На правах рукописи

БАДМАЕВА МАИНА ХАРЛАНОВНА

СОЦИАЛЬНО-ФИЛОСОФСКИЕ ПРОБЛЕМЫ И ПРИНЦИПЫ ПРИМЕНЕНИЯ СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Специальность 5.7.7. – Социальная и политическая философия

Автореферат
диссертации на соискание ученой степени
кандидата философских наук

Улан-Удэ – 2023

Работа выполнена на кафедре философии ФБГОУ ВО «Бурятский государственный университет имени Доржи Банзарова»

Научный руководитель: доктор философских наук, доцент **Золхоева Мария Валентиновна**

Оппоненты:

Цвык Ирина Вячеславовна доктор философских наук, доцент, федеральное государственное бюджетное учреждение высшего образования «Московский авиационный институт (национальный исследовательский университет)», профессор кафедры философии.

Мешкова Людмила Николаевна кандидат философских наук, доцент, федеральное государственное бюджетное образовательное учреждение высшего образования «Пензенский государственный университет», доцент кафедры «Изобразительное искусство и культурология».

Ведущая организация: Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский Томский государственный университет» (г. Томск)

Защита состоится 30 июня 2023 года в 10 часов на заседании диссертационного совета Д. 24.2.279.02 по защите диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук в ФГБОУ ВО «Бурятский государственный университет имени Доржи Банзарова» по адресу: 670000, г. Улан-Удэ, ул. Смолина, 24 «а», ауд. 0209, конференц-зал.

С диссертацией можно ознакомиться в научной библиотеке ФГБОУ ВО «Бурятский государственный университет имени Доржи Банзарова» и на сайте организации www.bsu.ru.

Автореферат разослан «__» _____2023 г.

Ученый секретарь диссертационного
совета, кандидат философских наук, доцент Багаева К. А.

Общая характеристика работы

Актуальность темы диссертационного исследования. 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, обоснованно ставит перед собой задачу выработать аргументированное и взвешенное отношение к внедрению систем искусственного интеллекта, рассмотрев, 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 соотносятся 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.

Социально-философский дискурс не раз обращался к теме 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 суще-

ственным образом отстают 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.

Эти недостатки могут быть преодолены разработкой 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.

Таким образом, актуальность нашего исследования обусловлена необходимостью более глубокого и адекватного отражения в социально-философской рефлексии современных реалий воздействия искусственного интеллекта на общество и человека, потребностью выявить содержание базовых принципов взаимодействия человека с системами искусственного интеллекта посредством анализа его наиболее важных характеристик; выявления возможных социальных, этических последствий и перспектив взаимодействия человека с ИИ для достижения целей гуманистического, безопасного и эффективного развития современного социума.

Степень научной разработанности проблемы. В настоящее время существует значительное количество отечественных и зарубежных исследований технического, юридического, социологического и философского характера, в которых рассмотрен искусственный интеллект и проблемы его применения. Несмотря на это, все еще отмечается дефицит социально-

философских исследований, посвященных рассмотрению разнообразных последствий применения искусственного интеллекта в современном обществе.

Рассмотрим круг работ, очерчивающих проблемное поле исследования, разделив их на несколько тематических блоков.

Исследования, посвященные влиянию техники на человека. В самом широком смысле, искусственный интеллект можно рассматривать как часть философии техники. К этому пулу исследователей относятся классические труды У. Бека[1], В. Г. Горохова[2], М. Кастельса[3], Г. Маркузе[4], М. Маклюэна[5], Л. Мэмфорда[6], Х. Ортеги-и-Гассета[7], В. М. Розина[8], Ф. Рело[9], А. Ридлера[10], М. А. Розова[11], С. А. Смирнова[12], В. С. Степина[13], П. Флоренского[14], Ф. Фукуямы[15], Ю. Хабермаса[16], М. Хайдеггера[17], И. В. Черниковой и Д. В. Черниковой[18] и др.

Так, Ж. Бодрийяр[19], Л. Мэмфорд, Д. Нейсбит[20], Ф. Фукуяма резко критиковали технологизацию и внедрение искусственного интеллекта без ос-

---

[1] Бек У. Общество риска. На пути к другому модерну. М.: Прогресс-Традиция, 2000. 383 с.
[2] Горохов В. Г. Место и роль философии техники и современной философии и её органическая связь с философией науки // Философия науки. 2011. № 1. С. 181-199.
[3] Кастельс М. Информационная эпоха: экономика, общество и культура. М.: ГУ ВШЭ, 2000. 608 с.
[4] Маркузе Г. Одномерный человек // Исследование идеологии Развитого Индустриального Общества. М.: Reefl-book, 1994. 368 с.
[5] Маклюэн М. Понимание медиа: внешнее расширение человека. – М.: Жуковский: «Канон-пресс-Ц», 2003. 464 с.
[6] Мэмфорд Л. Техника и природа человека // Новая технократическая волна на Западе. М.: Прогресс-Традиция, 1986. С. 225-239.
[7] Ортега-и-Гассет Х. Размышления о технике // Избранные труды пер. с исп.; сост., предисл. и общ. ред. А. М. Руткевича. М.: Весь Мир, 1997. С. 164-232.
[8] Розин В. М. Понятие и современные концепции техники. М.: Институт философии РАН, 2006. 255 с.
[9] Рело Ф. Техника и ее связь с задачею культуры. СПб.: Типография министерства путей сообщения, 1885. 27 с.
[10] Ридлер А. Германские высшие учебные заведения и запросы двадцатого столетия. СПб.: типография Р. Голике, 1900. 30 с.
[11] Степин В. С., Горохов В. Г., Розов М. А. Философия науки и техники. М., 1996. 380 с.
[12] Смирнов С. А. Человек перехода / Отв. за вып. П. А. Носова. Новосибирск, 2006. 177 с.
[13] Степин В. С. Научное познание и ценности техногенной цивилизации // Вопросы философии. 1989. № 10. С. 3-18.
[14] Флоренский П. Органопроекция // Русский космизм: антология философской мысли. М.: Педагогика-пресс, 1993. С. 149-162.
[15] Фукуяма Ф. Наше постчеловеческое будущее: Последствия биотехнологической революции; пер. с англ. М. Б. Левина. М.: АСТ : ЛЮКС, 2004. 349 с.
[16] Хабермас Ю. Техника и наука как «идеология». М.: Праксис, 2007. 208 с.
[17] Хайдеггер М. Вопрос о технике // Время и бытие: статьи и выступления: пер. с нем. М., 1993. 49 с.
[18] Черникова Д. В., Черникова И. В. Образовательные и этические аспекты вызовов технонауки в пространстве университета // Высшее образование в России. 2021. Т. 30, № 11. С. 42-51.
[19] Бодрийяр Ж. Система вещей / Пер. с фр. С. Н. Зенкина. М.: «Рудоми-но», 1999. 224 с.
[20] Нейсбит Д. Высокая технология, глубокая гуманность: Технологии и наши поиски смысла / пер. с англ. А. Н. Анваера. М.: Транзит Книга, 2005. 381 с.

мысления социальных и этических последствий. Отечественные исследователи В. А. Кутырев[21], И. В. Черникова и Д. В. Черникова[22] полагают, что человеческая идентичность подвержена опасности из-за необдуманного внедрения высоких технологий. А. Нордманн[23], Э. Тоффлер[24], Ф. Фукуяма, Ю. Хабермас исследуют использование новых технологий, задаваясь, главным образом, вопросом о вызванных им возможных изменениях человеческой природы. Д. В. Иванов[25], М. Кастельс, И. С. Мелюхин[26], А. И. Ракитов[27] анализируют социокультурные последствия использования информационных технологий. В. Г. Горохов рассматривает возможные результаты внедрения новых технологий с позиций этических норм и вводит новое понятие «нано-этика»[28].

Исследования, посвященные определению и классификации ИИ. К этому пулу исследований относятся работы таких авторов, как А. Ю. Алексеев[29], В. Архипов[30], Р. Брукс[31], Н. Бостром[32], Р. Бродхэрст[33], В. В. Васильев[34], Н. Винер[35], Б. Герцель[36], П. Готовцев[37], Х. Де Гарис[38], И. Дубровский[39], А. Е. Евст-

---

[21] Кутырев В. А. Культура и технология: борьба миров. М.: Прогресс-Традиция, 2001. 240 с.
[22] Черникова Д. В., Черникова И. В. Образовательные и этические аспекты вызовов технонауки в пространстве университета // Высшее образование в России. 2021. Т. 30, № 11. С. 42-51.
[23] Nordmann A., et al.. Synthetic Biology at the Limits of Science // Synthetic Biology. Character and impact. Heidelberg u. a.: Springer, 2015. P. 3-7.
[24] Тоффлер Э. Шок будущего. М.: АСТ, 2002. 557 с.
[25] Иванов Д. В. Виртуализация общества. СПб., 2002. 96 с.
[26] Мелюхин И. С. Информационное общество: истоки, проблемы, тенденции развития. М.: Издательство Московского университета, 1999. 206 с.
[27] Ракитов А. И. Наш путь к информационному обществу // Теория и практика общественно-научной информации. М.: ИНИОН, 1989. С. 50-68.
[28] Горохов В. Г. Социальные проблемы нанотехнологии // Высшее образование в России. 2008. № 3. С. 84-98.
[29] Алексеев А. Ю. Комплексный тест Тьюринга: философско-методологические и социокультурные аспекты. М.: ИИнтеЛЛ, 2013. 304 с.
[30] Архипов В. В., Наумов В. Б. О некоторых вопросах теоретических оснований развития законодательства о робототехнике: аспекты воли и правосубъектности // Закон. 2017. № 5. С. 157-170.
[31] Stone P., Rodney B., et al. Artificial intelligence and life in 2030 // One-hundred-year study on artificial intelligence: Report of the 2015–2016. Stanford, Stanford University. URL: http://ai100.stanford.edu/2016-report (дата обращения: 13.10.2022)
[32] Бостром Н. Искусственный интеллект. Этапы. Угрозы. Стратегии / пер. с англ. С. Филина. М.: Манн, Иванов и Фербер, 2016. 496 с.
[33] Broadhurst R., Brown P et al. 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.
[34] Васильев В. В. Трудная проблема сознания. М.: Прогресс-Традиция, 2009. 272 с.
[35] Wiener N. The human use of human beings: cybernetics and society. Boston: Houghton Mifflin, Second Edition Revised, NY : Doubleday anchor, 1954. 344 p.
[36] Goertzel B. Artificial General Intelligence: Concept, State of the Art, and Future Prospects // Journal of Artificial General Intelligence. 2014. Vol. 5(1). Pp. 1-46.

ратов[40], В. Карпов[41], Р. Курцвейл[42], Ш. Легг[43], Дж. Маккарти[44], К. Макниш[45], Р. Мерфи[46], М. Мински[47], П. М. Морхат[48], В. Наумов[49], А. В. Незнамов[50], Н. Нильсон[51], П. Норвиг[52], Р. Пенроуз[53], М. Райан[54], С. Рассел[55], О. В. Ревинский[56], А. В. Резаев , В. Ручкина[57], Д. Серль[58], В. Н. Синельникова[59], Н. Д. Трегубова [60], А. Тьюринг[61], А. Турчин[62], В. Фулин[63], М. Хэнлэйн[64].

---

[37]Готовцев П. М., Ройзенсон Г. В. Характеристика проектов стандартов на этичный искусственный интеллект IEEE / П. М. Готовцев, Г. В. Ройзенсон // 390 Этика и «цифра». – 2020. – URL: https://ethics.cdto.center/ieee (дата обращения: 12.04.2022).

[38] Де Гарис Х. Искусственный мозг: подход с развитым модулем нейронной сети / Х. Де Гарис // World Scientific. – 2010. – 400 с.

[39] Дубровский Д. И. Искусственный интеллект и проблема сознания / Д. И. Дубровский // Философия искусственного интеллекта: материалы всерос. междисциплинар. конф., М., МИЭМ, 17-19 янв. 2005 г. – М.: ИФ РАН, 2005. – С. 26-31.

[40] Евстратов А. Э., Гученков И. Ю. Пределы применения искусственного интеллекта (правовые проблемы) // Правоприменение. 2020. Т. 4, № 4. С.13-19.

[41] Карпов В. Э., Готовцев П. М., Ройзенсон Г. В. Машинная этика // 390 Этика и «цифра». 2020. URL: https://ethics.cdto.center/3_4 (дата обращения: 23.04.2022).

[42] Kurzweil R. The Age of Intelligent Machines. Cambridge, MA : MIT Press, 1990. 565 p.

[43] Legg S., Hutter M. A collection of definitions of intelligence // Advances in artificial general intelligence: concept, architectures and algorithms. Amsterdam : IOS Press., 2007. Vol.157. Pp. 17–24.

[44]McCarthy J. What is Artificial Intelligence? // Stanford University. 2007. URL: http://www-formal.stanford.edu/jmc/whatisai. (дата обращения: 21.09.2022)

[45] MacNish C., et al. Logics in Artificial Intelligence // European Workshop JELIA '94, York, UK, September 5-8, 1994. 413 p.

[46] Murphy R. F. Artificial Intelligence Applications to Support K-12 Teachers and Teaching // A Review of Promising Applications, Opportunities, and Challenges. RAND Corporation. URL: https://www.rand.org/content/dam/rand/pubs/perspectives/PE300/PE315/RAND_PE315.pdf (дата обращения: 30.03.2021).

[47] Мински М. Фреймы для представления знаний. М.: Мир, 1979. 151 с.

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

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

[50] Незнамов А. В. О концепции регулирования технологий искусственного интеллекта и робототехники в России // Закон. 2020. № 1. С. 171-185.

[51] Нильсон Н. Искусственный интеллект: методы поиска решений / Пер. с англ. В. Л. Стефанюка; под редакцией С. В. Фомина. М.: Мир, 1973. 272 с.

[52] Norvig P. Paradigms of Artificial Intelligence Programming: Case Studies in Common Lisp // Morgan Kaufmann. 1991. 948 p.

[53] Пенроуз Р. Новый ум короля: О компьютерах, мышлении и законах физики / Пер. с англ. под. ред. В.О. Малышенко. 3-е изд. М.: Издательство ЛКИ, 2008. С. 328.

[54] Ryan M. In AI we trust: Ethics, artificial intelligence, and reliability // Science and Engineering Ethics. 2020. Vol. 26. Pp. 2749-2767.

[55] Рассел С., Норвиг П. Искусственный интеллект: современный подход. 2-е изд. М.: Вильямс, 2006. 1408 с.

[56] Синельникова В. Н., Ревинский О. В. Права на результаты искусственного интеллекта // Вестник Российской академии интеллектуальной собственности и Российского авторского общества. 2017. № 4. С. 24-27.

[57] Ручкина Г. Ф. Искусственный интеллект, роботы и объекты робототехники: к вопросу о теории правового регулирования в Российской Федерации // Банковское право. 2020. № 1. С. 7-18.

[58] Серль Дж. Р. Сознание, мозг и программы // Аналитическая философия: Становление и развитие: Антология / Общ. ред. и сост. А. Ф. Грязнов. М., 1998. 528 с.

[59] Синельникова В. Н., Ревинский О. В. Права на результаты искусственного интеллекта // Вестник Российской академии интеллектуальной собственности и Российского авторского общества. 2017. № 4. С. 24-27.

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

[61] Turing A. Computing machinery and intelligence // Mind. 1950. Vol. 59. Pp. 433-460.

Впервые тема искусственного интеллекта в виде идеи «мыслящих машин» была введена в 1950 г. в работе британского математика А. Тьюринга «Вычислительные машины и разум»[65]. 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[66].

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»[67].

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[68].

В дальнейшем представители аналитической философии (Н. Блок[69] , Д. Деннет[70], Т. Нагель[71], Х. Патнэм[72], Дж. Сёрл[73], Дж. Фодор[74], Д. Чалмерс[75] и

---

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

[63] Фулин В. А. Универсальный искусственный интеллект и экспертные системы. СПб.: БХВ-Петербург. 2009. 240 с.

[64] Kaplan A., Haenlein M. On the interpretations, illustrations, and implications of artificial intelligence. URL: https://www.sciencedirect.com/science/article/abs/pii/ S0007681318301393 (дата обращения: 14.02.2021)

[65] Тьюринг А. Может ли машина мыслить. М.: Едиториал УРСС, Ленанд. 2016. 128 с.

[66]McCarthy J. What is Artificial Intelligence? // Stanford University.2007. URL: http://www-formal.stanford.edu/jmc/whatisai. (дата обращения: 21.09.2022)

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

[68] Синельникова В. Н., Ревинский О. В. Права на результаты искусственного интеллекта // Вестник Российской академии интеллектуальной собственности и Российского авторского общества. 2017. №4. С. 17-27.

[69] Block N. Troubles with functionalism. Minnesota Studies in the Philosophy of Science // Troubles with functionalism, Minnesota Studies in the Philosophy of Science. 1978. Pp.261-325.

[70] Деннет Д. Виды психики. На пути к пониманию сознания / пер. А. Веретенникова. М.: Идея-Пресс, 2004. 79 с.

[71] Нагель Т. Каково быть летучей мышью? // Глаз разума / пер. с англ. М. А. Эскиной. Самара : «Бахрах-М», 2003. С. 349-360.

[72] Putnam H. The Project of Artificial Intelligence // Renewing Philosophy. Cambridge, MA : Harvard University Press, 1992. P. 1-18.

др.) в своих трудах различали сильный ИИ и слабый ИИ как два основных подхода к искусственному интеллекту, сложившиеся в современной науке.

Разнообразие многочисленных подходов к пониманию ИИ представлено в недавнем исследовании Ш. Легг и М. Хаттер «Коллекция определений ИИ»[76].

Исследования, посвященные проблемам и принципам применения искусственного интеллекта. Проблемы применения искусственного интеллекта обсуждались в трудах Р. Г. Апресяна[77], П. М. Готовцева[78], А. А. Гусейнова[79], В. Э. Карпова[80], А. В. Разина[81], Г. В. Ройзензона[82], В. А. Цвык и И. В. Цвык[83], Б. Г. Юдина[84] и др. Эти 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 оказывает 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. М. Готовцев, В.Э. Карпов, А. В. Разин, Г. В.

---

[73] Серль Дж. Р. Сознание, мозг и программы // Аналитическая философия: Становление и развитие: Антология / Общ. ред. и сост. А.Ф. Грязнов. М., 1998. 528 с.

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

[75] Чалмерс Д. Сознающий ум. В поисках фундаментальной теории. М.: Либроком, 2019. 512 с.

[76] Legg S., Hutter M. A collection of definitions of intelligence // Advances in artificial general intelligence: concept, architectures and algorithms. Amsterdam : IOS Press., 2007. Vol.157. Pp. 17-24.

[77] Апресян Р. Г. Этика и дискуссии об искусственном интеллекте / XI международная конференция «Теоретическая и прикладная этика: Традиции и перспективы - 2019. К грядущему цифровому обществу. Опыт этического прогнозирования (100 лет со дня рождения Д. Белла - 1919-2019)» / Отв. ред. В. Ю. Перов. СПб: ООО «Сборка», 2019. С. 169-170.

[78] Готовцев П. М., Ройзенсон Г. В., Характеристика проектов стандартов на этичный искусственный интеллект IEEE // 390 Этика и «цифра». 2020. URL: https://ethics.cdto.center/ieee (дата обращения: 12.04.2022).

[79] Гусейнов А. А. Размышления о прикладной этике / Доклад на основе статьи: Размышления о прикладной этике // Ведомости НИИПЭ, Вып. 25: Общепрофессиональная этика. Тюмень : НИИПЭ, 2004. 148 с.

[80] Карпов В. Э., Готовцев П. М., Ройзенсон Г. В. Машинная этика //390 Этика и «цифра». 2020. URL: https://ethics.cdto.center/3_4 (дата обращения: 23.04.2022).

[81] Разин А. В. Этика искусственного интеллекта // Философия и общество. 2019. №1. С. 57-73.

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

[83] Цвык В. А., Цвык И. В. Социальные проблемы развития и применения искусственного интеллекта // Вестник РУДН. Серия: Социология. 2022. №1. С. 58-69.

[84] Юдин Б. Г. Социальные технологии, их производство и потребление // Эпистемология и философия науки. 2012. Вып. 31, № 1. С. 55-64.

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[85], А. Джобин, М. Йенка и Е. Вайена [86], В. Э. Карпов[87], М. Кэролан[88], В. А. Лаптев[89], П. М. Морхат[90], А. В. Попова[91], М. Райан[92], С. Рассел[93], Таддео[94], Э. Тополь[95], Л. Флориди[96], Т. Хагендорф[97], Э. Юдковски[98], Н. А. Ястреб[99].

Так, в трудах Л. Флориди обсуждаются вопросы, связанные с соблюдением конфиденциальности личных данных[100]. С. Паркенсон и Э. Харпер пытались выявить препятствия и трудности, возникающие в процессе включения конфиденциальных данных пользователей в наборы больших данных[101]. Обоснование необходимости и проблем обеспечения прозрачно-

---

[85] Bossmann Dzh. Top 9 Ethical Issues in Artificial Intelligence. URL: https://hr-portal.ru/article/9-glavnyh-eticheskih-problem-iskusstvennogo-intellekta (дата обращения: 22.03.2022).

[86] Jobin A., et al. Artificial Intelligence: The Global Landscape of Ethics Guidelines // Nature Machine Intelligence. 2019. Vol.1. Pp. 389-399.

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

[88] Carolan M. Automated agrifood futures: robotics, labor and the distributive politics of digital agriculture // J. Peasant Stud. 2020. Vol. 47. Pp.184-207.

[89] Лаптев В. А. Электронные доказательства в арбитражном процессе // Российская юстиция, № 2. 2017. С. 56-59.

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

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

[92] Ryan M. Ethics of using AI and big data in agriculture: the case of a large agriculture multinational // ORBIT Journal. 2019. Vol.2(2). 27 p.

[93] Russell S. Human-Compatible Artificial Intelligence // Human-Like Machine Intelligence. Oxford: Oxford University Press, 2021. Pp 3-23.

[94] Taddeo M. Is cybersecutiry a public good? // Minds and machines. 2019. Vol. 29, № 3. Pp. 349-354.

[95]Тополь Э. Будущее медицины: Ваше здоровье в ваших руках. М.: Альпина нон-фикшн, 2016. 491 с.

[96]Floridi L. The end of an era: from self-regulation to hard law for the digital industry // Philosophy & Technology. 2021. Vol. 34, № 4. Pp. 612-622.

[97] Hagendorff T. The Ethics of AI Ethics: An Evaluation of Guidelines // Minds & Machines. 2020. Vol. 30. Pp. 99-120.

[98] Юдковски Э. Систематические ошибки в рассуждениях, потенциально влияющие на оценку глобальных рисков. Новые технологии и продолжение эволюции человека? // Трансгуманистический проект будущего. М., 2008. С. 182-225.

[99] Ястреб Н. А. Индустрия 4.0: киберфизические системы и интернет вещей // Человек в технической среде: сборник научных статей / Под ред. доц. Н.А. Ястреб. Вологда : ВолГУ, 2015. С. 136-141.

[100]Mittelstadt B. D., Allo P. et al. The ethics of algorithms: Mapping the debate // Big Data and Society. 2016. Vol. 3(2). Pp. 1-21.

[101]Harper E. M., Parkerson S. Powering Big Data for Nursing Through Partnership. URL: https://pubmed.ncbi.nlm.nih.gov/26340243/ (дата обращения: 14.03.2022)

сти процессов, связанных с данными пользователей, рассматриваются в работах Дж. Баррел[102].

Преодоление и недопущение социальной несправедливости и предвзятого отношения – предмет исследований Т. Панч[103]. Российско-французский философ А. Гринбаум в труде «Машина-доносчица» поднимает вопросы, касающиеся «нравственности» узкого искусственного интеллекта, его ответственности перед человеком[104]. 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 с ИИ поднимались еще во времена А. Тьюринга. По сей день в трудах многих ученых, исследователей обсуждается возможность причинения вреда человечеству со стороны сильного ИИ, звучит тревога и опасения за будущее, за безопасность человеческой цивилизации в целом.

Определенные результаты в данном контексте достигнуты в отечественном обществознании. Так, труды П. М. Морхата посвящены правовой регламентации процессов жизненного цикла систем ИИ[105]. Л. В. Баева, Храпов С. А. исследуют риски и последствия цифровых технологий, в том числе ИИ, в области образования[106].

В. Э. Карпов сосредоточил внимание на этических аспектах применения ИИ в жизни современного общества, базовых принципах взаимодействия ИИ с человеком.

Несмотря на внимание ученых к отдельным проблемам, возникающим в отношениях человека с ИИ, указанные трудности, как правило, не иссле-

---

[102] Burrell J. How the machine ‘thinks’: Understanding opacity in machine learning algorithms // Big Data and Society. 2016. Vol. 3(1). Pp. 1-12.

[103]Panch T., Mattie H. et al. Artificial intelligence and algorithmic bias: implications for health systems // Journal of global health. 2019. Vol. 9(2). Pp. 23-32.

[104]Гринбаум А. Машина-доносчица: как избавить искусственный интеллект от зла. М.: ТрансЛит, 2017. 76 с.

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

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

дуются ими в комплексе, в целостности, на уровне социально-философского познания, что не позволяет выработать адекватную и безопасную стратегию взаимодействия общества с искусственным разумом и требует выработки философских принципов применения систем ИИ в современном мире.

Объектом исследования являются системы ИИ (слабого или узкого ИИ), созданные человеком для решения определенных практических задач.

Предмет исследования – проблемы применения систем ИИ, способы и пути их решения в современном обществе.

Цель данной работы – выявить и раскрыть содержание философских принципов взаимодействия систем ИИ с человеком для повышения безопасности и снижения рисков их использования в различных сферах общественной жизни.

Задачи исследования:

1. Выделить и охарактеризовать существующие в научной и философской литературе теоретические подходы к пониманию сущности искусственного интеллекта и классификации ИИ в контексте его практического применения в социальной реальности.
2. Определить наиболее значимые области использования ИИ в социальной практике.
3. Систематизировать существующие в настоящее время документы, регулирующие этические и правовые аспекты применения ИИ, выявить их содержание и основные идеи с целью установления возможных рисков и угроз для безопасного и эффективного использования человеком систем ИИ.
4. Обосновать необходимость исследования проблем применения систем ИИ с позиций социальной философии.
5. Выделить проблемы, возникающие вследствие использования систем ИИ в различных сферах общественной жизни, отражающие основное содержание и отдельные стороны негативного воздействия указанных систем на человека и общество.

6. Определить совокупность проблем, вызванных применением систем ИИ, указывающих на причины их отрицательного воздействия на жизнь социума.

7. Раскрыть содержание и значение философских принципов, нацеленных на выявление условий эффективного и безопасного применения систем ИИ в современном обществе.

8. Выявить смысл и соотношение философских принципов, раскрывающих отдельные стороны, аспекты негативного влияния ИИ на жизнь человека и общества.

Новизна научной работы состоит в следующем:

1. Показано, что сущность искусственного интеллекта в контексте проблем его применения, отражающая существующий уровень его использования в современном социуме, характеризуется системной связью составляющих его элементов, а также нацеленностью на решение определенных, узконаправленных задач, поставленных человеком, что позволяет определить его, в соответствии с действующей классификацией ИИ, как «узкий» (или слабый) ИИ.

2. Определены основные сферы применения ИИ и аргументирован тезис о глубоком проникновении, встраивании его в основы, фундамент современного социума, что стало причиной возникновения риска новой экзистенциальной угрозы, обусловленной возможной утратой человеком привычного для него места в мире.

3. Выделены и систематизированы документы, регулирующие этико-правовые аспекты разработки, внедрения и применения ИИ, изданные различными социальными институтами и организациями современного общества, содержание которых позволяет выявить основные риски и угрозы для безопасного и эффективного ИИ.

4. Обоснована необходимость социально-философского исследования проблем применения систем ИИ, позволяющего подойти к их решению с позиций целостности социальной жизни и присущей философии общества на-

целенности на индивида, его многогранные потребности и стремление к достижению социального идеала.

5. Выявлены социальные проблемы, возникающие вследствие применения систем ИИ, отражающие основные проявления отрицательного влияния ИИ на человека: проблемы причинения вреда, социальной несправедливости и нарушения автономии человека.

6. Определен круг проблем, раскрывающих причины возникновения нежелательных для человека и общества последствий взаимодействия с ИИ: проблемы непрозрачности, отсутствия ответственности и нарушения конфиденциальности.

7. Выявлено и раскрыто содержание философских принципов прозрачности, ответственности, конфиденциальности, отражающих условия безопасного и эффективного использования ИИ в различных областях жизни современного общества.

8. Определен смысл требований, составляющих содержание принципов социальной справедливости, автономии человека и непричинения вреда, представляющих отдельные стороны, грани безопасного, эффективного применения систем ИИ в различных областях жизни социума, а также обосновано основополагающее значение принципа непричинения вреда во всей совокупности рассмотренных выше философских принципов.

Основные положения, выносимые на защиту:

1. Существующий в настоящее время на уровне технологии и применяемый в различных областях общественной жизни искусственный интеллект представляет собой узкий (слабый) искусственный интеллект, созданный человеком для решения определенных узконаправленных задач и представляющий собой системы, включающие аппаратный комплекс, программное обеспечение и набор данных. Особенностью ИИ является способность осуществлять сбор данных, их интерпретацию в виде рассуждений, принимать решения на основании имеющейся информационной базы практически

в любой отрасли деятельности человека, что определяет постоянно расширяющийся спектр его применения.

2. Системы ИИ находят применение в ключевых сферах жизни социума (государственное управление, общественная безопасность, транспорт, сельское хозяйство, энергетика, здравоохранение, образование, правосудие, банковская, финансовая деятельность, военная сфера и др.), перечень которых непрерывно расширяется вследствие перманентного процесса их совершенствования, что привело к возникновению потенциальной угрозы пересмотра базовых мировоззренческих представлений о месте человека в мире, определяющем его роль и предназначение.

3. Экзистенциальная угроза, сформировавшаяся вследствие фундаментального встраивания систем ИИ в жизнь социума потребовала разработки и принятия целого комплекса документов, существующих в форме национальных стратегий, нормативно-правовых актов, этических руководств, рекомендаций и стандартов, изданных государственными институтами, коммерческими структурами, международными и неправительственными организациями и регулирующих этико-правовые аспекты использования ИИ.

4. Процесс применения ИИ требует глубокого и всестороннего социально-философского осмысления, фокусирующегося на человеке, его многогранных потребностях и стремлении к лучшей жизни. Исследование с теоретических позиций данной дисциплины необходимо для выявления рисков, вызываемых применением систем ИИ, для существования человека, сохранения его места и роли в современном мире.

5. Обобщение и осмысление существующего опыта взаимодействия человека с ИИ позволило выделить комплекс проблем их применения, отражающих основное содержание и различные стороны, аспекты отрицательного влияния ИИ на человека: проблемы причинения вреда, социальной несправедливости и нарушения автономии. Проблема причинения вреда приводит к риску отчуждения человека от самого себя, от собственной сущности и предназначения, искажению его особого статуса в мире как единственного

существа, способного к интеллектуальной деятельности, нацеленной на изменение, преобразование окружающей реальности. Проблемы социальной несправедливости и нарушения автономии призваны раскрыть отдельные грани и описать разнообразные проявления причиняемого человеку ущерба, а также последствия, возникающие из-за некорректного внедрения, применения систем ИИ.

6. Отдельную группу трудностей, обусловленных применением ИИ, составили проблемы, раскрывающие причины формирования негативных последствий взаимодействия человека и ИИ: проблемы непрозрачности, отсутствия ответственности и нарушения конфиденциальности. Проблема непрозрачности заключается в том, что принцип работы систем ИИ становится в процессе его совершенствования все более непрослеживаемым, необъяснимым и неинтерпретируемым. Проблема отсутствия ответственности состоит в том, что природа ИИ не позволяет установить субъект ответственности, на которого однозначно можно было бы возложить вину в случае причинения вреда человеку системой ИИ. Проблема нарушения конфиденциальности возникает из-за угрозы утечки потока персональных данных или потери контроля над этими данными.

7. Для успешного разрешения существующих проблем применения систем ИИ и предотвращения возникновения новых затруднений раскрыто содержание философских принципов взаимодействия человека с системами ИИ, представляющих требования об открытости и доступности (принцип прозрачности), подотчетности систем ИИ человеку (принцип ответственности), о наложении запрета на нарушение личных границ (принцип конфиденциальности), выступающие условиями эффективного и безопасного применения систем ИИ.

8. Выявлен и сформулирован смысл требований об исключении любой предвзятости и дискриминации со стороны ИИ, о самостоятельности человека в принятии решений, составляющих содержание философских принципов социальной справедливости и автономии, в совокупности с вышеназван-

ными принципами раскрывающих суть центрального философского принципа непричинения вреда человеку, призванного обеспечить безопасное для человека и общества, полезное и эффективное применение ИИ.

Теоретическая база и методология исследования. В качестве теоретической основы были использованы труды отечественных и зарубежных ученых: П. В. Алексеева[107], А. Гринбаума[108], А. Гринфилда[109], В. А. Кутырева[110], Х. Ортега-и-Гассет[111], М. Райана[112], Л. Флориди[113], М. Хайдеггера[114], Ф. Фукуямы[115], И. В. Черниковой и Д. В. Черниковой[116], Б. Г. Юдина[117], К. Ясперса [118] и др. Исследование представленных в них положений позволило сформировать понятийный аппарат диссертации, выявить основные проблемы, тенденции и закономерности развития ИИ, вызовы и пути преодоления возникающих трудностей, обусловленных эволюцией и расширением сферы применения ИИ в различных областях социальной реальности.

Осмысление взаимодействия человека и техники в современном обществе стало предметом исследования работ 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. М. Хайдеггер указывает на необходимость определения сущности техники, чтобы выявить, к каким последствиям должно быть готово общество при широком ее применении. При этом

---

[107] Алексеев П. В. Социальная философия. М.: ООО «ТК Велби», 2003. 256 с.

[108] Grinbaum A. et al. Ethics in Robotics Research // IEEE Robotics and Automation Magazine. 2017. № 24. Pp. 139-145.

[109]Гринфилд А. Радикальные технологии: устройство повседневной жизни. М.: Издательский дом «Дело» РАНХиГС, 2019. 424 с.

[110] Кутырев В. А. Культура и технология: борьба миров. М.: Прогресс-Традиция, 2001. 240 с.

[111] Ортега-и-Гассет Х. Размышления о технике // Избранные труды пер. с исп.; сост., предисл. и общ. ред. А. М. Руткевича. М.: Весь Мир, 1997. С. 164-232.

[112] Ryan M. Ethics of using AI and big data in agriculture: the case of a large agriculture multinational // ORBIT Journal. 2019. Vol. 2(2). 27 p.

[113] Floridi L., Cowls J. et al. How to design AI for social good: Seven Essential factors // Sci. Eng. Ethics. 2020. Vol.26. Pp. 1771-1796.

[114] Хайдеггер М. Вопрос о технике // Время и бытие : статьи и выступления : пер. с нем. М., 1993. 49 с.

[115] Фукуяма Ф. Наше постчеловеческое будущее: последствия биотехнологической революции / Пер. с англ. М. Б. Левина. М.: АСТ, 2004. С. 364.

[116] Черникова Д. В., Черникова И. В. Образовательные и этические аспекты вызовов технонауки в пространстве университета // Высшее образование в России. 2021. Т. 30, № 11. С. 42-51.

[117] Юдин Б. Г. Социальные технологии, их производство и потребление // Эпистемология и философия науки. 2012. Вып. 31, № 1. С. 55-64.

[118] Ясперс К. Смысл и назначение истории: пер.с нем. М.: Политиздат, 1991. 527 с.

он рассматривает технику не в качестве простого, нейтрального и подчиненного человеку инструмента, предназначенного для достижения его целей и удовлетворения потребностей. По мнению философа, техника перестраивает самого человека, изменяя его сущность и подвергая риску его онтологический статус в мире. К. Ясперс 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»[119]. 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.

Социально-философскими проблемами, вызванными применением систем искусственного интеллекта, занимались А. Гринбаум[120], М. Райан, М. Форд[121], Л. Флориди[122] и др. В трудах М. Форда обсуждаются проблемы роста безработицы, профессиональной поляризации, необходимости перестройки структуры управления организации в связи с внедрением технологии ИИ.

---

[119] Ясперс К. Смысл и назначение истории: пер.с нем. М.: Политиздат, 1991. С. 137.

[120] Grinbaum A. et al. Ethics in Robotics Research // IEEE Robotics and Automation Magazine. 2017. № 24. Pp. 139-145.

[121] М. Форд. Роботы наступают: развитие технологий и будущее без работы / пер.с англ. С. Чернина. М.: Альпина нон-фикшн, 2016. 572 с.

[122] Ryan M. Ethics of using AI and big data in agriculture: the case of a large agriculture multinational // ORBIT Journal. 2019. Vol. 2 (2). 27 p.

Методологической базой исследования выступили системный, деятельностный подходы, герменевтический метод исследования и принцип историзма. Системный подход был применен при выделении комплекса основных социально-философских проблем применения ИИ и раскрытии содержания базовых принципов его использования в социальной практике. Также системный подход позволил подойти к выявлению сущности искусственного интеллекта как системного образования, составляющие которого образуют целостность, не сводимую к простой совокупности, сумме его частей.

Деятельностный подход и наследие философии техники применялись в оценке влияния систем ИИ на сущность человека, его идентичность как разумного существа, способного к целеполагающей, природопреобразующей, творческой деятельности, ориентированной на принятые индивидом ценности, нормы, идеалы.

Герменевтический метод был использован при интерпретации нормативно-правовых документов, регулирующих процессы разработки и внедрения систем ИИ. Также данный метод помог проследить процесс трансформации толкования сущности систем ИИ в сфере науки и практического применения этих систем.

Принцип историзма позволил рассмотреть системы ИИ в динамике их изменения, становления во времени, в связи с конкретно-историческими условиями их существования.

В диссертации также применяются традиционные общенаучные методы (анализ, синтез, дедукция, индукция и др.).

Теоретическое и прикладное значение. Диссертация является одним из первых в современном обществознании исследований, позволивших доказать необходимость рассмотрения проблем искусственного интеллекта с позиций социально-философского понимания места и роли человека в мире. Избранная теоретическая позиция, основанная на принципах гуманизма, позволила выявить и раскрыть содержание философских принципов примене-

ния систем ИИ, что составляет личный вклад соискателя в развитие концептуальных положений о безопасном, надежном и эффективном ИИ.

Полученные соискателем 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 автор диссертации вводит в научный оборот около ста работ (в том числе на английском и китайском языках), отражающих международный опыт применения систем ИИ, а также наиболее актуальные проблемы ИИ, стратегии и пути их преодоления.

Не менее важным теоретическим результатом является проведенная соискателем работа по выявлению и классификации основных проблем, обусловленных ИИ, осуществленная исходя из системного понимания указанных проблем и их отношения к центральному принципу непричинения вреда, составляющему сущность и ядро концепции безопасного и надежного искусственного интеллекта.

Выводы работы могут быть использованы для создания 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 найти применение при разработке просветительских программ для IT-специалистов, создающих новые программные продукты, для людей, которые применяют ИИ в промышленности, здравоохранении, транспорте, образовании и др. сферах общественной жизни, для сотрудников коммерческих структур, реализующих применение ИИ в малом и среднем бизнесе, для служащих государственных структур и общественных деятелей, поскольку фундаментальные философские принци-

пы применения ИИ являются универсальными по своему характеру, дающими общее руководство по применению человеком любых систем ИИ.

Выводы исследования могут стать составной частью учебных курсов, посвященных актуальным проблемам социальной философии, спецкурсов, лекций и семинаров по машинной этике, философии техники, учебных и методических пособий по соответствующим дисциплинам, учебных курсов для средней школы в качестве источника информации и средства повышения общей осведомленности учащихся о проблемах и возможностям применения ИИ. Результаты диссертации также могут быть использованы при разработке широкого спектра научно-исследовательских программ социально-экономического развития.

Апробация работы. В процессе подготовки диссертации некоторые тезисы настоящего исследования были изложены и обсуждены в ряде публикаций: в монографии, научных сборниках и журналах.

Структура диссертационной работы. Структуру диссертации составляют введение, три главы, включающие восемь параграфов, заключение, список литературы и приложение.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обосновывается актуальность темы диссертационного исследования, описывается степень её разработанности, выделяются объект и предмет исследования, формулируются цель и необходимые для её достижения задачи. Также уточняются методологические основания и методы исследования, формулируются научная новизна и положения, выносимые на защиту. Обосновывается теоретическая и практическая значимость исследования, раскрываются степень достоверности и апробация полученных результатов, излагается основная структура исследования.

Глава I «Искусственный интеллект: сущность, области применения, особенности исследования» посвящена анализу содержания понятия и

классификации ИИ, определению основных областей применения ИИ и выявлению своеобразия источниковой базы исследования проблем ИИ.

В параграфе 1.1 «Определение и классификация ИИ» осуществлено исследование существующих теоретических подходов к определению понятия «искусственный интеллект». Отмечено, что термин «искусственный интеллект» является недостаточно определенным, семантически расплывчатым, и рефлексия над дефиницией данного понятия продолжается по сей день. Большая часть имеющихся определений не является универсальными, не покрывает тех значений понятия, которые существуют в современном обществе.

Информация, представленная в разнообразных источниках, посвященных рассмотрению сущности и особенностей искусственного интеллекта, позволяет сделать вывод о том, что существующий в настоящее время на уровне технологии, применяемый в различных областях общественной жизни и требующий регулирования со стороны государственных, общественных организаций, общества в целом искусственный интеллект (ИИ) – это слабый (узкий) искусственный интеллект, созданный человеком для решения определенных, узконаправленных задач. Узкий ИИ представлен системами, элементами которых являются аппаратные комплексы, программное обеспечение, наборы данных. Поэтому в рамках данного исследования, посвященного узкому ИИ и связанным с его применением рисков, понятия «искусственный интеллект (ИИ)» и «система искусственного интеллекта» используются как равнозначные, равнообъемные.

В параграфе 1.2. «Основные сферы применения и особенности исследования ИИ» был рассмотрен ряд областей, в которых системы ИИ получили широкое распространение. Это транспорт, государственное управление, образование, оборона и национальная безопасность, сельское хозяйство, промышленность и энергетика. Безусловно, приведенный перечень не полон, поскольку в современном мире практически не осталось сфер деятельности человека, куда еще не проникли системы ИИ. Тем не менее, знакомство с

указанными выше областями внедрения систем ИИ позволило выявить специфическую особенность исследования ИИ, которой, по нашему мнению, является своеобразие источниковой базы проблем ИИ. Помимо научных и философских трудов здесь обязательно должны рассматриваться документы регламентирующего характера, содержащие информацию об актуальных трудностях и препятствиях, с которыми человек сталкивается в процессе практического применения систем ИИ в различных сферах жизни общества, а также правовых нормах, моральных принципах регулирования ИИ, правилах и требованиях, уже сформулированных различными социальными институтами и организациями в ответ на вызовы новой цифровой реальности.

В параграфе 1.3. «Этическое и правовое регулирование ИИ» исследовано содержание значительного пула документов по этическому и правовому регулированию систем ИИ, отобранных на основе критериев надежности, новизны и разнообразия. Источником большинства документов выступили государственные учреждения, академические организации, научно-исследовательские институты, частные компании, межгосударственные или наднациональные организации, профессиональные ассоциации (полный перечень приведен в Приложении к диссертации). Рассмотренные документы, заявляя о необходимости учета этического аспекта при развитии ИИ, в то же время лишены целостности, единства в оценке и интерпретации уровня и перспектив развития ИИ, не учитывают содержание других источников и потому не могут служить общим, универсальным руководством для безопасного взаимодействия систем ИИ с человеком и обществом. Так, национальные стратегии определяют самые общие теоретические подходы и самые общие требования к процессу внедрения ИИ. Этические стандарты и кодексы, адресованные отдельным заинтересованным сторонам (например, профессиональным сообществам, государственному сектору, частным компаниям, разработчикам, исследователям и т.д.), сконцентрированы на осмыслении результатов практического применения ИИ в некоторых узких и специфических областях жизни социума.

Выработка стратегии безопасного использования ИИ на уровне академической науки в настоящее время предусматривает выявление основополагающих принципов и проблем, но их смысл, содержание остается не раскрытым, что затрудняет применение выделенных принципов на практике.

Правовое регулирование, по оценкам специалистов, существенно отстает от процесса внедрения ИИ в различные области социальной реальности. Существующие документы позволяют сделать обоснованный вывод об отсутствии в настоящее время полноценного подхода к регулированию ИИ с позиции права. Фрагментарный, разрозненный характер принятых документов, расплывчатость, многозначность используемых терминов и декларативный характер основных положений свидетельствуют о необходимости дальнейшего исследования и обобщения опыта применения ИИ в современном мире.

Все вышеперечисленное подтверждает важность и значимость избранной соискателем цели и указывает на актуальность исследования проблем применения систем ИИ и выявления принципов их безопасного и эффективного применения в современном социально-философском дискурсе.

Глава II «Социально-философские проблемы применения ИИ в современном обществе» посвящена выявлению основных социально-философских проблем, возникающих при проектировании, разработке и внедрении систем ИИ. Указанные проблемы образуют систему взаимосвязанных, взаимообусловленных трудностей, выступающих различными формами проявления центральной проблемы использования систем ИИ: проблемы причинения вреда человеку. Для достижения цели нашего исследования они разделены на две группы, одна из которых посвящена рассмотрению возможных проявлений вреда, наносимого человеку системами ИИ, вторая - осмыслению причин возникновения негативных эффектов взаимодействия человека с ИИ.

В параграфе 2.1. «Разнообразие и комплексный, социально-философский характер проблем применения ИИ» фиксируется, что проблема угрозы безопасности или причинения вреда, сформулированная А. Азимовым в качестве первого закона роботехники, выступает в роли центральной, системообразующей социально-философской проблемы применения технологии ИИ, по отношению к которой остальные трудности выступают раскрывающими возможный источник, причину возникновения вреда, либо его разновидности, варианты реализации.

Как следует из рассмотренных выше источников, к первой группе выделенных в диссертации проблем могут быть отнесены проблемы нарушения конфиденциальности, отсутствия ответственности, непрозрачности как порождающие, вызывающие возможное причинение ущерба, вреда человеку вследствие использования технологии ИИ, а ко второй - проблемы нарушения автономии человека и социальной несправедливости как представляющие аспекты, формы, грани, разновидности этого вреда.

Безусловно, предложенный перечень проблем применения ИИ не является полным, окончательным, исчерпывающим. В силу повсеместного проникновения систем ИИ в различные сферы общественной жизни, перманентного совершенствования технологии и постоянно растущего 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 в данном параграфе 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. В диссертации подчеркивается, что человек остается по отношению к искусственному разуму оригиналом, творцом, созда-

телем 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 отыскания путей решения, способов преодоления уже существующих и только формирующихся в данной области проблем и препятствий.

В параграфе 2.2. «Основные проявления отрицательного влияния ИИ на человека» проанализированы основные проявления негативного влияния ИИ на человека, представленные проблемами причинения вреда, социальной несправедливости и нарушения автономии человека. Указанные трудности исследованы нами во взаимосвязи, что позволило выявить роль и значение каждой из названных проблем.

Выделенные в данном параграфе проблемы упоминаются в целом ряде работ, посвященных определению конкретных источников рисков, исходящих от ИИ, однако указанные источники содержат лишь поверхностное и неполное их описание, не вскрывают их взаимосвязь и взаимообусловленность. Так, международные стандарты безопасности ИИ касаются лишь отдельных аспектов причинения вреда (например, необъяснимости или неуправляемости систем ИИ). При этом в них отсутствует обоснованная оценка выделенных рисков и необходимая для практического применения таксономия источников рисков, без чего разработка соответствующих стандартов

не может привести к достижению целей безопасного применения систем ИИ. Поскольку документы дают лишь краткое описание источников рисков, практически невозможно сформировать общее понимание потенциальных трудностей и нежелательных для человека последствий, обусловленных существующими угрозами.

На наш взгляд, проблема причинения вреда, возникающая в процессе взаимодействия человека с системами ИИ, приводит к риску отчуждения человека от самого себя, от собственной сущности и предназначения, возможной утрате и искажению его особого статуса в мире как единственного существа, способного к интеллектуальной деятельности, нацеленной на изменение, преобразование окружающей реальности. Проблемы социальной несправедливости и нарушения автономии призваны уточнить смысл, раскрыть отдельные грани и описать разнообразные проявления причиняемого человеку ущерба, а также последствия, возникающие из-за некорректного внедрения, применения систем ИИ.

В параграфе 2.3. «Причины формирования негативных последствий взаимодействия человека и ИИ» раскрыто содержание проблем непрозрачности, нарушения конфиденциальности и отсутствия ответственности. Существующие источники, в которых упоминаются названные проблемы, как правило, фиксируют их наличие, но не исследуют суть и значение указанных трудностей. Герменевтический анализ и сопоставление целого ряда документов правового и этического характера, а также результатов научных исследований последних лет позволили нам выявить сущность и взаимосвязь данных проблем с центральной проблемой причинения вреда человеку вследствие применения им систем ИИ.

Проблема непрозрачности заключается в том, что принцип работы систем ИИ становится в процессе его совершенствования все более непрослеживаемым, необъяснимым и неинтерпретируемым. Человеку оказывается принципиально недоступен порядок его работы, а также последовательность выстраиваемой им цепочки рассуждений. Непрозрачность в работе систем

ИИ чаще всего обусловлена сложностью этих систем и использованием метода глубокого обучения, а также высокой степенью вероятности возникновения ошибок в алгоритмах их деятельности.

Проблема отсутствия ответственности состоит в том, что природа ИИ не позволяет установить ответственное лицо, субъект ответственности, на которого однозначно можно было бы возложить вину в случае причинения вреда человеку системой ИИ. Также данная проблема заключается в невозможности определить меру ответственности тех или иных лиц в инцидентах, произошедших с участием систем ИИ, в силу недостаточности и несовершенства правового регулирования развития и применения ИИ, поспешного и зачастую непродуманного внедрения его систем в различные сферы жизни современного общества.

Проблема нарушения конфиденциальности возникает из-за угрозы утечки потока персональных данных или потери контроля над этими данными, которые могут быть скомпрометированы, вопреки воле и желанию человека опубликованы в открытых источниках и в дальнейшем использоваться для совершения мошеннических действий.

Указанные проблемы, по нашему мнению, выступают потенциальными источниками, причинами причинения вреда человеку, угрожающими его безопасности.

Наша позиция заключается в том, что в разработке и развитии ИИ важно выделять критические точки этого развития с позиций социальной, этической, аксиологической разрешимости, допустимости рассматриваемых технологий. Необходимо сосредоточиться на том, каких результатов мы ожидаем от технологий, а не на том, что технологии ожидают от нас. Разработка, внедрение систем ИИ – это междисциплинарная задача, которая требует подлинного объединения, синтеза технических, технологических, социологических, философских знаний. Она требует иных форм проектной работы, иного уровня осознания стоящих перед человеком проблем и задач. При этом ведущую, направляющую роль в подобных исследованиях должна играть соци-

альная философия, предметом постижения которой является общество, рассмотренное с позиций целостности и системности, наиболее общих законов его динамики, осмысления фундаментальных причин событий и процессов, основных направлений развития социума, необходимых для выявления места и роли человека в мире, реализации заложенного в нем творческого, созидательного потенциала. Именно социально-философское познание, научность которого, по словам П.В. Алексеева должна «сливаться с гуманистичностью», позволяет в процесс обсуждения проблем применения ИИ подойти к пониманию, выявлению сущности базовых принципов, требований, норм, регулирующих и обеспечивающих безопасное и эффективное применение искусственного разума в любых отраслях и сферах жизни социума.

В Главе III «Философские принципы разработки, внедрения и применения систем искусственного интеллекта» на основании результатов исследования особенностей и проблем применения систем ИИ предпринята попытка сформулировать философские принципы эффективного, безопасного и надежного ИИ, учет которых разработчиками и пользователями систем ИИ, по нашему мнению, способен предотвращать, снижать возможность возникновения и эскалации социально-этических проблем применения ИИ во всех отраслях жизни современного социума. Это принципы прозрачности, непричинения вреда, автономии, конфиденциальности, ответственности, социальной справедливости.

В параграфе 3.1. «Условия эффективного и безопасного применения систем ИИ» соискатель сосредоточивает внимание на философских принципах прозрачности, ответственности и конфиденциальности, отражающих соответственно требования об открытости и доступности, подотчетности систем ИИ человеку, о наложении запрета на нарушение его личных границ.

Открытость, доступность систем ИИ конкретизирует содержание принципа прозрачности. Последний подразумевает необходимость предоставления пользователю всей полноты информации о системе ИИ, принципах

и методах ее работы, используемых ею данных, возможных негативных последствиях и угрозах, выявленных ограничениях в ее деятельности и других возможных негативных последствиях ее применения.

Требование подотчетности систем ИИ раскрывает сущность принципа ответственности, согласно которому устанавливается обязательность выполнениях всех необходимых условий корректного использования ИИ для любых лиц или организаций, причастных к разработке и эксплуатации систем ИИ на всем протяжении их жизненного цикла. Причем обязательность сопровождается определением конкретного лица, на которое возлагается вся, в том числе, правовая ответственность за недочеты, неисправности, сбои, возникающие в работе системы.

Наконец, философский принцип конфиденциальности раскрывается в хорошо известных требованиях соблюдения правил работы с личными, персональными данными пользователей систем ИИ, что необходимо для защиты и соблюдения границ личного пространства человека, обеспечения безопасности и защищенности его от нежелательного вторжения третьих лиц.

В параграфе 3.2. «Принципы справедливости, автономии и непричинения вреда человеку в контексте использования ИИ» рассмотрены требования, обеспечивающие исключение возможной предвзятости и дискриминации человека со стороны ИИ, а также гарантирующие самостоятельность, независимость человека в принятии решений, что составляет суть принципов справедливости и автономии, выступающих важнейшими проявлениями принципа непричинения вреда человеку, как центрального, ведущего во всей совокупности правил и норм взаимодействия человека и ИИ.

Социальная несправедливость в контексте применения систем ИИ может возникать из-за объективно существующих различий индивидов, разницы их экономических, политических статусов, принадлежности к разным группам, выделенным по возрасту, полу, национальной принадлежности и т.д. Системы ИИ не должны усугублять уже существующие в обществе предрассудки и заблуждения. Напротив, их применение должно быть органи-

зовано таким образом, чтобы все группы населения имели свободный и равный доступ к преимуществам и выгодам, предоставляемым новыми технологиями.

Автономия в рассматриваемом контексте предполагает защиту тезиса о главенствующем, центральном положении человека во всей совокупности его отношений с искусственным разумом, поддержку людей в принятии ими взвешенных и обоснованных решений в соответствии исключительно с их собственными целями и задачами, позитивную свободу человека, его неотъемлемое право защищать себя от неоправданного принуждения, обмана или манипуляций со стороны систем ИИ.

Указанные принципы в совокупности с требованиями, рассмотренными в параграфе 3.2., позволяют подойти к пониманию центрального, основополагающего принципа применения ИИ, принципа непричинения вреда. Его содержание раскрывает тезис о недопущении нежелательных, негативных следствий во взаимодействии человека и ИИ, а также дает современную интерпретацию, толкование, понимание того, что собой представляет корректный, безопасный, эффективный ИИ, осуществляющий помощь и поддержку человеку в решении им собственных проблем, в достижении им собственных, самостоятельно определенных целей развития.

Изучая сущность и раскрывая содержание основных философских принципов применения систем ИИ, в работе рассмотрены и сформулированы рекомендации, необходимые для успешной реализации указанных выше принципов. Среди них инклюзивность и разнообразие командных ролей, обучение и осведомленность об этических ценностях, непрерывное планирование, выполнение и мониторинг фундаментальных принципов в жизненном цикле систем ИИ, начиная с разработки и заканчивая их применением на практике. Не менее важной рекомендацией является стандартизация, поскольку требование единого стандарта предназначено для достижения функциональной совместимости и совместной работы между производителями, недопущения отраслевой монополии и ограничения прав пользователей.

Нельзя забывать о постоянном контроле со стороны общественности, информировании ее о реальных трудностях и проблемах, возникающих в процессе использования систем ИИ, поскольку именно общество, его социальные группы и институты формулируют цели и задачи той деятельности, для которой проектируются, создаются и используются системы узкого ИИ.

В **заключении** работы подведены итоги проведенного исследования, резюмируются основные положения, дается оценка соответствия полученных результатов сформулированным в начале исследования цели и задачам, намечены направления дальнейшей работы соискателя.

Основные положения диссертации отражены в следующих публикациях автора:

Статьи в российских журналах, включенных в Перечень рецензируемых научных изданий, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук (далее – Перечень ВАК):

1. Бадмаева М.Х. Повседневная жизнь человека в умном городе / М.Х Бадмаева // Вестник Бурятского государственного университета. Философия. – 2020. – Вып.4. – С. 31-38.

2.Бадмаева М.Х. Этика искусственного интеллекта: принцип ответственности Г. Йонаса / М.Х Бадмаева // Вестник Бурятского государственного университета. Философия. –2022. – Вып.1. – С. 67-79.

3. Бадмаева М.Х. К вопросу об особенностях и проблемах применения систем ИИ в сельском хозяйстве / М.Х Бадмаева // Вестник Бурятского государственного университета. Философия. – 2022. – Вып.3. – С. 75-82.

4. Бадмаева М.Х. Искусственный интеллект: друг или враг? / М.Х Бадмаева, М.В. Золхоева // Евразийский юридический журнал. – 2023. – №2. – С. 504-507.

Монографии:

Бадмаева М.Х. Системы искусственного интеллекта: преимущества и ограничения // Социальные и культурные процессы в современном обществе:

монография / Л. Л. Абаева, Ж. А. Аякова, К. А. Багаева [и др.]; науч. ред. Д. Ш. Цырендоржиева. - Улан-Удэ: Издательство Бурятского госуниверситета, 2022. – С. 75-98.

Прочие публикации (в изданиях, не включенных в Перечень ВАК и международные базы данных):

Бадмаева М.Х. Проблемы применения искусственного интеллекта в «умных городах» мира / М.Х. Бадмаева // Социальные процессы в современном российском обществе: проблемы и перспективы. - Иркутск, 24 апр. 2020 г. / ФГБОУ ВО «ИГУ»; [отв. ред. О. Б. Истомина]. – Иркутск : Издательство ИГУ, 2020. – С. 324-331.

Бадмаева М.Х. Человек в цифровой действительности (опыт Сингапура) / М.Х. Бадмаева, М.В. Золхоева // Социальные процессы в современном российском обществе: проблемы и перспективы. - Иркутск, 23 апр. 2021 г. / ФГБОУ ВО «ИГУ»; [отв. ред. О. Б. Истомина]. – Иркутск : Издательство ИГУ, 2021. – С. 350-355.

Badmaeva M. H. Problems of Using Artificial Intelligence in the Field of Medicine: Socio-Philosophical Analysis // Maina H. Badmaeva, Ksenia A. Bagaeva, Oyuna B. Balchindorzhieva, Maria V. Zolkhoeva, Erzhena D. Chagdurova // European Proceedings of Social and Behavioural Sciences: International Scientific Conference. – Vol. 126. – № 12. – P. 97-108.