated and Connected Driving. URL: <https://bmdv.bund.de/SharedDocs/EN/publications/report-ethics-commission.pdf?blob=publicationFile>], Доклад об этике робототехники от ЮНЕСКО [Report of COMEST on robotics ethics. URL: <https://unesdoc.unesco.org/ark:/48223/pf0000253952>].

Появление такого рода источников — это ответ человеческого сообщества на еще одно важнейшее свойство систем искусственного интеллекта, отличающее их от всех ранее созданных человеком технологий: их способность действовать автономно.

Системы ИИ уже сегодня могут выполнять сложные интеллектуальные задачи (например, управление автомобилем) без активного вовлечения человека. Создание более универсальных продуктов ИИ, основанных на машинном обучении, со временем неизбежно приведет к появлению еще более автономных, а значит, еще более не подконтрольных человеку систем. Эта особенность превращает системы ИИ в потенциальный источник экзистенциальной угрозы для человечества в целом, которая будет постоянно нарастать вследствие того, что системы ИИ в перспективе смогут совершенствовать собственное программирование и, как результат, стать обладателями когнитивных способностей, превышающих познавательные возможности человека.

Безусловно, этот негативный футуристический сценарий больше похож сегодня на сюжет фантастического рассказа. Однако и без такого рода прогнозов очевидно, что технологии ИИ требуют тщательного контроля и регулирования уже сейчас.

Поэтому не удивительно, что возможности ИИ и проблемы, создаваемые им, привели к появлению целого ряда нормативных и регламентирующих документов, этических руководств, стандартов, законодательных актов, деклараций и т.д. в самых разных отраслях и сферах общественной жизни. Мировое сообщество в лице правительств, международных организаций, научно-исследовательских институтов, бизнес-корпораций отреагировало таким образом на проблемы и риски, порождаемые ИИ, сформулировало свой ответ на вызов, брошенный ИИ глобальному человечеству.

Общее количество публикаций, посвященных регулированию проблем применения систем ИИ, постоянно увеличивается. При этом большая часть существующих в настоящее время документов издана после 2016 г. В. Дигнум приводит результаты исследования,

свидетельствующего о том, что в настоящее время выпущено уже около 600 политических рекомендаций, руководств или стратегических отчетов, связанных с этическим регулированием систем ИИ [Dignum, 2022].

Таким образом, нами изучен обширный свод документов по этическому и правовому регулированию систем ИИ, существующих в мировой практике, из них 90, на наш взгляд, наиболее значимы и содержательны. Перечень документальных источников, отобранных с учетом критериев надежности, новизны и разнообразия, представлен в Приложении. Большинство документов было подготовлено частными компаниями и государственными учреждениями, за ними по мере убывания следуют академические организации, научно-исследовательские институты, межгосударственные или наднациональные организации, профессиональные ассоциации и т. д.

Попытки провести аналогичную работу предпринимались ранее другими исследователями. Отметим обзор Хагендорфа в *Minds and Machines*, им изучено 22 руководящих принципа [Hagendorff, 2020, p. 99–120], и публикацию в *Nature*, авторы А. Джобин, М. Йенка и Е. Вайена исследовали 84 источника руководящих этических принципов ИИ [Jobin, Lenca, Vayena, 2019, p. 389–399].

По нашему мнению, изучение принятых и действующих в мире документов, регламентирующих, упорядочивающих применение систем ИИ, необходимо для понимания того, что в действительности происходит в сфере внедрения и применения ИИ, какие риски, опасности, трудности и препятствия возникают в процессе взаимодействия человека с ИИ, каковы перспективы развития систем ИИ в ближайшем и отдаленном будущем.

1.3 Этическое и правовое регулирование ИИ

Аналитический обзор документов, в которых содержатся этические и правовые требования к регулированию ИИ, начнем с национальных стратегий, которые выступают в роли документов программного характера, задают вектор будущего развития, способствуют фор-

мированию основ комплексного правового регулирования общественных отношений в рассматриваемой сфере социальной действительности. В большинстве современные программные документы содержат положения о необходимости этического аспекта при развитии индустрии ИИ. Речь, в частности, идет о стратегиях, принятых в целом ряде стран, среди которых Россия, Китай, Франция, Финляндия, Швеция и др. Отметим, что в Южной Америке и Африке в настоящее время отсутствуют национальные документы рассматриваемой тематики, но они взаимодействуют по этому поводу с Организацией экономического сотрудничества и развития (ОЭСР).

Российская национальная стратегия искусственного интеллекта представлена в указе Президента РФ от 10 октября 2019 г. № 490 «О развитии искусственного интеллекта в Российской Федерации» [URL: <https://base.garant.ru/72838946/>]. В стратегии установлены основные этические принципы развития и использования технологий ИИ, основные задачи развития ИИ, прописаны действия государственной власти для воплощения этих задач. В стратегии утверждается, что нормативно-правовую базу необходимо создавать поэтапно. Эта база должна быть способна обеспечить формирование и функционирование комплексной системы регулирования общественных отношений в сфере развития и использования технологий ИИ. В стратегии также утверждается необходимость соблюдения основных этических принципов развития и применения ИИ, в частности, принципов безопасности, защиты прав и свобод человека, прозрачности, технологического суверенитета, целостности инновационного цикла, поддержки конкуренции и разумной бережливости.

Американская национальная стратегия развития ИИ «О сохранении американского лидерства в области искусственного интеллекта» начала действовать 11 февраля 2019 г. [Executive Order on Maintaining American Leadership in Artificial Intelligence. URL: <https://www.whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/>]. Стратегия содержит пять основных принципов развития ИИ: внедрение технологи-

ческих прорывов; просвещение работников в области разработки и применения ИИ; соблюдение прав и свобод личности; укрепление доверия к системам ИИ. В стратегии утверждается, что формирование адекватной правовой базы напрямую зависит от уровня просвещенности должностных лиц, принимающих решения в области ИИ.

Национальная стратегия КНР 2017 г. «План развития искусственного интеллекта нового поколения» устанавливает этические принципы применения ИИ [State Council, Notice of Issuing New Generation Artificial Intelligence Development Plan. URL: http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm]. Развитие ИИ является национальным приоритетом КНР. Указанная стратегия КНР преследует цель привести страну к мировому лидерству в области ИИ к 2030 г. Она направлена на содействие научным исследованиям, просвещение специалистов в этой области, разработку стандартов, этических руководств, на решение вопросов безопасности. В стратегии определены основные этапы внедрения ИИ: удержание конкурентноспособной позиции в области ИИ на мировом рынке до 2020 г.; активное использование ИИ на практике во всех секторах экономики до 2025 г.; обеспечение мирового лидерства в области ИИ до 2030 г. Документ направлен на установление тесного взаимодействия между органами власти, частным сектором и академическими кругами. Разработке этических принципов Китай не уделяет первостепенного внимания, однако в стратегии тем не менее они упоминаются.

В национальной стратегии Германии 2018 г. утверждается, что внедрение систем ИИ должно сопровождаться соблюдением «европейских ценностей, таких как неприкосновенность человеческого достоинства, уважение частной жизни и принцип равенства» [Bundesregierung, Eckpunkte der Bundesregierung für eine Strategie Künstliche Intelligenz. URL: https://www.bmbf.de/files/180718_Eckpunkte_KI-StrategiefinalLayout.pdf]. В Стратегии выделяется 12 сфер применения ИИ.

Национальная стратегия развития искусственного интеллекта Швеции ставит своей приоритетной целью лидерство в развитии новых цифровых технологий, в том числе ИИ [Government offices of Sweden, National Approach to artificial intelligence. URL: <https://www.regeringen.se/4aa638/contentassets/a6488cceb6f418e9ada18bae40bb71f/nationalapproach-to-artificial-intelligence.pdf>]. Под лидерством следует понимать создание оптимальных условий для разработки и внедрения ИИ с целью процветания и благосостояния страны.

В национальной стратегии Дании отражено определение ИИ, обозначены важные сферы (здравоохранение, энергетика и промышленность, сельское хозяйство, транспорт) [Ministry of Finance and Ministry of Industry, Business and Financial Affairs. National Strategy for Artificial Intelligence. URL: https://eng.em.dk/media/13081/305755-gb-version_4k.pdf]. В качестве одной из приоритетных задач в стратегии обозначено установление принципов ответственной разработки и использования ИИ.

Канада в настоящее время является одним из лидеров в области развития и внедрения искусственного интеллекта. При этом основное внимание уделено научным исследованиям, а не созданию нормативной базы управления [Hirsh, 2018]. Панканадская стратегия ИИ [Pan-Canadian Artificial Intelligence Strategy, Invest in Canada. URL: <https://www.investcanada.ca/why-invest/pancanadian-artificial-intelligence-strategy>] издана канадским Институтом перспективных исследований (CIFAR) [CIFAR is a Canadian-based global research organization. URL: <https://cifar.ca/>], который частично финансируется государством. Если исходить из стратегии, то Канада стремится увеличить число исследований в области ИИ и создать взаимосвязанные и взаимодействующие друг с другом научные центры. Для того чтобы добиться намеченных планов, Канада начала реализовывать программу AI&Society.

Эстонская национальная стратегия развития ИИ стремится придать статус правосубъектности практически всем алгоритмам, которые связаны с ИИ [Estonia accelerates artificial intelligence

development. URL: <https://e-estonia.com/estonia-accelerates-artificial-intelligence/>. Однако из содержания документа не ясно, по каким критериям может быть присвоен статус «цифрового субъекта»; что затрудняет развитие систем ИИ, особенно учитывая новые технологии, имеющие полиморфный и непостоянный характер. В стратегии сосредоточивается внимание на роли ИИ в государственном управлении, а также на правовом регулировании процессов применения ИИ в отдельных секторах, среди которых разработка автоматизированных транспортных средств, роботов-доставщиков и т. д.

Национальная стратегия Франции [Villani, 2018] была представлена на конференции «Искусственный интеллект для человечества» (2018 г., Париж) президентом Франции Э. Макроном. В стратегии уделяется большое внимание этическим принципам, связанным с развитием технологий ИИ, в частности, принципу прозрачности технологий ИИ. С. Виллани, автор национальной стратегии и депутат парламента Франции, полагает, что принцип прозрачности должен быть поддержан на государственном уровне. Соблюдение данного принципа поможет, по его мнению, разработать более доступные для понимания системы ИИ. Стратегия содержит анализ актуальных проблем, которые касаются новых рисков, связанных с внедрением ИИ в производство. Также в документе утверждается, что ИИ должен быть изначально разработан с этическими нормами в целях предотвращения социально-экономических последствий его применения. До 2022 г. Франция определила следующие приоритетные направления деятельности в рассматриваемой сфере: усиление позиций Франции в ЕС; формирование политики открытых данных; адаптация национальных нормативных норм к европейским; определение основных этических проблем, связанных с разработкой и внедрением ИИ.

В национальной стратегии Великобритании для достижения лидерства в области развития ИИ рекомендовано создание Совета по искусственному интеллекту, который будет целенаправленно руководить развитием ИИ, регулировать государственную политику в

этой сфере, а также содействовать внедрению ИИ в промышленность [Select Committee on Artificial Intelligence. AI in the UK: ready, willing and able. House of Lords, UK. URL: [https://www.scirp.org/\(S\(351jmbntvnsjt1aadkozje\)\)/reference/referencespapers.aspx?referenceid=2884461](https://www.scirp.org/(S(351jmbntvnsjt1aadkozje))/reference/referencespapers.aspx?referenceid=2884461)]. Важнейшими задачами в указанном документе являются поддержка экспорта и привлечение иностранных инвестиций в область ИИ.

Японский стратегический совет по технологиям искусственного интеллекта представил стратегию развития технологий ИИ, в которой разработаны дорожные карты индустриализации трех ведущих областей жизни общества — медицина, социальное обеспечение, социальная мобильность [Strategic Council for AI technology. URL: <http://www.nedo.go.jp/content/100865202.pdf>]. В стратегии устанавливаются этапы развития и основные приоритеты государства в области ИИ. Обозначены семь принципов ИИ для его этичного применения: принцип антропоцентричности, принцип просвещения (грамотности в области ИИ), принцип конфиденциальности, принцип безопасности, принцип честной конкуренции, принцип справедливости, подотчетности и прозрачности, принцип инноваций. В 2019 г. было издано два документа: «Социальные принципы искусственного интеллекта, ориентированные на человека» [Social Principles of Human-Centric AI. URL: <https://www.cas.go.jp/jp/seisaku/jinkouchinou/pdf/humancentricai.pdf>], «Стратегия искусственного интеллекта — 2019» (AI Strategy 2019. AI for everyone: People, Industries, Regions and Governments) [AI Strategy 2019. AI for Everyone: People, Industries, Regions and Governments. URL: <https://www8.cao.go.jp/cstp/english/humancentricai.pdf>].

Сингапурская национальная стратегия вышла в свет в 2019 г. [National Artificial Intelligence Strategy. URL: <https://www.smartnation.gov.sg/why-Smart-Nation/NationalAIStrategy>]. Она, как и многие подобные документы, определяет своей целью достижение мирового лидерства в области развития ИИ. В стратегии выделяются приоритетные области внедрения ИИ: логистика, медицина, образование,

туризм и др. Особое внимание уделяется принципу доверия ИИ, защите прав и интересов граждан. Национальная стратегия Сингапура тесно взаимосвязана с Консультативным советом по этическому использованию ИИ и данных (Advisory Council on the Ethical Use of AI and Data) с целью разработки необходимых этических кодексов в области ИИ. В Сингапуре также создана Рамочная модель управления искусственным интеллектом, в которой представлены принципы понятности, прозрачности и справедливости при внедрении и использовании ИИ [Model Artificial Intelligence Governance Framework (Second Edition). 2020. Infocomm Media Development Authority Singapore. URL: <https://www.imda.gov.sg/-/media/Imda/Files/Infocomm-Media-Landscape/SG-Digital/Tech-Pillars/Artificial-Intelligence/Primer-for-second-edition-of-the-Model-Framework.pdf?la=en>]. В данном документе отмечается, что добиться идеальной объяснимости, прозрачности и справедливости трудно, однако необходимо стремиться к соблюдению указанных требований. Также здесь утверждается, что решения ИИ должны быть антропоцентричными или ориентированными на человека. К 2025 г. Сингапур поставил задачу достичь полной автоматизации иммиграционного контроля, а также повсеместно распространить датчики для систем ИИ.

Национальная стратегия Республики Корея «Среднесрочный и долгосрочный генеральный план подготовки к интеллектуальному информационному сообществу: управление Четвертой промышленной революцией» была издана еще в 2016 г. [Mid-to Long-term Master Plan in Preparation for the Intelligent Information Society: Managing the Fourth Industrial Revolution. URL: https://english.msit.go.kr/cms/english/pl/policies2/_icsFiles/afieldfile/2017/07/20/Master%20Plan%20for%20the%20intelligent%20information%20society.pdf].

В этой стратегии искусственный интеллект был рассмотрен наряду с такими технологиями, как «интернет вещей», анализ больших данных и т. д. Также в документе утверждалось, что необходимо принять рамочное законодательство, устанавливающее необходимые этические и правовые нормы в сфере ИИ. В 2019 г. была

утверждена новая национальная стратегия на ближайшие 10 лет, в которой позиция по отношению к разработке и внедрению ИИ существенно пересмотрена [Eun-jin, 2021]. Она сводится к тому, что прекращается жесткий контроль научно-исследовательских проектов в области ИИ, поскольку такой чрезмерный надзор тормозит развитие ИИ. Соответственно, было решено увеличить финансирование в этой области и дать больше свободы авторам опытно-конструкторских работ. Также в стратегии утверждается, что военнослужащие и государственные работники должны быть просвещены в области ИИ и иметь соответствующее образование. Кроме того, планируется ввести в начальных и средних школах дисциплины по искусственному интеллекту.

Помимо принятия стратегий развития в названной сфере государственным сектором опубликован ряд документов, посвященных выделению базовых принципов этичного применения ИИ. Например, правительство Канады разработало директиву об автоматизированном принятии решений, в которой особое внимание уделяется справедливости и прозрачности при использовании искусственного интеллекта государственными органами власти [Government of Canada, “Directive on Automated Decision-Making”, 2019. URL: <https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592>]. Однако данный документ не имеет никакой юридической силы в отношениях государства и частного бизнеса [Government of Canada. Responsible use of artificial intelligence (AI). URL: <https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai.html>].

Австралия опубликовала концепцию этики ИИ, в которой на основе результатов соответствующих научных исследований предлагается набор инструментов для реализации на практике этичного ИИ [Dawson et al., 2019]. Счетная палата США в 2021 г. издала рекомендации о системе ответственности государственных структур, применяющих в своей деятельности технологию ИИ [Artificial Intelligence. An Accountability Framework for Federal Agencies and

Other Entities.GAO, 2021. URL: <https://www.gao.gov/assets/gao-21-519sp.pdf>].

Изучая географию документов, устанавливающих этические принципы применения систем ИИ, можно сделать вывод о безусловном лидерстве в этом вопросе экономически развитых стран. Так, на США и Великобританию приходится более трети всех документов по этическому регулированию систем ИИ, за ними следуют Япония, Германия, Франция и Финляндия. В Азии Китай находится на первом месте, за ним — Япония, Южная Корея и Сингапур. В Океании лидирует Австралия, за которой следует Новая Зеландия. В африканских странах Южная Африка первой установила принципы этичного применения ИИ через Research ICT Africa (африканский аналитический центр, специально созданный для устранения стратегического пробела, имеющего место в развитии устойчивого информационного общества и цифровой экономики) [Research ICT Africa (RIA). URL: [https:// researchictafrica.net/people/](https://researchictafrica.net/people/)]. Другие же решают похожие проблемы через институт членства в ОЭСР.

Аналогичного заключения придерживается в своем исследовании Л. Тиджон, который сформулировал и предложил несколько критериев, свидетельствующих, по его мнению, о готовности государств к внедрению ИИ [The Different Faces of AI Ethics Across the World: A Principle-Implementation Gap Analysis Lionel Nganyewou Tidjon and Foutse Khomh, Senior Member, IEEE. URL: <https://arxiv.org/pdf/2206.03225.pdf>]. Среди них, в частности, способность правительственных структур быстро адаптироваться к инновациям и перейти к их практическому внедрению в социальные процессы. Также имеет значение оценка самого процесса внедрения инструментов ИИ, а именно: обучены ли они на высококачественных и репрезентативных данных, имеется ли соответствующая инфраструктура для предоставления решений ИИ и содействия их внедрению. В итоге рейтинг готовности к внедрению ИИ двадцати рассмотренных автором государств в 2021 г. выглядит следующим образом: США, Сингапур, Великобритания, Финляндия, Нидерланды.

ды, Швеция, Канада, Германия, Дания, Республика Корея, Франция, Япония, Норвегия, Австралия, Китай, Люксембург, Ирландия, Тайвань, ОАЭ и Израиль. Кроме того, по мнению Л. Тиджона, США, Великобритания и Германия являются наиболее благоприятными государственными образованиями для развития ИИ, инвестиций, бизнеса и исследований, поскольку они лучше других подготовлены с точки зрения уровня технологий ИИ, управления ИИ, а также имеющихся баз данных и инфраструктуры.

Анализ отобранных нами документов по целевой аудитории, к которой они обращены, позволил сделать вывод о том, что большинство из них не имеет универсального характера, но адресовано сразу нескольким заинтересованным сторонам (например, профессиональным сообществам, государственному сектору, частным компаниям, разработчикам, исследователям и т. д.). Есть примеры документов, принятых на национальном уровне и нацеленных на решение проблем в некоторых областях применения систем ИИ. Так, Австралия в 2019 г. сформулировала девять этических принципов ИИ в медицине [Ethical Principles for AI in Medicine. The Royal Australian and New Zealand College of Radiologists. URL: <https://www.ranzcr.com/documents/4952-ethical-principles-for-ai-in-medicine/file>], а Германия в 2017 г. разработала этические принципы для эксплуатации беспилотного транспорта [Ethics Commission. Automated and Connected Driving. URL: https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission-automated-and-connecteddriving.pdf?__blob=publicationFile].

Обращаясь к юридическим документам, существующим в настоящее время в сфере регулирования ИИ, рассмотрим опыт Российской Федерации. В нашей стране распоряжением правительства от 19 августа 2020 г. № 2129-р утверждена Концепция развития регулирования отношений в сфере технологий искусственного интеллекта и робототехники до 2024 г. [СЗ РФ. 2020. № 35. Ст. 5593]. В данной концепции определен комплекс правовых перспектив в указанной области, а также основные направления работы в рамках

создания многоаспектной правовой платформы в сфере робототехники и ИИ.

В контексте обсуждаемого вопроса интерес также вызывает Федеральный закон от 24 апреля 2020 г. № 123-ФЗ «О проведении эксперимента по установлению специального регулирования в целях создания необходимых условий для разработки и внедрения технологий искусственного интеллекта...» [URL: <https://base.garant.ru/400797482/>]. Основная цель документа — обеспечение высшим исполнительным органом власти Москвы необходимых условий для разработки, внедрения, реализации и оборота технологий ИИ. Благодаря ему в Москве был дан старт таким проектам, как создание и эксплуатация беспилотных автомобилей на дорогах общего пользования, использование роботов-курьеров службами сервиса и т. д.

21 мая 2021 г. было издано постановление правительства РФ № 767, согласно которому утверждены правила получения господдержки при реализации пилотных проектов по внедрению искусственного интеллекта в различных отраслях экономики [URL: <https://base.garant.ru/400797482/>]. Это постановление принято для реализации национальной стратегии России в сфере развития искусственного интеллекта. С этой же целью сегодня издаются дорожные карты развития технологий ИИ, которые содержат технологические задачи, сроки и этапы разработки продуктов в конкретной сфере, но эти документы уже не являются актами правового регулирования.

Необходимо упомянуть еще один закон «Об экспериментальных правовых режимах в сфере цифровых инноваций в Российской Федерации» [Законопроект № 922869-7. 2020. URL: <https://sozd.duma.gov.ru/bill/922869-7>]. Согласно этому документу, в области нормативного регулирования ИИ поставлены следующие задачи: обеспечение нормативных условий для доступа к данным; разработка более упрощенного административно-правового и нормативно-технического порядка тестирования и внедрения разработок в области ИИ и т. д.

В США активно происходит обсуждение различных вопросов правового регулирования ИИ. Так, уже принятый «Закон о защите конфиденциальности» (2016 г.) регламентирует передачу персональных данных между ЕС и США [Eu-U.S and swiss-U.S. privacy shield, 2016. URL: <https://www.privacyshield.gov/program-overview>]. В последние годы обсуждалось несколько законопроектов в этой сфере, среди которых проект закона о будущем ИИ (Future of AI Act, 2017); проект закона о Комиссии по национальной безопасности в области ИИ, одобренный Конгрессом США (National Security Commission Artificial Intelligence Act, 2018); проект закона о применении ИИ в государственном секторе (AI in Government Act, 2018), проект «Билля о правах ИИ» (The Blueprint for an AI Bill of Rights, 2022) и др. В мае 2025 г. Палата представителей одобрила инициативу, запрещающую местным органам власти регулировать искусственный интеллект и автоматизированные системы принятия решений на ближайшие 10 лет, что в будущем, по мнению авторов проекта, позволит выработать единый федеральный подход к решению проблем применения ИИ и «...позволит американским компаниям не проиграть в глобальной гонке технологий» [Палата представителей США приняла законопроект, который запретит регулирование ИИ на 10 лет. URL: <https://habr.com/ru/companies/bothub/news/912366/>].

В Канаде законодательное регулирование ИИ представлено «Законом о защите личной информации и электронных документов» (2000 г.) [Personal information protection and electronic documents act, 2000. URL: <https://lawslois.justice.gc.ca/ENG/ACTS/P-8.6/index.html>]. Во Франции принят закон «О цифровой республике» [Loi pour une Republique numerique. 2016. URL: <https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000033202746&categorieLien=id>]. В январе 2020 г. Национальное собрание Франции внесло предложение об издании Хартии искусственного интеллекта и алгоритмов [Proposition de loi constitutionnelle relative à la Charte de l'intelligence artificielle et des algorithmes. URL: <http://www.assemblee-nationale>].

fr/dyn/15/textes/l15b2585_proposition-loi]. Это предложение рассматривалось парламентским комитетом по конституционному законодательству. Авторы законопроекта включили в его содержание ряд таких вопросов, так регулярный аудит систем ИИ, а также предотвращение злонамеренных манипуляций с ИИ. В Великобритании, ОАЭ, Германии существуют законы о защите данных [Data protection act. 2018. URL: <https://www.hhs.gov/hipaa/for-professionals/index.html>; Data protection law. 2021. URL: <https://u.ae/en/about-the-uae/digital-uae/data/dataprotection-laws>; Bundesdatenschutzgesetz. 2009. URL: <https://www.datenschutz-wiki.de/BDSG>], Европейский Союз принял «Общее положение о защите данных» [General data protection regulation, 2016. URL: <https://gdpr-info.eu/>], Китай — Закон о защите личной информации [Personal information protection law, 2021. URL: <https://www.china-briefing.com/news/the-prc-personalinformation-protection-law-final-a-full-translation/>], Закон о кибербезопасности [Cyber security law, 2016. URL: <https://www.tradecommissioner.gc.ca/china-chine/cybersecuritycybersecuritechina-chine.aspx?lang=eng>], Сингапур — Закон о защите от лжи и манипуляций в интернете [Protection from online falsehoods and manipulation act, 2019. URL: <https://sso.agc.gov.sg/Acts-Supp/18-2019>].

Б. П. Немиц в своем исследовании «Конституционная демократия и технология в эпоху ИИ» размышляет о том, какие аспекты могут регулироваться законом, а какие — этическими кодексами [Nemitz, 2018, p. 1–14]. Важнейшей проблемой в области регулирования ИИ является ответственность за действия систем ИИ. В этом отношении некоторые исследователи придерживаются позиции, что правовое регулирование становится необходимостью, в особенности когда вопрос касается ответственности за действия систем ИИ и причиненный ими вред. Другим аспектом проблемы ответственности является вопрос о том, кто и как будет нести ответственность за вредоносное использование ИИ. В настоящее время речь в большинстве случаев идет о привлечении к ответственности за подобное использование ИИ либо разработчика, либо исполнителя [Аверин-

ская, Севостьянова, 2019, с. 94–96]. При этом ни в одном государстве мира не существует нормативно-правового регулирования, однозначно определяющего степень ответственности каждого из участников процессов взаимодействия общества с системами ИИ. Определенные усилия в сфере регулирования ИИ предпринимаются бизнес-корпорациями, особенно теми из них, которые в своей деятельности активно используют системы ИИ.

В 2018 г. компании Google и SAP опубликовали руководящие принципы в сфере ИИ. Google сформулировала семь этических принципов, которыми необходимо руководствоваться при разработке и внедрении систем ИИ: полезность; справедливость; безопасность; ответственность; конфиденциальность; поддержание стандартов высокого мастерства; ограничение применения вредоносных программ. Google также создала специальный инструмент, который позволяет влиять на работу тех инструментов, которые порождают несправедливые или дискриминационные решения для людей [Visually probe the behavior of trained machine learning models, with minimal coding. URL: <https://pair-code.github.io/what-if-tool/>].

«Руководящий комитет SAP по этике ИИ» состоит из руководителей высшего звена всех подразделений организации. SAP гарантирует, что ее программное обеспечение разработано в соответствии с принципами этики. Также в компании создали внешнюю «консультативную группу по этике ИИ», в которую вошли «эксперты из академических кругов, политики и бизнеса, чья специализация находится на стыке этики и технологий — в частности, технологии ИИ» [Там же]. SAP выступает за разработку глобального кодекса поведения в отношении этических методов ведения бизнеса с использованием ИИ, которым могли бы руководствоваться любые компании [SAP. European Prosperity Through Human-Centric Artificial Intelligence; The Intelligent Enterprise: 2018, p. 28. URL: www.sap.com/documents/2018/01/3e67a134-ee7c-0010-82c7-eda71af511fa.html].

Этический комитет, созданный Microsoft, называется AETHER (AI Ethics in Engineering and Research). Он состоит из экспертов по

прикладной этике, инженеров и представителей, назначенных руководителями основных подразделений» [Putting principles into practice at Microsoft. URL: <https://www.microsoft.com/en-us/ai/our-approach?activetab=pivot1:primaryr5>]. Комитет дает советы и разрабатывает рекомендации по инновациям в области ИИ для всех подразделений Microsoft. В частности, были созданы рабочие группы, занимающиеся такими проблемами, как «предвзятость и справедливость ИИ» и «прозрачность ИИ». Microsoft была первой крупной компанией, которая публично приняла государственное регулирование в отношении ИИ, и одной из первых крупных компаний, выступивших за новое законодательство в области ИИ [Microsoft. The future computed: Artificial Intelligence and its role in society: chapter 2. URL: news.microsoft.com/cloudforgood/_media/downloads/the-future-computed-english.pdf].

Amazon Web Services долгое время пыталась оставаться в стороне от дискуссий об этичности ИИ. По словам исполнительного директора AWS Питера Стански, «клиенты сами должны решить, этично ли использовать инструменты AWS» [AWS is ethical about AI but 'we just don't talk about it' say APAC execs. URL: <http://cdn.computerworld.com.au/article/661203/aws-ethical-about-ai-we-just-don-t-talk-about-it-say-apac-execs/>]. Однако несколько спорных вопросов, связанных с ИИ, вынудили платформу изменить свою позицию. Сначала компании пришлось отказаться от использования созданного ею на основе технологии ИИ инструмента рекрутинга, поскольку он демонстрировал предвзятое отношение к претендентам на ту или иную позицию по расовому и половому признакам. В дальнейшем еще один инструмент, предназначенный для распознавания лиц (Rekognition), также подвергся критике по указанной причине. Ведущие исследователи ИИ, представляющие промышленные предприятия и научное сообщество, осудили этот инструмент за гендерные и расовые предубеждения, а сотрудники и акционеры AWS призвали компанию прекратить продажу данного продукта правоохранительным органам. В конце концов после двух

лет осуждений Amazon неохотно согласился с критикой и приостановил продажи [We are implementing a one-year moratorium on police use of Rekognition. URL: <https://blog.aboutamazon.com/policy/we-are-implementing-a-one-year-moratorium-on-police-use-of-rekognition/>].

Впоследствии AWS предложила рекомендации для будущего законодательного регулирования программных продуктов с целью распознавания лиц для правоохранительных органов, в числе которых обязательная проверка результатов работы систем ИИ человеком, уровень достоверности не ниже 99 %, обязательные отчеты о прозрачности, публичное уведомление всякий раз, когда видеонаблюдение и распознавание лиц используются совместно) [Some Thoughts on Facial Recognition Legislation. URL: <https://aws.amazon.com/ru/blogs/machine-learning/some-thoughts-on-facial-recognition-legislation/>]. Таким образом, AWS поддержала идею законодательного регулирования приложения ИИ.

IBM также была вовлечена в дебаты о необходимости регулирования ИИ из-за проблем с распознаванием лиц. Свой подход к государственному регулированию ИИ компания обозначила как «точное регулирование», что означает требование контролировать только определенные, конкретные приложения и всякий раз выяснять, где и по какой причине в процессе использования продукта возникли негативные для общества последствия. Google была последней компанией, присоединившейся к группе тех, кто требовал нового законодательного регулирования инструментов ИИ. В их отчете об управлении ИИ отмечается, что они «с нетерпением ожидают взаимодействия с правительствами, отраслевыми специалистами и гражданским обществом по вопросам, связанным с ИИ» — например, посвященным обсуждению новых видов вооружения или полицейского наблюдения [Google. Perspectives on Issues in AI Governance. URL: <https://ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf>].

Intel обсуждает вопросы регулирования ИИ в контексте обеспечения конфиденциальности. В 2017 г. Intel заявила, что регулирую-

щие органы должны осуществлять надзор и вмешиваться в случае необходимости, но в рамках существующих правовых норм. Компании, в свою очередь, должны в своей деятельности руководствоваться Принципами справедливой информационной практики (сформулированными ОЭСР) и иметь возможность продемонстрировать регулирующим органам, что эти требования ими соблюдаются [Intel. Artificial Intelligence: The Public Policy Opportunity. URL: <https://blogs.intel.com/policy/files/2017/10/Intel-Artificial-Intelligence-Public-Policy-White-Paper-2017.pdf>].

Facebook компании Meta инициировал создание независимого наблюдательного совета, который наделен правом принимать решение в каждом конкретном случае оценки содержания того или иного контента (например, на предмет возможного разжигания ненависти, дискриминации, нарушения конфиденциальности и т. д.) [Независимый надзорный совет. URL: <https://www.oversightboard.com/>].

Международные организации (ОЭСР, ООН, Европейская комиссия) вносят свой вклад в формирование системы регулирования ИИ. ООН заявила о целом ряде отдельных инициатив, однако до настоящего времени все еще не принято ни одного полноценного договора [United Nations Activities on Artificial Intelligence (AI), 2019. URL: https://www.itu.int/dms_pub/itu-s/opb/gen/S-GEN-UNACT-2019-1-PDF-E.pdf]. ЮНЕСКО в 2019 г. разработаны рекомендации об этических аспектах искусственного интеллекта [Elaboration of a Recommendation on the ethics of artificial intelligence. URL: <https://en.unesco.org/artificial-intelligence/ethics#drafttext>]. ОЭСР 22 мая 2019 г. приняты рекомендации, в которых было сформулировано пять принципов ответственного управления ИИ [OECD moves forward on developing guidelines for artificial intelligence (AI). URL: <https://www.oecd.org/going-digital/ai/oecd-moves-forward-on-developing-guidelines-for-artificial-intelligence.htm>]. Среди этих основополагающих принципов ведущее место занимает принцип принесения пользы ИИ человечеству для его устойчивого развития и роста благосостояния. Остальные требования — это верховенство прав и

свобод человека, демократические ценности, справедливость, прозрачность и ответственность. Помимо формулирования принципов члены организации разработали общие рекомендации в области развития ИИ для правительств стран, входящих в организацию: а) необходимо содействовать государственным и частным инвестициям по вопросам исследований и разработки ИИ; б) необходимо обеспечить надлежащие политические условия для развертывания систем ИИ; в) важно обеспечить просвещение профессиональных кадров в области ИИ для развития навыков в новых условиях.

Указанные принципы развития ИИ одобрила так называемая «Большая двадцатка» в документе «G20 Ministerial Statement on Trade and Digital Economy», в нем, помимо прочего, приведены разработанные странами-участницами G20 рекомендации по улучшению внедрения систем ИИ в социальные процессы, среди которых необходимость разработки цифровой политики, направленной на уменьшение рисков, связанных с развитием цифровой экономики, верховенство принципа антропоцентричности при внедрении ИИ и учет рисков возникновения социальных проблем [G20 Ministerial Statement on Trade and Digital Economy. URL: <https://www.mofa.go.jp/files/000486596.pdf>].

Руководство, созданное инженерной ассоциацией IEEE, возможно, является сегодня одним из наиболее полных и содержательных документов, посвященных регулированию ИИ. Оно нацелено на то, чтобы гарантировать заинтересованность всех сторон, участвующих в проектировании и разработке автономных и интеллектуальных систем, а также на то, чтобы они учитывали приоритет этических соображений для развития этих технологий на благо человечества» [IEEE. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. IEEE Standards Association. URL: <https://standards.ieee.org/industryconnections/ec/autonomous-systems.html>]. Отметим, что в указанном руководстве заявлено, что данный документ является исходной информацией для предстоящей «Серии P7000 стандартов» IEEE по этике ИИ. Этот проект, в свою очередь, обещает

предоставить измеримый, поддающийся сертификации стандарт прозрачности автономных систем.

Следующий документ — рекомендации от группы экспертов высокого уровня комиссии ЕС по искусственному интеллекту, призванные заложить основы нормативно-правовой базы регулирования ИИ в государствах ЕС. В изданных группой экспертов высокого уровня «Этических рекомендациях для надежного ИИ» подчеркивается, что ИИ должен быть «ориентированным на человека» и «заслуживающим доверия» [High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI. URL: <https://digitalstrategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>].

Совет Европы также представил документы, направленные на регулирование применения ИИ. Отметим здесь Европейскую этическую хартию по использованию ИИ в судебных системах, которая утверждает, что ИИ способен повысить эффективность судебной системы, однако его внедрение не должно нарушать основные права и свободы человека [European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment. URL: <https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c>]. Также указанный документ интересен тем, что он содержит всеобъемлющий обзор применения ИИ в судебных системах Европы.

В другом документе Совета Европы — руководстве о защите данных при применении ИИ — упоминаются следующие этические принципы использования ИИ: конфиденциальность, справедливость, ответственность, прозрачность, безопасность данных и управление рисками [The Guidelines on Artificial Intelligence and Data Protection. URL: <https://rm.coe.int/lignesdirhttps://rm.coe.int/guidelines-on-artificial-intelligence-and-data-protection/168091f9d8>].

Помимо ранее указанных Европейский Союз разработал один из самых известных и обсуждаемых документов по применению ИИ — резолюцию Европейского парламента 2015/2103 «Нормы гражданского права о робототехнике» [Civil Law Rules on Robotics, European Parliament resolution of 16 February 2017 with recommendations to the

Commission on Civil Law Rules on Robotics (2015/2103(INL)). OJ C. 2018; 252; 18.7: 239–257. URL: <http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML+TA+P8-TA-2017-0051+0+DOC+PDF+V0//EN>]. В нем прописано, что регулирование ИИ должно основываться на уважении прав человека, законах робототехники А. Азимова, нормах международного права. В Резолюции определены некоторые сферы социальной действительности, в которых регулирование ИИ чрезвычайно важно и способно привести к существенным позитивным изменениям: рынок труда, социальная сфера, государственное управление и др.

В Пекинских принципах ИИ, представляющих позицию китайского государства по вопросам этики ИИ, содержится ссылка на ключевые концепции и ценности китайского общества [Beijing AI Principles. URL: <https://www.baai.ac.cn/news/beijing-ai-principles-en.html>]. Например, здесь указано, что ИИ должен приносить пользу всему человечеству, уважать частную жизнь, достоинство, свободу, автономию и права человека, быть максимально справедливым, уменьшая возможную дискриминацию и предубеждения, обеспечивать прозрачность, объяснимость и предсказуемость принятия решений [Там же]. Рассматривая в их контексте такой проект, как система социального кредитного рейтинга Китая, можно понять обеспокоенность общественности, СМИ, политических деятелей и ученых, которые усматривают в нем своеобразный аналог оруэлловского социального контроля, фактически уничтожающего любую конфиденциальность.

Однако в реальности принцип защиты конфиденциальности личных данных четко сформулирован в законодательных актах Китая: более 100 приложений было запрещено правительством за нарушение конфиденциальности пользовательских данных и еще десятки получили предписания о необходимости внесения изменений, связанных со сбором и хранением данных¹.

¹ В 2019 г. Министерство общественной безопасности КНР приостановило работу 100 приложений, не соответствующих стандартам конфиденциаль-

Исходя из анализа вышеперечисленных документов можно сделать вывод, что в настоящее время в мире еще не выработан полноценный международный подход к регулированию ИИ. Можно говорить лишь об отдельных разрозненных документах, заявлениях, рекомендациях, формулирующих требования, принципы, правила, содержание которых расплывчато, многозначно и существует только на уровне деклараций.

Еще одна группа источников — этические стандарты, кодексы, декларации, источником которых выступают международные и неправительственные организации по всему миру.

Как известно, Стандарты издаются на основе консенсуса, признания специалистами соответствия той или иной продукции установленным требованиям. Продукция, как правило, должна соответствовать условиям безопасности для природной среды и для здоровья человека.

Этические стандарты, как и этические кодексы, не являются законодательными документами и служат в качестве руководств или «мягких законов», не имеющих обязательной юридической силы. Их появление обусловлено стремлением действовать во благо человечества, в данном случае в контексте взаимодействия человека с системами и технологиями ИИ.

Международная организация по стандартизации ISO выступает наиболее авторитетным в мировом сообществе разработчиком разного рода стандартов, технических отчетов, рекомендаций. ISO в 2017 г. создан отдельный технический комитет по ИИ, членами которого стали 30 стран [ISO/IEC JTC 1/SC 42 Artificial intelligence. URL: <https://www.iso.org/ru/committee/6794475.html?view=participation>].

ности персональных данных. В этом же году Министерство общественной безопасности проверило 683 приложения. Также в 2019 г. Министерство промышленности и информационных технологий КНР опубликовало перечень из 41 приложения, которые к концу 2019 г. должны быть изменены так, чтобы соответствовать требованиям к данным. В 2018 г. в Шаньдуне возбудили дело о нарушении прав на персональные данные против 11 компаний.

Некоторые стандарты, изданные данным комитетом, уже вступили в силу: например, стандарт ISO/IEC 20546:2019, содержащий набор основных терминов и определений, имеющий целью улучшение коммуникации и понимания между людьми в области ИИ. В настоящее время в разработке этой организации находится стандарт ISO/IEC AWI TR 24368 Information technology — Artificial intelligence — Overview of ethical and societal concerns, в котором описываются этические и социальные проблемы в области ИИ.

Этическое регулирование применения технологий ИИ находится в центре внимания многочисленных неправительственных организаций, подготовивших соответствующие документы. Наиболее известны сегодня Торонтская декларация [The Toronto Declaration. 2018. URL: <https://www.torontodeclaration.org/declaration-text/english/>], Универсальные рекомендации по ИИ [The Public Voice. Universal Guidelines for Artificial Intelligence, 2018. URL: <http://thepublicvoice.org/ai-universal-guidelines>], Азиломарские принципы ИИ [Asilomar AI Principles. URL: <https://www.artificial-intelligence.blog/news/asilomar-ai-principles>], Монреальская декларация [Montréal Declaration: Responsible AI. 2018. URL: https://monoskop.org/images/d/d2/Montreal_Declaration_for_a_Responsible_Development_of_Artificial_Intelligence_2018.pdf] и 10 этических принципов ИИ Глобального союза [UNI Global Union. Top 10 principles for ethical artificial intelligence. URL: http://www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf].

В «Декларации Торонто» уделяется особое внимание принципам равенства, недискриминации и инклюзивности [The Toronto Declaration. 2018. URL: <https://www.torontodeclaration.org/declaration-text/english/>]. В ней также затрагиваются такие принципы, как прозрачность и подотчетность систем ИИ. Документ был составлен в 2018 г. Amnesty International (глобальное движение, выступающее за защиту прав человека) и Access Now (институт Нью-Йоркского университета, занимающийся исследованием социальных последствий ИИ).

«Универсальные рекомендации по ИИ» включают те же принципы, но дополнительно обращается внимание на качество данных, общественную безопасность, запрет на секретное профилирование [The Public Voice. Universal Guidelines for Artificial Intelligence, 2018. URL:// thepublicvoice.org/ai-universal-guidelines].

«Азилмарские принципы искусственного интеллекта» были озвучены на конференции «The Asilomar Conference on Beneficial AI», организованной Бостонским некоммерческим исследовательским институтом «Future of Life» [Asilomar AI Principles. URL: <https://www.artificial-intelligence.blog/news/asilomar-ai-principles>]. Этот институт считает своей миссией поиск путей и способов смягчения экзистенциальных рисков, с которыми сталкивается человечество, в том числе рисков, исходящих от ИИ. Около 100 исследователей, представляющих разные отрасли научного знания (экономика, право, философия и т. д.), стали участниками конференции, в ходе которой был принят документ, содержащий основные этические принципы применения ИИ, в числе которых доверие, открытость, ответственность, безопасность, защита данных, устойчивость, польза для всего человечества. Этот документ также содержит рекомендации для ученых, ведущих научные исследования в рассматриваемой сфере, и государственных деятелей, определяющих политику в области развития ИИ и решения возникающих в связи с этим социальных проблем. На конференции, кроме того, обсуждались вопросы, связанные с правовыми рисками использования систем ИИ и их моральным статусом.

«Монреальская декларация ответственного развития ИИ», составленная по итогам форума в Монреальском университете в 2017 г., помимо выделения этических принципов регулирования ИИ направлена на установление широкого диалога между общественностью, экспертами и лицами, принимающими решения на правительственном уровне [Montréal Declaration: Responsible AI. URL: https://monoskop.org/images/d/d2/Montreal_Declaration_for_a_Responsible_Development_of_Artificial_Intelligence_2018.pdf].

Наконец, Глобальный союз UNI (федерация профсоюзов) сформулировал «10 основных принципов этического ИИ», которые, по мнению создателей документа, должны быть приняты во внимание профсоюзами работников в сфере материального производства в качестве требований к применению технологий ИИ [UNI Global Union (no year) Top 10 principles for ethical artificial intelligence. URL://www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf].

Таким образом, проблема регулирования ИИ с позиций этики и права сегодня широко обсуждается и в сфере разработки и применения систем ИИ, и в науке, и на уровне государственных органов власти, крупных бизнес-корпораций, международных, общественных, исследовательских, профессиональных и пр. объединений. Это свидетельствует о высокой степени актуальности рассматриваемой тематики, заинтересованности общества в развитии, распространении, совершенствовании систем ИИ, а также пристальном внимании социума к проблемам безопасности, ответственности, ограничений и запретов в процессах взаимодействия человека с искусственным разумом.

Глава II

СОЦИАЛЬНО-ФИЛОСОФСКИЕ ПРОБЛЕМЫ ПРИМЕНЕНИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В СОВРЕМЕННОМ ОБЩЕСТВЕ

2.1 Разнообразие и комплексный социально-философский характер проблем применения ИИ

При рассмотрении проблем применения систем ИИ вначале обратимся к первому закону робототехники, сформулированному писателем-фантастом А. Азимовым в рассказе «Хоровод», изданном в далеком 1942 г.: робот не может причинить вред человеку или своим бездействием допустить, чтобы человеку был причинён вред.

Распространяя действие первого закона на системы ИИ, подчеркнем, что последние создаются человеком для решения важных для него задач, достижения поставленных им целей, но не для продуцирования дополнительных трудностей, препятствий и проблем. Тем более они ни при каких условиях не должны превращаться в угрозу для человека, его здоровья, безопасности, жизни, для его настоящего и будущего. Напротив, они призваны облегчить жизнь человека, освободив его от рутинных и неинтересных либо чрезвычайно трудоемких операций, для созидательных и раскрывающих его безграничный творческий потенциал занятий.

Тем не менее, исследуя актуальные области применения ИИ, мы выяснили, что в процессе проектирования, внедрения и распространения систем ИИ неизбежно возникает ряд специфических проблем, перечень которых по мере развития сферы применения систем ИИ постоянно расширяется.

Традиционно эти проблемы принято называть этическими, однако, на наш взгляд, они выходят за пределы собственно этической

проблематики. Мы характеризуем их как социально-философские, поскольку они охватывают широкий комплекс вопросов, затрагивающих в целом взаимодействие человека с искусственным интеллектом в различных сферах общественной жизни, порождая помимо собственно этических, нравственных проблем экзистенциальные, аксиологические и другие. Кроме того, их действие ухудшает состояние и положение человека в мире. Следовательно, они требуют осмысления и разрешения не только из этических соображений, но главным образом с точки зрения социальной философии.

Так, экзистенциальная проблема утраты человеком смысла бытия может возникнуть, например, в процессе применения систем ИИ в современном производстве. Человеку, потерявшему работу из-за замены его роботом, зачастую сложно найти другие формы успешной социализации, и он ощущает свою ненужность, невостребованность в социуме, отсюда появляются вопросы о предназначении, ценности и смысле его жизни. Аксиологический аспект воздействия технологий искусственного интеллекта на миропонимание современного человека связан с тем, что искусственный интеллект воспринимается в обществе в качестве абсолютной, безусловной ценности, вытесняя на обочину бытия традиционные ценности человеческой жизни.

Нельзя не отметить особый класс проблем, порожденных относительно новизной феномена ИИ, все еще существующим недостатком изученности опыта применения его систем в социальной практике и, как следствие, недостаточной осведомленностью человека о реальных и потенциальных возможностях ИИ.

Проблемы применения ИИ не раз становились предметом научной дискуссии. В итоге учеными получены значимые результаты и опубликованы обзоры наиболее важных, по их мнению, вопросов. Так, Дж. Боссмэн выделил девять основных проблем в области ИИ: безработица, неравенство, человечность, искусственная глупость, расистские роботы, безопасность, злые джинны, сингулярность и права роботов [Bossmann Dzh., 2016]. М. Райан, Дж. Анто-

нию отмечают уже семнадцать подобных вызовов [Antoniou, Ryan, Brooks, Jiya, Macnish, Stahl, 2021, p. 16]. Т. Хагендорф в своем фундаментальном исследовании различных этических руководств, выпущенных до 2020 г., обнаружил проблемы непрозрачности, нарушения конфиденциальности, дискриминации, угрозы безопасности, отсутствия подотчетности (ответственности) [Hagendorff, 2020, p. 99–120]. В докладе «Искусственный интеллект и жизнь в 2030 г.», подготовленном в Стэнфордском университете в 2016 г., в числе насущных проблем применения систем ИИ указываются: угроза безопасности, конфиденциальности личных данных, проблема нарушения уголовной и гражданской ответственности, безработица, сертификация систем ИИ [Stone, Brooks, Brynjolfsson, Calo, Etzioni, Hager, 2016]. Эти же проблемы упоминаются в резолюции Европейского парламента «Нормы гражданского права» [Delvaux, 2019]. В Национальной стратегии РФ обозначена проблема причинения вреда, а также указано, каким образом и вследствие каких причин системы ИИ могут причинить вред человеку [О развитии искусственного интеллекта в Российской Федерации: указ Президента РФ от 10.10.2019 № 490 (вместе с «Национальной стратегией развития искусственного интеллекта на период до 2030 г.»). URL: <https://base.garant.ru/72838946/>]. И. Н. Мосечкин утверждает, что самообучающиеся системы ИИ представляют опасность для общества, и рассматривает некоторые инциденты с участием ИИ, требующие, по его мнению, законодательных мер регулирования [Мосечкин, 2019, с. 461–476].

Институтом инженеров электротехники и электроники (Institute of Electrical and Electronics Engineers, IEEE) были опубликованы «Рекомендации по этически обоснованному проектированию», в документе содержится перечень социальных и этических проблем внедрения систем ИИ: проблема определения ответственности, непрозрачности, недостаточной осведомленности; проблема неуниверсальности и конфликтности этических норм, зависящая от характера задач и различий между сферами применения; проблема

достижения доверия между людьми и системами ИИ; проблема непредвиденного поведения систем ИИ; трудность в обеспечении безопасности работы систем ИИ; проблема конфиденциальности персональных данных; проблема обеспечения прозрачности в работе систем ИИ; проблема юридической ответственности в случае нанесения системами ИИ вреда человеку [Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. Version 2/ The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. URL: http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html].

Д. Балкин исследовал проблему распределения ответственности между людьми и системами ИИ [Balkin, 2015, p. 45–60]. Он размышляет над тем, что усложнение и распространение систем ИИ приводит к вопросу о наделении этих систем юридической ответственностью, субъектностью. Иначе невозможно понять, кто будет отвечать перед законом, если в результате использования системы ИИ возникнут негативные последствия и будет нанесен вред, ущерб кому-либо из участников инцидента.

Системы ИИ активно внедряются в так называемые «умные города», и «широкое использование технологий в повседневной жизни приводит к значительным изменениям в материальной и духовной жизни человека» [Бадмаева, 2020, с. 34]. М. Шрекер [Scherer, 2015, p. 353] и С. Рид [Reed, 2008, p. 201] обратили внимание на проблему распределения ответственности в контексте применения систем ИИ в умных городах. Они пишут о том, что системы ИИ должны быть всегда регулируемы, подотчетными и снабженными механизмами компенсации сбоев.

Речь идет именно о таком виде регулирования, которое стоит на страже человеческих ценностей, обеспечивая непричинение вреда жизни человека, устойчивое развитие и защиту окружающей среды, соблюдение требования социальной справедливости. Эти общественные ценности, на их взгляд, должны иметь безусловный приоритет перед экономическими ценностями, соображениями прибыли и выгоды.

О том, что экономические цели корпораций и интересы политических элит не должны быть приоритетной задачей для ИИ, также писали Л. Флориди, Дж. Коулс, Т. Кин и М. Таддео [Floridi, Cowls et al., 2020, p. 1771–1796]. Они убеждены, что ИИ является общим ресурсом, призванным работать исключительно на благо каждого человека и общества в целом.

В центре внимания В. Майер-Шёнбергер и К. Кукьер (2014) — проблема непрозрачности, ставшая ощутимой, по мнению авторов, в связи с анализом больших данных, поскольку операции с ними невозможно контролировать так, как это было возможно в случае с простым компьютерным кодом.

Дж. Баррелл, исследуя проблему непрозрачности, попытался выяснить причины ее появления [Etzioni, Etzioni, 2017]. При этом он различает преднамеренную непрозрачность, возникающую, например, в тех случаях, когда правительство или корпорации, преследуя некие собственные цели, принимают решение оставить последовательность действий системы ИИ недоступной для общественности, общественного мнения, и непрозрачность, обусловленную масштабами применения системы ИИ, когда принципы машинного обучения или количество задействованных акторов делают алгоритм ее действий непрозрачным даже для экспертов.

А. Гринфилд (2019), Т. Ахмад, А. Тередесей и К. Эккерт рассмотрели проблему непрозрачности систем ИИ в связи с вопросом о недоверии к ИИ [Ahmad, Teredesai, Eckert, 2020, p. 690–699]. Они считают, что системы ИИ уже стали неотъемлемой частью повседневной жизни человека, но при этом принцип работы их остается окутанным завесой тайны. Таким образом, непрозрачный характер принятия решений системами ИИ подрывает доверие к ним со стороны общества и его институтов.

Указанные проблемы обостряются в процессе использования сравнительно недавно появившегося чата GPT 4, основанного на ИИ. Новый чат-бот с искусственным интеллектом GPT 4 — это генеративная языковая модель, запущенная коммерческой компанией

OpenAI 14 марта 2023 г. для развития ранее разработанных чат-ботов GPT. Она называется генеративной, поскольку способна создавать новые данные, а не только анализировать существующие. ChatGPT 4, основанный на МО, не только распознает шаблоны, но и использует их для создания новых данных.

В настоящее время чат-бот пишет стихи и песни, готовит слайды в определенном стиле, успешно проходит собеседование по поводу приема на работу в Google, пишет маркетинговые кампании для определенной демографической группы, комментарии к онлайн-играм и создает изображения с высоким разрешением, выстраивает прогнозы, финансовую аналитику, генерирует простейший код, общается на многих языках мира. Продукты, полученные GPT 4, почти неотличимы от контента, созданного человеком, так как эта система обучается на базе всей информации, доступной в настоящее время в интернете. Поэтому, на первый взгляд, GPT4 выглядит полноценным партнером человека в творческой работе по созданию инновационных продуктов.

Преимущества этого нового инструмента искусственного интеллекта по сравнению с его предшественниками очевидны, однако важно видеть и понимать оборотную сторону этого явления и те проблемы, которые оно порождает. Так, в недавнем мультидисциплинарном исследовании, объединившем 43 статьи экспертов из разных областей, утверждается, что применение ChatGPT может привести к репутационным и юридическим рискам, использованию оскорбительного контента или контента, защищенного авторским правом, потере конфиденциальности, мошенничеству, транзакции и распространению ложной информации. Угрозы, исходящие от Chat GPT4 и подобных ИИ-ботов, связаны с проблемой непрозрачности, могут способствовать дискриминации и предубеждениям, нарушению авторских прав, плагиату или сфабрикованному ботами неаутентичному текстовому контенту, проблеме распространения поддельных медиа [Dwivedi, Kshetri, Hughes, Slade, Jeyaraj, 2023].

В сложных видах деятельности, имеющих творческий и инновационный характер, GPT 4 сегодня не способен создавать действительно оригинальный, новый продукт. Музыкант Ник Кейв недавно получил от одного из своих поклонников текст песни, который был написан Chat-GPT на основе конкретной подсказки: «в стиле Ника Кейва». Разочарованный Кейв ответил поклоннику: «Написание хорошей песни — это не мимикрия, не репликация и не стилизация...» [Cain, 2023]. Л. К. Джена, С. Гоял считают, что гораздо больше шансов стать по-настоящему творческими имеют системы с формами эмоционального интеллекта [Jena, Goyal, 2022].

Бум распространения чат-бота вновь актуализировал опасения, касающиеся страха растущей зависимости людей от подобных программ и постепенного делегирования им человеком решения все более сложных и значимых задач [Baird, Mapuring, 2021]. Например, в развивающихся странах, где не хватает экспертов в тех или иных отраслях знаний, Chat GPT 4 может заменить их, генерируя необходимую пользователям информацию. Широкое использование чат-бота в образовании уже в настоящее время требует перестройки и адаптации учебного процесса, работы преподавателей, научных сотрудников и руководства университетов к условиям быстроменяющейся цифровой среды. По мнению ряда исследователей, подобные технологии вполне могут использоваться и используются студентами для написания научных работ, что порождает проблемы с плагиатом и академической честностью [Stoker-Walker, 2022]. Помимо очевидных, лежащих на поверхности результатов применения указанной технологии необходимо учитывать риск возникновения более глобальных и отдаленных во времени негативных последствий. Речь идет, например, о возможном отказе учащихся от углубленного изучения и критического анализа различных аспектов изучаемого предмета вследствие использования ими Chat GPT 4, позволяющего получить быстрый результат без реального погружения и овладения новыми навыками и знаниями. С. Коннор пишет, что это может привести к подавлению критического мышления,

творческого потенциала, абсолютно необходимых для развития человека [O' Conner, 2023].

Проблема непрозрачности в связи с применением Chat GPT 4 выражается в том, что от разработчика и тем более пользователя скрытыми остаются механизмы работы бота, системы ранжирования и рекомендации. GPT 4, как и многие другие системы ИИ, основанные на МО, может использовать данные, представляющие собой сложившиеся в обществе предубеждения и заблуждения, способствуя усугублению проблемы социальной несправедливости. Д. Винбергер писал, что «предвзятость — это первородный грех машинного обучения» [Weinberger, 2019]. Как известно, знания мы обретаем вследствие формулирования объяснений и проверки их на соответствие реальности, тогда как система опирается на всю совокупность данных, перенесенных человеком в сеть. Следовательно, избежать в полной мере предвзятости на данном этапе развития технологии ИИ не представляется возможным. Chat GPT 4 фактически строит свои рассуждения на основе информации, которая далеко не всегда надежна и достоверна. Не будучи способным к подлинному пониманию, он занимается сопоставлением, компилированием больших объемов данных, находящихся в открытых для обучения источниках. Сгенерированная GPT 4 информация требует тщательной проверки, поскольку его программное обеспечение создает неточный и ложный контент, который может выглядеть вполне достоверным, но на деле является ложным результатом логически некорректных выводов.

Данная система, помимо прочего, нивелирует значимость фактов и выводит на поверхность глобальных информационных потоков любую информацию, в том числе и конфиденциальную, что порождает проблемы нарушения конфиденциальности и управления данными, а также проблему подотчетности в том случае, когда организация не знает и не может определить, кто должен нести ответственность за недостоверную информацию, полученную от ChatGPT 4 и активно используемую в работе сотрудниками компании.

Доступность данной технологии для массового потребителя и разработчиков через языки программирования (например, Python), облачные сервисы (например, Amazon Web Services, Microsoft Azure, Google) позволяет применять ее так же широко, как, например, программы Excel и Access. Эта легкость и простотакратно увеличивают риск нанесения вреда, ущерба человеку вследствие использования информации, сгенерированной данной системой, которая не имеет ничего общего с действительным решением поставленных человеком задач.

По нашему мнению, указанные проблемы, порожденные применением ИИ, требуют социально-философского осмысления, поскольку по своему смыслу и содержанию они соответствуют основной интенции социально-философского исследования. Социальная философия традиционно изучает с позиций целостности и системности универсальные, сущностные черты общества, наиболее общие законы его динамики, глубинные причины тех или иных событий и процессов, перспективы развития социума. Но все эти знания нужны для уяснения места и роли человека, выступающего в качестве творца мира культуры, мира ценностей и норм, необходимых для совместного бытия больших групп людей, функционирования социальных институтов, существования социальной реальности в целом. Иначе говоря, в фокусе социально-философского исследования всегда находится человек как центральный элемент мироздания, создатель и носитель всех свойств и качеств социального, без которого говорить об обществе не имеет смысла и не представляется возможным. Следовательно, проблемы, возникающие в процессе взаимодействия человека с системами ИИ, должны быть введены в круг вопросов, исследуемых социальной философией, поскольку обсуждение и решение этих вопросов затрагивает основы бытия человека в мире, его положение по отношению к другим составляющим социальной реальности, а также ближайшие и отдаленные перспективы его развития. Как отмечает П. В. Алексеев, «...подлинно философские трактовки социальной философии, ее задач и предмета фокусируются

на индивиде, на его многогранных запросах и обеспечении лучшей жизни человека. Именно эти интересы...должны просвечивать все исследования по социальной философии. Научность социально-философского познания должна сливаться с гуманистичностью — таков ведущий принцип познания в сфере социальной философии» [Алексеев, 2003, с. 8].

Социально-философское исследование также необходимо для выявления комплексной природы проблем ИИ, для поиска решений этих проблем не только с теоретических позиций этики, аксиологии, философской антропологии и других отраслей знаний о человеке и обществе, но и понимания целостности социальной жизни людей, реализации гуманистических идеалов, деятельностной природы, целеполагающей функции человека, его способности к природопреобразующей деятельности и обусловленного этим центрального положения человека в социальной действительности.

Обобщая вышесказанное, отметим, что проблема угрозы безопасности или причинения вреда, сформулированная А. Азимовым в качестве первого закона робототехники, выступает в роли центральной, системообразующей социально-философской проблемы применения технологии ИИ, по отношению к которой остальные представляются как раскрывающие возможный источник, причину возникновения вреда либо его разновидности, варианты реализации.

Как следует из рассмотренных нами и упомянутых ранее источников, к наиболее часто обсуждаемым вопросам и трудностям применения ИИ авторы и исследователи относят проблемы нарушения автономии человека, социальной несправедливости, представляющие аспекты, формы, грани, разновидности этого вреда, и проблемы нарушения конфиденциальности, отсутствия ответственности, непрозрачности, порождающие, вызывающие возможное причинение ущерба, вреда человеку вследствие использования технологии ИИ.

Безусловно, предложенный перечень проблем применения ИИ не является полным, окончательным, исчерпывающим. В силу повсеместного проникновения систем ИИ в различные сферы об-

ственной жизни, перманентного совершенствования технологии и постоянно возрастающего влияния их на жизнь и деятельность человека этот перечень может изменяться, пополняться, уточняться. Кроме того, очевидно, что социальные проблемы могут быть вызваны не только применением ИИ. Иные технологии, не связанные с ИИ, также могут выступать в роли источника этих проблем. Однако использование ИИ способно усугубить действие других причин, а также стать самостоятельной причиной ухудшения положения человека в мире.

Рассмотренный пример Chat GPT 4 подтверждает тезис о принципиально «узком» характере современного ИИ, который, несмотря на свои выдающиеся характеристики, все же не может стать полноценной заменой человеку. Так, человек остается по отношению к искусственному разуму оригиналом, творцом, создателем действительно нового. Его сущность заключается в неисчерпаемости, открытости, бесконечности, загадочности и необъятности. Технологии воспроизводят и используют результаты деятельности человека, но превзойти его непостижимую с позиций рациональности природу они не в состоянии. При этом созданный GPT 4 контент в виде фейковых новостей, пропаганды, дезинформации уже сегодня может вводить людей в заблуждение, усугубляя существующие в обществе предубеждения, разжигая вражду, ненависть, подрывая социальную сплоченность, доверие людей друг к другу и к результатам развития новых технологий, технологий будущего. Все это обосновывает необходимость системного, целостного, взвешенного исследования воздействия ИИ на общество и человека средствами и методами гуманистически ориентированной социально-философской интерпретации процессов взаимодействия человека с искусственным разумом, открывающей возможность нахождения путей решения, способов преодоления уже существующих и только формирующихся в данной области проблем и препятствий.

2.2 Основные проявления отрицательного влияния ИИ на человека

Причинение вреда. Проблема причинения вреда вызывает беспокойство у ученых, общественных деятелей, мирового сообщества в целом. Британский физик С. Хокинг в 2014 г. сказал: «Успешное создание искусственного интеллекта станет самым большим событием в истории человечества. К сожалению, оно может оказаться последним, если мы не научимся избегать рисков... Когда искусственный интеллект начнет управлять финансовыми рынками, научными исследованиями, людьми и разработкой оружия, это будут вещи, недоступные нашему пониманию. Если краткосрочный эффект искусственного интеллекта зависит от того, кто им управляет, то долгосрочный эффект — от того, можно ли будет им управлять вообще» [Хокинг, 2014].

Обсуждаемая проблема зачастую сводится к намеренному использованию ИИ в целях, представляющих угрозу для жизни, здоровья и благополучия человека и общества. Так, в системе обеспечения правопорядка применение ИИ зачастую усугубляет существующие негативные тенденции.

Некоторые системы ИИ используются в разработке автономного оружия, специально создаются для решения военных задач. В 2015 г. на международной конференции по искусственному интеллекту исследователи в области робототехники и ИИ, общественные деятели, ученые призвали прекратить дальнейшее развитие вооружения, основанного на ИИ, которое может действовать без строгого человеческого контроля. Открытое письмо подписало более 20 тыс. человек [Открытое письмо в ООН с призывом запретить разработку вооружений с ИИ. URL: <https://robogeek.ru/iskusstvennyi-intellekt/otkrytoe-pismo-v-oon-s-prizyvom-zapretit-razrabotku-vooruzhenii-s-ii#>].

Тем не менее в настоящее время существуют разные виды автономного вооружения на основе ИИ. Помимо тех, которые подконтрольны человеку и предоставляют ему возможность отмены реше-

ния системы, обсуждается идея создания полностью автономного оружия, принципиально неподвластного человеческому контролю [Etzioni, Etzioni, 2017]. Есть попытки регулирования процесса введения автономного и полуавтономного оружия: в одном из документов министерства обороны США указано, что командиры и операторы должны иметь возможность осуществлять необходимый уровень человеческого суждения относительно использования силы [Sharkey, 2014, p. 305–324]. Но что именно подразумевается под выражением «необходимый уровень человеческого суждения», остается неясным.

Автономное вооружение, таким образом, представляется еще более опасным в сравнении с традиционным оружием, поскольку его применение существенно сокращает время на принятие и осмысление важных стратегических решений, минимизирует контроль над ситуацией со стороны человека и лишает его возможности самостоятельно принимать решения, сокращая автономию, ограничивая свободу. Непрозрачный характер приводит к отсутствию взаимопонимания между машиной и человеком, а логика принятия решений ИИ может оставаться непонятной для последнего.

Угрозы постепенно возникают во всех сферах: в медицине — ненадлежащая деятельность автономного медицинского оборудования; в агросекторе — разрушения, причиненные беспилотной сельскохозяйственной техникой (например, трактор Spirit [AT400 Spirit. URL: <http://robotrends.ru/robopedia/at400-spirit>] или устройство Lettuce Bot [Lettuce Bot: Roomba for Weeds. URL: <https://modernfarmer.com/2013/05/lettuce-bot-roomba-for-weeds/>]); в обороне — ошибка в системе «свой — чужой» военных устройств для подавления противника.

Международные стандарты безопасности ИИ в основном находятся в стадии разработки и обычно касаются только таких ее отдельных аспектов, как объяснимость или управляемость систем ИИ [ISO/IEC AWI TS 6254; Information Technology-Artificial Intelligence-Objectives and Approaches for Explainability of ML Models and AI

Systems. International Electrotechnical Commission; International Organization for Standardization: Geneva, Switzerland, 2021]. В них не рассматривается и не предлагается структурированный процесс для полной оценки рисков в области ИИ или полная таксономия источников рисков, которые были бы необходимы для разработки соответствующих стандартов, либо приводится лишь краткое описание источников риска, что затрудняет общее понимание связанных с ними трудностей. Существует всего несколько исследований, посвященных определению и описанию конкретных источников рисков, исходящих от ИИ, но и они содержат лишь поверхностное и неполное описание [Steimers, Bömer, 2021, p. 239–251].

К сожалению, применение систем ИИ в военных и преступных целях не исчерпывает весь потенциальный вред, наносимый человеку. Проблема причинения вреда имеет более сложный, комплексный характер, проявляющийся посредством множества социально-философских проблем, раскрывающих отдельные грани причиняемого человеку вреда и породившие их причины.

До недавнего времени считалось, что технологии ИИ заменяют человека в выполнении относительно простых, рутинных работ, но в действительности оказалось, что они могут заменить человека, выполняя практически любую работу, алгоритм которой можно вычислить и предсказать. Не зря еще на заре технологий ИИ Н. Винер предостерегал, что цифровые вычислительные машины — это нечто, что коренным образом отличается от всех предыдущих механических средств и способно разрушить привычный социальный порядок [Dyson, 2012]. Конечно, следует учитывать, что разные технологии, автоматизирующие работу, имеют различный потенциальный эффект. Тем не менее некоторые из них способны привести к потере человеком работы, которую он ценит и считает важной и значимой, лишит его рабочего места, законного заработка и средства обеспечить себя и своих близких всем необходимым. Более того, технологии ИИ могут реализовать это незаметно, убрав определенные задачи в алгоритмах действий работников, и участие

человека станет не обязательным, не выгодным и излишним для работодателя.

Уникальная особенность ИИ как технологии заключается в том, что он обладает уровнем интеллекта, беспрецедентным по сравнению с другими предшествующими ему в истории развития технологиями (информационными и коммуникационными), что проявляется в способности ИИ принимать решения, решать проблемы и даже думать в узкоспециализированном понимании этого слова. Системы ИИ, подробно изучая описание наших действий и повторяя за нами, могут овладеть необходимым навыком, качественно выполняя гораздо больше когнитивных задач, чем среднестатистический работник-человек. Тем самым системы ИИ обретают возможность влиять на характер происходящих изменений, порождая новые, неожиданные, противоречивые и негативные последствия для занятости людей, для их самооценки, самореализации, понимания своего места и роли в мире.

Можно предположить, что если системы ИИ берут на себя более сложные и важные задачи, то последствия могут быть непредсказуемыми. Пример GPT 4 демонстрирует, что профессии, которые ранее считались прерогативой высококвалифицированных кадров, становятся теперь уязвимыми для автоматизации системами ИИ, что создает риски трудоустройства, профессиональной востребованности специалистов, занятых в творческих и интеллектуальных видах деятельности.

Системы ИИ активно внедряются в области живописи и музыки. Известен пример, когда ИИ сочинил целое музыкальное произведение [Iamus, a music-making computer, could be the next Mozart. URL: <https://www.vice.com/en/article/pgg8yy/iamus-a-music-making-computer-could-be-the-next-mozart>]. Тогда как относиться к тому творческому продукту, который осуществляется системами ИИ? Можно ли результат их «творчества» воспринимать как нечто, заслуживающее внимания? Гораздо больше опасений вызывает то, что подобные программные продукты основываются на генетическом программи-

ровании, которое нацеливается на постоянное обучение до тех пор, пока не начнет решать задачу, близкую к идеальному варианту выполнения. Тогда возникает вопрос «Возможно ли, что алгоритмы научатся в ближайшем будущем создавать за нас, например, правовые концепции или решать задачу в области государственного управления?».

Обратимся к еще одному примеру — системе ИИ Alpha Fold, предназначенной для автоматизации и ускорения процесса прогнозирования структуры белка в разработке новых методов лечения болезней человека [Hassabis, Revell, 2021, p. 44–49]. Эта система может решить быстро и точно задачу, которую обычно выполняли специалисты, обучавшиеся для этого годами. Подобная ситуация создает значительные риски для них использовать имеющиеся знания и умения, демонстрировать свое мастерство и навыки, а значит, быть и ощущать себя профессионалами, нужными, востребованными в обществе людьми, следующими своему предназначению, своей миссии.

Проблема деквалификации актуализирует социально-политический аспект применения систем ИИ, вопросы о том, каким образом общество и государство должны реагировать, если большие массы людей окажутся внезапно лишенными источника дохода без перспективы найти себе другое занятие, чтобы удовлетворять базовые потребности, иметь возможности и условия для личной и социальной реализации. М. Форд отмечает, что в настоящее время политическое регулирование не способно минимизировать последствия структурных изменений, вызванных внедрением цифровых технологий, уменьшить растущее неравенство в контексте сокращения доли труда в национальном доходе. При этом он подчеркивает повсеместное присутствие ИИ, тот факт, что с каждым днем остается все меньше профессий, где человека не заменила бы машина [Форд, 2016].

Кроме очевидной экономической угрозы материальному благополучию людей существует еще один не менее важный вопрос, свя-

зывающий потерю работы с экзистенциальной угрозой утраты смысла жизни, о чем, в частности, пишут Ким и Шеллер-Вольф [Kim, Scheller-Wolf, 2019, p. 319–337]. Более того, зафиксировав данную взаимосвязь, авторы указывают на сохранение актуальности проблемы смыслоутраты даже при компенсации потери дохода, возникшей вследствие автоматизации. Безработные страдают от потери навыков, социальной изоляции и отсутствия цели в жизни, утрата работы оказывает долгосрочное негативное воздействие на их социальное самочувствие [Clark, Oswald, 1994, p. 648–659]. В более поздних исследованиях появился такой феномен, как «смерть от отчаяния», возникающий в связи с отсутствием оплачиваемой работы [Case, Deaton, 2020]. Это означает, что автоматизация может содержать потенциальную экзистенциальную угрозу для человечества, поскольку без работы гораздо больше людей имеют шанс столкнуться с хроническими заболеваниями и, утратив надежду, сделать непоправимый и неправильный выбор против самой жизни.

Несколько иная ситуация складывается, если человек занимает роль оператора, осуществляющего контроль за эффективностью работы систем ИИ [Langlois, 2020, p. 174]. На первый взгляд, такой вид работы может показаться более щадящим вариантом по сравнению с первым сценарием. Однако человек, постоянно выполняя повторяющиеся и фрагментарные по характеру действия, отчуждается от результатов своей работы и утрачивает возможность видения целостной картины. Такой характер работы дает мало возможностей для развития разнообразных навыков, для раскрытия и реализации собственного потенциала.

Таким образом, системы ИИ, выполняя значимую, интересную, творческую часть работы человека, способны отчуждать его от непосредственной производительной деятельности, что негативно влияет на родовую сущность человека как производящего существа, единственного в мире субъекта природопреобразующей деятельности, направленной на адаптацию, приспособление мира к удовлетворению потребностей человека, занимающего главное, центральное

положение в известной и доступной ему реальности. Согласно Марксу, рабочие с появлением машин сталкиваются с технологическим отчуждением и борются с безработицей. В связи с применением систем ИИ работники отчуждаются от продукта собственной деятельности, а в процессе жесткой конкуренции на рынке труда ИИ все больше превосходит их способности и возможности. Человек также отчуждается от самого себя, от своей родовой сущности, поскольку все меньше ему удается решать задачи, соответствующие его творческому общественно-производственному потенциалу. Он перестает быть целью и становится средством достижения целей, поставленных не им. Системы ИИ усугубляют процесс отчуждения, сталкивая человека с чем-то чуждым ему по своему происхождению и субстрату, но в то же время сверхъестественно знакомым по своим способностям. Если бы системы ИИ соотносились с механическим ткацким станком или паровой машиной, то экономическое положение человека не подвергалось бы такой значимой угрозе. Если раньше рабочий беспокоился об утрате работы в своем профессиональном секторе, то у него всегда была возможность переобучиться и снова обрести свою безусловную ценность с экономической точки зрения. Использование систем ИИ в современном мире чрезвычайно усложняет, проблематизирует эту возможность, стремительно сокращая перечень областей деятельности человека, не подлежащих автоматизации.

Таким образом можно сделать вывод, что системы ИИ в отличие от других технологий в перспективе могут стать не ограниченными в своих возможностях и способными обучаться и осуществлять производственный процесс во всех областях жизни социума. Соответственно, возникает риск отчуждения человека от самого себя вследствие его взаимодействия с системами ИИ, наделенными способностью к интеллектуальной деятельности и на ее основе материальной деятельности, направленной на изменение, преобразование окружающей реальности исходя из целей и задач человека, человеческой цивилизации. Именно это отличает технологии ИИ от всех

иных разработанных ранее технологий, что обуславливает широкий перечень проблем, рисков и опасностей, возникающих в процессе развития человека и общества, их взаимодействия с системами ИИ. В краткосрочной перспективе, как было установлено ранее, применение систем ИИ грозит безработицей, а в долгосрочной перспективе возникает угроза не только полной безработицы, но и настоятельной необходимости защищать особый статус человека в мире.

По отношению к проблеме причинения вреда все остальные вторичны, подчинены, раскрывают и уточняют смысл главной проблемы. Они призваны выявить источник, причины возникновения нежелательных для человека последствий, а также описать разнообразные проявления возможного ущерба для человека и последствия, возникающие в случае некорректного внедрения, применения систем ИИ или ошибок, допущенных создателями таких систем в процессе их разработки.

Социальная несправедливость. Помимо указанных проблем в процессе исследования результатов применения ИИ выяснилось, что системы ИИ могут быть предвзятыми в отношении определенных индивидов или групп, подвергая их несправедливой дискриминации по этническим, гендерным или религиозным основаниям, что позволяет сделать вывод о наличии проблемы социальной несправедливости. Ее обсуждению посвящен целый ряд работ, появившихся в последнее время.

Одна группа исследователей по теме справедливости относится к тем, кто считает, что справедливость возможно алгоритмизировать и ввести в системы как измеряемый показатель [Michael, Van Kleek, Binns, 2018, p. 1–14], другая — к тем, кто уверен, что справедливость нельзя толковать с точки зрения четко определенных количественных показателей [Holstein, Wortman, Daumé, Dudík, Wallach, 2019, p. 1–16].

Несправедливость в контексте применения ИИ принято обозначать в литературе как алгоритмическую предвзятость [Ryan, Antoniou, 2021]. Справедливость выражается главным образом в

предотвращении или смягчении последствий нежелательной предвзятости и дискриминации. Источники определяют несправедливость систем ИИ как отсутствие разнообразия [Green Digital Working Group. Position on Robotics and Artificial Intelligence. URL: <https://felixreda.eu/wp-content/uploads/2017/02/Green-Digital-Working-Group-Position-on-Robotics-and-Artificial-Intelligence-2016-11-22.pdf>], интерпрации [SAP's guiding principles for artificial intelligence. URL: <https://www.sap.com/products/leonardo/machine-learning/ai-ethics.html#guidingprinciples>] и равенства [The Future Society. Science, Law and Society (SLS) Initiative. The Future Society. URL: <https://web.archive.org/web/20180621203843/http://thefuturesociety.org/science-lawsociety-sls-initiative/>]. Представители государственного сектора по проблеме несправедливости систем ИИ уделяют особое внимание их влиянию на рынок труда [UNI Global. 10 Principles for Ethical AI. URL: <http://www.thefutureworldofwork.org/opinions/10-principles-for-ethical-ai/>].

Принято считать, что системы ИИ могут помочь при справедливом принятии решений и способствовать предотвращению предвзятости [Kleinberg, Mullainathan, Manish, 2018, p. 1–23]. Однако на практике принятие решений системами ИИ зачастую воспроизводит существующие в обществе системные, институциональные и социальные предубеждения. Такие предубеждения могут привести к дискриминации, несправедливости и проблемам нарушения конфиденциальности и прав человека.

Так, в большинстве европейских государств в сфере обеспечения правопорядка применяются системы ИИ для профилирования людей, прогнозирования вероятного преступного поведения людей или мест будущих преступлений, а также для оценки предполагаемого риска причастности отдельных лиц к совершению преступлений (например, список преступников 600 (Нидерланды)) [Pushback against AI policing in Europe heats up over. URL: <https://www.globaltimes.cn/page/202110/1237232.shtml>]. Полученные результаты используются для слежки, проведения обысков или до-

просов лиц, которым системой ИИ был присвоен высокий рисковый рейтинг.

Оценивая работу подобных систем, специалисты пришли к неожиданным выводам. Так, Дж. Ангвин, Дж. Ларсон, С. Матту в своем исследовании выяснили, что системы ИИ предвзяты в автоматических оценках возможности повторения преступления осужденными [Angwin, Larson, Mattu, Kirchner, 2016].

Системы искусственного интеллекта применяются также в сфере образования и трудоустройства для отбора кандидатов при поступлении в вузы, для найма работников, например, с просмотром резюме и для целевой рекламы вакансий [Tristram, 2019, p. 215–221]. М. Боген и А. Рик отметили, что риск систематической ошибки был зафиксирован на каждом из этих этапов процесса найма [Bogen, Rieke, 2018]. Применение систем ИИ в указанном направлении, безусловно, имеет преимущества в виде автоматизации рутинной работы и сокращения временных затрат. Однако возникают опасения по поводу того, насколько адекватной является оценка системами ИИ навыков соискателей, на каких основаниях и фактах принимается решение, не подвергается ли угрозе безопасность социума и отдельного человека.

В финансах и банковском деле алгоритмы ИИ составляют основу множества различных приложений, таких как прогнозирование рынка для торговли, управление рисками для оценки кредитоспособности, распределение кредитов и определение ставок по ипотечным кредитам [Longbing, 2022, p. 1–38]. Учеными выявлены многочисленные случаи, когда решения по кредитным заявкам были несправедливыми и предвзятыми по отношению к заемщикам из числа национальных и гендерных меньшинств. Речь идет, в частности, о более высоком проценте отказов в ипотеке и кредите для латиноамериканских и чернокожих заемщиков в США [Bartlett, Morse, Stanton, Wallace, 2022] или более низких кредитных лимитах для женщин по сравнению с мужчинами, имеющими равные кредитные характеристики [Haridasari, 2019].

Рекомендательные системы в поисковиках, основанные на ИИ, помогают пользователям ориентироваться в сети, предлагают интересующую их информацию. Было установлено, что рекомендательные системы усиливают различные виды предвзятости, упорно предлагая материалы с чрезмерным участием мужчин, представителей европеоидной расы и молодых пользователей [Ribeiro, Henrique, Benevito et al., 2018], или предлагая в качестве лучших только определенные предприятия и маркетплейсы [Dash, Chakrabatov, Ghosh, Mukherjee, Krishna, 2021, p. 873–884]. Это увеличивает дисбаланс сил между доминирующими на рынке крупными платформами и более мелкими, которые не имеют доступа к равным объемам высококачественных потребительских данных, жизненно важных для выхода на рынок.

На наш взгляд, такая концентрация власти в распоряжении небольшого числа компаний, которые разрабатывают большинство приложений ИИ и действуют при этом для получения ими максимальной прибыли, а не в интересах развития общества, является дополнительной угрозой для институтов демократии. Поскольку рекомендательные системы поиска нацелены на привлечение внимания и потому определяют большую часть информации, которая становится известна пользователям, высок риск искаженного восприятия последними информации, что в перспективе способно лишить людей возможности осознанно участвовать в здоровом политическом и социальном дискурсе. Так, системы искусственного интеллекта негативно влияют на участие избирателей в выборах, подрывая основы демократического мироустройства посредством назойливо предлагаемой теми или иными группировками недобросовестной целевой рекламы.

Особенно очевидной становится социальная несправедливость в контексте европейского законодательства, согласно которому мужчины и женщины должны пользоваться равными правами на рынке труда и иметь равный доступ в сфере предложения товаров и

услуг². Также должно быть гарантировано равное отношение к людям, независимо от их расового или этнического происхождения [Там же]. Аналогичным образом должно соблюдаться равенство в сфере труда и занятости, если речь идет о людях с инвалидностью, о людях с различными религиозными предпочтениями, убеждениями, о людях различных возрастных групп и сексуальной ориентации [Там же].

С. Вахтер, Б. Митльштат, С. Рассел [Wachter, Mittelstadt, Russel, 2021, p. 105–567], изучив целый ряд европейских документов, пришли к выводу, что в отношении ИИ чрезвычайно сложно определить, были ли в действительности нарушены права людей в том или ином случае, поскольку алгоритмы ИИ осуществляют свою деятельность с такой скоростью, масштабом и уровнем сложности, которые в принципе не сопоставимы с уровнем человеческого интеллекта, не поддаются пониманию [Burrell, 2016, p. 1–12]. Например, потребителям сложно оценить, была ли им предложена лучшая цена. Они даже не имеют возможности узнать о том, что определенные рекламные объявления им вовсе не были предложены. Масштабы и эффекты автоматизированного принятия решений не позволяют людям понять, что они оказались в обделенном, ущербном положении, в результате у них не будет явных оснований для предъявления иска в соответствии с законом о недискриминации. Таким образом, использование систем ИИ обуславливает обсуждение вопроса о справедливости на фундаментально новом уровне, что предполагает выполнение требований о методах обнаружения возможной дискриминации и обязательном представлении доказательств, свидетельствующих о том, что права людей не были ущемлены.

Перед разработчиками и пользователями ИИ возникает вопрос «Может ли справедливость быть автоматизирована?». ИИ дискри-

² Council of the European Union. Council Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services. 2004.

минирует людей так, что становится очень сложно, практически невозможно обнаружить, расследовать и предотвратить случаи дискриминации. Справедливость часто зависит от контекста и не может (и, возможно, не должна) быть принципиально автоматизирована. Поэтому традиционные (например, правовые) процедуры обеспечения справедливости в принятии решений должны быть более последовательными, постоянными и четкими в своих определениях. Даже если появятся стандартные показатели и пороговые значения, позволяющие в каких-либо ситуациях отличить справедливое от несправедливого, проблема несправедливости останется, ибо она никогда не сможет охватить все возможные случаи нарушения нормы.

Еще один аспект рассматриваемой проблемы возникает из-за сложного характера современных систем машинного обучения, действия которых осуществляются по принципу «черного ящика», когда проследить за целым рядом факторов, составляющих алгоритм деятельности ИИ, не представляется возможным.

Проблема заключается в том, что человек не знает, каким образом системы ИИ, основанные на нейронных сетях, приходят к тем или иным результатам. Например, вышеупомянутый алгоритм COMPAS, оценивающий вероятность совершения насильственного преступления, как оказалось, дискриминирует преступников с темным цветом кожи. Однако, как показало исследование, система в действительности не учитывала расовые признаки в качестве входных данных. Вместо этого она получила конфиденциальные данные о расовой принадлежности в качестве следствия из другой доступной ей информации о месте жительства потенциального преступника.

Аналогичным образом две программы ИИ, которые независимо друг от друга научились распознавать изображения лошадей, использовали при этом совершенно разные подходы [Lapuschkin, Wäldchen, Binder, Montavon, Samek, Müller, 2019, p. 1096]. Первая программа ИИ правильно сосредоточилась на особенностях живот-

ного, вторая строила выводы на основании набора пикселей в нижнем левом углу каждого изображения лошади. ИИ приходил к верным заключениям исходя из совершенно ошибочных посылок.

В сфере здравоохранения существует специфический риск применения технологии ИИ, заключающийся в том, что они могут усугубить существующее неравенство, обусловленное, в частности, этнической принадлежностью, социально-экономическим статусом, возрастом пациента. Сбор данных для системы ИИ при этом может привести к получению некорректных сведений вследствие языковых барьеров, а также того обстоятельства, что люди будут предоставлять неверную или неполную информацию из-за испытываемого ими недоверия к автоматизированным системам. Ошибки приложений ИИ, в свою очередь, в конечном итоге могут негативно сказаться на клинических результатах пациентов, на состоянии их здоровья. Человеческое измерение здравоохранения предполагает уникальное отношение между профессионалом и пациентом, пронизанное особыми ценностями и обязанностями. Это отношение рассматривается учеными как требующее особого подхода, ориентированного на пациента, предполагающего автономию пациента, способствующую его осознанному выбору, который человек имеет право и может сделать в соответствии с разделяемой им системой ценностных установок.

В сфере сельского хозяйства внедрение ИИ приводит к увеличению цифрового разрыва между отдельными фермами и странами. Одни активно используют ИИ и получают представляемые им преимущества, другие в силу отставания в технологическом развитии лишены такой возможности. Задача общества, государства состоит в том, чтобы обеспечить справедливую и инклюзивную пользу от ИИ для всех производителей и пользователей. С другой стороны, как отмечает М. Райан, проблема заключается в том, что ИИ приносит экономическую выгоду агробизнесу и технологическим компаниям, но не самим фермерам, которые зачастую получают экологический и социальный ущерб от применения систем ИИ [Ryan, 2020,

р. 49–69]. В. А Цвык и И. В. Цвык отмечают, что, несмотря на явные преимущества применения ИИ в сельском хозяйстве, он может способствовать отчуждению человека от животного мира и окружающей среды [Цвык В. А., Цвык И. В., 2022, с. 58–69].

Справедливость связана также с солидарностью. Последняя обсуждается чаще всего в связи с последствиями использования систем ИИ на рынке труда, в контексте обеспечения социальной защиты и равного отношения ко всем претендентам на вакантные рабочие места. Солидарность выражается в том, чтобы добиться справедливого перераспределения преимуществ, предоставляемых ИИ, осуществляющим подбор сотрудников на имеющиеся вакансии среди претендентов из всех без исключения социальных групп. А. Джобин, М. Йенка и Е. Вайена, в частности, описывают существующие на рынке труда практики сбора данных, ориентированные на отдельных, соответствующих неким определенным требованиям лиц, которые своим избирательным механизмом фактически наносят ущерб социальной солидарности и действуют по принципу «радикального индивидуализма» [Jobin, Lenca, Vayena, 2019, p. 389–399].

По нашему мнению, внедрение и использование систем ИИ во всех сферах общественной жизни должно быть на основе справедливого, равного, беспристрастного отношения ко всем участникам разнообразных процессов социального взаимодействия. Подобные требования выдвигают сегодня многочисленные международные организации, комитеты по этике, научные институты, разрабатывающие собственные этические кодексы, а также ученые, участвующие в работе научных конференций, например ACM «Справедливость, подотчетность и прозрачность» [ACM FAccT Conference. URL: <https://facctconference.org>]. По мере того как все больше и больше решений делегируется человеком системам ИИ, общество, его институты должны проследить, чтобы эти решения были свободны от любой предвзятости и дискриминации.

Нарушение автономии. Данная проблема, специально выделяемая исследователями систем ИИ, заключается в том, что последние могут посягать на автономию человека, ущемляя свободу принятия им решения и уязвляя таким образом человеческое достоинство. У людей отнимают саму возможность самостоятельно, без помощи ИИ принимать осознанные и независимые решения, что можно трактовать как утрату человеком его права на самоопределение. Л. Н. Мешкова утверждает, что ИИ и подобные ему технологии способны существенно влиять на свободу человека, порождая новые зависимости и ограничения [Мешкова, 2020, с. 53–60]. Системы ИИ, например, могут прямо или косвенно навязывать людям определенный образ жизни, применяя репрессивные механизмы наблюдения или стимулирования [Montréal Declaration: Responsible AI. URL: https://monoskop.org/images/d/d2/Montreal_Declaration_for_a_Responsible_Development_of_Artificial_Intelligence_2018.pdf]. В результате срывается процесс целеполагания, теряются навыки самостоятельности, совершается менее аутентичный выбор, принцип человеческой автономии фактически нарушается. Привлекательные на первый взгляд перспективы использования систем ИИ (расширение когнитивных способностей человека, освобождение его от монотонного труда, повышение качества жизни и обслуживания и др.) сопровождаются прокламациями о неминуемых угрозах и рисках для самостоятельности и автономии человека, неизбежно возникающих вследствие внедрения ИИ в жизнь общества.

Проблема нарушения автономии особенно очевидна в сфере медицины, где системы ИИ преобразуют базисные структуры врачевания, меняют социальный контекст оказания медицинской помощи. Преобразования имеют системный характер, поскольку кардинально затрагивают все аспекты деятельности врача, изменяя диагностические процедуры [Abdulkareem, Leiner, Petersen, 2019, p. 133], практики принятия врачебных решений [Shortliffe, Sepúlveda, 2018, p. 2199–2200], порядок проведения необходимых манипуляций (например, хирургических операций [Там же], по-

слеоперационного ухода [Hasimoto, Rosman, Witkowski, 2019, p. 21]).

Системы ИИ в области медицины уже сегодня претендуют на роль автономных, свободно действующих субъектов, выполняя ассистирующую и консультирующую функции. Роботы-ассистенты в некоторых клиниках занимаются уходом за больными после тяжелых операций [Bian, Xiang, Tong, Feng, Weng, 2020, p. 23], а программы на основе ИИ консультируют врачей, помогая им в принятии решений [Обзор российских систем принятия врачебных решений (СППВР) // WEBIOMED 30 января 2021 г. URL: <https://webiomed.ai/blog/obzor-rossiiskikh-sistempodderzhki-priniatiia-vrachebnykh-reshenii/>]. В области инновационной медицины уже ведутся дискуссии о том, как разработать и внедрить полностью автономную систему ИИ. Так, на заседании президиума РАН по развитию робототехники в медицине, в частности, было заявлено о необходимости создания роботов, полностью заменяющих хирургов [О внедрении робототехники в отечественную медицину — проблемы и пути решения: заседание Президиума РАН. URL: http://www.ras.ru/news/news_release.aspx?ID=2eaccb0cd887-4d02-bb10-7caae48b083b&print=1]. О. Р. Чепьюк выражает беспокойство по поводу того, что четвертая промышленная революция может стать непосредственной угрозой автономии человека, когда на «умных заводах» и в «умных» отраслях решения будут приниматься ИИ [Чепьюк, 2020].

Проблема нарушения автономии, в свою очередь, порождает проблему деквалификации сотрудников и проблему доверия. У врачей, которые все больше опираются на ассистирующие системы ИИ, со временем меняются методы принятия решений (имеется в виду обоснованное принятие решения), стиль обучения, снижается качество медицинских знаний, вырабатывается стереотипное, небрежное, обезличенное отношение к пациентам. Этот феномен деквалификации (постепенной утраты навыков) выявил Т. Хофф в ходе исследования работы врачей первичного звена [Karpicke, Blunt, 2011, p. 772–775]. У начинающих специалистов навыки и вовсе не бу-

дут развиваться, как считает ученый, и врач не сможет адекватно выполнять свои профессиональные обязанности. Опасения по поводу проблемы деквалификации и изменения природы человека выражает М. В. Заладина, утверждая, что чем внушительнее становится влияние ИИ на человеческую природу и окружающую среду, тем труднее его контролировать, тем больше опасность для человека [Заладина, 2020]. Л. В. Баева утверждает, что роботизация и замещение человека приведут к значительной трансформации ценностных приоритетов человека, изменятся его воззрения о долге, справедливости, необходимости и т. д. Автора интересует вопрос о том, меняют ли эти технологии саму генетическую и духовную сущность личности [Баева, 2012, с. 110–119]. К тому же системы на основе глубокого обучения не вполне прозрачны, действия их принципиально непонятны для врача, не говоря уже о том, что они еще более непонятны пациентам. Можно сказать, что подобные системы в определенном смысле обладают субъектностью, поскольку им делегируется часть врачебных полномочий. Это, в свою очередь, вновь затрагивает вопрос о том, кто несет ответственность: человек или машина. Некоторые компании даже пытаются запустить проекты, подобные «Moral Machine», чтобы наделить ИИ полноценной субъектностью [Whitby, 2015, p. 223–233].

Внедрение систем ИИ, все больше посягающих на автономию врачей, порождает проблему доверия. Она сводится к необходимости поиска ответа на вопросы о том, будет ли пациент доверять ИИ так же, как врачу, будут ли врачи доверять системам ИИ, принцип работы которых может быть непонятен человеку. Непрозрачность решений систем ИИ угрожает автономии пациента, который может быть ограничен в принятии свободных и осознанных решений по поводу своего здоровья на основе неполного информирования. К тому же если пациент остается неосведомленным о том, с кем он взаимодействует, с врачом или искусственным разумом, то выходит, что система обманывает пациента и нарушает его права. Недоверие системам ИИ выразили врачи в IBM во время клинического

испытания Watson Oncology, посчитав рекомендации проекта недостаточно компетентными и даже опасными [Bioethics briefing note: artificial intelligence (AI) in healthcare and research. URL: <https://www.abhi.org.uk/resource-hub/file/9335>]. Таким образом, чрезмерное доверие системам ИИ может привести к нерешаемым проблемам в сфере оказания медицинских услуг. В связи с этим возрастает необходимость оценки последствий использования таких систем.

Проблема нарушения автономии человека в значительной степени проявляется из-за самообучаемости систем ИИ, их способности независимо от действий производителя/пользователя привести к неправильной последовательности действий (или бездействию). Дело в том, что системы с высокой степенью собственной автономности могут демонстрировать неожиданное поведение и принять неправильное решение, причинив тем самым ущерб субъекту, ограничив его автономию, его самостоятельность в принятии решений. Более того, риски, создаваемые высокоавтоматизированными системами ИИ, в свою очередь, могут усугубляться человеческим фактором (например, свойственным среднестатистическому человеку временем реакции на ситуацию, ее истолкование). Не случайно И. В. Понкин и А. И. Редькина утверждают, что правовое положение систем ИИ должно определяться мерой и природой их автономности от человека [Понкин, Редькина, 2018, с. 91–109]. Как правило, высокая степень самостоятельности ИИ ограничивает возможности контроля и влияния на него со стороны человека, а значит, и автономию человека. Поэтому разработчикам систем ИИ следует предусмотреть, чтобы действия человека всегда имели приоритет при их использовании. Лучший способ обеспечить необходимый уровень автономии человека — вовлечь будущих пользователей, а также экспертов в предметной области в сам процесс разработки системы ИИ. Степень автоматизации должна соответствовать контексту приложения и предоставлять пользователям все необходимые возможности управления. В конечном итоге это приведет к созда-

нию искусственного интеллекта, ориентированного на человека. Системы ИИ уже сегодня широко используются во многих приложениях, связанных с безопасностью, причем в таких отраслях деятельности, которые сопряжены с чрезвычайно высоким уровнем риска для человека, его жизни и здоровья (например, авиация или работа атомных электростанций). Здесь особенно важно, чтобы средства управления системой были понятны и подконтрольны людям, вели себя в работе так же, как на этапе проектирования.

Проблема нарушения автономии проявляется и в невозможности выбирать людям ту работу, которая способствует саморазвитию, нравственному совершенствованию и позволяет придерживаться собственных ценностей, жить в соответствии с теми принципами, которые они действительно разделяют. В этом контексте Л. Флориди считает, что постепенно разрастающееся делегирование полномочий системам ИИ снижает автономию человека [Floridi, Cowls, Beltrametti et al., 2018, p. 689–707]. Это означает, что в случае активного участия ИИ в процессе решения задачи последний может ограничивать человека в стремлении следовать высоким духовным целям своей деятельности, что приводит к потере важных, существенных его качеств и ограничивает возможности развития человека. ИИ также может сократить автономию и в том случае, если доступ к передовой технологии будет закрыт для большинства членов общества. Автономия работника, выполняющего функции оператора систем ИИ, может быть нарушена посредством неправомерного ограничения искусственным разумом объема информации, доступной для просмотра и использования [Kellogg, Valentine, Christin, 2020, p. 366–410]. К тому же если работа оператора сама по себе является обыденной и скучной, то это заставляет его ощущать себя «рабом машины» и по этой причине также испытывать чувство несвободы.

По нашему мнению, куда больший риск для автономии человека возникает из-за слежки и манипуляций со стороны систем ИИ. М. Фуко обозначал данную проблему ростом «общества слежки»,

что заставляет людей постоянно переживать по этому поводу [Abrams, 2020, p. 241–258]. Когда человек находится под наблюдением, он, как правило, ощущает скованность и действует менее свободно. Использование систем ИИ для наблюдения за работниками, как мы предполагаем, будет иметь аналогичные последствия и может стать для работодателей средством усиления контроля за ними. Например, использование камер с искусственным интеллектом для наблюдения за водителями службы доставки Amazon заставляет их чувствовать себя зажатыми, ограниченными и неспособными действовать автономно [Asher-Schapiro, 2021]. Некоторые системы ИИ внедряются для мониторинга онлайн-совещаний с целью осуществления наблюдения за вовлеченностью работников в процесс обсуждения, что может привести к стрессу, депрессии и прочим негативным последствиям ощущения собственной несвободы [Pardes, 2020].

Слежка за рабочими с помощью систем ИИ и работа в качестве оператора на службе у машины может привести также к профессиональной поляризации, к разделению людей на неквалифицированных и привилегированных работников. Последним при этом достается более интересная и увлекательная работа. Эти опасения отсылают к проблеме социальной несправедливости, возникающей при распределении доли преимуществ и объема тяжелой, рутинной работы на каждом рабочем месте, что подрывает солидарность в коллективе, наблюдается раскол между теми, кто получает выгоду от внедрения ИИ, и теми, кто ее не имеет. Например, в работе любого колл-центра ИИ может использоваться для мониторинга и оценки звонков, осуществляемых и поступивших каждому оператору в отдельности. Такое усиленное наблюдение, с одной стороны, может быть воспринято операторами как навязчивое и ограничивающее их автономию. С другой стороны, подобное использование систем ИИ может обеспечить успех организации, помогая ей лучше обучать и управлять деятельностью операторов. Однако неизбежно будут формироваться микрогруппы внутри организации, по-разному оце-

нивающие роль ИИ: одни будут видеть в ИИ ненужное и мешающее работе устройство, а другие — удачный способ повышения результативности работы исполнителей. Люди, принадлежащие к первой группе и испытывающие дискомфорт от давления со стороны систем ИИ, могут даже принять решение об увольнении либо утратить чувство солидарности, ответственности и принадлежности к единому коллективу, сопричастности его успехам и результатам.

Области применения ИИ в будущем будут только расширяться, и человеку необходимо позаботиться о том, чтобы при разработке интеллектуальных систем любого уровня сложности человек не утратил своего центрального положения в мироздании, чтобы в основе любых технологий неизменным оставался подход, ориентированный на человека, на достижение им целей и задач своего развития, на реализацию его подлинного предназначения в мире.

2.3 Причины формирования негативных последствий взаимодействия человека и ИИ

Непрозрачность. Исследуя причины возникновения негативных эффектов от взаимодействия с искусственным разумом, прежде всего обратим внимание на проблему непрозрачности систем ИИ, которые из-за постепенной эволюции и усложнения превращаются в своего рода «черные ящики» [Isak, Lee, 2017, p. 77–98]. Принцип работы этих систем становится все более непрослеживаемым, необъяснимым и неинтерпретируемым. Понять действия или мотивы принятия решения такими непрозрачными системами ИИ становится практически невозможно. Система не демонстрирует ход своей работы, последовательность выстраиваемой ею цепочки рассуждений. Для человека (в том числе разработчика системы) остается неизвестным то, на основании каких данных был совершен тот или иной вывод, получено то или иное заключение. М. Ананни и К. Кроуфорд составлен подробный обзор множества случаев, связанных с проблемой непрозрачности [Ananny, Crawford, 2018, p. 973–989].

Ежегодно в мире проходят международные конференции, посвященные обсуждению прозрачности, справедливости, подотчетности систем ИИ (например, FAT-ML [Fairness Accountability and Transparency in Machine Learning. URL: <http://www.fatml.org/>], семинар ICML [Workshop on Human Interpretability in Machine Learning. URL: <https://sites.google.com/view/whi2018/>]).

В настоящее время известны многочисленные прецеденты нарушения прав людей, возникшие в результате непрозрачной работы систем ИИ. Непрозрачной может быть либо система в целом, либо отдельные решения, принятые системой. Некоторые исследователи обеспокоены тем, что требование к прозрачности систем ИИ находится на нереально высоком уровне, поскольку даже решения, принимаемые человеком, далеко не всегда могут быть понятны и прозрачны для окружающих [Zerilli, Knott, Maclaurin, Gavaghan, 2019, p. 661–683]. Они полагают, что требовать прозрачности всего алгоритма бесполезно, поскольку сам процесс считается трудоемким, дорогим и не всегда можно раскрыть логику принятия решений. Авторы приходят к выводу, что важнее подумать о том, в каких случаях объяснения действий системы полезны для человека. Они считают, что целесообразнее будет установление требований, направленных на раскрытие сведений о логике принятия конкретных решений, касающихся входных данных, статистического анализа, проверки кода и т. д.

На наш взгляд, непрозрачность в работе систем ИИ может быть продиктована сложностью алгоритмов ИИ, обуславливающей низкий уровень интерпретируемости результатов деятельности системы, а также вероятностью возникновения ошибок в ее работе.

При этом если от системы требуется получение информации высокой степени точности, то проблема непрозрачности должна быть максимально устранена. Если же некоторые погрешности не оказывают значительного негативного влияния на результат работы системы, то требование прозрачности можно считать излишним. Например, для системы ИИ, подбирающей целевую рекламу для

того или иного пользователя социальной сети, достаточно будет относительно низкого уровня интерпретируемости, поскольку последствия ее неправильной работы не столь значительны, они не способны нанести заметный ущерб человеку, потребителю этой рекламы. В то же время интерпретируемость работы системы на основе ИИ, осуществляющей диагностику состояния здоровья человека, должна быть предельно высокой, поскольку любые ошибки могут навредить здоровью пациента, и здесь непрозрачность в действиях и решениях системы недопустима.

В середине 1990-х годов система ИИ, внедренная в алгоритмы работы одной из клиник и обученная прогнозировать, каких пациентов клиники следует госпитализировать, а каких лечить амбулаторно, сделала вывод, что пациенты с пневмонией и астмой имеют более низкий риск смерти и поэтому не подлежат госпитализации. С медицинской точки зрения это решение абсолютно неверно, поскольку пациентов с астмой и пневмонией требуется не просто госпитализировать, но и поместить в отделение интенсивной терапии [Caruana, Lou, Gehrke, Koch, Sturm, Elhadad, 2015, p. 17–21]. Столкнувшись с подобной проблемой, в клинике приняли решение отказаться от данной системы ИИ, и в настоящее время исследователи проводят работу, для того чтобы клинические системы на основе ИИ стали более прозрачными.

Беспилотные транспортные средства, создание которых было вызвано стремлением снизить количество смертей в результате дорожно-транспортных происшествий, из-за непрозрачности принимаемых ими решений могут, как свидетельствует опыт их применения, оказаться еще одной угрозой жизни и здоровью человека. Так, беспилотный автомобиль Uber убил женщину в Аризоне [McFarland, 2018]. Это был первый известный случай со смертельным исходом, связанный с полностью автономным транспортным средством. Беспилотные автомобили должны верно оценивать дорожную ситуацию и принимать правильные решения. Однако в данном случае оказалось, что программное обеспечение автомобиля

ошибочно распознано возникший перед ним объект как пластиковый пакет, а не человека, переходящего дорогу. Непрозрачность работы алгоритма, управляющего движением автомобиля, послужила причиной гибели человека. Только прозрачная система, решения которой известны и понятны человеку, одобрены им, может предотвратить возникновение подобных ситуаций. Исследования беспилотных транспортных средств, чтобы объяснить алгоритм их действия людям, уже ведутся, но каковы будут результаты, предугадать практически невозможно [Bojarski, del Testa, Dworakowski, Firner, Flepp et al., 2016, p. 1–9; Haspiel, Meyerson, Robert Jr. et al., 2018].

В правосудии создаваемые системы ИИ могут помочь, например, в оценке риска возникновения рецидивов преступлений. Здесь важно иметь гарантии, что прогнозирование риска рецидива осуществляется с позиций справедливости и не ущемляет права человека. Известный пример дела 2013 г. «Лумис против Висконсина», когда обвиняемый получил максимальное наказание, свидетельствует о существующей в сфере правосудия опасности принятия необъективных решений вследствие применения систем ИИ [Lightbourne, 2017, p. 327–343]. Задействованная в данном случае система ИИ COMPAS, принимая во внимание пол, расу, место жительства и другие критерии, присвоила Лумису высокий рисковый рейтинг повторения преступления. Но используемые ею алгоритмы составляли коммерческую тайну, а сам процесс установления причинно-следственных связей между данными анкеты, обрабатываемой COMPAS, и предлагаемым ею решением не был очевиден для судьи. Непрозрачность в данной области могла привести к негативным последствиям для осужденного. О возможности принятия необъективного решения Верховный суд заявил в принятом им решении. Но в настоящее время ситуация практически не изменилась, и автоматизированное принятие решений в правовой системе не стало более прозрачным [Detecting bias in black-box models using transparent model distillation. URL: https://www.aies-conference.com/2018/contents/papers/main/AIES_2018_paper_96.pdf].

В сфере финансовых услуг системы ИИ могут найти применение, например, в процедуре оценки кредитных рейтингов. Проблемы возникают тогда, когда система не объясняет, почему клиенту было отказано в кредите. Такие кредитные бюро, как Equifax и Experian, работают над многообещающими исследовательскими проектами, для того чтобы генерировать автоматические коды причин и решения о кредитных рейтингах на основе ИИ стали более прозрачными [Equifax Launches NeuroDecision Technology. URL: <https://investor.equifax.com/news-events/press-releases/detail/203/equifax-launches-neurodecision-technology>].

Принципиально непрозрачная природа систем ИИ обусловлена тем, что они построены на основе искусственных нейронных сетей методом глубокого обучения и способны к самостоятельному извлечению информации из массива данных. Результат работы такой системы может оказаться совершенно непредсказуемым. Например, чат-бот Microsoft под названием Tay, сконструированный для общения с людьми в возрасте от 18 до 24 лет, уже через сутки начал проявлять признаки некорректного поведения, допуская нацистские и расистские высказывания [Чат-бот от Microsoft за сутки научился ругаться и стал расистом. URL: <https://www.interfax.ru/world/500152>].

В Национальной стратегии развития искусственного интеллекта, принятой в РФ на период до 2030 г., указано, что алгоритмы работы нейронных сетей могут быть чрезмерно сложными для интерпретации, соответственно, результаты их работы могут быть даже отмечены человеком [О развитии искусственного интеллекта в Российской Федерации: указ Президента РФ от 10.10.2019 № 490 (вместе с «Национальной стратегией развития искусственного интеллекта на период до 2030 г.»). URL: <https://base.garant.ru/72838946/>]. Следовательно, данные, которые получены с помощью систем ИИ, должны быть проверены человеком, а механизм принятия решения должен быть прозрачным.

Учитывая, что системы ИИ будут со временем выполнять гораздо более сложные задачи и оказывать большее влияние на наш мир,

чем технологии предыдущих поколений, следует признать, что проблема непрозрачности обостряется.

Помимо прочего, непрозрачность работы систем ИИ порождает проблему доверия и угрозы достоинству личности, что может в конечном счете приостановить дальнейшее внедрение систем ИИ в различные сферы жизни социума и нивелировать объективные преимущества, которыми они в действительности обладают. Трудно и опасно доверить важные решения системе, принцип работы которой скрыт и недоступен.

В школьном округе Хьюстон (США) с помощью системы ИИ было проведено оценивание эффективности работы преподавателей и влияния их на обучающихся. Результаты оценивания впоследствии стали основанием для увольнения ряда педагогов, признанных недостаточно квалифицированными [Gacuta, Selvadurai, 2019, p. 193–216]. В данном случае угроза достоинству личности заключалась в том, что оценивание компетентности специалистов осуществлялось на основе данных, обработанных системой ИИ без раскрытия аргументов, послуживших основанием для принятия негативных для карьеры преподавателей решений. Люди должны понимать и знать, на основании чего принимаются такие решения. Но алгоритм работы систем ИИ не позволяет получить эту информацию. В настоящее время проблема интерпретации непрозрачных решений широко обсуждается в самых разных областях, в том числе специалистами в области права и социальной философии, однако консенсус все еще не достигнут.

Отсутствие ответственности. Проблема отсутствия ответственности является наиболее распространенной вследствие некорректного применения систем ИИ. В современной литературе неоднократно приводились случаи причинения вреда человеку системами ИИ, в которых виновные не были установлены и наказаны, т. е. ни на кого не возложена ответственность за гибель и страдания людей. Так, на заводе Volkswagen робот-манипулятор прижал к плите работника, последний не смог выбраться и скончался на месте [Попова, 2019,

с. 69–73]. Boeing 737 попал в аварию из-за неверных показаний датчика и системы контроля полета [Авария Boeing 737 Max глазами разработчика ПО. URL: <https://habr.com/ru/post/449564/>]. Беспилотный автомобиль не снизил скорость и сбил человека [Чем опасно внедрение технологий с искусственным интеллектом. Рядом с нами появились электронные личности // Российская газета. 2020. № 18(8072)].

Проблема отсутствия ответственности состоит прежде всего в том, что в подобных описанных ситуациях не представляется возможным установить ответственное лицо, субъекта ответственности, на которого однозначно можно было бы возложить вину за произошедшее.

Ю. Тихомиров считает, что в подобных ситуациях «... ответственность прежде всего лежит на создателе (изготовителе), отдельно выделяется лицо, осуществляющее обучение ИИ, оператор, собственник и третьи лица, повлиявшие на решение ИИ» [Юридическая концепция роботизации..., 2019, с. 89]. О. Я. Ястребов, сформулировавший концепцию опережающего регулирования, считает проблему отсутствия ответственности одной из основных [Ястребов, 2017, с. 280].

Вопрос об ответственности систем ИИ в настоящее время нашел отражение в специальных документах компаний-разработчиков систем ИИ. Например, Google опубликовано руководство для регулирования систем ИИ, предусматривающее ограничение доступа к отдельным приложениям и технологиям [Google. AI at Google: Our Principles. URL: <https://ai.google/principles/>].

В то же время очевидно, что слишком жесткое или непродуманное регулирование может замедлить развитие систем ИИ. Р. Кларк предлагает в связи с этим альтернативные подходы к регулированию ИИ, такие как саморегулирование, отраслевое регулирование, механизмы совместного регулирования и формальное право [Roger, 2019, р. 398–409]. Он убежден, что регулирование должно быть тщательно разработано и оценено с учетом проблем технического и

политического характера. К. Форбс отмечает, что некоторые из предложенных видов регулирования активно используются сегодня для стимулирования процессов внедрения ИИ, особенно это касается практики саморегулирования [Forbes, 2021, p. 297–300].

Вопрос об ответственности систем ИИ особенно обостряется, когда они применяются там, где деятельность их может напрямую влиять на жизнь, здоровье и благополучие людей. Речь идет, например, о применении ИИ в сфере здравоохранения, где ИИ сегодня используется в качестве консультантов (в системах поддержки принятия решений), выступающих в виде самостоятельных автономных субъектов (роботов). Недавно в Китае компанией iFlytekCo. Ltd. был разработан робот Xiaoyi (Сяо И), который успешно прошел экзамен (456 баллов из 600 возможных), позволяющий получить лицензию. Врачом он пока не работает, но находит и анализирует информацию о пациентах. Компания IBM разработала суперкомпьютер Watson, который располагает данными около 100 млн пациентов [IBM Pitched Its Watson Supercomputer as a Revolution in Cancer Care: It's Nowhere Close. URL: <https://www.statnews.com/2017/09/05/watson-ibm-cancer/>]. Эта система проанализировала 15 млн страниц биомедицинской литературы, более 200 учебников по медицине, 300 медицинских журналов. Компания-разработчик уверена, что система действительно улучшает качество медицинской диагностики и гарантирует лучшие результаты лечения, будучи способной в каждом конкретном случае установить наиболее точный медицинский диагноз. Однако даже на такую супермашину нельзя переложить ответственность за ошибочное и, возможно, опасное для жизни и здоровья пациента решение. В онкологических центрах, сотрудничающих на этапе запуска с Watson, признали несостоятельность проекта и пришли к выводу, что поглощение такого объема данных — это не то же самое, что их осмысление и разумное использование.

Кто же будет нести моральную и правовую ответственность в случае принятия неправильного решения: лечащий врач, использу-

ющий систему ИИ, или эксплуатирующая ее медицинская компания? Может быть, ответственность берет на себя производитель системы ИИ, из-за недостатков которой произошел инцидент? Как разграничить ответственность между различными субъектами оказания конкретной медицинской услуги? В настоящее время специалисты ведут бурные дискуссии о статусе ИИ, в том числе юридическом [Nechkin, 2020, р. 78–85]. В ходе обсуждений не вызывало сомнений стремление участников дискуссий к переформатированию этических-правовых оснований взаимодействия врачей и пациентов в случае использования в процессе диагностики и лечения систем ИИ.

Проблема отсутствия ответственности и необходимость ее решения очевидны при исследовании применения беспилотных транспортных средств. Кто будет виновным, если произойдет аварийная ситуация: владелец беспилотного транспортного средства, разработчик технологических основ беспилотного транспортного средства или его изготовитель? Многие считают, что внутри беспилотного транспортного средства всегда должен присутствовать человек, который в любую секунду может и готов взять на себя управление и ответственность, особенно при аварийной ситуации. Этой позиции придерживается большинство стран, разрабатывающих беспилотные транспортные средства. Она закреплена в положениях Венской конвенции о дорожном движении (1968 г.), в которой устанавливается, что каждое транспортное средство или состав транспортных средств, находящихся в движении, обязаны иметь водителя [Конвенция о дорожном движении..., с. 8]. Однако в этом случае транспортное средство считается беспилотным.

В настоящее время ни одно государство не готово в полной мере использовать для транспортных перевозок беспилотные автомобили, управляемые ИИ. Одна из причин — отсутствие необходимого законодательного регулирования применения беспилотных транспортных средств. Нормативные правовые акты, существующие сегодня, разрознены, слабо коррелируют с существующим законодательством отдельных государств и международными правовыми актами.

В Германии в середине 2017 г. отдельным законодательным актом были внесены изменения в действующий Закон о дорожном движении, предусматривающие допуск высокоавтоматизированных автомобилей на дороги общего пользования [Восьмой закон о внесении изменений в Закон о дорожном движении от 16 июня 2017 г. URL: <http://robopravo.ru/uploads/s/z/6/g/z6gi0wkwhv1o/file/5MZOclyT.pdf>]. В документе содержится 20 требований к беспилотному транспорту, их производителям и водителям. Например, требование к ценности человеческой жизни означает, что система автопилота в любой внештатной ситуации должна быть нацелена на сохранение жизни людей. При этом подчеркивается, что автопилот не должен совершать определенный выбор, чья именно жизнь должна быть спасена. Напротив, он должен действовать в интересах всех участников аварии в равной степени. В случае аварии встроенный в автомобиль «черный ящик» определяет, под чьим управлением находилось транспортное средство и, следовательно, уточняет, несет ли ответственность за аварийную ситуацию водитель или производитель беспилотного транспортного средства.

Также в законе утверждается, что водитель может отвлекаться во время движения, но при этом система гарантирует ему «адекватный запас времени» в том случае, если человеку придется взять управление на себя. Очевидно, что водитель не будет иметь права спать, но остается неясным, что законодатель считает «адекватным». Во время юридических прений по данному документу были упомянуты временные промежутки продолжительностью 2–5 секунд. Однако следует иметь в виду, что указанное положение будет применяться как к городскому движению со скоростью 30 км/ч, так и к движению по шоссе со скоростью 130 км/ч.

Настораживает и то, что производители автомобилей не берут на себя прямую ответственность в случае возникновения аварийной ситуации. Эта ответственность целиком возлагается на водителя, который, по сути, непосредственно не управляет автомобилем. Подчеркнуто, что немецкий законодатель на данном этапе технологического

развития не считает возможным менять действующие правила дорожного движения в части ответственности водителя автотранспортного средства. Подобные изменения, по мнению законодателя, можно будет внести только после достижения последней стадии автоматизации.

В Норвегии в январе 2018 г. также был принят закон «Об испытаниях высокоавтоматизированных автотранспортных средств», позволяющий осуществлять проверку машинам с разным уровнем автоматизации на дорогах общего пользования (даже при полном отсутствии в машине водителя) [URL:<https://lovdata.no/dokument/NL/lov/20171215112?q=Lov%20om%20utpr%C3%B8ving%20av%20selvkj%C3%B8rende>].

В России регулирование беспилотного транспортного средства впервые отражено в постановлении правительства РФ «О проведении эксперимента по опытной эксплуатации на автомобильных дорогах общего пользования высокоавтоматизированных транспортных средств» (ноябрь 2018 г.) [URL: [http:// publication.pravo.gov.ru/File/GetFile/0001201811270008?type=pdf](http://publication.pravo.gov.ru/File/GetFile/0001201811270008?type=pdf)]. В данном нормативном акте указано, что в России планируется проведение правового эксперимента, в рамках которого высокоавтоматизированные автомобили получат возможность двигаться по дорогам общего пользования наравне с другими участниками движения (на территории г. Москвы и Республики Татарстан). В документе также отмечено, что разрешение могут получить только юридические лица, а ответственность за причинение вреда каждым автомобилем должна быть застрахована на 10 миллионов рублей. Водители этих транспортных средств должны иметь стаж не менее трех лет. При этом в их биографии должен отсутствовать факт лишения прав на вождение по какой-либо причине. Но самым примечательным и спорным положением является пункт о полной ответственности собственника автомобиля за все происшествия (кроме тех, которые произошли по вине других участников). При этом в постановлении указывается, что тестируемые беспилотники должны управляться водителем во время проведения эксперимента. Таким образом, регулирование

беспилотных и обычных серийных автомобилей ничем не отличается. Видимо, водитель нужен для того, чтобы его в любой ситуации можно было бы назначить субъектом ответственности за возможный вред, который причинит беспилотник.

Мы полагаем, что деятельность, связанная с использованием ИИ, нуждается в специальном правовом регулировании, поскольку речь идет о жизни и здоровье человека. Поскольку системы ИИ могут быть потенциально опасны для людей, необходимо строго лицензировать деятельность, связанную с их производством и эксплуатацией. Для этого необходимо на законодательном уровне четко определить, за кем должна быть закреплена ответственность в случае возникновения нештатных ситуаций, а также ответить на вопрос, возможен или правомерен переход ИИ из статуса объекта в статус субъекта правоотношений.

Европейский парламент рассматривает законопроект о присуждении роботизированным системам ИИ статуса «электронной особы» [Европарламентарий может предоставить роботам юридический статус. URL: <https://ria.ru/world/20170114/1485715425.html?injj=1>]. Авторы законопроекта полагают, что возрастающая автономия позволяет считать их не вещью, а самостоятельными субъектами правоотношений. Соответственно, этот статус позволит возложить на них и юридическую ответственность за совершаемые ими действия. В документе указано, что «обычные нормы ответственности становятся неэффективными, появляется необходимость создать новые нормы, которые в первую очередь определяют, как аппараты могут быть привлечены к ответственности — полной или частичной — за свои действия или бездействие» [Там же]. Документ касается роботов, автономных транспортных средств, дронов и т. п. устройств. Данный законопроект многие раскритиковали, около 150 экспертов из 14 стран написали открытое письмо Европейскому парламенту, содержащее официальное осуждение содержания документа [Ксе-нофонта, 2018]. По их мнению, машина не может быть ответственной за причиненный вред, поскольку ошибки были допущены

человеком, создавшим алгоритм ее действий, или появились в результате сбоя работы этого алгоритма. В настоящее время не существует системы ИИ, которая была бы полностью автономной.

Однако попытки превратить системы ИИ в автономного агента можно обнаружить и в других отраслях современной экономики. Например, в 2016 г. в российских и зарубежных СМИ появилась новость, что система ИИ, принадлежащая Сбербанку, в ближайшие годы научится принимать около 80% всех решений с помощью ИИ [Греф: Сбербанк сможет принимать 80% решений искусственным интеллектом. URL: <https://ria.ru/economy/20160908/1476449735.html>].

Таким образом, проблема отсутствия ответственности возникла вследствие невозможности однозначно определить субъекта и меру ответственности в инцидентах, произошедших с участием систем ИИ, в силу несовершенства существующей законодательной базы, относительной новизны феномена ИИ, поспешного и зачастую непродуманного внедрения указанных систем в различные сферы жизни современного общества.

Правовая регламентация ответственности в случае применения систем ИИ, на наш взгляд, должна быть распределена между всеми участниками процесса разработки, внедрения и использования этих систем. Следовательно, необходимо установить границы ответственности изобретателей, разработчиков, производителей технологий ИИ. Также определенную долю ответственности должны взять на себя пользователи систем ИИ, поскольку их деятельность может привести к созданию ситуаций повышенной опасности, возникших вследствие специфических свойств технологий и низкого уровня контроля за процессом со стороны человека. Применение мер уголовной ответственности непосредственно по отношению к системам ИИ нецелесообразно, так как они не обладают должным уровнем самосознания. Добавим также, что вопрос об ответственности в каждом конкретном случае, конечно, должен решаться в судебном порядке.

Нарушение конфиденциальности. Не менее значимой является проблема нарушения конфиденциальности, также возникающая в

случае применения систем ИИ. Сегодня они широко используются для сбора, обработки и защиты личных данных. Это необходимо, например, для поиска и отслеживания людей в больших и густонаселенных городах, биометрической аутентификации по голосу и лицу человека, для поиска преступников по поведенческому шаблону и т.д. Указанная проблема возникает из-за угрозы разглашения персональных данных или потери контроля над ними при использовании систем ИИ. Эти данные могут быть опубликованы в открытых источниках и в дальнейшем использоваться мошенниками.

Проблема нарушения конфиденциальности имеет комплексный характер. Во-первых, это проявляется в том, что данные могут собираться или передаваться без явного и осознанного согласия пользователя. Например, во время вспышки лихорадки Эбола в Западной Африке в 2014 г. пришлось экстренно рассекречивать данные и анализировать записи звонков мобильных телефонов пациентов. Принятые меры позволили эпидемиологам предотвратить распространение болезни. Важно отметить, что дальнейшая публикация данных была приостановлена из-за опасений по поводу нарушения конфиденциальности и использования рассекреченных данных промышленными конкурентами.

Другим примером является проект социального кредитования в Китае, введенный в 2014 г. [Chorzempa, Triolo, Sacks, 2018]. Он направлен на регулирование делового и частного поведения граждан, обеспечение объективности в принятии решений о мерах их наказания или поощрения. Главная цель — укрепление доверия граждан к органам власти. В документе «Пекинские принципы ИИ» специально подчеркивается, что «акторы должны иметь достаточное информированное согласие о влиянии системы на их права и интересы» [Beijing AI Principles. URL: <https://www.baai.ac.cn/news/beijing-aiprinciples-en.html>]. Казалось бы, речь идет о защите конфиденциальных данных, но реализация проекта социального кредитования противоречит принципу конфиденциальности, поскольку собираются сведения о повседневной жизни человека: совершенных

им покупках, поведении в соцсетях, выбираемых им компьютерных играх и т. д., затем полученная информация переводится в числовые данные по утвержденным правительством правилам, для того чтобы провести оценку положительных или отрицательных сторон личности. Составленный рейтинг социального доверия в дальнейшем учитывается при совершении индивидом разнообразных действий — вплоть до бронирования отелей или поездок на отдельных видах транспорта, которые становятся недоступны пользователю из-за введенных ограничений. Результатом проекта должно стать повышение прозрачности деятельности всех организаций, органов власти, усиление социального контроля, вовлечение широкой общественности в процесс управления государством и, наконец, достижение главной цели — повышение уровня доверия со стороны общества к органам власти. Однако уже в 2020 г. стало очевидно, что проект оценивается в обществе как нежелательный синтез, «брак между коммунистическим надзором и капиталистическим потенциалом» [Botsman, 2017].

Проблема заключается в том, что на практике информированное согласие сводится к тому, что пользователем достаточно формально принимаются условия системы. Он в действительности не осознает, какие именно его персональные данные будут использоваться, в каких ситуациях они будут применены. Помимо этого, данные пользователя, сведения об его активности успешно монетизируются различными бизнес-структурами, но он сам об этом практически ничего не знает и никакой компенсации не бывает. К тому же ответственность за сохранность персональных данных возлагается на самого же пользователя. Соглашаясь на обработку персональной информации, он берет на себя ответственность и за возможные негативные последствия. Многие думают, что без предоставления ими данных о себе система не сможет полноценно функционировать, но на самом деле приложениями используется лишь малая часть тайно собранных данных [Большие данные в социальных и гуманитарных науках., 2019]. В результате человек оказывается

совершенно беззащитным перед алгоритмами ИИ. Без специального просвещения, образования в этой области он не понимает, на что именно соглашается, принимая условия системы. Он не имеет представления, какие данные о нем получит система ИИ и как они будут в дальнейшем использованы. Он не знает, будут ли нарушены в данном случае его права.

Практика передачи человеком системам ИИ конфиденциальных личных данных в современном мире продолжается, принимает новые формы, но последствия и риски совершенно не оценены и слабо осознаются.

Второй аспект проблемы нарушения конфиденциальности заключается в появлении риска раскрытия собранных данных внешним агентам или попыток деанонимизации [Narayanan, Shmatikov, 2008, p. 111–125]. Отдельные пользователи систем ИИ могут стать мишенью для разнообразных вредоносных агентов, что неизбежно приведет к нарушению их прав (например, в том случае, если личная информация используется таким образом, что под угрозой оказывается индивидуальная автономия человека).

Особенно это актуально, когда человек вынужден взаимодействовать с системой ИИ, например в сфере оказания неотложной медицинской помощи. Здесь существует риск того, что данные пользователей могут быть переданы компаниям, которые применяют технологии ИИ для маркетинга товаров и услуг, для извлечения прибыли, создания и реализации каких-либо продуктов. Эти сведения могут использовать страховые фирмы или крупные технологические компании.

В мае 2017 г. был зафиксирован факт передачи данных о здоровье людей, собранных компанией Google, что вызвало огромный резонанс. Возник вопрос о том, как, в каких целях может быть вообще использована персональная информация. Google тогда объявила о стратегическом партнерстве с Чикагским университетом и Медицинским университетом Чикаго в США [Wood, 2017]. Целью партнерства стала разработка новых инструментов машинного обу-

чения для прогнозирования медицинских событий (например, срочной госпитализации). При реализации этой цели университет предоставил Google доступ к сотням тысяч историй болезни «обезличенных» пациентов. Спустя некоторое время, в июне 2019 г., один из пациентов М. Динерштейн подал групповой иск против Университета и Google от имени всех людей, чья конфиденциальная информация была раскрыта [Minssen, Gerke, Shachar, 2019]. В результате расследования стало очевидно, что технологии ИИ, прогнозирующие появление различных заболеваний, могут создать излишнюю стигматизацию отдельных лиц, подвергать людей агрессивному маркетингу со стороны фармацевтических компаний и других фирм, представляющих широкий спектр коммерческих медицинских услуг.

Таким образом можно сделать вывод о том, что сегодня существует риск переоценки преимуществ использования ИИ, появления завышенных ожиданий от его потенциальных возможностей и, как следствие, внедрения продуктов и услуг с недоказанной безопасностью и эффективностью. Вполне может быть, что сама технология искусственного интеллекта может не соответствовать стандартам научной достоверности и точности, которые в настоящее время применяются к медицинским технологиям. На наш взгляд, даже чрезвычайная ситуация не может быть оправданием для использования непроверенных технологий. Например, цифровые технологии, разработанные и применяемые на ранних этапах пандемии COVID-19, не всегда соответствовали каким-либо объективным стандартам эффективности, оправдывающим их использование [Gasser, Lenca, Scheibner, Sleight, Vayena, 2020, p. 425–434]. Технологии ИИ были внедрены в качестве ответных мер на пандемию без достаточных доказательств и клинических испытаний. Указанные примеры позволяют сделать вывод о необходимости правового регулирования подобной деятельности и тщательного анализа возникающих проблем.

И, наконец, третий аспект проблемы нарушения конфиденциальности дает о себе знать на этапе получения выводов, которые система ИИ может сделать из имеющихся в ее распоряжении данных. Пользователи, как правило, не знают о содержании этих выводов и могли бы возражать против такого использования их личных данных в том случае, если бы их проинформировали.

Например, в Китае система QR-кодов была создана на основе цифровой платежной системы Alipay, платформы для мобильных и онлайн-платежей, которая была значительно расширена за счет создания инфраструктуры цифровой идентификации для хранения данных о здоровье граждан [Mozur, Zhong, Krolik, 2020]. В результате возник риск превращения этой системы в форму дополнительного контроля за действиями людей и даже применения в отношении отдельных лиц карательных мер со стороны государства в случае передачи собранных данных госорганам.

Подтверждением этого стало, например, признание правительством Сингапура в начале 2021 г. о том, что данные, полученные из приложения для отслеживания контактных лиц COVID-19 (Trace Together), могут быть доступны «для целей уголовного расследования», несмотря на предыдущие заверения в том, что этого не произойдет [Illmer, 2021].

Похожая ситуация складывается в сельскохозяйственном секторе. Фермеры беспокоятся, какие именно типы данных извлекает ИИ и как они будут распространяться и использоваться [Ryan, 2019, p. 27]. Многие из них жертвуют своей конфиденциальностью, соблазняясь преимуществами, которые сулит внедрение ИИ в процесс производства. Однако около 78% фермеров обеспокоены тем, что их данные передаются и продаются корпорациями [Stock, Gardezi, 2021, p. 193–203]. Агробизнес может недобросовестно использовать эти данные для перепродажи продукции фермерам на невыгодных для них условиях, для того чтобы привязать применение систем ИИ к поставкам семян и техники, а также использовать эти данные для скупки земли по очень низким ценам.

Из-за неправомерного доступа к базе данных систем ИИ возникает вопрос об ответственности за возможный вред, причиненный этими системами, который вновь отсылает к правовому регулированию процесса получения, хранения и передачи персональных данных в соответствии с уголовным, административным, трудовым и гражданским правом.

Глава III

ФИЛОСОФСКИЕ ПРИНЦИПЫ РАЗРАБОТКИ, ВНЕДРЕНИЯ И ПРИМЕНЕНИЯ СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

3.1 Условия эффективного и безопасного применения систем ИИ

Системы ИИ, широко применяемые в современном мире, помимо благ и преимуществ приносят в жизнь человека новые риски, поскольку их бездумное внедрение потенциально может увеличить количество несчастных случаев, а пострадавшие рискуют остаться без компенсации из-за сложности применения существующих режимов юридической ответственности, из-за нерешенности множества вопросов о статусе систем ИИ в современном социуме.

Иначе говоря, взаимодействие человека с этими системами нельзя назвать нейтральным, принципиально не способным вызвать негативные последствия. Напротив, оно таит в себе угрозы безопасности как личности, так и общества, мира в целом.

Реакцией социума на возникшую и постоянно возрастающую угрозу со стороны ИИ стали многочисленные попытки выработать перечень основополагающих принципов взаимодействия человека с искусственным разумом, необходимых для того, чтобы наиболее эффективным способом предотвращать и снижать возможность возникновения и эскалации социально-философских, этических проблем в процессе разработки, внедрения и применения систем ИИ во всех отраслях жизни современного общества. Эти принципы обусловлены проблемами применения систем ИИ и потому имеют коррелирующее с ними наименование: принципы непричинения

вреда, прозрачности, ответственности, конфиденциальности, справедливости и автономии. Необходимость решения задач обусловила разделение этой совокупности на две группы: 1) принципы, раскрывающие условия безопасного применения систем ИИ); 2) принципы, конкретизирующие различные грани, аспекты, стороны безопасного и эффективного применения ИИ.

Безусловно, каждый из названных принципов имеет множество собственных нерешенных проблем (например, проблема с определением самого понятия, реализацией принципа на практике и т. д.). Не вызывает сомнения и тот факт, что содержание каждого из них подлежит дальнейшему исследованию и уточнению, а перечень — пополнению.

Тем не менее выбор указанных шести принципов определен выводами и содержанием значительного количества источников различного происхождения, посвященных регулированию и применению систем ИИ, подтверждающих своим существованием актуальность и несомненную важность этих фундаментальных принципов взаимодействия человека с системами ИИ. В этих источниках зачастую используются различные термины для обозначения одних и тех же по содержанию норм и требований, поэтому для анализа в представленных ниже таблицах, предваряющих рассмотрение основных принципов применения систем ИИ, нами собраны близкие по значению термины, встречающиеся в такого рода документах. Отметим, что в своде документов авторы, как правило, лишь формулируют названные принципы, но не раскрывают их содержание, выявление и разъяснение которого стало целью проведенного нами исследования.

Мы предполагаем, что могут возникнуть вопросы, касающиеся неполноты перечня философских принципов. Например, почему устойчивость и экологическая ответственность не упомянуты в представленном списке. Мы обеспокоены экологическими проблемами и тем, что системы ИИ оказывают и будут оказывать влияние на окружающую среду, поэтому решили включить принцип эколо-

гической ответственности или устойчивости в принцип непричинения вреда. Подобным образом мы объединяли ряд принципов в один общий, если обнаруживали тождественность смыслов, общность содержания. Так, «подотчетность» была включена в принцип «ответственность». По той же логике японский принцип ИИ о добросовестной конкуренции, утверждающий, что не должно быть недобросовестного сбора данных и посягательства на суверенитет, был включен нами в принцип конфиденциальности.

В данном параграфе предметом исследования стали принципы, раскрывающие условия безопасного применения систем ИИ. Первый в названной группе — принцип прозрачности. Это требование, согласно которому основа конкретного решения системы ИИ всегда должна быть наглядна для пользователя этой системы. При этом должна быть обеспечена возможность выяснения причин в случае возникновения в ее работе какого-либо нарушения, сбоя. Принцип «прозрачности» заключается в утверждении, что системы ИИ должны разрабатываться и внедряться таким образом, чтобы можно было осуществлять надзор за их деятельностью.

Прозрачность	Объяснимость, открытость, отслеживаемость, предсказуемость, отчетность, понятность, интерпретируемость, открытые данные и алгоритмы, право на информацию, уведомление о принятии ИИ решения о личности, регулярная отчетность, проверяемость, инспектируемость
--------------	--

Прозрачность характеризует то, что представляют собой данные, которыми располагает система ИИ, где осуществляется сбор этих данных, что с ними происходит в процессе работы системы ИИ, а также то, как они ею используются [Digital Curation Centre. The University of Edinburgh The Role of Data in AI: Report for the Data Governance Working Group of the Global Partnership of AI. URL: <https://www.research.ed.ac.uk/en/publications/the-role-of-data-in-ai>]. На

международном форуме по борьбе с мошенничеством в государственном секторе принцип прозрачности обозначен как требование обеспечения объяснимости процесса работы систем ИИ и полученных результатов, причем как для самих экспертов, так и для широкой общественности [Office, 2020]. Участники форума пришли к выводу, что открытость алгоритма дает возможность отследить, по какому принципу, на каком основании было принято то или иное решение.

В Национальной стратегии развития искусственного интеллекта РФ на период до 2030 г. принцип прозрачности является одним из основных условий развития и внедрения систем ИИ [О развитии искусственного интеллекта в Российской Федерации: указ Президента РФ от 10.10.2019 № 490 (вместе с «Национальной стратегией развития искусственного интеллекта на период до 2030 г.»)]. URL: <https://base.garant.ru/72838946/>. В данном документе принцип прозрачности трактуется как объяснимость работы ИИ, процесса достижения им результатов, отсутствие дискриминации в доступе пользователей к продуктам, которые созданы с помощью ИИ.

Принцип прозрачности был также упомянут во многих этических рекомендациях, в том числе европейской комиссии по надежному ИИ [High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI. URL: <https://digitalstrategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>], Европейского парламента (2021 г.) [European Parliament Report on artificial intelligence in education, culture and the audiovisual sector Committee on Culture and Education. URL: https://www.europarl.europa.eu/doceo/document/A-9-2021-0127_EN.html], ЮНЕСКО в отношении проблем образования [Miao, Holmes, Huang, Zhang, 2021], в Пекинском консенсусе ЮНЕСКО (2019 г.) [Beijing Consensus on Artificial Intelligence and Education // International Conference on Artificial Intelligence and Education, Planning Education in the AI Era: Lead the Leap, Beijing. 2019. 70 p.] и «Принципах ответственного управления надежным ИИ» ОЭСР (2021 г.) [OECD's Principles for responsible stewardship of

trustworthy AI. URL: <https://oecd.ai/en/ai-principles>]. Однако описание принципа прозрачности в этих отчетах и руководствах различается. Например, если Европейская комиссия понимает под прозрачностью объяснимость элементов, относящихся к системе ИИ [High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI. URL: <https://digitalstrategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>], то проект ЮНЕСКО 2020 г. под прозрачностью имеет в виду связь с ответственностью и подотчетностью. В документе прозрачность признается необходимым условием выполнения принципа ответственности. В Европейской этической хартии принцип прозрачности считается важнейшим при использовании искусственного интеллекта в судебной и правоохранительной системах, а основными составляющими данного принципа объявлены объяснимость, понятность, доступность [Европейская этическая хартия об использовании искусственного интеллекта в судебных системах и окружающих их реалиях. URL: <https://rm.coe.int/rulethical-charter-en-version-17-12-2018-mdl-06092019-2-/16809860f4>].

Что касается доступности, то здесь неизбежно возникает целый ряд проблем. Даже если удастся объяснить содержательную составляющую алгоритмов ИИ, то по экономическим, юридическим или политическим причинам система ИИ все равно может стать непрозрачной.

Кроме того, продукты глубокой нейронной сети зависят от количества и уровней вычислений, которые связаны между собой столь сложным образом, что ни один ввод или вычисление не может по определению являться доминирующим фактором [Miller, 2019, p. 1–38]. Иначе говоря, надзор за деятельностью систем ИИ в отдельных случаях сложно осуществлять, поскольку эти системы могут принимать в значительной степени автономные решения. Такие системы ИИ представлены в самых разных областях, например в медицине в качестве диагностических инструментов для обнаружения диабетической ретинопатии [Abramoff, Lavin, Birch, Shah, 2018, p. 39], в сети Интернет в виде рекомендательных алгоритмов

YouTube [Bishop, 2018, p. 69–84], в предиктивном полицейском управлении при вынесении уголовных приговоров [Brayne, Christin, 2020, p. 608–624].

Высшая экспертная группа по искусственному интеллекту европейской комиссии признает, что существуют технические ограничения, для того чтобы система была прозрачной, и иногда невозможно изложить суть того, как система пришла к тому или иному решению. Вследствие этого при реализации принципа прозрачности появляется так называемая проблема «черного ящика», когда системы становятся неинтерпретируемыми и необъяснимыми, а также непредсказуемыми. Следовательно, они могут навредить людям, например, вследствие предвзятости решений исходя из анализа больших данных, который осуществляется порой по непонятным и недоступным человеку закономерностям.

Все это способно подорвать общественное доверие к ИИ, что, в свою очередь, может поставить под сомнение легитимность государственного сектора в целом. Иначе говоря, понятие прозрачности означает, помимо прочего, юридическую прозрачность, т. е. открытость кода программы, несмотря на утвержденные в законодательстве режимы интеллектуальной собственности, коммерческой тайны и т. п. В соответствии с указанными правовыми нормами никто не может иметь право доступа к исходному коду и технической документации продукта [Burrell, 2016, p. 1–12]. Даже оператор алгоритма не знает, как данные обрабатываются и как принимаются те или иные решения. Отсюда возникает необходимость в принудительном раскрытии принципов работы алгоритма в тех случаях, когда принятое ИИ решение способно оказывать существенное влияние на жизнь людей (наложение штрафов, диагностика болезней, назначение лечения) [Kaminski, 2019, p. 1529–1616]. Однако даже при наличии доступа к исходному коду алгоритма и коэффициентам, при высокой точности и надёжности применяемых алгоритмов все еще отсутствует гарантия объяснимости вследствие особенностей работы некоторых систем ИИ (в частности, основанных на искус-

ственных нейронных сетях), т. е. как они генерируют решения или прогнозы на основе имеющихся в их распоряжении входных данных, о чем мы уже упоминали.

По нашему мнению, надзор за деятельностью системы ИИ должен обеспечиваться в форме публичного раскрытия информации о рассматриваемой системе, ее процессах, прямом и косвенном влиянии на права человека и мерах, принимаемых для выявления и ослабления неблагоприятного воздействия системы на субъекта этих прав в случае возникновения негативной для пользователя ситуации. Кроме того, полученная информация должна быть полной и достоверной, позволяющей специалистам проводить содержательную оценку работы системы ИИ.

Например, если человеку отказано в получении кредита, то он вправе затребовать от кредитной организации информацию о причине отказа в удовлетворении кредитной заявки. Иначе говоря, человек не должен становиться жертвой автоматизированной обработки данных с использованием ИИ. При ее осуществлении должно быть полное информирование субъекта данных, а также возможность ознакомления с принятым системой решением, и человек имеет право оспаривать его.

Особое место принцип прозрачности занимает в Кодексе этики для разработчиков робототехники, который считается сегодня одним из наиболее комплексных и доработанных документов в области ИИ [Civil Law Rules on Robotics...]. В данном документе авторы, рассматривая принцип прозрачности, различают техническое и организационное направление. Последнее касается разработчиков и их обязанности предоставлять необходимую информацию пользователям, оповещать об угрозах и рисках, которые могут создавать системы ИИ. Техническое направление касается вопроса о предоставлении доступа к алгоритму в случае причинения вреда системой ИИ, чтобы обеспечить возможность отследить причины возникновения сбоя в работе системы.

В «Общем регламенте по защите данных ЕС» к вопросу прозрачности и подотчетности алгоритмов относится положение о том, что субъект вправе получать информацию «о наличии системы автоматизированного принятия решения, включая профилирование, а также о значимости и предполагаемых последствиях такой обработки для субъекта данных» [General Data Protection Regulation (GDPR). URL: <http://www.gdpr.eu/doc/gdpr%20reglament.odt>. Ст. 13–15]. Это положение интерпретируют по-разному: либо пользователю раскрывается вся логика принятия решения [Malgeri, Comandé, 2017, p. 243–265], либо пользователя уведомляют в самых общих чертах о процессе принятия решений алгоритмом [Wachter, Mittelstadt, Floridi, 2017, p. 76–99].

Некоторые авторы полагают, что прозрачность может относиться либо в целом к системе, либо к принятию отдельных решений [Koene, Clifton, Hatada, Webb, Richardson, 2019]. При этом они утверждают, что обеспечение полной прозрачности работы алгоритма не имеет смысла и слишком ресурсозатратно, к тому же не ведет к значимому объяснению логики принятия решений. В литературе даже упоминается термин «принцип относительной прозрачности» [Бахтеев, Тарасова, 2020, с. 249–254].

В документе «Этически обусловленное проектирование» (IEEE) [Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. Version 2 / The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. URL: http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html] и Белой книге ЕС [European Commission. White paper on artificial intelligence: A European approach to excellence and trust, 2020. URL: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf] прозрачность определяется как один из восьми всеобъемлющих принципов, как способность обнаруживать основу решения, принятого системами ИИ. IEEE рекомендуют разработать «новые стандарты, описывающие измеримые, поддающиеся проверке уровни прозрачности, позволяющие объек-

тивно оценивать работу системы». В «Руководящих принципах по этике для надежного ИИ» прозрачность имеет решающее значение для установления и поддержания доверия пользователей к системам ИИ [High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI. URL: [https:// digitalstrategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai](https://digitalstrategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai)]. Это означает, что процессы должны быть прозрачными, возможности и назначение систем искусственного интеллекта — открытыми, а решения, насколько это возможно, — объяснимыми для тех, кого они прямо или косвенно затрагивают. Без такой информации решение не может быть должным образом обжаловано.

В Азиломарских принципах под прозрачностью понимается возможность выяснения причин при сбоях в системе, требование того, чтобы процесс работы происходил под наблюдением специалиста, если он желает уточнить детали [Asilomar AI Principles. URL: <https://www.artificial-intelligence.blog/news/asilomar-ai-principles>]. Иными словами, любой процесс принятия решений системами ИИ, оказывающий существенное влияние на права человека, должен быть идентифицируемым. Люди должны иметь возможность понять, как принимаются решения и как они проверяются. Это, конечно, не значит, что пользователю должны выдавать информацию о коэффициентах нейронной сети, на которой построен ИИ. Это означает, что общие принципы работы должны быть понятными и ясными, чтобы была возможность общественного контроля за работой этих систем.

Интересно, что в рассматриваемом документе отдельно формулируется требование к языку описания работы систем ИИ, использование которых, по мнению авторов, можно объяснить с помощью четких и доступных терминов.

Разъяснение смысла этого требования, по нашему мнению, содержат «Принципы подотчетных алгоритмов» FAT-ML, в которых прозрачность означает обеспечение того, что решения систем ИИ могли быть изложены конечным пользователям и другим заинтере-

сованным сторонам с помощью нетехнических терминов [Fairness, Accountability, and Transparency in Machine Learning. URL: <https://www.fatml.org/>]. При этом заинтересованным сторонам должна быть предоставлена информация, которая позволяет объяснить общее функционирование системы, конкретное использование данных в системе и отдельные решения, принимаемые системой.

На наш взгляд, необходимые технические и организационные меры должны быть направлены на обеспечение прозрачности с самого начала разработки системы ИИ, а не только после завершения. Проектировщики систем и заинтересованные стороны должны изначально подумать о том, как сделать информацию о стандартах принятия решений доступной и понятной для самых различных категорий пользователей, а также, например, для регулирующих и контролирующих органов власти.

При разработке систем ИИ следует учитывать требования к прозрачности, сложность процесса принятия решений. В зависимости от информационной цели и аудитории стандарты принятия решений могут быть представлены в виде общих принципов высокого уровня или более конкретных и подробных стандартов. Даже если добиться соответствующего уровня прозрачности не удастся, необходимо стремиться к возможному уменьшению непрозрачности в работе системы ИИ. Так, IBM в 2018 г. признал, что хотя предубеждения и дискриминация никогда не могут быть полностью устранены в системах ИИ, необходимо приложить усилия для их уменьшения или ослабления [IBM. Principles for trust and transparency. URL: https://www.ibm.com/blogs/policy/wp-content/uploads/2018/05/IBM_Principles_OnePage.pdf].

Особенно важно сделать прозрачными риски, связанные с работой системы ИИ. Так, риски, обусловленные автоматизированным принятием решений, рассматривались в трудах Т. Араужо [Araujo, Helberger, Kruijkemeier, De Vreese, 2020, p. 611–623] и Дж. Бахнера [Bahner, Hüper, Manzey, 2008, p. 688–699] (например, нарушения конфиденциальности, дискриминация и предвзятость в принятии

решений). Отметим, что степень обеспечения прозрачности в данном контексте в значительной степени зависит от сферы применения систем ИИ. Например, в сфере медицины требование прозрачности представляется важным и необходимым. Пациент должен обладать максимально подробной и полной информацией о том, какие операции будут проводиться и какие последствия при этом могут быть. Это чрезвычайно важно, чтобы оценить возможные риски, возникающие вследствие использования ИИ.

С юридической точки зрения до сих пор не ясно, какие средства правовой защиты будут у людей против непрозрачных систем ИИ. Дискриминационная практика, возникающая в результате использования таких технологий, недостаточно описана, например, в законе о защите данных [Wachter, Mittelstadt, 2019, p. 494–620].

В целях обеспечения прозрачности в работе систем ИИ важно, чтобы деятельность их разработчиков тоже была понятной и доступной для заказчиков и потенциальных пользователей посредством предоставления подробных отчетов, в которых содержится описательная и сводная информация о системе ИИ с точки зрения ее внедрения и использования. Например, разработчики алгоритмов уголовного правосудия и инструментов оценки криминального риска должны публиковать сведения о том, где используются эти инструменты (т.е. в каких юрисдикциях), сколько решений было принято системой и какими они были. Там, где это возможно, следует оценивать качество принятия решений системами ИИ путем сравнения их с результатами, полученными вследствие применения других подходов к принятию решений в данной области. Такие отчеты могут публиковаться с разной периодичностью в зависимости от важности и своевременности информации.

Возможно, решению проблемы прозрачности могло бы содействовать закрепление требования к разработчикам предоставлять информацию о последовательности операций, которые осуществляются системой ИИ, и результатах, к которым они приводят. В описательной схеме операций, которую можно составить в виде

блок-схемы, необходимо отразить главные вариации решений, к которым система ИИ приведет на определенных стадиях своих вычислений. Здесь важно не допустить предельного упрощения при объяснении технического содержания работы системы или, напротив, предельного усложнения, когда даже экспертам будет сложно понять принцип работы ИИ. Такого обеспечения трудно достичь в случае систем, основанных на методах машинного и глубокого обучения, так как алгоритм может обучать и выявлять при обработке данных такие закономерности, которые могут быть неизвестны нам.

Интеграция прозрачности в процесс проектирования и внедрения систем искусственного интеллекта — непростая задача. Скорость технологического развития, многочисленные аспекты концепции прозрачности, неопределенность в отношении того, где требуется прозрачность, как лучше всего общаться с различными заинтересованными сторонами и как интегрировать прозрачность в значимые и организационно реалистичные меры подотчетности — все это существующие проблемы, ожидающие практической реализации.

Вероятно, можно было бы повысить прозрачность работы любых систем ИИ, если бы это было закреплено в нормативных правовых актах, разного рода конвенциях, требованиях и т. д. Людям тогда, например, было бы проще оспаривать решения, принятые ИИ и причинившие им какой-либо ущерб. Мы считаем, что в любом случае, особенно в сферах, где жизнь человека зависит от принятого системой решения, окончательный вердикт всегда остается за человеком.

Проектирование и разработка систем ИИ высокого риска должны осуществляться таким образом, чтобы изначально обеспечить необходимую прозрачность их функционирования, позволяющую пользователю интерпретировать результаты работы систем ИИ на любом этапе. Для достижения этой цели в процессе разработки необходимо создание междисциплинарных групп специалистов в области программирования, философии, права, психологии и т. д., что позволит принимать более продуманные, этичные, а также со-

ответствующие существующим в обществе юридическим, политическим, культурным нормам и требованиям решения.

Важно отметить, что не системы ИИ, а люди должны быть ответственными за их внедрение и обеспечить прозрачность, т. е. объяснить, что они решают и делают. Принцип ответственности в данном случае коррелирует с понятием подотчетности, когда разработчику необходимо предоставить пользователю нужную информацию. Технологии и технологическое образование должны помочь пользователям и разработчикам ИИ противостоять любым вызовам. Кроме того, обществу нужно задуматься о том, следует ли допускать высокую степень автоматизации в случаях, когда важна мгновенная скорость решения (например, в военном деле), на что может не хватить временного ресурса.

Философский принцип прозрачности подразумевает, что по мере возможности во всех случаях применения систем ИИ человеку, пользователю:

- должна быть предоставлена адекватная информация о характере и функциональности системы ИИ, о возможных вариантах изменения функций системы, а также сведения о данных, которые обрабатывает система ИИ;

- должна быть обеспечена открытость в отношении логической схемы принятия системой ИИ решений;

- известно о потенциальных рисках и угрозах принимаемых системой ИИ решений;

- эти данные должны быть переданы на доступном и понятном языке;

- должна быть предоставлена информация о механизме подачи жалоб и претензий, этапах прохождения этой юридической процедуры, точных полномочиях контактных лиц, ориентировочных сроках и ожидаемых результатах.

Без реализации принципа прозрачности невозможно осуществить ведущий принцип — непричинение вреда. Соответственно, от этого зависит дальнейшее развитие и внедрение систем ИИ.

Несоблюдение принципа прозрачности и отсутствие достаточной степени объяснимости, в свою очередь, создает проблему, связанную с ответственным использованием ИИ. Здесь важно уметь объяснить принципы принятия решения. Прозрачность обуславливает реализацию требования ответственности, сознательного поведения. Эта взаимосвязь прозрачности и ответственности подводит к следующему принципу — ответственности.

Ответственность	Подотчётность, моральная и правовая ответственность, добросовестное принятие решений
-----------------	--

По мере совершенствования систем ИИ и увеличения их влияния на жизнь социума возникает необходимость переосмысления понятия ответственности. В научной литературе обсуждение понятия часто сводится к обвинению кого-либо в безответственности, однако мы полагаем, что это не способствует раскрытию содержания термина. Некоторые исследователи поставили перед собой задачу сформулировать новую этику, соответствующую технологическим реалиям современности (Р. Г. Рополь, А. Маттиас, Г. Йонас). Они полагают, что новая этика ответственности должна быть нацелена на сохранение будущего человечества, на осмысление отдаленных последствий осуществляемой им технической, технологической деятельности, на предотвращение, недопущение непредсказуемых негативных ее результатов.

Ответственность, рассматриваемая сквозь призму последствий, приводит к изменению временного горизонта, а также объекта и субъекта ответственности. Г. Рополь, в частности, пришел к выводу, что человек, совершающий действие, оказывающее определенное воздействие на нравственно значимые ценности, может быть привлечен к ответственности, может быть субъектом ответственности [Ropohl, 1993, p. 149–176]. Для ученого «быть ответственным» означает, что субъекты должны оправдывать свои действия перед пострадавшей стороной. При этом, по его мнению, ответственность может быть наложена ретроспективно (после) или перспективно

(заранее), поэтому необходимо в обязательном порядке учитывать временной аспект.

С ним согласна Х. Арендт, которая также считает, что человек, меняя мир, должен нести (не только принимать) ответственность за изменения вследствие его деяний [Arendt, 2003]. Иными словами, новатор несет ответственность за свои действия, которые ведут к переменам, даже в том случае, если последствия проявятся значительно позже, то есть ответственность должна быть пролонгирована в будущее, на весь период результатов последствий деятельности новатора.

Еще одна проблема, возникающая в связи с толкованием ответственности, — это проблема субъекта ответственности. Например, в геномной инженерии, где широко обсуждается вопрос кризиса человеческой идентичности вследствие биологических трансформаций человека, сложно определить, кто выступает субъектом ответственности, на кого она должна быть возложена. Более того, сложно и в отношении понимания морали как одного из сущностных признаков человека, своего рода моральной способности человека [Хабермас, 2002].

Ответственность в профессиональной этике часто обсуждается применительно к коллективу, как групповая ответственность. При этом Х. Ленк, в частности, полагает, что ответственность здесь зачастую слишком абстрактно описана. Фактически она возложена на разработчика, инженера, непосредственного исполнителя, имеет место и чрезмерное делегирование ответственности инженеру [Ленк, 1996]. Решая эту проблему, некоторые исследователи считают, что в данном случае ответственность инженера становится ролевой. Он отвечает за свои действия только перед начальством, организацией, но не перед обществом в целом. Коллективный характер ставит под сомнение саму реализацию моральной ответственности, провоцирует безнаказанность, вседозволенность, что едва ли может способствовать достижению социального благополучия и отвечать критериям прогрессивного развития общества.

Ряд теоретиков в отношении этого вопроса придерживаются позиции, согласно которой приписать коллективу моральную ответственность можно, так как, взаимодействуя с коллективом, мы больше не взаимодействуем с отдельным индивидом (Э. Штрекер [Stroker, 1984, p. 64], Х. Ленк (1996)).

С ними не согласны другие исследователи, полагающие, что невозможно возложить ответственность на коллектив, поскольку нельзя игнорировать личностный аспект, решения и поступки каждого отдельного индивида [Neimaier, 1990, p. 43–54]. Возможно, решением может послужить то, что возложение моральной ответственности должно быть реализовано, но без разделения на части. Другими словами, ответственность приписывается не целому коллективу, а именно тем лицам, которые причастны.

Вопрос о том, в какой степени люди могут или должны нести ответственность за поведение ИИ, стал одним из самых обсуждаемых в сфере ИИ [Special Interest Group on Artificial Intelligence..., 2018; MI Garage Ethics Framework - Responsible AI. MIGarage. URL: <https://www.migarage.ai/ethics-framework/>; Accenture UK. Responsible AI and robotics. An ethical framework. URL: <https://www.accenture.com/gb-en/company-responsible-ai-robotics>]. В настоящее время существуют разные подходы к установлению ответственности за вред, причиненный системами ИИ. Например, по мнению Р. В. Душкина, автономная система ИИ, причинившая вред ее пользователю, должна сама нести за это ответственность [Этичное применение искусственного интеллекта. URL: <https://ethics.cdto-center/>]. Обычно провинившийся человек возмещает свой ущерб в виде выплаты штрафа или отбывания в тюрьме. Соответственно система тоже должна быть оштрафована. Например, беспилотному автомобилю требуется обслуживание, которое стоит денег. Система зарабатывает их своей функциональностью, собирая с пассажиров деньги, которые накапливаются и являются для нее ценностью. Если система причинит вред использующему ее человеку, то должна оплатить штраф.

А. А. Щитова также считает, что система ИИ потенциально может быть ответственной и возмещать ущерб материально. Следовательно, по мнению ученого, ее можно наделить определенным правом собственности в отношении материального имущества [Щитова, 2021]. Но если система ИИ не владеет этим имуществом, то следует ее привлечь к обязательным работам, наложив запрет заниматься определенными видами деятельности на установленный срок.

А. В. Незнамов считает возложение ответственности непосредственно на систему ИИ своего рода крайним вариантом, которого следует избегать, равно как и другой крайности, когда ответственность не несет никто, когда виновника обнаружить не удастся в принципе [Незнамов, 2020, с. 171–185].

В 2004 г. опубликована работа А. Маттиаса, в которой автор обсуждает «разрыв ответственности» во взаимодействии с обучаемыми системами ИИ [Mathias, 2004, p. 175–183]. Он указал на то, что ни один человек не может быть законно обвинен или признан виновным в нежелательных результатах действий, опосредованных системами ИИ. Проблема заключается в том, что поведение системы ИИ, использующей сложные алгоритмы, слишком автономно и слишком непредсказуемо, чтобы кто-либо из людей, внесших в ее разработку и создание свой вклад, мог в дальнейшем быть ответственным за результаты работы этой системы. В этом случае выявить субъекта ответственности не представляется возможным.

Также существуют работы, в которых предприняты попытки уточнить распределение ответственности между разными субъектами [Guidelines for artificial intelligence. Deutsche Telekom. URL: <https://www.telekom.com/en/company/digital-responsibility/details/artificial-intelligence-aiguide-524366>]. Речь идет и об юридической ответственности [Responsible bots: 10 guidelines for developers of conversational AI. URL: <https://www.microsoft.com/en-us/research/publication/responsible-bots/>].

Ответственность, связанная с применением ИИ, часто рассматривается не только в научных работах, но и в различных кодексах и

руководствах по этике ИИ и других видах мягкого права. Однако в большинстве случаев трактовка ответственности чрезмерно абстрактна и потому неприменима к конкретным ситуациям. Кроме того, содержание понятия ответственности в такого рода документах, как правило, предусматривает ее нормативный или профессиональный аспект, что сужает понимание ответственности и исключает из ее толкования любые другие смыслы. Например, в Азиломарских принципах ответственность описана буквально одним предложением, что не раскрывает содержания принципа ответственности: «разработчики продвинутых систем ИИ играют ключевую роль в формировании нравственных последствий использования ИИ, неправильного использования ИИ и действий ИИ; они имеют возможность и несут обязанность влиять на такие последствия» [Asilomar AI Principles. URL: <https://www.artificial-intelligence.blog/news/asilomar-ai-principles>].

Отвечая на вопрос, может ли субъект ответственности быть представлен системой ИИ, мы считаем, что вспомогательные системы ИИ (те, которые основаны не на машинном обучении) не могут быть привлечены к ответственности, поскольку не обладают ни свободой воли, ни сознанием. Соответственно, ответственность в этом случае должна быть возложена на людей, которые разрабатывают и используют технологии ИИ.

Проблемы с возложением ответственности возникают в случае с самообучающимися системами ИИ, которые обладают наибольшей автономностью по сравнению с предыдущими, их принцип работы не до конца понятен даже самим разработчикам. Более того, даже они не имеют достаточного контроля над действиями таких систем ИИ.

Рассмотрим аварию с беспилотным автомобилем, в которой машина сбивает пешехода. На кого возложить ответственность? Кто виноват в случившемся?

Другой пример с чат-ботом Tay от Microsoft Twitter, который после взаимодействия с пользователями начал высказывать расистские и женоненавистнические комментарии [URL: <https://www>.

techinsider.ru/technologies/237098-robot-obshchayas-v-tvittere-za-sutki-stal-rasistom-i-chelovekonenavistnikom/]. В разбирательстве этого дела участвовали разработчики, дизайнеры, компании и пользователи. Суароз-Гонзало, анализируя этот инцидент, отмечает, что ответственность должна быть возложена в основном на дизайнеров, разработчиков, а не на пользователей Twitter, взаимодействовавших с ботом [Suárez-Gonzalo, Mas-Manchón, Guerrero-Solé, 2019, p. 1–14].

Р. Спарроу уверен, что в подобных ситуациях необходимо вовсе препятствовать использованию любых автономных систем, особенно связанных с вооружением [Sparrow, 2007, p. 62–77]. По его мнению, разработка все более автономных систем вооружения неэтична еще и потому, что в данном случае трудно выявить лицо, ответственное за гибель людей в ходе войны [Sparrow, 2009, p. 25–29]. Исчезает сама возможность привлекать к ответственности конкретных лиц, виновных в совершении преступлений, противоправных деяний.

А. Гринбаум предлагает рассматривать системы ИИ в качестве «цифровых особ», которые в отличие от всех других видов техники алгоритмически наделены способностью к машинному обучению, соответственно, обладают определенной степенью автономии и способностью принимать самостоятельные решения [Гринбаум, 2017]. Указанные системы не обладают сознанием, но при этом не находятся под полным контролем человека и совершают действия, не вполне прозрачные для него. Эти системы основаны на многослойных технологиях обучения (например, на глубоком обучении), где выдача строгой математической модели невозможна. Даже эксперт не способен объяснить связь между исходными данными и конечным результатом. Следовательно, такого рода системы могут вызывать особые опасения, например в области медицины, когда жизнь и здоровье человека находятся под контролем машины. Работая с подобными системами, люди должны быть осведомлены о том, что они не подконтрольны человеку. Задачей разработчика подобных систем является необходимость избавить других и себя от

иллюзии детерминизма в процессе разработки и взаимодействия с такого рода системами ИИ и четко определить те виды деятельности, которые человек готов делегировать указанным системам, не беспокоясь о возможных последствиях принятых ими решений.

Обобщая вышесказанное, можно сделать вывод, что философский принцип ответственности устанавливает требование, согласно которому вовлеченные в создание ИИ-систем стороны обязаны нести моральную и юридическую ответственность за все действия, совершенные такими устройствами на всех этапах их жизненного цикла. При этом необходимо понимать, что цель исследования, проектирования, создания, развертывания и применения ИИ-систем состоит в том, чтобы помочь людям принимать решения и выполнять определенные функции, но не заменить человека, на которого всегда в конечном итоге будет возложена правовая и моральная ответственность за принятые решения и совершенные либо несовершенные деяния.

Так, в Азиломарских принципах утверждается, что разработчики и создатели продвинутых систем ИИ должны нести полную ответственность за действия, за последствия использования и за возможные злоупотребления, допущенные этими системами. Аналогичное требование можно найти в Монреальской декларации [Montréal Declaration: Responsible AI. URL: https://monoskop.org/images/d/d2/Montreal_Declaration_for_a_Responsible_Development_of_Artificial_Intelligence_2018.pdf].

На наш взгляд, окончательное решение всегда должен принимать человек, оно должно быть свободным и осознанным. При этом ответственность также берет на себя человек, но ни в коем случае системы ИИ. Ответственность за преступление или правонарушение — лица, которые уполномочивают ИИ на совершение преступления или правонарушения или проявляют халатность, позволяя ИИ совершать противоправные деяния. Таким образом, содержание философского принципа ответственности может быть сформулировано следующим образом. Заинтересованные стороны (разработчи-

ки, пользователи) должны нести правовую и моральную ответственность за то, чтобы системы ИИ использовались в надлежащих условиях и четко выполняли ряд поставленных им задач. При этом должно быть определено конкретное лицо, на которого может быть возложена ответственность в случае возникновения сбоев в работе систем ИИ.

Оценка и испытание систем ИИ на этапе разработки и внедрения должны быть осуществлены человеком. Иными словами, именно он должен в конечном итоге подтвердить безопасность и эффективность систем ИИ для пользователей.

Организации, внедряющие системы ИИ, должны предусмотреть выполнение обязательных работ по разбору, анализу и выплате компенсации пострадавшим лицам и организациям в случае, если системой ИИ причинен вред человеку.

Принцип конфиденциальности

Конфиденциальность	Приватность, защита данных, личная, частная информация, защита от утечки персональных данных
--------------------	--

Системы ИИ аккумулируют и используют конфиденциальные данные о человеке, такие как имя, адрес пребывания, цифровой след в сети, а также информацию, которая собирается через различные приложения. Здесь возникает вопрос «Каким образом использование систем ИИ отразится на нас: приведет ли это к лучшему пониманию общественных процессов, более взвешенному принятию решений или же завершится эксплуатацией, контролем со стороны властных структур?».

К. Керри отметил, что использование ИИ может привести к нарушению конфиденциальности, выводя анализ личных данных на новый уровень мощности и скорости [Kerry, 2020]. Шеннон Валлор пришел к выводу, что технологии наблюдения, которые нарушают конфиденциальность, в долгосрочной перспективе могут препятствовать моральному и культурному росту [Shannon, 2016].

Реган утверждает, что конфиденциальность присуща индивидууму как личности и важна для его саморазвития и установления отношений с другими членами социума [Priscilla, 1995]. Э. Блаустейн характеризует конфиденциальность как сохранение независимости, достоинства и целостности человека [Bloustein, 1984, p. 156–202]. Таким образом, конфиденциальность является важнейшим условием сохранения личного достоинства человека, что предполагает осознание им собственной ценности, уникальности.

Также предпринимались попытки более точно определить конфиденциальность, но чаще всего она связана с защитой данных [Internet Society. Artificial Intelligence&Machine Learning: Policy Paper. Internet Society. URL: <https://www.internetsociety.org/resources/doc/2017/artificial-intelligence-andmachine-learning-policy-paper/>; Asilomar AI Principles. URL: <https://www.artificial-intelligence.blog/news/asilomar-ai-principles>; Microsoft. Responsible bots: 10 guidelines for developers of conversational AI. URL: https://www.microsoft.com/en-us/research/uploads/prod/2018/11/Bot_Guidelines_Nov_2018.pdf] и их безопасностью [Dawson, 2019; Sony. Sony Group AI Ethics Guidelines. URL: <https://www.sony.com/en/SonyInfo/sonyai/guidelines.html>].

В настоящее время в обществе достигнуто согласие относительно необходимости контроля за соблюдением конфиденциальности на законодательном уровне, что выгодно отличает рассматриваемый принцип от всех остальных. В то же время в сфере бизнеса многие компании выступают за смягчение правового регулирования в области конфиденциальности, поскольку излишне жесткие требования защиты персональных данных ограничивают развитие этих компаний. Многие страны, напротив, настаивают на сохранении существующих требований к обезличенности общедоступных данных. Но проблема заключается в том, что это далеко не всегда гарантирует их конфиденциальность. Так, в сфере действия правоохранительных органов и правосудия нередко закрытые в соответствии с законом персональные данные оказываются доступны всем заинтересованным сторонам. По сетевым следам специалистам не

трудно получить информацию о пользователе и его действиях в сети. Это означает, что юридически конфиденциальность данных подлежит защите, однако на деле эти нормы постоянно нарушаются.

На наш взгляд, юридический контроль должен быть направлен непосредственно на то, в каких конкретных случаях будут использоваться данные. В. Майер-Шёнбергер, философ в области больших данных, утверждает: «Разумнее было бы отменить практику индивидуального управления конфиденциальностью и заменить ее расширенной подотчетностью, которая предъявлялась бы к пользователям данных, повышая их ответственность за свои действия. Компании, работающие с данными, больше не смогли бы приводить в свое оправдание то, что человек разрешил их использовать. Напротив, им пришлось бы оценивать потенциальные опасности, с которыми могут столкнуться люди при вторичном применении их данных» [Майер-Шёнбергер, Кукьер, 2014].

Особенно важно получать согласие пользователей на обработку их личных данных и не преступать при этом так называемый порог согласия. Социальное напряжение часто возрастает при возникновении экстренных ситуаций: когда люди находятся на грани жизни и смерти, в условиях катастрофы или пандемии. Например, во время вспышки лихорадки Эбола в 2014 г. срочное рассекречивание и анализ записей звонков пользователей мобильных телефонов в регионе позволили эпидемиологам остановить быстрое распространение опасной болезни. Но публиковать данные запретили из-за риска нарушения конфиденциальности пользователей, а также по причине потенциальной ценности данных для промышленных конкурентов.

В Монреальской декларации принцип конфиденциальности является одним из ключевых, а разработка систем ИИ, по мысли создателей декларации, должна предотвращать риски неправомерного использования персональных данных и защищать их целостность и конфиденциальность [Montréal Declaration: Responsible AI. URL: https://monoskop.org/images/d/d2/Montreal_Declaration_for_a_Responsible_Development_of_Artificial_Intelligence_2018.pdf]. Согласно до-

кументу, люди всегда должны иметь право на «цифровое отключение» в своей личной жизни, ИИ должен сам предлагать эту возможность через регулярные промежутки времени, не побуждая людей постоянно оставаться на связи. Их мысли и эмоции, переживания должны быть строго защищены от ИИ, способного не только помочь, но и причинить вред. Люди должны иметь полный контроль над информацией о себе и своих предпочтениях. ИИ не должен создавать профили индивидуальных предпочтений, чтобы влиять на поведение людей без их свободного и информированного согласия.

Положение об информированном согласии входит в перечень основных принципов программы «Beijing AI Principles». В базовом документе указанной программы отмечается, что «стороны должны иметь достаточное информированное согласие о влиянии системы на их права и интересы» [Beijing AI Principles. URL: <https://www.baai.ac.cn/news/beijing-ai-principles-en.html>]. Однако, несмотря на это, сегодня в Китае правительство активно использует данные граждан, собирая информацию об их повседневной жизни: досуге, покупательской активности, а также активности в сети [Chorzempa, Triolo, Sacks, 2018]. На практике пользователь, предоставивший информированное согласие, не осведомлен о том, как именно его данные будут использоваться, для каких задач и в каких ситуациях. Возможно, одни зарабатывают и получают незаконную прибыль, а другие не получают никакой компенсации, а, напротив, еще и платят за установку приложений и устройств. Добавим к сказанному и то обстоятельство, что информированное согласие на обработку данных автоматически возлагает всю ответственность за дальнейшее использование информации на самого пользователя. Например, подобное происходит в тех случаях, когда люди, предоставляющие такое согласие, считают, что оно является условием работы приложения, без которого нужный им процесс не запустится [Большие данные в социальных и гуманитарных науках..., 2019].

В Азиломарских принципах под конфиденциальностью подразумевается, что человек должен иметь доступ к персональным дан-

ным и их изменению, а также возможность отслеживать их и контролировать [AsilomarAIPrinciples. URL: <https://www.artificial-intelligence.blog/news/asilomar-ai-principles>].

Последнее вполне осуществимо благодаря механизму «динамического согласия» [Kaye, Whitley, Lund, Morrison, Teare, Melham, 2015, p. 141–146] (в некоторых источниках «дифференцированного согласия» [Contract for the Web. URL: <https://9nrane41lq4966uwmljcfggvwpengine.netdna-ssl.com/wp-content/uploads/Contract-for-the-Web-3.pdf>]), предоставляющего пользователю возможность отслеживать и корректировать свои предпочтения в отношении конфиденциальности более детально. Человек вправе разрешать или запрещать сбор данных. К примеру, он может рассчитаться за пользование программой и при этом не предоставлять системе свои данные. Может показаться, что в данном случае пользователь оказывается не в лучшем положении, лишается каких-либо преимуществ, но на самом деле именно такой тип отношений свидетельствует о признании ценности личных данных и праве пользователя распоряжаться ими по своему усмотрению.

Таким образом, в процессе работы системы ИИ с персональными данными существуют риски нарушения конфиденциального характера этих данных. Однако при должном регулировании они могут быть успешно устранены. Важно повысить ответственность перед законом всех лиц, отвечающих за разработку и дальнейшее использование систем ИИ в различных отраслях и сферах общественной жизни.

ЕС рассматривает в качестве главного нормативного акта в сфере защиты личных данных регламент европейского парламента и Совета Европейского союза 2016 г. «О защите физических лиц при обработке персональных данных и о свободном обращении таких данных, а также об отмене Директивы 95/46/ЕС». Он определяет процесс обработки личных данных, который проводится полностью либо частично при помощи автоматизированных средств. Регламентом предусмотрен комплекс мер, направленных на защиту пользо-

вателей от каких-либо действий с их биометрическими данными, предпринимаемых третьими лицами или организациями без их прямого согласия.

В российском законодательстве отсутствуют акты, регулирующие деятельность по организации и осуществлению видеонаблюдения, использованию биометрических данных. Существуют разрозненные рамочные упоминания о защите данных в отдельных законах, но только малая часть их имеет отношение к системам ИИ (например, ст. 152.1 ГК РФ, ст. 3 ФЗ РФ от 27.07.2006 г. № 152-ФЗ «О персональных данных»).

Принцип конфиденциальности, на наш взгляд, является одним из самых труднореализуемых на практике, поскольку это может противоречить действию других принципов. Например, повышение прозрачности автоматических решений может привести к невозможности реализации принципа конфиденциальности. При этом в каждом отдельном случае требуется разработать конкретные правила профилирования, предусмотреть гарантии эффективной защиты данных, определить перечень и меру ответственности тех лиц, кто отвечает за работу с персональными данными пользователей. Здесь важен правильный выбор в пользу сотрудничества, а не чрезмерной специализации систем ИИ. Например, эксперт по обеспечению справедливости алгоритмов ИИ и эксперт по конфиденциальности данных должны иметь возможность сотрудничать для разработки модели машинного обучения, которая одновременно является и справедливой, и обеспечивающей конфиденциальность личных данных пользователя. Не менее важно, чтобы решения ИИ контролировались человеком, когда они могут затронуть интересы отдельных лиц или группы. Особой проблемой является проверка качества и эффективности человеческого контроля, а также выбор момента его вмешательства (до того как решение будет принято машиной или после завершения рабочего процесса). Подводя итог вышесказанному, можно сформулировать философский принцип конфиденциальности как обращение к разработчикам и пользователям с требованием:

– предотвращать риски неправомерного сбора, хранения и использования пользовательских данных и защищать их целостность и конфиденциальность на протяжении всего жизненного цикла: от начала и до конца ее обработки;

– обеспечить в процессе разработки систем ИИ надлежащие механизмы управления данными на всех этапах жизненного цикла проектируемых ИИ-систем, в том числе касающиеся сбора данных, контроля использования данных на основе информированного согласия и разрешений, обеспечения личных прав пользователей на владение собственными данными и получение доступа к ним, а также процедуры раскрытия информации о применении и использовании данных;

– усилить юридическую ответственность лиц, отвечающих за разработку и внедрение в социальную практику систем ИИ, в частности, за работу с персональной информацией пользователей, установить определенные правила профилирования, чтобы гарантировать надежную защиту данных человека;

– обеспечить максимальную анонимизацию или неидентифицируемость данных, что вполне возможно в процессе сбора данных, обработки информации.

3.2 Принципы справедливости, автономии и непричинения вреда человеку в контексте использования ИИ

Обозначенные принципы конкретизируют различные грани, аспекты, стороны безопасного и эффективного применения ИИ.

Принцип справедливости

Справедливость	Недискриминация, равенство, солидарность, разнообразие, инклюзивность, честность, исключение предрассудков, беспристрастность, стандарты для всех, минимизация предвзятости
----------------	---

В философском энциклопедическом словаре справедливость определяется как «понятие о должном, соответствующее определениям о сущности человека и его неотъемлемых правах... содержит в себе требования соответствия между практической ролью различных индивидов (социальных групп) в жизни общества и их социальным положением, между их правами и обязанностями, между деянием и воздаянием, трудом и вознаграждением, преступлением и наказанием, заслугами людей и их общественным признанием [Ильичев, Федосеев, Ковалев, 1983, с. 622]. Справедливость в отношении систем ИИ означает предотвращение и устранение предвзятости, равное отношение к группам и индивидам. Философский принцип справедливости призывает к тому, что системы ИИ должны быть разработаны и обучены таким образом, чтобы не создавать, не усиливать и не воспроизводить дискриминацию, основанную на социальных, половых, этнических, культурных, религиозных или иных различиях, а, напротив, способствовать ее устранению либо сокращению.

Внедрение систем ИИ должно помочь предотвратить отношения доминирования между группами и людьми, основанные на различиях во власти, богатстве или знаниях. Развитие ИИ должно способствовать получению социальных и экономических выгод для всех посредством сокращения существующего социального неравенства и социальной уязвимости отдельных лиц, социальных групп и слоев общества. ИИ не должен отдавать предпочтение, предоставлять исключительные права, преференции каким-либо группам, подгруппам или индивидам. Недавние исследования показали, что алгоритмы машинного обучения способны различать расу и пол человека. Это касается автоматического анализа лица и набора данных в отношении фенотипических подгрупп [Buolamwini, T. Gebru, 2018]. Учитывая это, необходимо исключить, в частности, любую дискриминацию по демографическим показателям (возраст, пол, этническая принадлежность и т. д.).

В Стандарте рассмотрения алгоритмических предубеждений (IEEE p7003) отдельным лицам или организациям, создающим

системы ИИ, предписывается обязательное соблюдение правил, помогающих избежать непреднамеренных, необоснованных и необъяснимо различающихся результатов работы ИИ с данными пользователей [Koene, Dowthwaite, Seth, 2018, p. 38–41]. Это практическое руководство для разработчиков позволяет определить, когда им следует отступить, остановиться, чтобы оценить внезапно возникшие в работе системы проявления предвзятости, социальной несправедливости.

Авторы Монреальской декларации призывают к тому, что ИИ должен избегать использования полученных данных для блокировки отдельных лиц в профиле пользователя, фиксации личной идентичности или ограничения фильтром, что препятствовало бы их личностному росту и развитию, особенно в таких областях, как образование, правосудие или бизнес [Montréal Declaration: Responsible AI. URL: https://monoskop.org/images/d/d2/Montreal_Declaration_for_a_Responsible_Development_of_Artificial_Intelligence_2018.pdf]. Для соблюдения принципа справедливости, как считают авторы документа, важна особая осторожность как на этапах разработки, так и внедрения ИИ-систем. Иными словами, еще на этапе предварительной обработки данных необходимо избежать какой-либо дискриминации. При выявлении признаков дискриминации необходимо рассмотреть вопрос о корректирующих мерах по ограничению или нейтрализации этих рисков, а также задуматься о повышении осведомленности заинтересованных сторон. ИИ не должен разрабатываться или использоваться с целью ограничения свободного выражения идей или возможности услышать различные мнения, поскольку и то, и другое являются неотъемлемыми условиями демократического общества.

Учитывая быстроменяющийся характер и темп смены технологий, методов ИИ, общество обязано знать и обсуждать возникающие в связи с этим проблемы. Это особенно актуально для поддержки наиболее незащищенных слоев населения и групп, для которых в силу определенных социальных особенностей ИИ, как и

любые другие технологии эпохи информатизации, представляет опасность.

ИИ может привести как к лучшему, более справедливому и более эффективному обществу, так и, напротив, способствовать обогащению небольшой группы цифровой элиты. В научном обзоре Б. Митльштат, П. Алло, М. Таддео отмечено, что большая часть рассмотренной ими литературы посвящена тому, как дискриминация возникает вследствие необъективной, предвзятой обработки данных и столь же предвзятого принятия решений [Mittelstadt, Allo, Taddeo, Wachter, Floridi, 2016, p. 1–21]. Ф. Паскуале, анализируя дискриминационную практику применения систем ИИ, тоже выражает тревогу и беспокойство: «если общество не будет привержено принципам справедливой обработки данных, то цифровая дискриминация будет только усиливаться» [Pascuale, 2015, p. 145]. Поэтому в области ИИ инклюзивная политика должна быть направлена на достижение двух целей: гарантировать, что развитие этих технологий не приведет к увеличению социального и экономического неравенства, и использовать ИИ для уменьшения уже существующего неравенства.

Принцип справедливости неразрывно связан с принципом прозрачности систем ИИ, чтобы избежать несправедливости. Ф. Паскуале подчеркивает, что отсутствие прозрачности, характерное для многих систем ИИ, затрудняет оценку его внутренней работы [Там же]. Нам нужно знать, какие именно параметры были заданы в качестве главных, основных, как данные были собраны и как они повлияли на результат. Непрозрачность в работе систем ИИ не позволяет получить эту информацию, что впоследствии неизбежно приводит к обострению социальной дискриминации. В ряде случаев в системах ИИ используются достаточно простые модели, поэтому реализовать принцип прозрачности сравнительно легко, если только сами компании или организации не блокируют раскрытие используемых ими алгоритмов. Если речь идет о более сложных моделях (в частности, алгоритмах глубокого обучения), то обеспечить

прозрачность процедур становится гораздо сложнее. В данном случае этому препятствует не нежелание компании или сложность корреляций, а природа и структура самих алгоритмов, недоступных для понимания, для корректной интерпретации.

Поучительным примером является использование судами инструмента «профилирование правонарушителя для осуществления альтернативных санкций (COMPAS)» в нескольких штатах США. Данная система применяется также для оценки риска совершения нового преступления обвиняемым. ProPublica подвергла резкой критике COMPAS за предвзятое отношение к подсудимым с темным цветом кожи [Angwin, Larson, Mattu, Kirchner, 2016]. Однако владельцы системы ИИ не признали свою вину и не раскрыли в полной мере алгоритм, используемый для расчета оценки риска совершения повторного преступления [Dieterich, Mendoza, Brennan, 2016]. В данный момент остается недоказанным, действительно ли COMPAS является дискриминационным или нет. Учитывая значимость той сферы деятельности, в которой система COMPAS используется, едва ли можно считать сложившуюся ситуацию нормальной или приемлемой для общества и его институтов.

Авторы концепции «объяснимого ИИ» пытаются решить эту проблему, приняв такие меры, как отслеживаемость, аудит и прозрачное информирование о возможностях системы [Wierzynski, 2018]. И это все, что можно сделать человеку для увеличения доступности результатов работы сложных алгоритмов систем ИИ.

Для реализации принципа справедливости необходимо обеспечить прозрачность в отношении использования личных данных человека. Члены общества должны знать о том, какие именно особенности, признаки, свойства людей не могут служить критерием различения их, разделения на иерархические подклассы, подмножества в том или ином контексте, поскольку это может противоречить существующему в общественном сознании представлению о справедливом обществе. Другими словами, члены общества должны быть осведомлены о том, какие параметры учитывает алгоритм ИИ и как

они влияют на результат его работы. Обобщая вышесказанное, сформулируем основные положения философского принципа справедливости:

- системы ИИ изначально должны разрабатываться таким образом, чтобы исключить предвзятое, несправедливое отношение по признакам гендерной принадлежности, уровню заработной платы, возрасту, расовой принадлежности и другим характеристикам;

- должно быть обеспечено широкое использование или равный доступ к технологиям ИИ, которые приносят благо обществу, вне зависимости от уровня благосостояния и социального благополучия пользователей;

- внедрение и использование систем ИИ не должно усугублять существующие в обществе формы дискриминации и предвзятости. Напротив, применение систем ИИ призвано создать условия для нивелирования и искоренения подобных явлений;

- необходимо исключить передачу полномочий от человека системам ИИ в ситуациях нравственного выбора, поскольку только человек как физическое или юридическое лицо может быть полноправным субъектом ответственности;

- в процессе разработки, внедрения и использования систем ИИ следует четко определить субъекта ответственности на каждом из этапов жизненного цикла этих систем ИИ.

Принцип автономии

Автономия	Человеческий контроль над технологией, управляемость, право человека на самоопределение, расширение возможностей человека, право выбора за человеком, а не за машиной
-----------	---

Прогресс в развитии искусственного интеллекта, открывающий новые, ранее не существовавшие перспективы для человека, одновременно создает значительные риски для его автономии, его права и возможности совершать поступки в соответствии с собственными целями и ценностями. Учитывая распространенность технологий

ИИ — от поисковых систем до автоматических сканеров лица и тела — и их влияние на жизнь людей, мы пришли к выводу, что человеческая идентичность и достоинство сегодня формируются не только в социальном, но и в социотехническом аспекте.

Сохранение автономии при взаимодействии с системами ИИ — важнейшая составляющая благополучия современного человека. Обман, манипуляции или принуждение с помощью систем ИИ, получившие в современном обществе небывалый размах и невиданное ранее разнообразие, позволяют предположить, что технологии ИИ могут навредить автономии человека и уже сейчас служат орудием совершения мошеннических действий, введения в заблуждение с целью нанесения вреда, ущерба людям во всем мире.

Попытка Cambridge Analytica манипулировать избирателями — лишь один из примеров подобного воздействия [Susser, Roessler, Nissenbaum, 2019]. Другой пример — эксперимент Facebook по «эмоциональному заражению», в ходе которого пользователей склоняли к принятию определенных решений под воздействием эмоций [Experimental evidence of massive-scale emotional contagion through social networks. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4066473/>]. Попытки мошенников похитить финансовые средства клиентов банковской системы, телекоммуникационных компаний посредством технологий ИИ стали настоящим испытанием для абсолютного большинства граждан современных государств.

Таким образом, появляется уверенность в том, что системы ИИ могут нанести ущерб человеческой автономии, препятствовать достижению его целей и совершению менее аутентичного выбора, снижать компетенции человека в области автономии, подрывать автономию личности, делая ее более уязвимой для цифровых мошенников, грабителей, преступников. В литературе достаточно часто в контексте обсуждения автономии человека встречается проблема правосубъектности и автономности систем ИИ. По нашему мнению, системы ИИ могут способствовать или препятствовать автономии человека, но при этом сами они не обладают подлинной субъектностью,

не являются моральными агентами и потому не могут самостоятельно угрожать человеческой автономии.

Тем не менее активное распространение систем ИИ и их использование в различных мошеннических действиях актуализируют вопрос «Можно ли в процессе взаимодействия с системами ИИ считать выбор пользователя подлинным, не страдает ли, не ущемлена ли в этом случае способность человека к самоопределению?». Отвечая на вопрос, начнем с теоретического обсуждения проблемы автономии именно человека, а не систем ИИ, а также исследуем вопрос о потенциальном воздействии систем ИИ на человеческую автономию.

В этом же контексте мыслят специалисты авторитетных организаций, занимающихся развитием и управлением процессов создания и использования систем ИИ. Так, группа экспертов высокого уровня (HLEG) европейской комиссии считают «автономию» одним из четырех ключевых этических принципов в документе «Руководство по надежному ИИ» [High Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI. URL: <https://digitalstrategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>]. В других программных документах, в том числе Монреальский декларации по ответственному развитию искусственного интеллекта [Montréal Declaration: Responsible AI. URL: https://monoskop.org/images/d/d2/Montreal_Declaration_for_a_Responsible_Development_of_Artificial_Intelligence_2018.pdf] и Белой книге европейской комиссии по искусственному интеллекту [ECWP. On Artificial Intelligence - A European approach to excellence and trust. White Paper COM. URL: https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en], в равной степени подчеркивается необходимость защиты автономии. Организация экономического сотрудничества и развития (ОЭСР) называет автономию одной из основных ценностных установок, ориентированных на человека [OECD. Recommendation of the Council on Artificial Intelligence. URL: <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>].

В определении автономии мы полагаемся на трактовку, данную М. Райаном и Л. Деси [Ryan, Deci, 2000, p. 68–78], которая согласуется как с аналитической [Frankfurt, 1971, p. 5–20; Friedman, 2003], так и с феноменологической точкой зрения [Pfander, 1967; Ricoeur, 1966]. М. Райан и Л. Деси рассматривают автономию как чувство готовности и воли в действиях, которые «одобряются или будут одобрены самим человеком». Иными словами, автономия предполагает действие в соответствии со своими целями и ценностями, а не просто независимость или контроль. Это положение необходимо подчеркнуть, поскольку в процессе взаимодействия с ИИ люди могут одобрять или не одобрять передачу контроля системе ИИ.

Иначе говоря, человеческая автономия в технологических системах требует готовности, желания и стремления совершить действие; отсутствия внешнего давления или принуждения; возможности контролировать действия, предпринимаемые системой ИИ; отсутствия обмана или преднамеренной дезинформации.

Безусловно, это далеко не полная и не достаточная концептуализация автономии человека в системах ИИ, но она формирует необходимую основу, обеспечивающую условия для устранения большого количества ключевых противоречий, возникающих в контексте проблемы автономии.

Рассмотрим, как эту проблему решают другие исследователи. А. Лайтинен выделяет следующие условия реализации автономии человека: способность к самоопределению, уважение и поддержание человеческой автономии, признание и уважение со стороны других членов общества и т.д. Ученый отмечает, что различные материальные, экономические, правовые, культурные и информационные ресурсы также могут рассматриваться в качестве необходимых составляющих автономии человека [Laitinen, Sahlgren, 2021]. Так, способность человека к самоопределению позволяет выработать такое мировоззрение, при котором человек формирует собственные оценки, ценности и действует в соответствии с ними. Эта же способность порождает самоуважение и уважение права на са-

моопределение других индивидов. Социальное признание и социальное взаимодействие являются важными факторами, способствующими развитию и реализации человеческих способностей. Далее А. Лайтинен выделяет несколько видов препятствий, которые подрывают или ограничивают автономию человека: прямое вмешательство, принуждение, угрозы, неприкрытая сила, манипуляции, идеологическая обработка, обман, подталкивание, патернализм и т.д.

Понятие «самоуважение» Джона Ролза, на наш взгляд, также отражает сущностную сторону автономии человека. Самоуважение, согласно воззрениям Ролза, является неотъемлемой частью развития социальной справедливости, относится к первичным благам, распространять которые призвана его теория справедливости [Rawls, 1971]. Самоуважение, по Ролзу, обеспечивает индивиду как «чувство собственной ценности», так и «твердую убежденность в том, что его представление о своем благе, его план жизни заслуживают осуществления» [Там же]. Также самоуважение основано на ощущении себя равноправным членом общества, на одинаковых условиях разделяющим с другими ответственность за фундаментальные суждения по социальным и политическим вопросам.

Действительно, без твердой убежденности в себе и своем жизненном плане, без ощущения себя равным среди равных жить в мире и достигать своих целей намного труднее. Самоуважение обеспечивает равенство и независимость личности по отношению к другим членам общества, обеспечивая и активно формируя в процессах взаимодействия с системами ИИ ее моральное, политическое и культурное развитие. И наоборот, если паттерны лишения прав, дискриминации встроены (сознательно или случайно) в системы ИИ, то последние оказываются способны препятствовать автономии человека, лишая его самоуважения, уверенности в себе и своем праве.

Признавая значимость автономии человека и обсуждая эту проблему в контексте взаимодействия людей с системами ИИ, авторы отдельных документов провозглашают принцип автономии как утверждение так называемой позитивной свободы, права на са-

моопределение демократическими средствами, права устанавливать и развивать отношения с другими людьми, свободы отзывать согласие или свободы использовать предпочтительную платформу/технологии [Internet Society. Artificial Intelligence & Machine Learning: Policy Paper. URL: <https://www.internetsociety.org/resources/doc/2017/artificial-intelligence-andmachine-learning-policy-paper/>].

В других источниках в трактовке принципа автономии также указывается свобода от технологических экспериментов, манипуляций или наблюдения [European Group on Ethics in Science and New Technologies..., 2018]. Обращаясь к вопросу об автономии человека, AI HLEG указывает, что люди должны иметь возможность принимать обоснованные решения в отношении систем ИИ [High Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI. URL: <https://digitalstrategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai/>]. Необходимо, чтобы автономия человека не вызывала неоправданного принуждения, обмана или манипуляций со стороны систем ИИ. Люди вправе обладать необходимой информацией и инструментами, для того чтобы без ущерба для себя и других взаимодействовать с такими системами. Компания Google, которая сформулировала семь этических принципов применения ИИ, считает, что технологии ИИ должны быть подотчетными человеку, подлежать надлежащему руководству и контролю со стороны человека [Google. AI at Google: Our Principles. URL: <https://ai.google/principles/>].

В области здравоохранения принцип сохранения автономии человека при использовании систем ИИ может быть признан одним из часто обсуждаемых [Karimian, Petelos, Evers, 2022, p. 539–551]. Проблемы, связанные с этим принципом, заключаются в отсутствии ориентированности на пациента [Blease, Kaptchuk, Bernstein, Mandl, Halamka et al., 2019, p. 1–24] и в отсутствии совместного принятия решений человеком и системой ИИ [Liyanage, Liaw, Jonnagaddala, Schreiber, Kuziemsy et al., 2019, p. 41–46].

Принцип автономии особенно важен при разработке и внедрении автономных систем вооружения, которые могут выбирать и атаковать цели без вмешательства человека или самостоятельно инициировать атаки. Эти системы вызывают опасения по поводу потери человеком контроля над вооружением.

В Национальной стратегии Дании принцип автономии подразумевает возможность принимать осознанные и независимые решения без искусственного интеллекта, лишаящего человека права на самоопределение [The Danish National Strategy for artificial intelligence. URL: <https://en.digst.dk/strategy/the-danish-national-strategy-for-artificial-intelligence/>]. В других документах подчеркивается, что необходимы «контроль человека над системами ИИ и знания об автономных системах» [ECAA. Statement on Artificial Intelligence, Robotics., 2018], что принципы автономии трансформируются в «полномочия человека принимать решения» [Floridi, Cowls, 2022, p. 535–545]. Эти выводы близки идеям Дж. Фьелд, Н. Ахтен, Х. Хиллигосс, А. Надь, М. Срикумар, утверждающих в своей работе, что автономия обычно обеспечивает теоретическое обоснование принципов «человеческого контроля над технологиями» [Fjeld, Achten, Hilligoss, Nagy, Srikumar, 2020].

Несмотря на дискуссию об автономии человека и во многом совпадающие позиции ее участников, какие-либо единые рекомендации для разработчиков или пользователей, способствующие устранению потенциальных рисков, угрожающих человеческой автономии, разработаны не были. Отсутствие единства в данном вопросе усугубляется сложностью технического ландшафта и большой неопределенностью социальных последствий внедрения ИИ. В совокупности все это способно нивелировать уже предпринятые меры ученых и специалистов в области ИИ для успешного управления системами ИИ и поставить вопрос о целесообразности и безопасности их использования в различных сферах общественной жизни.

На наш взгляд, задача систем ИИ в рассматриваемом аспекте — это поддержка людей в принятии обоснованных решений в соответ-

ствии с их собственными целями. Сегодня существует реальная угроза, когда системы ИИ потенциально могут быть использованы для влияния на поведение человека в интересах третьих лиц или организаций посредством скрытых манипуляций, обмана, слежки и принуждения, нарушающих индивидуальную автономию человека [EU Member States sign up to cooperate on Artificial Intelligence//European Commission. URL: <https://ec.europa.eu/jrc/communities/en/community/humaint/news/eu-member-states-sign-cooperate-artificial-intelligence>]. Способность принимать решение самостоятельно утрачивается, если системам ИИ регулярно передается все больше и больше задач, включая принятие решений в социальных, медицинских или финансовых вопросах. Такой патернализм предполагает «благонамеренные» посягательства на автономию человека против его воли, своего рода насильственное «причинение добра». Подобное использование систем ИИ может привести к необоснованному искажению подлинных убеждений, мотивов, ценностных установок и целей человека.

Философский принцип автономии пользователей может быть сформулирован как требование к разработчикам систем ИИ обеспечить в процессах взаимодействия с ИИ:

- реализацию права человека самостоятельно принимать осознанные и независимые решения по любым вопросам в соответствии с собственными целями и ценностями,
- реализацию права человека обладать всей необходимой информацией для успешного взаимодействия с ИИ- системами,
- возможность для человека на любом этапе взаимодействия полностью контролировать процесс,
- защиту человека от влияния третьих лиц или организаций, оказываемого посредством психологических или иных манипуляций, обмана, шантажа, угроз, слежки, принуждения и т. п.

В свете вышесказанного принцип автономии, по нашему мнению, должен стать одним из основных при разработке, внедрении и оценке функциональности систем ИИ. Выполнению этого принципа

будет способствовать право не подчиняться решению, основанному исключительно на автоматизированной обработке. Тем более, если это решение приводит к правовым последствиям для пользователей. Защита людей должна быть важнее всего. Прежде всего цель состоит в том, чтобы уменьшить ущерб до такой степени, чтобы избежать его вообще. Государственный сектор должен нести ответственность за обеспечение безопасного для автономии человека внедрения и распространения систем ИИ. Автономное принятие решений людьми является выражением позиции общества, согласно которой в центре внимания при решении любых социальных вопросов и проблем всегда должен быть человек с его правом на развитие и потребностью в защите. При этом речь не идет о вседозволенности и отсутствии общественных норм и регуляторов, поскольку максимум личной свободы выбора уравнивается свободой других и безопасностью всех членов общества.

Принцип автономии призван помочь людям, пользователям систем ИИ, стать более независимыми, самостоятельными, а не пытаться манипулировать, деквалифицировать, незаконно контролировать и неуважительно относиться к автономии других. Окажется ли влияние систем ИИ на автономию человека положительным или отрицательным, покажет будущее. Очевидно то, что потенциальное влияние систем ИИ на человеческую автономию чрезвычайно глубоко и многогранно.

Принцип непричинения вреда, объединивший несколько близких по смыслу требований: безопасность, защита жизни, благодеяние, устойчивость и т. д., означает, что деятельность, связанная с исследованием, проектированием, созданием, внедрением и применением систем ИИ, должна исключать любое непреднамеренное причинение вреда человеку.

Непричинение вреда	Защита жизни, безопасность, физическая и психическая неприкосновенность человека, устойчивость (экологическая безопасность), защи-
--------------------	--

	ценность человека, благодеяние, благосостояние человечества, обратимость, минимизация ущерба, предотвращение несчастных случаев, управление рисками
--	---

Системы ИИ должны проектироваться, разрабатываться и внедряться, во-первых, с целью предотвращения или решения проблем, отрицательно влияющих на жизнь человека и благополучие окружающей среды, во-вторых, должны представлять собой идеи и разработки, одобренные обществом и признанные экологически безопасными.

По нашему мнению, принцип непричинения вреда имеет особое значение, поскольку не только отражает отдельные грани, аспекты, стороны безопасного и эффективного применения ИИ, но и играет ведущую роль во всей совокупности представленных принципов. Он является системообразующим по отношению к ним, поскольку принципы раскрывают его сущность, представляют собой различные стороны этого принципа. При несоблюдении любого из пяти требований автоматически невозможно безопасное для человека применение систем ИИ. И наоборот, только путем соблюдения принципов прозрачности, конфиденциальности, справедливости, автономии и ответственности можно добиться осуществления принципа непричинения вреда человеку системами ИИ.

Для выполнения рассматриваемого принципа важно, чтобы разработчики гарантировали фальсифицируемость ключевых требований и гипотез при запуске систем ИИ в безопасных, защищенных контекстах. Нужно также учитывать необходимость исключительно постепенного внедрения систем ИИ, это предоставляет возможность пошаговой, поэтапной проверки их деятельности. Пользователи должны быть готовы к быстрой остановке или изменению данных систем в случае обнаружения опасных или нежелательных последствий.

Принцип непричинения вреда, отсылающий к первому закону робототехники А. Азимова, не исключает того, что вред человеку все-таки может быть причинен вследствие применения системы ИИ, однако основная идея данного принципа заключается в том, что непричинение вреда человеку должно быть заложено разработчиком в саму систему ИИ в качестве целеполагания, в качестве основной цели ее деятельности. И в том случае, если ИИ случайным образом нанес ущерб человеку, разработчики и пользователи должны иметь возможность найти оптимальное решение для исключения подобных негативных последствий в будущем. Здесь речь не идет об использовании систем ИИ для каких-либо видов современного оружия. Данная тема требует отдельных исследований и далеко выходит за пределы обсуждения рассматриваемой проблемы.

В Азиломарских принципах непричинение вреда интерпретируется посредством следующих положений: 1) должна быть обеспечена безопасность и защищенность систем ИИ на протяжении всего срока эксплуатации; 2) необходимо обязательно выявлять причины причинения вреда системой искусственного интеллекта; 3) цели и поведение систем ИИ и человека должны быть в обязательном порядке синхронизированы [Asilomar AI Principles. URL: <https://www.artificial-intelligence.blog/news/asilomar-ai-principles>].

В рекомендациях по искусственному интеллекту, составленных Организацией экономического сотрудничества и развития (ОЭСР), системы ИИ должны быть нацелены на принесение пользы людям, действие в рамках закона, безопасное функционирование в течение всего срока эксплуатации [OECD's Principles for responsible stewardship of trustworthy AI. URL: <https://oecd.ai/en/ai-principles>].

Принцип «предотвращения вреда» в AIHLEG предусматривает, что «хотя ИИ может принести много пользы, он также может причинить вред» [High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI. URL: <https://digitalstrategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>]. Этот вред может быть как материальным (безопасность и здоровье людей, в том чис-

ле гибель людей, материальный ущерб), так и нематериальным (утрата права на неприкосновенность частной жизни, ограничение права на свободу выражения мнений, человеческое достоинство, дискриминация, например, при трудоустройстве), может относиться к широкому спектру рисков.

В ECWP комментируются эти положения: «нормативная база должна концентрироваться на том, как свести к минимуму различные риски потенциального вреда» [European Commission. White paper on artificial intelligence: A European approach to excellence and trust. URL: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf]. Риски могут быть вызваны недостатками в разработке технологии ИИ, быть связаны с проблемами непрозрачности и качества данных.

Существует позиция, согласно которой системы ИИ способствуют повышению общественного благосостояния, справляясь со многими задачами гораздо лучше, чем человек, могут даже взаимодействовать с живыми существами, составлять им компанию и заботиться о них. Однако, по нашему мнению, следует задуматься над тем, что эффект обеспечиваемых системами ИИ «преимуществ» в значительной степени сомнителен. Беспрецедентные изменения, которые осуществляют в нашем мире системы ИИ, как было показано ранее, неизбежно порождают разнообразные социально-философские проблемы. Философский принцип непричинения вреда, по нашему мнению:

– требует разработать основу для понимания и проверки любых социальных изменений, вызванных ИИ, прежде чем они будут приняты в виде безусловной ценности и названы «преимуществами». Важно сосредоточиться на более целостном изучении последствий применения технологий ИИ в глобальном контексте. Кроме того, следует задаться вопросом, кто определяет, что в каждом конкретном случае будет полезно для человека, применяющего системы ИИ, а что нанесет ему вред;

– требует, чтобы системы ИИ проходили строгий процесс проверки безопасности и производительности. В настоящее время су-

существует множество неопровержимых доказательств того, что эти системы не являются объективными или ценностно нейтральными [O'Neil, 2016] или даже надежными и безопасными [Qayyum, Qadir, Bilal, Al-Fuqaha, 2020, p. 156–180];

– призывает разработчиков при создании систем ИИ с целью обеспечения их безопасного использования консультироваться с пользователями, которые будут непосредственно взаимодействовать с этими технологиями, подвергаться их влиянию.

Возможно, понимание целевых установок пользователя, условий эксплуатации систем ИИ, применяемых методов управления и оценка последствий взаимодействия пользователя с системой ИИ позволят в перспективе разработать указанные системы более безопасными. Быть может, своевременное обеспечение этих условий помогло бы избежать трагической катастрофы двух авиалайнеров Boeing 737 Max, когда пилоты пытались ликвидировать сбой неисправных датчиков системы ИИ, но все их усилия были тщетны из-за того, что отсутствовали так называемые «дополнительные функции безопасности», которые данная компания не предоставляла в исходном комплекте, а продавала отдельно [Boeing во время аудита обнаружила ошибки в обновленном ПО для самолетов модели 737 MAX. URL: <https://habr.com/ru/news/t/484434/>]. Таким образом, при оценке уровня рисков должны учитываться известные и потенциальные риски в краткосрочной и долгосрочной перспективе;

– означает, что системы ИИ должны использоваться для расширения человеческих возможностей, защиты благополучия и безопасности каждого члена общества. Необходимо принимать такие решения, которые не препятствуют процессу реализации всех потенциальных творческих возможностей человека для достижения гармонии во всех сферах жизни общества: экономической, социальной, духовной и политической;

– формулирует для разработчиков систем ИИ требование учета сложившихся и принятых в данном обществе ценностей, прав и свобод граждан, существующих механизмов и способов раскрытия

творческого и интеллектуального потенциала людей, а также имеющих культурно-исторических особенностей коллективного бытия.

Человек и его потребности должны быть единицей измерения допустимости использования ИИ. Человек, его права и свободы должны рассматриваться как наивысшая ценность;

– предполагает оценку влияния применения ИИ на его пользователей, общество и окружающую среду.

Ответственность акторов ИИ должна учитывать влияние этих систем на общество и граждан на каждом этапе их жизненного цикла, включая неприкосновенность частной жизни, этичное, безопасное и ответственное использование персональных данных. В случае если внедрение систем ИИ может привести к неприемлемым для человека и общества последствиям, должны быть приняты меры для предотвращения подобных негативных случаев в будущем. Использование систем ИИ, способных целенаправленно причинять вред окружающей среде, жизни и благополучию человека, принципиально недопустимо.

ЗАКЛЮЧЕНИЕ

В современном мире технологии искусственного интеллекта существенным образом влияют на происходящие в обществе процессы, на самые разные аспекты его жизнедеятельности. Их влияние выражается, в частности, в трансформации экономических, политических или иных социальных стандартов и норм.

Нами были рассмотрены существующие определения понятий «искусственный интеллект», «системы искусственного интеллекта», выбрана наиболее всеобъемлющая, отражающая современный уровень развития систем ИИ дефиниция; представлены существующие в научной литературе подходы к классификации искусственного интеллекта; установлена специфическая особенность источниковой базы исследования проблем ИИ; выявлены основные социально-философские проблемы, риски и угрозы применения систем искусственного интеллекта, возникающие в ходе их разработки и внедрения. В монографии подробно рассмотрена реакция мирового сообщества на возникшие трудности, исследован имеющийся опыт этического и правового регулирования систем ИИ, систематизирован обширный свод документов, содержащих правовые и этические принципы регулирования отношений в сфере ИИ, раскрыто содержание основных социально-философских принципов применения систем ИИ, необходимых как для их разработки, внедрения, так и успешного, безопасного применения в любых сферах общественной жизни.

В результате проведенного исследования были сделаны следующие выводы. В настоящее время, характеризующееся бурным ростом и развитием, широким распространением ИИ-систем в социальной практике, все еще нет единого общепринятого определения понятия «искусственный интеллект». Учитывая имеющийся уровень технологических достижений в области информационных систем и искусственного интеллекта, позволяющий с позиции научного исследования обсуждать искусственный интеллект только

в виде «узкого» ИИ, а также содержание целого комплекса документов, посвященных проблемам их этического и правового регулирования, наиболее приемлемым, раскрывающим суть изучаемого феномена представляется понятие «системы искусственного интеллекта», определенное экспертной группой высокого уровня европейской комиссии как «...программные (и, возможно, также аппаратные) системы, разработанные людьми, которые, имея сложную цель, действуют в физическом или цифровом измерении, воспринимая свою среду посредством сбора данных, интерпретируя собранные структурированные или неструктурированные данные, рассуждений на основе знаний или обработки информации, полученной из этих данных, и принятия решения о наилучших действиях для достижения поставленной цели. Системы искусственного интеллекта могут либо использовать символические правила, либо изучать цифровую модель, а также могут адаптировать своё поведение, анализируя влияние предыдущих действий на окружающую среду» [High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI. 8 April 2019. URL: <https://digitalstrategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>].

Результатом обобщения опыта применения систем ИИ стало выделение имеющих наиболее универсальный характер социально-философских проблем, с которыми сталкиваются разработчики и пользователи систем ИИ в ходе их внедрения в разнообразные социальные и производственные процессы. Это проблемы непрозрачности, нарушения конфиденциальности, социальной несправедливости, нарушения автономии, отсутствия ответственности, причинения физического или морального вреда. Эти проблемы мы рассматриваем как социально-философские, поскольку они охватывают широкий комплекс вопросов взаимодействия человека и искусственного интеллекта, содержание которых выходит далеко за пределы предмета науки этики.

Возникновение вышеназванных социально-философских проблем потребовало необходимой реакции со стороны государственных

структур, промышленных и деловых кругов, ученых и общественных, предпринявших многочисленные попытки упорядочить отношения человека и искусственного интеллекта посредством публикации документов, содержащих этические и правовые требования, нормы, принципы, предназначенные для разработчиков и пользователей систем ИИ. Большинство указанных документов носит декларативный характер, представляет собой стратегические документы, определяющие направления развития и финансового обеспечения проектов по внедрению и использованию систем ИИ, этические руководства, которые своей целью имеют регулирование отдельных систем, нашедших применение в различных сферах общественной жизни. Сформулированные в них этические и правовые принципы разрознены, не в полной мере обоснованы, их содержание не раскрыто или имеет многозначную интерпретацию. Наиболее значимыми в рамках данного исследования видятся требования, касающиеся непричинения вреда, прозрачности, ответственности, конфиденциальности, справедливости и автономии человека.

Нами были выявлены основные философские принципы регулирования разработки и применения систем ИИ, раскрыто их содержание и предназначение. Мы считаем, что они необходимы для предотвращения, снижения риска возникновения и эскалации социально-философских, этических проблем применении систем ИИ во всех отраслях жизни современного социума. Это принципы прозрачности, ответственности, автономии, конфиденциальности, социальной справедливости и непричинения вреда человеку. Указанные принципы, понимаемые нами как наиболее общие требования, предъявляемые к разработчикам и пользователям систем ИИ, необходимо отличать от регламентаций другого уровня, таких как международные нормы в области прав человека, внутренние или региональные нормы, профессиональные нормы и т. д., поскольку они выступают по отношению к последним в качестве универсальных, всеобщих требований, имеющих значение для всех без исключения

индивидов и групп, взаимодействующих с системами ИИ во всех областях их применения. Эти фундаментальные философские принципы, по нашему мнению, могут стать основой нормотворческой деятельности в сфере ИИ и руководством для дальнейшего этического регулирования процессов взаимодействия человека с искусственным интеллектом.

Рассмотрение проблем ИИ на уровне социально-философского исследования позволяет сделать вывод об актуальности избранной темы, ее масштабности, сложности и многогранности. Сам объект исследования находится в стадии становления, развития, демонстрирует появление новых свойств и состояний, вследствие чего открываются новые перспективные направления дальнейшего изучения систем ИИ и информационного общества в целом. Исследование этих вопросов невозможно ограничить рамками технических наук и технологического дискурса. Требуется проведение множества междисциплинарных, философских исследований, рассматривающих процессы информатизации сквозь призму проблем человека, с позиции признания человека высшей ценностью, без чего преимущества, предлагаемые системами ИИ, все больше вызывают недоверие в обществе и ставят под сомнение достижимость широко декларируемых в любом государстве целей улучшения качества жизни человека, прогресса в развитии социальных систем и гармонизации общественных отношений.

ЛИТЕРАТУРА

Нормативная правовая информация

1. Восьмой закон о внесении изменений в закон о дорожном движении от 16 июня 2017 г. URL: <http://robopravo.ru/uploads/s/z/6/g/z6gj0wkwhv1o/file/5MZOclyT.pdf>

2. Дорожная карта развития «сквозной» цифровой технологии «Нейротехнологии и искусственный интеллект». URL: <https://digital.gov.ru/ru/documents/6658/> (дата обращения: 19.11.2022); Дорожная карта развития «сквозной» цифровой технологии «Компоненты робототехники и сенсорики». URL: <https://digital.gov.ru/ru/documents/6666/> (дата обращения: 11.12.2022).

3. Законопроект № 922869-7. URL: <https://sozd.duma.gov.ru/bill/922869-7> (дата обращения: 20.04.2020).

4. О внедрении робототехники в отечественную медицину — проблемы и пути решения: заседание президиума Российской академии наук. URL: <http://www.ras.ru/news/newsrelease.aspx?ID=2eaccb0cd887-4d02-bb10-7caae48b083b&print=1> (дата обращения: 23.02.2022).

5. О внесении изменений в Гражданский кодекс Российской Федерации в части совершенствования правового регулирования отношений в области робототехники: проект федерального закона / Д. С. Гришин (Grishin Robotics) (не вносился в Госдуму). URL: <https://www.dentons.com/ru/insights/alerts/2017/january/27/dentons-develops-first-robotics-draft-law-in-russia> (дата обращения: 20.08.2022).

6. О проведении эксперимента по опытной эксплуатации на автомобильных дорогах общего пользования высокоавтоматизированных транспортных средств: постановление правительства РФ от 26 ноября 2018 г. № 1415. URL: <http://publication.pravo.gov.ru/File/GetFile/0001201811270008?type=pdf> (дата обращения: 01.10.2012).

7. О проведении эксперимента по установлению специального регулирования в целях создания необходимых условий для разработки и внедрения технологий искусственного интеллекта в субъекте Российской Федерации — городе федерального значения Москве

и внесении изменений в статьи 6 и 10 Федерального закона «О персональных данных»: федер. закон от 24.04.2020 № 123-ФЗ // СЗ РФ. 2020 № 17. Ст. 2701.

8. О развитии искусственного интеллекта в Российской Федерации: указ Президента РФ от 10.10.2019 № 490 (вместе с «Национальной стратегией развития искусственного интеллекта на период до 2030 г.»). URL: <https://base.garant.ru/72838946/> (дата обращения: 14.05.2022).

9. Об основах охраны здоровья граждан в Российской Федерации: федер. закон от 21.11.2011 № 323-ФЗ // Собрание законодательства РФ. 2011. № 48. Ст. 6724.

10. Об утверждении Концепции развития регулирования отношений в сфере технологий искусственного интеллекта и робототехники на период до 2024 г.: распоряжение правительства РФ от 19 августа 2020 г. № 2129-р // СЗ РФ. № 35, 2020 г. Ст. 5593.

11. Об утверждении Правил государственной регистрации медицинских изделий: постановление правительства РФ от 27.12.2012 № 1416 // Собрание законодательства РФ. 2013. № 1. Ст. 14.

12. Acm code of ethics and professional conduct. URL: <https://www.acm.org/binaries/content/assets/about/acmcode-of-ethics-booklet.pdf> (access 15.03.2022).

13. AI Strategy 2019. AI for Everyone: People, Industries, Regions and Governments. URL: <https://www8.cao.go.jp/cstp/english/humancentricai.pdf> (access 21.06.2021).

14. Artificial Intelligence at Google: Our Principles. URL: <https://ai.google/principles/> (access 30.08.2022).

15. Artificial Intelligence Initiative Act. 2019. URL: <https://www.congress.gov/bill/116th-congress/senate-bill/1558/text> (access 12.09.2022).

16. Artificial intelligence principles and ethics. URL: <https://www.digitaldubai.ae/initiatives/ai-principles-ethics> (access 23.09.2022).

17. Artificial Intelligence. An Accountability Framework for Federal Agencies and Other Entities. URL: <https://www.gao.gov/assets/gao-21-519sp.pdf> (access 25.05.2022).

18. Asilomar AI Principles. URL: <https://www.artificial-intelligence.blog/news/asilomar-ai-principles> (access 12.12.2021).

19. Australia's AI Ethics Principles. URL: <https://www.industry.gov.au/data-and-publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles> (access 30.01.2022).

20. Automated and Electric Vehicles Act. URL: <https://www.legislation.gov.uk/ukpga/2018/18/contents/enacted> (access 03.06.2022).

21. B. BS8611, Robots and robotic devices, guide to the ethical design and application of robots and robotic systems, British Standards Institute, 2016.

22. Beijing AI Principles. URL: <https://www.baai.ac.cn/news/beijing-ai-principles-en.html> (access 12.09.2022).

23. Bundesdatenschutzgesetz. URL: <https://www.datenschutz-wiki.de/BDSG> (access 12.05.2022).

24. Bundesregierung, Eckpunkte der Bundesregierung für eine Strategie Künstliche Intelligenz. URL: https://www.bmbf.de/files/180718Eckpunkte_KI-StrategiefinalLayout.pdf (access 13.08.2022).

25. CAHAI — Ad hoc Committee on Artificial Intelligence. URL: <https://www.coe.int/en/web/artificial-intelligence/cahai> (access 19.04.2021).

26. Civil Law Rules on Robotics, European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)) // OJ C 252, 2018. P. 239–257. URL: <http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML+TA+P8-TA-2017-0051+0+DOC+PDF+V0//EN> (access 09.04.2022).

27. Cyber security law. URL: <https://www.tradecommissioner.gc.ca/china-chine/cybersecuritycybersecuritechina-chine.aspx?lang=eng> (access 30.04.2021).

28. Data protection act. URL: <https://www.hhs.gov/hipaa/for-professionals/index.html> (access 15.03.2022).

29. Elaboration of a Recommendation on the ethics of artificial intelligence. URL: <https://en.unesco.org/artificial-intelligence/ethics#drafttext> (access 27.01.2022).

30. Estonia accelerates artificial intelligence development. URL: <https://e-estonia.com/estonia-accelerates-artificial-intelligence/> (access 19.10.2022 г.); Kratid Eesti heaks (на эстонском языке). URL: <https://www.kratid.ee/> (access 19.10.2022).

31. Ethical design and use of automated decision systems. URL: <https://viewer.joomag.com/ciosc-standardcan-ciosc-101-2019/0672323001570024164> (access 16.05.2022).

32. Ethical Principles for AI in Medicine. The Royal Australian and New Zealand College of Radiologists. URL: <https://www.ranzcrc.com/documents/4952-ethical-principles-for-ai-in-medicine/file> (access 19.09.2022).

33. Ethics Commission. Automated and Connected Driving. URL: https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission-automated-and-connecteddriving.pdf?__blob=publicationFile (access 14.01.2022).

34. Eun-jin K. Korean Government to Repeal Regulations in AI Industry December 18, 2019. URL: <http://www.businesskorea.co.kr/news/articleView.html?idxno=39324> (access 10.03.2022).

35. European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment (Dec. 3/4, 2018), European Commission for the Efficiency of Justice (CEPEJ)(2018)14 // Adopted at the 31st plenary meeting of the CEPEJ. Strasbourg, 2018. URL: <https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c> (access 19.11.2022).

36. Executive Order on Maintaining American Leadership in Artificial Intelligence. URL: <https://www.whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/> (access 20.06.2022).

37. Executive Order on Maintaining American Leadership in Artificial Intelligence. URL: <https://www.whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/> (access 20.06.2022).

38. G20 Ministerial Statement on Trade and Digital Economy. URL: <https://www.mofa.go.jp/files/000486596.pdf> (access 20.05.2022).

39. General data protection regulation. URL: <https://gdpr-info.eu/> (access 17.02.2022).

40. Government of Canada, “Directive on Automated Decision-Making”, 2019.

41. Government of Canada. Responsible use of artificial intelligence (AI). URL: <https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsibleuse-ai.html> (access 21.07.2022).

42. Government offices of Sweden, National Approach to artificial intelligence. URL: <https://www.regeringen.se/4aa638/contentassets/a6488cceb6f418e9ada18bae40bb71f/nationalapproach-to-artificial-intelligence.pdf> (access 21.03.2022).

43. High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI. URL: <https://digitalstrategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai> (access 19.10.2022).

44. IEEE. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. IEEE Standards Association. URL: <https://standards.ieee.org/industryconnections/ec/autonomous-systems.html> (access 20.04.2022).

45. Intel. Artificial Intelligence: The Public Policy Opportunity. URL: <https://blogs.intel.com/policy/files/2017/10/Intel-Artificial-Intelligence-Public-Policy-White-Paper-2017.pdf> (access 16.02.2022).

46. Intelligent Robots Development and Promotion Act, Act № 9014, Mar. 28, 2008. Act № 13744. 2016, art. 2. URL: http://elaw.klri.re.kr/eng_service/lawView.do?hseq=39153&lang=ENG (access 10.06.2022).

47. ISO/IEC JTC 1/SC 42 Artificial intelligence. URL: <https://www.iso.org/ru/committee/6794475.html?view=participation> (access 20.12.2022).

48. Loi pour une Republique numerique. № 2016-1321. URL: <https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000033202746&categorieLien=id> (access 23.09.2022).

49. Microsoft. The future computed: Artificial Intelligence and its role in society: chapter 2. URL: [//news.microsoft.com/ cloudforgood/_media/downloads/the-future-computed-english.pdf](https://news.microsoft.com/cloudforgood/media/downloads/the-future-computed-english.pdf) (access 17.09.2022).

50. Mid-to Long-term Master Plan in Preparation for the Intelligent Information Society: Managing the Fourth Industrial Revolution. URL: https://english.msit.go.kr/cms/english/pl/policies2/_icsFiles/afieldfile/2017/07/20/Master%20Plan%20for%20the%20intelligent%20information%20society.pdf (access 20.02.2022).

51. Ministry of Finance and Ministry of Industry, Business and Financial Affairs. National Strategy for Artificial Intelligence. URL: https://eng.em.dk/media/13081/305755-gb-version_4k.pdf (access 11.07.2022).

52. Model Artificial Intelligence Governance Framework (Second Edition). Infocomm Media Development Authority Singapore. URL: <https://www.imda.gov.sg/-/media/Imda/Files/Infocomm-Media-Landscape/SG-Digital/Tech-Pillars/Artificial-Intelligence/Primer-for-second-edition-of-the-Model-Framework.pdf?la=en> (access 19.10.2022).

53. Montréal Declaration: Responsible AI. URL: https://monoskop.org/images/d/d2/Montreal_Declaration_for_a_Responsible_Development_of_Artificial_Intelligence_2018.pdf (access 16.08.2022).

54. National Artificial Intelligence Strategy. URL: [https://www. smartnation.gov.sg/why-Smart-Nation/NationalAISstrategy](https://www.smartnation.gov.sg/why-Smart-Nation/NationalAISstrategy) (access 12.12.2022).

55. Oceanis. URL: <https://ethicsstandards.org/repository/> (access 23.07.2022).

56. OECD moves forward on developing guidelines for artificial intelligence (AI). URL: <https://www.oecd.org/going-digital/ai/oecd-moves-forward-on-developing-guidelines-for-artificial-intelligence.htm> (access 17.08.2022).

57. Pan-Canadian Artificial Intelligence Strategy, Invest in Canada. URL: <https://www.investcanada.ca/why-invest/pancanadian-artificial-intelligence-strategy> (access 23.06.2022)

58. Personal information protection and electronic documents act. URL: <https://lawslois.justice.gc.ca/ENG/ACTS/P-8.6/index.html> (access 21.02.2022).

59. Proposition de loi constitutionnelle relative à la Charte de l'intelligence artificielle et des algorithmes. URL: http://www.assemblee-nationale.fr/dyn/15/textes/115b2585_proposition-loi (access 20.10.2022).

60. Protection from online falsehoods and manipulation act. URL: <https://sso.agc.gov.sg/Acts-Supp/18-2019> (access 11.03.2021).

61. Putting principles into practice at Microsoft. URL: <https://www.microsoft.com/en-us/ai/our-approach?activetab=pivot1:primaryr5> (access 12.09.2022).

62. Research ICT Africa (RIA). URL: <https://researchictafrica.net/people/> (access 21.09.2022).

63. SAP. European Prosperity Through Human-Centric Artificial Intelligence // The Intelligent Enterprise. 2018. P. 28. URL: <http://www.sap.com/documents/2018/01/3e67a134-ee7c-0010-82c7-eda71af511fa.html> (access 12.04.2022).

64. SAP's Guiding Principles for Artificial Intelligence. URL: <https://news.sap.com/2018/09/sap-guiding-principles-for-artificial-intelligence/> (access 21.07.2022).

65. Select Committee on Artificial Intelligence. AI in the UK: ready, willing and able // House of Lords, UK, Apr. 2018.

66. Social Principles of Human-Centric AI. URL: <https://www.cas.go.jp/jp/seisaku/jinkouchinou/pdf/humancentricai.pdf> (access 12.09.2022).

67. State Council, Notice of Issuing New Generation Artificial Intelligence Development Plan. Guo Fa, 2017. № 35. URL: http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.html (на китайском языке) (access 16.05.2022).

68. Strategic Council for AI technology. URL: <http://www.nedo.go.jp/content/100865202.pdf> (access 13.10.2022).

69. The Different Faces of AI Ethics Across the World: A Principle-Implementation Gap Analysis Lionel Nganyewou Tidjon and Foutse Khomh, Senior Member, IEEE. URL: <https://arxiv.org/pdf/2206.03225.pdf> (access 21.09.2022).

70. The Guidelines on Artificial Intelligence and Data Protection // Consultative Committee of the Convention for the protection of individuals with regard to automatic processing of personal data (Convention 108). URL: <https://rm.coe.int/lignesdirhttps://rm.coe.int/guidelines-on-artificial-intelligence-and-data-protection/168091f9d8> (access 18.11.2022).

71. The Public Voice. Universal Guidelines for Artificial Intelligence. URL: [//thepublicvoice.org/ai-universal-guidelines](https://thepublicvoice.org/ai-universal-guidelines) (access 12.01.2022).

72. The Toronto Declaration. URL: <https://www.torontodeclaration.org/declaration-text/english/> (access 21.01.2022).

73. UNI Global Union. Top 10 principles for ethical artificial intelligence. URL: [//www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf](https://www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf) (access 12.08.2022).

74. United Nations Activities on Artificial Intelligence (AI). URL: https://www.itu.int/dms_pub/itu-s/opb/gen/S-GEN-UNACT-2019-1-PDF-E.pdf (access 01.12.2022).

75. Villani C. For a meaningful artificial intelligence towards. A French and European strategy. URL: https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf (access 12.12.2022).

76. Visually probe the behavior of trained machine learning models, with minimal coding. URL: <https://pair-code.github.io/what-if-tool/> (access 21.09.2022).

Источники (опубликованные)

77. Азимов А. Я, Робот / А. Азимов. Санкт-Петербург: Издатель, 1991. 800 с.

78. Алексеев А. Ю. Комплексный тест Тьюринга: философско-методологические и социокультурные аспекты / А. Ю. Алексеев. Москва: ИИнтелЛ, 2013. 304 с.

79. Алексеев П. В. Социальная философия / П. В. Алексеев. Москва: Велби, 2003. 256 с.

80. Антимонов Б. С. Гражданская ответственность за вред, причиненный источником повышенной опасности / Б. С. Антимонов. Москва: Юрид. лит., 1952. 300 с.

81. Аристотель. Сочинения: в 4 томах / Аристотель. Москва: Мысль, 1984. Т. 4. 832 с.

82. Бодрийяр Ж. Система вещей / Ж. Бодрийяр; перевод с франц. С. Н. Зенкина. Москва: Рудоми-но, 1999. 224 с.

83. Болотова Л. С. Системы искусственного интеллекта: модели и технологии, основанные на знаниях: учебник / Л. С. Болотова. Москва: Финансы и статистика, 2012. С. 38–39.

84. Бостром Н. Искусственный интеллект. Этапы. Угрозы. Стратегии / Н. Бостром; перевод с англ. С. Филина. Москва: Манн, Иванов и Фербер, 2016. 496 с.

85. Васильев В. В. Трудная проблема сознания / В. В. Васильев. Москва: Прогресс-Традиция, 2009. 272 с.

86. Гринбаум А. Машина-доносчица: как избавиться искусственный интеллект от зла / А. Гринбаум. Москва: ТрансЛит, 2017. 76 с.

87. Де Гарис Х. Искусственный мозг: подход с развитым модулем нейронной сети / Х. Де Гарис // World Scientific. 2010. 400 с.

88. Деннет Д. Виды психики. На пути к пониманию сознания / Д. Деннет; пер. А. Веретенникова. Москва: Идея-Пресс, 2004.

89. Иванов Д. В. Виртуализация общества / Д. В. Иванов. Санкт-Петербург, 2002. С. 24.

90. Интеллектуальные системы управления. Теория и практика: учебное пособие. Москва: Радиотехника, 2009. 392 с.

91. Кутырев В. А. Культура и технология: борьба миров / В. А. Кутырев. Москва: Прогресс-Традиция, 2001. 240 с.

92. Ленк Х. Размышления о современной технике / Х. Ленк. Москва: Аспект-Пресс, 1996. 183 с.

93. Майер-Шёнбергер В. Большие данные: революция, которая изменит то, как мы живем, работаем и мыслим / В. Майер-Шёнбергер, К. Кукьер; перевод с англ. Москва: Манн, Иванов и Фербер, 2014. 240 с.

94. Маклюэн М. Понимание медиа: внешнее расширение человека / М. Маклюэн. Москва: Канон-пресс-Ц, 2003. 464 с.

95. Маркузе Г. Одномерный человек / Г. Маркузе // Исследование идеологии развитого индустриального общества. Москва: Reefbook, 1994. 368 с.

96. Мелюхин И. С. Информационное общество: истоки, проблемы, тенденции развития / И. С. Мелюхин. Москва: Изд-во Моск. унта, 1999. 206 с.

97. Мински М. Фреймы для представления знаний / М. Мински. Москва: Мир, 1979. 151 с.

98. Морхат П. М. Искусственный интеллект: правовой взгляд / П. М. Морхат // Институт государственно-конфессиональных отношений и права. Москва: Буки Веди, 2017. 257 с.

99. Морхат П. М. Право интеллектуальной собственности и искусственный интеллект / П. М. Морхат. Москва: ЮНИТИ-ДАНА, 2018. 121 с.

100. Мосечкин И. Н. Искусственный интеллект и уголовная ответственность: проблемы становления нового вида субъекта преступления / И. Н. Мосечкин // Вестник Санкт-Петербургского университета. Право. 2019. № 10(3). С. 461–476.

101. Мэмфорд Л. Техника и природа человека / Д. Мэмфорд // Новая технократическая волна на Западе. Москва: Прогресс-Традиция, 1986. С. 225–239.

102. Нейсбит Д. Высокая технология, глубокая гуманность: технологии и наши поиски смысла / Д. Нейсбит при участии Наны Нейсбит и Дугласа Филиппса; перевод с англ. А. Н. Анваера // Москва: Изд-во АСТ; Транзиткнига, 2005. 381 с.

103. Нильсон Н. Искусственный интеллект: методы поиска решений / Н. Нильсон; перевод с англ. В. Л. Стефанюка; под редакцией С. В. Фомина. Москва: Мир, 1973. 272 с.

104. Ортега-и-Гассет Х. Размышления о технике / Х. Ортега-и-Гассет // Избранные труды / перевод с исп.; сост., предисл. и общ. ред. А. М. Руткевича. Москва: Весь мир, 1997. С. 164–232.

105. Пенроуз Р. Новый ум короля: о компьютерах, мышлении и законах физики / Р. Пенроуз; перевод с англ.; под редакцией В. О. Малышенко. 3-е изд. Москва: Изд-во ЛКИ, 2008. С. 328.

106. Рело Ф. Техника и ее связь с задачей культуры / Ф. Рело. Санкт-Петербург: Типография Министерства путей сообщения, 1885. 27 с.

107. Ридлер А. Германские высшие учебные заведения и запросы двадцатого столетия / А. Ридлер. Санкт-Петербург: Типография Р. Голике, 1900. 30 с.

108. Розин В. М. Понятие и современные концепции техники / В. М. Розин. Москва: Изд-во Ин-та философии РАН, 2006. 255 с.

109. Ручкин В. Н. Универсальный искусственный интеллект и экспертные системы / В. Н. Ручкин, В. А. Фулин. Санкт-Петербург: БХВ-Петербург, 2009. 240 с.

110. Серль Дж. Р. Сознание, мозг и программы / Дж. Р. Серль // Аналитическая философия: становление и развитие: антология / под общей редакцией и составитель А. Ф. Грязнов. Москва, 1998. 528 с.

111. Смирнов С. А. Человек перехода / С. А. Смирнов; ответственный за вып. П. А. Носова. Новосибирск, 2006. 177 с.

112. Степин В. С. Философия науки и техники / В. С. Степин, В. Г. Горохов, М. А. Розов. Москва, 1996. 380 с.

113. Тополь Э. Будущее медицины: ваше здоровье в ваших руках / Э. Тополь. Москва: Альпина нон-фикшн, 2016. 491 с.

114. Турчин А. В. Футурология. XXI век. Бессмертие или глобальная катастрофа? / А. В. Турчин, М. А. Бахтин. Москва, 2013. URL: <https://libking.ru/books/nonf-/nonf-publicism/205876-aleksey-turchin-rossiyskaya-akademiya-nauk.html>. (дата обращения: 12.09.2022).

115. Тьюринг А. Может ли машина мыслить / А. Тьюринг. Москва: Едиториал УРСС, Ленанд, 2016. 128 с.

116. Филипова И. А. Правовое регулирование искусственного интеллекта: учебное пособие. 2-е изд. Нижний Новгород: Изд-во Нижегород. гос. ун-та, 2022. 275 с.

117. Философский энциклопедический словарь / Л. Ф. Ильичев, П. Н. Федосеев, С. М. Ковалев, В. Г. Панов. Москва: Советская энциклопедия, 1983. С. 622.

118. Форд М. Роботы наступают: развитие технологий и будущее без работы / М. Форд; перевод с англ. С. Чернина. Москва: Альпина нон-фикшн, 2016. 572 с.

119. Фукуяма Ф. Наше постчеловеческое будущее: последствия биотехнологической революции / Ф. Фукуяма; перевод с англ. М. Б. Левина. Москва: Изд-во АСТ; ЛЮКС, 2004. 349 с.

120. Хабермас Ю. Будущее человеческой природы / Ю. Хабермас. Москва: Весь мир, 2002. 144 с.

121. Хабермас Ю. Техника и наука как «идеология» / Ю. Хабермас. Москва: Праксис, 2007. 208 с.

122. Хайдеггер М. Вопрос о технике / М. Хайдеггер // Время и бытие: статьи и выступления: перевод с нем. Москва, 1993. 49 с.

123. Юдковски Э. Систематические ошибки в рассуждениях, потенциально влияющие на оценку глобальных рисков. Новые технологии и продолжение эволюции человека? / Э. Юдковски // Трансгуманистический проект будущего. Москва: URSS, 2008. С. 182–225.

Периодические издания

124. Аверинская С. А. Создание искусственного интеллекта с целью злонамеренного использования в уголовном праве Российской Федерации / С. А. Аверинская, А. А. Севостьянова // Закон и право. 2019. № 2. С. 94–96.

125. Апресян Р. Г. Этика и дискуссии об искусственном интеллекте / Р. Г. Апресян // Теоретическая и прикладная этика: традиции и перспективы — 2019. К грядущему цифровому обществу. Опыт этического прогнозирования (100 лет со дня рождения Д. Белла. 1919–2019): материалы XI Международной конференции (Санкт-Петербургский государственный университет, 21–23 ноября 2019 г.) / ответственный редактор В. Ю. Перов. Санкт-Петербург, 2019. С. 169–170.

126. Архипов В. В. Искусственный интеллект и автономные устройства в контексте права: о разработке первого в России закона

о робототехнике / В. В. Архипов, В. Б. Наумов // Труды СПИИРАН. 2017. № 6. С. 157–170.

127. Архипов В. В. О некоторых вопросах теоретических оснований развития законодательства о робототехнике: аспекты воли и правосубъектности / В. В. Архипов, В. Б. Наумов // Закон. 2017. № 5. С. 157–170.

128. Асеева И. А. Искусственный интеллект и большие данные: этические проблемы практического использования (аналитический обзор) / И. А. Асеева // Социальные и гуманитарные науки. Отечественная и зарубежная литература. Сер. 8. Науковедение. 2022. № 2. С. 89–98.

129. Бадмаева М. Х. Повседневная жизнь человека в умном городе / М. Х. Бадмаева // Вестник Бурятского государственного университета. Философия. 2020. Вып. 4. С. 34.

130. Баева Л. В. Цифровизация образовательного пространства: эмоциональные риски и эффекты / Л. В. Баева, С. А. Храпов // Вопросы философии. 2022. № 4. С. 16–24.

131. Баева Л. Социокультурные изменения в условиях развития высоких технологий / Л. Баева // Инноватика и экспертиза: научный журнал. 2012. № 2(11). С. 110–119.

132. Бахтеев Д. В. Применение искусственного интеллекта в деятельности арбитражных судов РФ: перспективные направления и проблемы / Д. В. Бахтеев, Л. В. Тарасова // Вестник Костромского государственного университета. 2020. Т. 26, № 4. С. 249–254.

133. Горохов В. Г. Место и роль философии техники и современной философии и её органическая связь с философией науки / В. Г. Горохов // Философия науки. 2011. № 1. С. 181–199.

134. Горохов В. Г. Социальные проблемы нанотехнологии / В. Г. Горохов // Высшее образование в России. 2008. № 3. С. 84–98.

135. Готовцев П. М. Характеристика проектов стандартов на этичный искусственный интеллект IEEE / П. М. Готовцев, Г. В. Ройзенсон // Этика и «цифра». 2020. URL: <https://ethics.cdto.center/ieee> (дата обращения: 12.04.2022).

136. Гусейнов А. А. Размышления о прикладной этике: доклад на основе статьи: размышления о прикладной этике / А. А. Гусейнов // Ведомости НИИПЭ. Вып. 25. Общепрофессиональная этика. Тюмень: Изд-во НИИПЭ, 2004.

137. Евстратов А. Э. Пределы применения искусственного интеллекта (правовые проблемы) / А. Э. Евстратов, И. Ю. Гученков // Правоприменение. 2020. Т. 4(2). С. 13–19.

138. Карпов В. Э. К вопросу об этике и системах искусственного интеллекта / В. Э. Карпов, П. М. Готовцев, Г. В. Ройзензон // Философия и общество. 2018. № 2(87). С. 84–105.

139. Козаев Н. Ш. Состояние уголовной политики и вопросы преодоления кризисных явлений в уголовном праве / Н. Ш. Козаев // Юридический вестник ДГУ. 2016. № 1. С. 98.

140. Лаптев В. А. Электронные доказательства в арбитражном процессе / В. А. Лаптев // Российская юстиция. 2017. № 2. С. 56–59.

141. Маслов С. Ю. Обратный метод установления выводимости в классическом исчислении предикатов / С. Ю. Маслов // ДАН СССР. 1964. Т. 159, № 1. С. 17–20.

142. Нагель Т. Каково быть летучей мышью? / Т. Нагель // Хоф-штадтер Д., Деннет Д. Глаз разума / перевод с англ. М. А. Эскиной. Самара: Бахрах-М, 2003. С. 349–360.

143. Наумов В. Б. Право в эпоху цифровой трансформации: в поисках решений / В. Б. Наумов // Российское право: образование, практика, наука. 2018. № 6(108). С. 4–11.

144. Незнамов А. В. О концепции регулирования технологий искусственного интеллекта и робототехники в России / А. В. Незнамов // Закон. 2020. № 1. С. 171–185.

145. Осипов Г. С. Искусственный интеллект: состояние исследований и взгляд в будущее / Г. С. Осипов // Новости искусственного интеллекта. 2001. № 1. С. 3–13.

146. Понкин И. В. Искусственный интеллект с точки зрения права / И. В. Понкин, А. И. Редькина // Вестник Российского университета дружбы народов. Сер. Юридические науки. 2018. Т. 22, № 1. С. 91–109.

147. Попова А. В. Этические принципы взаимодействия с искусственным интеллектом как основа правового регулирования / А. В. Попова // Правовое государство: теория и практика. 2020. № 3(61). С. 34–43.

148. Попова А. В. Этические принципы взаимодействия с искусственным интеллектом как основа правового регулирования / А. В. Попова // Правовое государство: теория и практика. 2020. № 3(61). С. 34–43.

149. Разин А. В. Этика искусственного интеллекта / А. В. Разин // Философия и общество. 2019. № 1. С. 57–73.

150. Ракитов А. И. Наш путь к информационному обществу / А. И. Ракитов // Теория и практика общественно-научной информации. Москва: ИНИОН, 1989. С. 50–68.

151. Резаев А. В. Искусственный интеллект и искусственная социальность: новые явления и проблемы для развития медицинских наук / А. В. Резаев, Н. Д. Трегубова // Эпистемология и философия науки. 2019. Т. 56, № 4. С. 183–199.

152. Ройзензон Г. В. Проблемы формализации понятия этики в искусственном интеллекте / Г. В. Ройзензон // XVI национальная конференция по искусственному интеллекту с международным участием КИИ-2018. Москва, 2018. С. 245–252.

153. Синельникова В. Н. Права на результаты искусственного интеллекта / В. Н. Синельникова, О. В. Ревинский // Копирайт. 2017. № 4. С. 24–27.

154. Степин В. С. Научное познание и ценности техногенной цивилизации / В. С. Степин // Вопросы философии. 1989. № 10. С. 3–18.

155. Флоренский П. Органопроекция / П. Флоренский // Русский космизм: антология философской мысли. Москва: Педагогика-пресс, 1993. С. 149–162.

156. Шевченко А. И. К вопросу о создании искусственного интеллекта / А. И. Шевченко // Искусственный интеллект. 2016. № 2. С. 7–15.

157. Юдин Б. Г. Социальные технологии, их производство и потребление / Б. Г. Юдин // Эпистемология и философия науки. 2012. Вып. 31, № 1. С. 55–64.
158. Ясперс К. Смысл и назначение истории. Москва, 1994. С. 139.
159. Ястреб Н. А. Индустрия 4.0: киберфизические системы и интернет вещей / Н. А. Ястреб // Человек в технической среде: сборник научных статей / под редакцией Н. А. Ястреб. Вологда: Изд-во ВолГУ, 2015. С. 136–141.
160. Abdulkareem M., Leiner T., Petersen S. E. Artificial intelligence will transform cardiac imaging – opportunities and challenges. *Front Cardiovasc Med*. 2019; 6: 133.
161. Abrams J. J. Pragmatism, artificial intelligence, and posthuman bioethics: Shusterman, Rorty, Foucault. *Human Studies*. 2020; 27(3): 241–258.
162. Ahmad M. A., Teredesai A., Eckert C. Fairness, accountability, transparency in AI at scale: Lessons from national programs. *Proceedings of the 2020 Conference on Fairness, Accountability and Transparency*. 2020; 27–30: 690–699.
163. Angwin J., Larson J., Mattu S., Kirchner L. Machine Bias. There is software that is used across the county to predict future criminals. And it is biased against blacks. URL: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> (дата обращения: 13.08.2022).
164. Araujo T., Helberger N., Kruijemeier S., De Vreese C. H. In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & SOCIETY*. 2020; 35: 611–623.
165. Arendt H. Collective Responsibility // Responsibility and Judgment. New York: Schocken books, 2003. 295 p.
166. Arkin R. C. Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture / Eds. P. Wang, B. Goertzel, S. Franklin // Artificial general intelligence, 2008: Proceedings of the first AGI conference. Washington DC: ISO Press, 2008. Pp. 51–62.

167. Bahner J. E., Hüper A. D., Manzey D. Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. *International Journal of Human-Computer Studies*. 2008; 66(9): 688–699.

168. Balkin J. B. The Path of Robotics Law. *California Law Review*. 2015; 6: 45–60.

169. Bian Y., Xiang, Tong B., Feng B., Weng X. Artificial intelligence–assisted system in postoperative follow-up of orthopedic patients: exploratory quantitative and qualitative study. *Med. Internet Res.* 2020; 22(5): 23.

170. Bishop S. Anxiety, panic and self-optimization: Inequalities and the YouTube algorithm. *Convergence*. 2018; 24(1): 69–84.

171. Blease C., Kaptchuk T. J., Bernstein M. H., Mandl K. D., Halamka J. D., Desroches C. M. Artificial intelligence and the future of primary care: exploratory qualitative study of UK general practitioners' views. *Journal of Med. Internet Research*. 2019; 21(3): 1–24.

172. Block N. Troubles with functionalism. Minnesota Studies in the Philosophy of Science. *Troubles with functionalism, Minnesota Studies in the Philosophy of Science*. 1978: 261–325.

173. Bloustein E. J. Privacy as an aspect of human dignity: An answer to Dean Prosser. *Philosophical Dimensions of Privacy: An Anthology*. 1984: 156–202.

174. Brayne S., Christin A. Technologies of crime prediction: The reception of algorithms in policing and criminal courts. *Social Problems*. 2020; 68(3): 608–624.

175. Broadhurst R., Brown P., Maxim D., Trivedi H., Wang J. Artificial Intelligence and Crime // Research Paper, Korean Institute of Criminology and Australian National University Cybercrime Observatory, College of Asia and the Pacific. Canberra, 2019. Pp. 1–70.

176. Brooks R. A. Elephants don't play chess. *Robotics and Autonomous Systems*. 1990; 6(1–2): 3–15.

177. Burrell J. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data and Society*. 2016; 3(1): 1–12.

178. Carolan M. Automated agrifood futures: robotics, labor and the distributive politics of digital agriculture. *Peasant Stud.* 2020; 47: 184–207.

179. Caruana R., Lou Y., Gehrke J., Koch P., Sturm M., Elhadad N. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. *Sydney, NSW, Australia: ACM.* 2015; 30: 17–21.

180. Case A., Deaton A. Deaths of despair and the future of capitalism. Princeton University Press. 2020.

181. Castro D., New J. The Promise of Artificial Intelligence. Center for data innovation. 2016: 3. URL: <https://www2.datainnovation.org/2016-promise-of-ai.pdf> (дата обращения: 23.09.2022).

182. Chorzempa M., Triolo P., Sacks S. China's social credit system: a mark of progress or a threat to privacy? *Policy Briefs PB18–14, Peterson Institute for International Economics.* 2018; PB18-14.

183. Clark A., Oswald A. Unhappiness and unemployment. *Economic Journal.* 1994; 104(424): 648–659.

184. De Garis H. The Artilect War: A bitter controversy concerning whether humanity should build godlike massively intelligent machines / Eds. P. Wang, B. Goertzel, S. Franklin // Artificial general intelligence, 2008: Proceedings of the first AGI conference. Washington DC: ISO Press, 2008. Pp. 362–373.

185. Dyson G. Turing's cathedral: The origins of the Digital Universe. New-York: Vintage. 2012. 464 p.

186. European Parliament Report on artificial intelligence in education, culture and the audiovisual sector Committee on Culture and Education. 2021. URL: https://www.europarl.europa.eu/doceo/document/A-9-2021-0127_EN.html (access 12.07.2022).

187. Executive Order on Maintaining American Leadership in Artificial Intelligence. URL: <https://www.whitehouse.gov/presidential->

actions/executive-order-maintaining-american-leadership-artificial-intelligence/ (access 20.06.2022).

188. Floridi L. The end of an era: from self-regulation to hard law for the digital industry. *Philosophy & Technology*. 2021; 34; 4: 612–622.

189. Floridi L., Cowls J. A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*. 2022; 1(1): 535–545.

190. Floridi L., Cowls J., Beltrametti M. et al. AI4People - An ethical framework for a good AI society. *Minds and Machines*. 2018; 28(4): 689–707.

191. Fodor J. A. In critical condition: Polemical essays on cognitive science and the philosophy of mind. Cambridge, MA: MIT Press, 1998. P. 17.

192. Frankfurt H. G. Freedom of the Will and the Concept of a Person. *Journal of Philosophy*. 1971; 68: 5–20.

193. Friedman M. Autonomy, Gender, Politics. New York: Oxford University Press, 2003. 263 p.

194. Gasser U., Lenca M., Scheibner J., Sleight J., Vayena E. Digital tools against COVID-19: Taxonomy, ethical challenges, and navigation aid. *Healthpolicy*. 2020; 2: 425–434.

195. Goertzel B. Artificial General Intelligence: Concept, State of the Art, and Future Prospects. *Journal of Artificial General Intelligence*. 2014; 5(1): 1–46.

196. Goertzel B. Should humanity build a global AI nanny to delay the singularity until it's better understood? *Journal of consciousness studies*. 2012; 19: 96–111.

197. Goertzel B. Should humanity build a global AI nanny to delay the singularity until it's better understood? *Journal of consciousness studies*. 2012; 19: 96–111.

198. Grinbaum A., Chatila R., Devillers L., Ganascia J. G. Ethics in Robotics Research. *IEEE Robotics and Automation Magazine*. 2017; 24: 139–145.

199. Grogger J., Ivandic R., Kirchmaier T. Comparing conventional and machine-learning approaches to risk assessment in domestic abuse cases. *Journal of Empirical Legal Studies*. 2020; 18(1): 90–130.

200. Hackman J. R., Oldham G. R. Motivation through the design of work: Test of a theory. *Organizational Behavior and Human Performance*. 1976; 16(2): 258.
201. Haefner K. Mensch und Computer im Jahre 2000. Basel - Boston - Stuttgart: Birkhauser, 1994. 402 p.
202. Hagedorff T. The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds & Machines*. 2020; 30: 99–120.
203. Haridasani G. A. Are Algorithms Sexist? URL: <https://www.nytimes.com/2019/11/15/us/apple-card-goldman-sachs.html> (access 21.05. 2022).
204. Hashimoto D. A., Rosman G., Witkowski E. R. Computer vision analysis of intraoperative video: automated recognition of operative steps in laparoscopic sleeve gastrectomy. *Ann Surg*. 2019; 270(3): 21.
205. Hassabis D., Revell T. With AI, you might unlock some of the secrets about how life works. *New Scientist*. 2021; 249(3315): 44-49.
206. Holstein K., Wortman V., Daumé J., Dudík M., Wallach H. Improving fairness in machine learning systems: What do industry practitioners need? *Proceedings of the 2019 CHI conference on human factors in computing systems*. 2019: 1–16.
207. Jobin A., Ienca M., Vayena E. Artificial Intelligence: The Global Landscape of Ethics Guidelines. *Nature Machine Intelligence*. 2019; 1: 389–399.
208. Kellogg K. C., Valentine M. A., Christin A. Algorithms at work: The new contested terrain of control. *Academy of Management Annals*. 2020; 14(1): 366–410.
209. Khrapov S. A., Lopatinskaya T. D., Kashkarov A. M. Artificial intelligence as cognitive and sociocultural phenomenon. *The Turkish Online Journal of Design, Art and Communication*. 2018: 107.
210. Kim T. W., Scheller-Wolf A. Technological unemployment, meaning in life, purpose of business, and the future of stakeholders. *Journal of Business Ethics*. 2019; 160(2): 319–337.

211. Langlois R. N. Cognitive comparative advantage and the organization of work: Lessons from Herbert Simon's vision of the future. *Journal of Economic Psychology*. 2020; 24(2): 174.

212. Legg S., Hutter M. A collection of definitions of intelligence / B. Goertzel, P. Wang (Eds.). *Advances in artificial general intelligence: concept, architectures and algorithms*. 2007; 157: 17–24.

213. Rawls J. A Theory of Justice. Cambridge, Massachusetts: Belknap Press, 1971. 562 p.

214. Ryan M., Antoniou J., Brooks L., Jiya T., Macnish K., Stahl B. Research and Practice of AI Ethics: A Case Study Approach Juxtaposing Academic Discourse with Organisational Reality. *Science and Engineering Ethics*. 2021; 27(2): 16.

215. Shortliffe E. H., Sepúlveda M. J. Clinical decision support in the era of artificial intelligence. *JAMA*. 2018; 320(21): 2199–2200.

216. Sparrow R. Killer robots. *Journal of Applied Philosophy*. 2007; 24(1): 62–77.

217. Sparrow R. Predators or plowshares? Arms control of robotic weapons. *IEEE Technology and Society*. 2009; 28(1): 25–29.

218. Walsh T., Levy N., Bell G., Elliott A., Maclaurin J. et al. The effective and ethical development of artificial intelligence // ACOLA. 2019. URL: https://acola.org/wp-content/uploads/2019/07/hs4_artificial-intelligence-report.pdf (access 30.01. 2022).

219. Zerilli J., Knott A., Maclaurin J., Gavaghan C. Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard? *Philosophy and Technology*. 2019; 32(4): 661–683.

Диссертации и авторефераты диссертаций

220. Заладина М. В. Социально-философский анализ духовного отчуждения: автореферат диссертации на соискание ученой степени кандидата философских наук / М. В. Заладина. Нижний Новгород, 2020. 25 с.

221. Маслов В. М. Высокие технологии и феномен постчеловеческого в современном обществе: автореферат диссертации на со-

искание ученой степени кандидата философских наук / В. М. Маслов. Нижний Новгород, 2014. 38 с.

222. Морхат П. М. Правосубъектность искусственного интеллекта в сфере права интеллектуальной собственности: гражданско-правовые проблемы: диссертация на соискание ученой степени доктора юридических наук / П. М. Морхат. Москва, 2019. С. 30–31.

223. Цуркан Е. Г. Социокультурная динамика и интернет-технологии: социально-философский анализ: автореферат диссертации на соискание ученой степени кандидата философских наук / Е. Г. Цуркан. Москва, 2021. 33 с.

224. Чепьюк О. Р. Экономическая бессубъектность как фактор дегуманизации социальных отношений: диссертация на соискание ученой степени доктора философских наук / О. Р. Чепьюк. Нижний Новгород, 2020. 371 с.

225. Щитова А. А. Правовое регулирование информационных отношений по использованию систем искусственного интеллекта: диссертация на соискание ученой степени кандидата юридических наук / А. А. Щитова. Москва, 2021. 225 с.

Интернет-источники

226. Греф: Сбербанк сможет принимать 80% решений искусственным интеллектом. URL: <https://ria.ru/economy/20160908/1476449735.html> (дата обращения: 19.07.2022).

227. Европарламентарий могут предоставить роботам юридический статус. URL: <https://ria.ru/world/20170114/1485715425.html?injj=1> (дата обращения: 19.08.2022).

228. Ефимова Е. Новое слово в живописи: искусственный интеллект «пишет» картины в уникальном стиле / Е. Ефимова. URL: <http://www.vesti.ru/doc.html?id=2905726> (дата обращения: 16.09.2022).

229. Искусственный интеллект как ключевой фактор цифровизации глобальной экономики. URL: <https://www.iksmedia.ru/news/5385191-Iskusstvennyj-intellekt-II-Artifici.html> (дата обращения: 29.09.2021).

230. Конев С. И. Этико-правовые проблемы регулирования искусственного интеллекта и робототехники в отечественном и зарубежном праве / С. И. Конев. URL: <https://cyberleninka.ru/article/n/etiko-pravovye-problemy-regulirovaniya-iskusstvennogo-intellekta-i-robototekhniki-v-otechestvennom-i-zarubezhnom-prave> (дата обращения: 19.09.2022).

231. Ксенофонтова А. Бесправный механизм: почему ученые выступили против присвоения роботам статуса «электронной личности» / А. Ксенофонтова. URL: <https://russian.rt.com/science/article/504118-roboty-evroparlament-yuridicheskoye-lico> (дата обращения: 17.05.2022).

232. Об испытаниях высокоавтоматизированных автотранспортных средств. URL: <https://lovdata.no/dokument/NL/lov/20171215112?q=Lov%20om%20utpr%C3%B8ving%20av%20selvkj%C3%B8rende> (дата обращения: 12.09.2022).

233. Серль Дж. Р. Разум мозга — компьютерная программа? / Дж. Р. Серль. URL: <https://psychosearch.ru/teoriya/psikhika/338-searle-john-razum-mozga-kompyuternaya-programma> (дата обращения: 12.03.2022).

234. Специальный комитет по искусственному интеллекту (CAHAI). URL: https://sk.ru/media/documents/CAHAI_AI_Research.pdf (дата обращения: 19.11.2022).

235. Чат-бот от Microsoft за сутки научился ругаться и стал расистом. URL: <https://www.interfax.ru/world/500152> (дата обращения: 16.09.2022).

236. ACM FAccT Conference. URL: <https://facctconference.org/> (access 04.01.2022).

237. Asher-Schapiro A. Amazon AI van cameras spark surveillance concerns // News.Trust.Org. 2021. URL: <https://news.trust.org/item/20210205132207-c0mz7/> (access 12.09.2022).

238. Boeing во время аудита обнаружила ошибки в обновленном ПО для самолетов модели 737 MAX. URL: <https://habr.com/ru/news/t/484434/> (дата обращения: 12.09.2022).

239. Bossmann Dzh. Top 9 Ethical Issues in Artificial Intelligence. URL: <https://hr-portal.ru/article/9-glavnyh-eticheskikh-problem-iskusstvennogo-intellekta> (access 22.03.2022).

240. Castro D., New J. The Promise of Artificial Intelligence. Center for data innovation. 2016. URL: <https://www2.datainnovation.org/2016-promise-of-ai.pdf> (access 23.09. 2022).

241. Dawson D., Schleiger E., Horton J., McLaughlin J., Robinson C. et al. Artificial Intelligence - Australia's Ethics Framework // Data 61 CSIRO. Australia, 2019. URL: <https://www.csiro.au/-/media/D61/Reports/Artificial-Intelligence-ethics-framework.pdf> (access 21.09.2022).

242. Dignum V. Responsible Artificial Intelligence – from Principles to Practice Virginia Dignum. URL: file:///C:/Users/oem/Downloads/Responsible_Artificial_Intelligence_--_from_Principi.pdf (access 12.09.2021).

243. Ellen M Harper , Sara Parkerson, Powering Big Data for Nursing Through Partnership. URL: <https://pubmed.ncbi.nlm.nih.gov/26340243/> (access 14.03.2022).

244. Equifax Launches NeuroDecision Technology. 2018. URL: <https://investor.equifax.com/news-events/press-releases/detail/203/equifax-launches-neurodecision-technology> (access 12.09.2022).

245. Government of Canada, “Directive on Automated Decision-Making”, 2019. URL: <https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592> (access 12.09.2022).

246. GPT 4 – OPENAI. URL: <https://openai.com/product/gpt-4> (access 02.04.2023).

247. Grossman L. 2045: The year man becomes immortal // Time. 2011. URL: <http://content.time.com/time/magazine/article/0,9171,2048299,00.html> (access 21.03.2022).

248. Iamus, a music-making computer, could be the next Mozart. URL: <https://www.vice.com/en/article/pgg8yy/iamus-a-music-making-computer-could-be-the-next-mozart> (access 21.01.2023).

249. McCarthy J. What is Artificial Intelligence? // Stanford University. 2007. URL: <http://www-formal.stanford.edu/jmc/whatisai>. (access 21.09.2022).

250. Pardes A. AI can run your work meetings now. Wired. URL: <https://www.wired.com/story/ai-can-run-work-meetings-now-headroom-clockwise> (access 14.09.2022).

251. Report of COMEST on robotics ethics. 2017. URL: <https://unesdoc.unesco.org/ark:/48223/pf0000253952> (access 31.11.2022).

252. Roger C. Regulatory alternatives for AI. ScienceDirect (2019). URL: <http://www.rogerclarke.com/EC/AIR-Final.pdf> (access 12.12.2022).

253. Search Enterprise AI. Artificial intelligence. URL: <https://searchenterpriseai.techtarget.com/definition/AI-Artificial-Intelligence> (access 15.05.2022).

254. Shakey the Robot – SRI International. URL: <https://www.sri.com/hoi/shakey-the-robot/> (access 12.05.2022).

255. Singularity Institute for Artificial Intelligence, Seed AI. General Intelligence and Seed AI. URL: http://singinst.org/ourresearch/publications/GISAI/paradigms/seedAI.html#glossary_crystalline (access 12.09.2021).

256. Work Fusion объединяет интеллектуальную автоматизацию с облачным хранилищем. URL: <https://robroy.ru/intellektualnuyu-avtomatizacziyu-s-oblachnyim-xranilishhem.html> (дата обращения: 12.02.2023).

ПРИЛОЖЕНИЕ

№	Название документа	Организация	Тип
1	California Consumer Privacy Act (CCPA) («Закон Калифорнии о конфиденциальности потребителей»), 2018	Законодательное собрание штата Калифорния, США	Закон
2	EU-U.S. and Swiss-U.S. Privacy Shield («Закон о защите конфиденциальности»), 2016	Министерство торговли США, Европейская комиссия и администрация Швейцарии	Закон
3	Personal information protection and electronic documents act («Закон о защите личной информации и электронных документов»), 2000	Правительство Канады	Закон
4	Data Protection Act («Закон о защите данных»), 2018	Палата лордов, Великобритания	Закон
5	General data protection regulation («Общее положение о защите данных»), 2016	ЕС	Закон
6	Fair Credit Reporting Act («Закон о достоверной кредитной отчетности»), 2018	Федеральная торговая комиссия США, Бюро финансовой защиты потребителей, США	Закон
7	Personal information protection law («Закон о защите личной информации»), 2021	ВСНП, КНР	Закон
8	Cyber security law («Закон о киберзащите»), 2016	ВСНП, КНР	Закон
9	Protection from online falsehoods and manipulation act («Закон о защите от лжи и манипуляций в интернете»)	Парламент, Сингапур	Закон

ПРИЛОЖЕНИЕ

10	Data protection law («Закон о защите данных»), 2021	Федеральный национальный совет, ОАЭ	Закон
11	Bundesdatenschutzgesetz («Федеральный закон о защите данных»), 2009	Парламент, Германия	Закон
12	Распоряжение правительства РФ от 19 августа 2020 г. № 2129-р «Об утверждении Концепции развития регулирования отношений в сфере технологий искусственного интеллекта и робототехники на период до 2024 г.», 2020	Правительство, РФ	Концепция
13	Федеральный закон от 24.04.2020 г. № 123-ФЗ «О проведении эксперимента по установлению специального регулирования в целях создания необходимых условий для разработки и внедрения технологий искусственного интеллекта в субъекте Российской Федерации - городе федерального значения Москве и внесении изменений в статьи 6 и 10 Федерального закона "О персональных данных"», 2020	Парламент, РФ	Закон
14	«Об экспериментальных правовых режимах в сфере цифровых инноваций в Российской Федерации» (Законопроект № 922869-7), 2020	Правительство РФ	Законопроект
15	Loi pour une Republique numerique. No 2016-1321 («Закон Франции о цифровой республике»), 2016	Национальное собрание Франции	Закон
16	Intelligent robots development and distribution promotion act («Закон о	Парламент Южной Кореи	Закон

ПРИМЕНЕНИЕ СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА:
СОЦИАЛЬНО-ФИЛОСОФСКИЕ ПРОБЛЕМЫ И ПРИНЦИПЫ

	развитии и распространении интеллектуальных роботов»), 2016		
17	Указ Президента РФ от 10.10.2019. № 490 «О развитии искусственного интеллекта в Российской Федерации» (вместе с «Национальной стратегией развития искусственного интеллекта на период до 2030 г.») // Собрание законодательства РФ. 2019. № 41. Ст. 5700; Распоряжение правительства РФ от 19 августа 2020 г. № 2129-р «Об утверждении Концепции развития регулирования отношений в сфере технологий искусственного интеллекта и робототехники на период до 2024 г.», 2020	Правительство РФ	Национальная стратегия РФ
18	State Council, Notice of Issuing New Generation Artificial Intelligence Development Plan («План развития искусственного интеллекта нового поколения») 2017	Государственный совет КНР	Национальная стратегия КНР
19	Bundesregierung, Eckpunkte der Bundesregierung für eine Strategie Künstliche Intelligenz («Основные положения Федеральной правительственной стратегии по искусственному интеллекту»), 2018	Федеральное правительство Германии	Национальная стратегия Германии
20	For a meaningful artificialintelligencetowards a french and european strategy («На пути к осмысленному искусственному интеллекту»), 2018	Парламент Франции	Национальная стратегия Франции
21	National Strategy forArtificial Intelligence («Национальная стра-	Правительство Дании, Министерство	Национальная

ПРИЛОЖЕНИЕ

	тегия по искусственному интеллектту»), 2020	финансов и Министертство промышленности, бизнеса и финансов	стратегия Дании
22	Pan-Canadian Artificial Intelligence Strategy («Панканадская стратегия в области искусственного интеллекта»), 2018	CIFAR (Canadian Institute for Advanced Research) - канадская государственная научно-исследовательская организация	Национальная стратегия Канады
23	AI Strategy 2019 AI for Everyone: People, Industries, Regions and Governments («Стратегия в области искусственного интеллекта, предназначенная для всех: людям, отраслям, регионам и правительству»), 2019	Совет по продвижению комплексной инновационной стратегии	Национальная стратегия Японии
24	National Approach to artificial intelligence («Национальный подход к искусственному интеллекту»), 2018	Государственные учреждения Швеции	Национальная стратегия Швеции
25	Eesti 2035 («Эстония 2035»), 2021	Государственная канцелярия и Министерство финансов	Национальная стратегия Эстонии
26	Australia's AI Ethics Principles, 2019	Правительство Австралии, Департамент промышленности, науки и ресурсов	Руководство
27	Artificial intelligence principles and ethics, 2020	Правительство ОАЭ	Руководство

ПРИМЕНЕНИЕ СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА:
СОЦИАЛЬНО-ФИЛОСОФСКИЕ ПРОБЛЕМЫ И ПРИНЦИПЫ

28	Directive on Automated Decision-Making («Директива по автоматизированному принятию решений»), 2019	Правительство Канады	Директива
29	Asilomar AI Principles («Азильмарские принципы искусственного интеллекта»), 2017	Институт будущего жизни – добровольная ассоциация ученых, предпринимателей, экспертов в Бостоне	Результаты международной научной конференции
30	Montreal Declaration: Responsible AI («Монреальская декларация по ответственному ИИ»), 2017	Монреальский университет	Декларация
31	Draft Recommendation on the Ethics of Artificial Intelligence («Первоначальный вариант проекта рекомендации об этических аспектах искусственного интеллекта»), 2020	Специальная группа экспертов по подготовке проекта рекомендации об этических аспектах искусственного интеллекта от ЮНЭСКО, Франция	Рекомендации
32	IEEE Ethically Aligned Design («Этически согласованный дизайн»), 2019	IEEE - Международная некоммерческая ассоциация специалистов в области техники	Трактат
33	Preliminary study on a possible standard-setting instrument on the ethics of artificial intelligence («Предварительное исследование возможности подготовки нормативного акта по вопросам этики применения искусственного интеллекта»), 2019	ЮНЭСКО	Программный документ и документ совещаний

ПРИЛОЖЕНИЕ

34	AI in the UK: ready, willing and able? («Пять всеобъемлющих принципов для кода ИИ»), 2018	Комитет по искусственному интеллекту Палаты лордов, Великобритания	Доклад
35	Terms of Reference («Круг полномочий»), 2020	Global Partnership on Artificial Intelligence, Франция и Канада	Рекомендации
36	Civil Law Rules on Robotics (2015/2103(INL)) («Резолюция 2015/2103 (INL) от Парламента ЕС»), 2017	Европейский парламент, Комиссии по гражданско-правовым нормам в отношении робототехники, ЕС	Резолюция
37	Strategy for automated and connected driving («Стратегия автоматизированного и подключенного к сети вождения»), 2019	Федеральное министерство транспорта и цифровой инфраструктуры, Германия	Стратегия
38	Automated vehicle trial guidelines («Руководство по испытаниям автоматизированных транспортных средств»), 2017	Национальная комиссия по автомобильному транспорту, Австралия	Руководство
39	Résolution visant à interdire l'utilisation, par la Défense belge, de robots tueurs et de drones armés («Резолюция о запрете применения автономных смертельных систем вооружения, Бельгия»), 2018	Бельгийская палата народных представителей, Бельгия	Резолюция
40	Auomated and connected driving: report («Рекомендация по беспилотным автомобилям от Комиссии по этике при Министерстве транспорта и цифровой инфраструктуры Германии»), 2017	Федеральное министерство цифровых технологий и транспорта является министерством на уровне кабинета министров, Этиче-	Доклад, рекомендация

ПРИМЕНЕНИЕ СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА:
СОЦИАЛЬНО-ФИЛОСОФСКИЕ ПРОБЛЕМЫ И ПРИНЦИПЫ

		ская комиссия, Германия	
41	Report of COMEST on robotics ethics («Доклад об этике робототехники от Юнеско»), 2017	ЮНЭСКО	Доклад, рекомендация
42	中国《人工智能标准化白皮书2018》发布完整版 (附下载) («Белая книга по стандартизации ИИ»), Китай, 2018	Национальная обща- я группа по стандартизации искус- ственного интел- лекта и Консульта- тивная группа экс- пертов, КНР	Реко- мента- ции
43	《新一代人工智能伦理规范》发布, («Этический кодекс искусствен- ного интеллекта нового поколения»), 2021	Национальный профессиональный комитет по управ- лению искусствен- ным интеллектом нового поколения, Министерство науки и технологий, КНР	Кодекс
44	Модельная конвенция робототех- ники и искусственного интеллек- та, 2018	Исследовательский центр проблем ре- гулирования робо- тотехники и искус- ственного интел- лекта, РФ	Концеп- ция
45	Developing Standards for Artificial Intelligence: Hearing Australia's Voice («Разработка стандартов для искусственного интеллекта: Слышать голос Австралии»), 2019	Standards Australia, неправительствен- ная, некоммерче- ская организация по стандартизации	Доклад
46	A Plan for Federal Engagement in Developing Technical Standards and Related Tools («Проект плана фе- дерального участия в разработке	Национальный ин- ститут стандартов и технологий (NIST) – лаборатория физи-	Стан- дарт

ПРИЛОЖЕНИЕ

	технических стандартов ИИ и связанных с ними инструментов»), 2019	ческих наук и регулирующее агентство Министерства торговли США	
47	European ethical Charter on the use of Artificial Intelligence in judicial systems and their environment («Европейская этическая хартия Совета Европы по использованию ИИ в судебных системах»), 2018	Европейская комиссия по эффективности правосудия (CEPEJ)	Хартия
48	Guidelines on artificial intelligence and data protection («Руководство по защите данных при использовании ИИ»), Strasbourg, 2019	Консультативный комитет конвенции о защите лиц в отношении автоматической обработки персональных данных, Главное управление по правам человека и верховенству закона, Франция	Конвенция
49	Declaration by the Committee of Ministers on the manipulative capabilities of algorithmic processes («Декларация Комитета министров о манипулятивных возможностях алгоритмов»), 2019	Группа экспертов Совета Европы	Декларация
50	Unboxing Artificial Intelligence: 10 steps to protect Human Rights («10 шагов для защиты прав человека при использовании ИИ»), 2019	Комиссар СЕ по правам человека	Рекомендации
51	Recommendation of the Council on Artificial Intelligence («Принципы ИИ и рекомендации по национальной политике Экспертной	Совет по искусственному интеллекту, ОЭСР	Рекомендации

ПРИМЕНЕНИЕ СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА:
СОЦИАЛЬНО-ФИЛОСОФСКИЕ ПРОБЛЕМЫ И ПРИНЦИПЫ

	группы по искусственному интеллекту»), 2019		
52	Addressing the impacts of Algorithms on Human Rights («Устранение воздействия алгоритмов на права человека»), 2019	Комитет экспертов по правам человека, по аспектам автоматизированной обработки данных и различных форм искусственного интеллекта, Комитет министров Совета Европы	Проект рекомендаций
53	Artificial Intelligence. Australia's Ethics Framework: A Discussion Paper Department of Industry Innovation and Science («Этические рамки искусственного интеллекта в Австралии»), 2019	Организация научных и промышленных исследований Common Well (CSIRO), под руководством Министерства инноваций и науки правительства Австралии	Дискуссионный документ
54	Work in the Age of Artificial Intelligence. Four Perspectives on the Economy, Employment, Skills and Ethics («Работа в эпоху искусственного интеллекта: четыре взгляда на экономику, занятость, навыки и этику»), 2018	Министерство экономики и занятости, Финляндия	Доклад
55	Tieto's AI Ethics Guidelines («Этические принципы в области искусственного интеллекта»), 2018	Tieto - скандинавская компания по предоставлению ИТ-услуг для промышленности и сферы обслуживания, Финляндия	Рекомендации
56	How Can Humans Keep the Upper Hand? Report on the Ethical Matters	Национальная комиссия по инфор-	Отчет

ПРИЛОЖЕНИЕ

	Raised by AI Algorithms («Как люди могут одержать верх? Отчет об этических вопросах, поднятых алгоритмами ИИ»), Франция, 2017	матике и свободам, Французское управление по защите данных (CNIL), Франция	
57	The Ethics of AI Ethics – An Evaluation of Guidelines («Этика ИИ – оценка руководящих принципов»), 2020	T. Hagendorff, Университет Тюбинген, Германия	Руководство
58	Understanding artificial intelligence ethics and safety («Понимание этики и безопасности искусственного интеллекта»), 2019	Д. Лесли, Институт Алана Тьюринга, США	Руководство
59	The global landscape of AI ethics guidelines («Руководство по этике ИИ в мире»), 2019	А. Джобин и др., Высшая техническая школа, Цюрих, Швейцария	Руководство
60	AI HLEG. “Ethics Guidelines for Trustworthy AI («Руководство по этике для надежного ИИ»), 2019	Европейская комиссия	Руководство
61	From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices («От “чего” к вопросу “как”: первоначальный обзор общедоступных инструментов, методов и исследований по этике ИИ для воплощения принципов в практику»), 2020	J. Morleyetal, Оксфордский институт интернета, Великобритания	Руководство
62	From Principles to Practice: A ninter disciplinary framework to operationalize AI ethics («От принципов к практике: междисциплинарная основа для практической реализации этики ИИ»), 2020	С. Халленслебен, AIEI Group, Германия	Руководство

ПРИМЕНЕНИЕ СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА:
СОЦИАЛЬНО-ФИЛОСОФСКИЕ ПРОБЛЕМЫ И ПРИНЦИПЫ

63	Building Ethics into AI: Lessons Learned from Pioneers in the Trenches («Внедрение этических норм в ИИ: уроки, полученные от первопроходцев»), 2019	К. Бакстеп, Salesforce Research, США	Руководство
64	Technical and Organizational Best Practices («Технические и организационные передовые практики»), 2021	FBPML (The Foundation for Best Practices in Machine Learning), Нидерланды	Руководство
65	Advancing AI ethics beyond compliance From principles to practice («Продвижение этики ИИ за пределы соблюдения требований. От принципов к практике»), 2020	В. Геринг, IBM, США	Руководство
66	Guidelines for AI Procurement («Руководство по закупкам с использованием ИИ»), 2019	World Economic Forum, Швейцария	Руководство
67	Guidance on AI an data protection («Руководство по ИИ и защите данных»), 2020	Information Commissioner Office, Управление комиссара по информации, Великобритания	Руководство
68	Guidance on the AI auditing framework («Руководство по системе аудита ИИ»), 2020	Information Commissioner Office, Управление комиссара по информации, Великобритания	Руководство
69	Building Ethics into Privacy Frameworks for Big Data and AI («Встраивание этики в рамки конфиденциальности для больших данных и искусственного интеллекта»), 2018	IAPP, США	Руководство

ПРИЛОЖЕНИЕ

70	Using Big Data to Improve the Quality of Care and Outcomes («Руководящие принципы использования больших данных в сестринском деле — использование больших данных для улучшения качества ухода и результатов»), 2018	HIMSS (Общество медицинских информационных и управленческих систем)	Руководство
71	Responsible use of artificial intelligence (AI) («Ответственное использование искусственного интеллекта»), 2019	Правительство Канады	Руководство
72	Artificial Intelligence — Australia's Ethics Framework («Искусственный интеллект — австралийская система этических норм»), 2019	CSIRO, Организация научных и промышленных исследований содружества, Австралия	Дискуссионное исследование
73	An Accountability Framework for Federal Agencies and Other Entities («Система подотчетности федеральных агентств и других организаций»), 2021	Счетная палата, США	Рекомендации
74	Universal Guidelines for Artificial Intelligence («Универсальные принципы для искусственного интеллекта»), 2018	The Public Voice, комиссия по интересам сообщества, созданная при Информационном центре электронной конфиденциальности (EPIC), Бельгия	Рекомендации

ПРИМЕНЕНИЕ СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА:
СОЦИАЛЬНО-ФИЛОСОФСКИЕ ПРОБЛЕМЫ И ПРИНЦИПЫ

75	African Observatory on Responsible Artificial Intelligence («Африканская обсерватория ответственного искусственного интеллекта»), 2022	Research ICT Africa, африканский аналитический центр, работающий над устранением стратегического пробела в развитии устойчивого информационного общества и цифровой экономики, ЮАР	Рекомендации
76	Top 10 principles for ethical artificial intelligence («10 принципов для этического искусственного интеллекта»)	UNI Global Union, глобальная федерация профсоюзов в сфере навыков и услуг, Швейцария	Рекомендации
77	The Toronto Declaration. Protecting the right to equality in machine learning («Декларация Торонто. Защита права на равенство в машинном обучении»), 2018	Amnesty International (глобальное движение, выступающее за защиту прав человека) и Access Now (институт Нью-Йоркского университета, занимающийся исследованием социальных последствий ИИ), Канада	Декларация
78	Artificial Intelligence and its role in society («Искусственный интеллект и его роль в обществе»), 2020	Microsoft, США	Рекомендации
79	SAP's Guiding Principles for Artificial Intelligence («Руководящие принципы для искусственного интеллекта»), 2018	SAP, Германия	Руководство

ПРИЛОЖЕНИЕ

80	Ethical Principles for Artificial Intelligence in Medicine («Этические принципы искусственного интеллекта в медицине»), 2019	The Royal Australian and New Zealand College of Radiologists, Австралия	Рекомендации
81	Automated and Connected Driving («Автоматизированное и подключенное к сети вождение»), 2017	Министерство транспорта и цифровой инфраструктуры, Германия	Доклад
82	Artificial Intelligence at Google: Our Principles («Искусственный интеллект Goggle: наши принципы»), 2018	GOOGLE	Рекомендации
83	Beijing AI Principles («Пекинские принципы искусственного интеллекта»), 2019	Пекинская академия искусственного интеллекта (BAAI), КНР	Рекомендации
84	B.S8611, «Robots and robotic devices, guide to the ethical design and application of robots and robotic systems», 2016	British Standards Institute, Великобритания	Стандарт
85	Ethical design and use of automated decision systems (CAN/CIOSC 101:2019), 2019 — настоящее время	CIO Strategy Council, Канада	Стандарт
86	ISO/IEC JTC 1/SC 42 Standards for Artificial Intelligence, 2017-настоящее время («Information technology — artificial intelligence — overview of trustworthiness in artificial intelligence»), 2020	ISO	Стандарт
87	Artificial Intelligence and User Trust (NISTIR 8332), 2021 (B. Stanton, T. Jensen et al. “Trust and artificial intelligence”), 2021	NIST, США	Стандарт
88	Artificial Intelligence and Public Standards, 2020	Committee on Standards in Public	Стандарт

ПРИМЕНЕНИЕ СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА:
СОЦИАЛЬНО-ФИЛОСОФСКИЕ ПРОБЛЕМЫ И ПРИНЦИПЫ

		Life UK, Велико- британия	
89	The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2019	IEEE SA	Стан- дарт
90	Code of Ethics and Professional Conduct (Acm code of ethics and professional conduct),2018	ACM, США	Стан- дарт

Оглавление

ПРЕДИСЛОВИЕ	3
Глава I ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ: СУЩНОСТЬ, ОБЛАСТИ ПРИМЕНЕНИЯ, ОСОБЕННОСТИ ИССЛЕДОВАНИЯ	
1.1 Определение и классификация ИИ	12
1.2 Основные сферы применения и особенности исследо- вания ИИ	25
1.3 Этическое и правовое регулирование ИИ	37
Глава II СОЦИАЛЬНО-ФИЛОСОФСКИЕ ПРОБЛЕМЫ ПРИМЕНЕНИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В СОВРЕМЕННОМ ОБЩЕСТВЕ	
2.1 Разнообразие и комплексный социально-философский характер проблем применения ИИ	61
2.2 Основные проявления отрицательного влияния ИИ на человека	72
2.3 Причины формирования негативных последствий вза- имодействия человека и ИИ	93
Глава III ФИЛОСОФСКИЕ ПРИНЦИПЫ РАЗРАБОТКИ, ВНЕДРЕНИЯ И ПРИМЕНЕНИЯ СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА	
3.1 Условия эффективного и безопасного применения систем ИИ	112

3.2 Принципы справедливости, автономии и непричинения вреда человеку в контексте использования ИИ	138
ЗАКЛЮЧЕНИЕ	157
ЛИТЕРАТУРА	161
Приложение	186

Научное издание

*Маина Харлановна Бадмаева
Мария Валентиновна Золхоева*

ПРИМЕНЕНИЕ СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА:
СОЦИАЛЬНО-ФИЛОСОФСКИЕ ПРОБЛЕМЫ И ПРИНЦИПЫ

Монография

Свидетельство о государственной аккредитации
№ 2670 от 11 августа 2017 г.

Редактор
Ж. В. Галсанова
Компьютерная верстка
Ж. В. Галсановой
Дизайн обложки
Э. Б. Дашиевой

Подписано в печать 13.10.2026. Формат 60х84 1/16.
Уч-изд. л. 9,16. Усл. печ. л. 11,86. Заказ 120.
Цена свободная

Издательство Бурятского госуниверситета им. Д. Банзарова
670000, г. Улан-Удэ, ул. Ранжурова, 4
rio@bsu.ru

Отпечатано в типографии Издательства БГУ
670000, г. Улан-Удэ, ул. Сухэ-Батора, 3а