Вестник Томского государственного университета. Философия. Социология. Политология. 2025. № 85. С. 5–20.
Tomsk State University Journal of Philosophy, Sociology and Political Science. 2025. 85. pp. 5–20.

ОНТОЛОГИЯ, ЭПИСТЕМОЛОГИЯ, ЛОГИКА

Научная статья
УДК 004.8:17
doi: 10.17223/1998863X/85/1

«ЧЕРНЫЙ ЯЩИК» ИЛИ ПРОЗРАЧНЫЙ АЛГОРИТМ: АНАЛИТИЧЕСКИЙ ОБЗОР ИСТОЧНИКОВ ПО ЭТИКЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Павел Николаевич Барышников

Пятигорский государственный университет, Пятигорск, Россия, pnbaryshnikov@pgu.ru

Аннотация. В статье анализируются современные исследования этических аспектов ИИ, результаты которых были опубликованы за последние 5 лет. Рассмотрены ключевые проблемы: прозрачность алгоритмов, распределение ответственности, дискриминация, защита данных. Выявлены противоречия между теорией и практикой, раскрыты причины доминирования технократического подхода. Особое внимание уделено военному применению ИИ и необходимости разработки практических механизмов реализации этических принципов.
Ключевые слова: этика искусственного интеллекта, распределение моральной ответственности, этические дилеммы, регулирование, объяснимый искусственный интеллект, затрудняющие факторы

Благодарности: исследование выполнено за счет гранта Российского научного фонда № 25-28-00254, https://rscf.ru/project/25-28-00254

Для цитирования: Барышников П.Н. «Черный ящик» или прозрачный алгоритм: аналитический обзор источников по этике искусственного интеллекта // Вестник Томского государственного университета. Философия. Социология. Политология. 2025. № 85. С. 5–20. doi: 10.17223/1998863X/85/1

ONTOLOGY, EPISTEMOLOGY, LOGIC

Original article

A “BLACK BOX” OR A TRANSPARENT ALGORITHM: AN ANALYTICAL REVIEW OF SOURCES ON THE ETHICS OF ARTIFICIAL INTELLIGENCE

Pavel N. Baryshnikov

Pyatigorsk State University, Pyatigorsk, Russian Federation, pnbaryshnikov@pgu.ru

Abstract. The rapid advancement of artificial intelligence (AI) technologies presents significant ethical challenges that demand attention from both the scientific community and society. This article reviews AI ethics research published over the past five years, identifying key issues and barriers to implementing ethical principles across various domains. The aim of the study is to identify key problem areas and limiting factors that hinder the implementation of ethical

principles in various areas of application of computational intelligent systems. The sampling methodology includes a systematic content analysis of source metadata with subsequent data clustering and thematic modeling implemented using NLP tools. The work pays special attention to the fundamental contradictions between theoretical provisions and their practical implementation, the dominance of the technocratic utilitarian approach over humanitarian expertise, as well as insufficient elaboration of mechanisms for distributing moral responsibility. In medicine, the opaque nature of machine learning models complicates clinical decision-making and accountability. In the legal field, AI used for risk assessment and judicial decisions raises concerns about algorithmic bias and fairness. A broader issue is the unclear distribution of moral responsibility in multi-stakeholder environments involving developers, users, and affected parties. Fairness and discrimination remain central concerns, as mathematical definitions of fairness often conflict and fail to reflect diverse cultural contexts. Data privacy is increasingly strained by AI's demand for personal information, and global disparities in regulation hinder consistent protection. Military AI poses distinct ethical dilemmas, particularly regarding lethal autonomous systems lacking meaningful human control. The study highlights a technocratic trend in AI ethics discourse, prioritizing quantifiable performance metrics over nuanced ethical reflection. Emerging challenges at the intersection of AI applications create complex feedback loops beyond the reach of existing frameworks. Despite ongoing efforts, unresolved issues include achieving transparency, defining collaboration guidelines, and developing adaptive ethical models. The article calls for interdisciplinary research, greater stakeholder inclusion, and international cooperation to build AI systems aligned with human values and social well-being.
Keywords: ethics of artificial intelligence, distribution of moral responsibility, ethical dilemmas, regulation, explainable artificial intelligence, complicating factors

Acknowledgments: The study is supported by the Russian Science Foundation, Project No. 24-28-00540, https://rscf.ru/project/24-28-00540

For citation: Baryshnikov, P.N. (2025) A "black box" or a transparent algorithm: an analytical review of sources on the ethics of artificial intelligence. *Vestnik Tomskogo gosudarstvennogo universiteta. Filosofiya. Sotsiologiya. Politologiya – Tomsk State University Journal of Philosophy, Sociology and Political Science.* 85. pp. 5–20. (In Russian). doi: 10.17223/1998863X/85/1

1. Введение. Параметры и области применения этики ИИ

1.0. Определение терминов

Прежде чем переходить к подробному обзору современных исследований по этике этических ИИ в различных областях человеческой деятельности, необходимо определить список понятий, которые будут использоваться в ходе дальнейших рассуждений.

Искусственный интеллект (ИИ) – технология, которая позволяет компьютерам и машинам имитировать человеческое обучение, понимание, решение задач, принятие решений, творчество и автономность. Приложения и устройства с ИИ могут видеть и распознавать объекты, понимать и реагировать на человеческий язык, обучаться на основе новой информации и опыта, предоставлять детализированные рекомендации пользователям и экспертам, а также способны действовать в автономном режиме[1].

[1] Это определение, которое является обобщением классических формулировок, нуждается в дополнениях и уточнениях. Дело в том, что в течение последних нескольких лет внимание исследователей сосредоточено на прорывах в области генеративных трансформеров. В обывательском представлении термины ИИ и GPT стали синонимами. Важно подчеркнуть, что сегодня широкое тематическое поле «Искусственный интеллект» включает в себя все эволюционные технологические этапы: от «Логического теоретика» Г. Саймона и А. Ньюмана, работающего с простыми списками, до последних мультимодальных генеративных LLM (Large Language Models – Большие языковые модели), обучающихся на больших объемах данных при помощи триллионов параметров.

Этика ИИ – междисциплинарная область, изучающая, как максимизировать пользу от искусственного интеллекта, минимизируя при этом риски и негативные последствия. Принципы этики ИИ реализуются через систему управления ИИ, которая включает механизмы контроля, направленные на обеспечение безопасности и соответствия этическим нормам инструментов и систем ИИ [1].

Этические вызовы ИИ – системные ограничения, препятствующие справедливому и ответственному использованию ИИ.

Этические критерии (далее в тексте: этические рамки, принципы, ограничения, фреймворки) – структурированные подходы, используемые для оценки моральных проблем и принятия решений. Теоретические основы критериев: утилитаризм, деонтология и этика добродетели. Утилитаризм фокусируется на результатах и стремится к наибольшему благу для наибольшего числа людей, в то время как деонтология подчеркивает долг и правила, а этика добродетели подчеркивает особый статус моральных добродетелей, приближающих человека к совершенству.

1.1. Общая характеристика предметной области

Для общей характеристики такой предметной области, как «Этика искусственного интеллекта», целесообразным будет опереться на нормативный документ, в основании которого заложены этические принципы и ценностные установки. Мы предлагаем в качестве иллюстрирующего документа использовать «Рекомендации по этике ИИ», принятые 23 ноября 2021 г. ЮНЕСКО [2]. Структура и общее содержание документа указывают на то, что с момента свершения так называемой третьей вычислительной (нейросетевой) революции [3] глобальные вызовы, связанные с ИИ, и мировой опыт по их преодолению были осмыслены, обобщены и структурированы. Потребовалось чуть более пяти лет для общественной реакции на появление прорывных ИИ-технологий и новых принципов управления информацией. Можно сказать, что сегодня постепенно вырабатываются правила по организации деятельности в эпоху ИИ и больших вычислений. При этом очевидно, что значимость ИИ и его влияние на общество, окружающую среду, экосистемы и человеческую жизнь в целом уже не оспаривается даже самыми убежденными ИИ-скептиками [4].

Собственно, документ структурирован согласно основным принципам экспозиции проблемы: авторы подчеркивают необходимость признания как положительных, так и отрицательных воздействий, которые ИИ может оказывать на различные сферы жизни. Очевидно, что дальнейшее развитие этой технологии допускает как улучшение качества жизни, так и потенциальные угрозы, такие как нарушение прав человека и социальное неравенство. Этические нормы должны укрепить нормативно-правовое поле для формирования точки баланса между выгодами и угрозами, которые неразрывно связаны с внедрением технологий ИИ в глобальные производственные и коммуникационные процессы. В тексте документа отмечается, что возрастает роль международного сотрудничества в области регулирования ИИ, которое должно привлекать различные заинтересованные стороны, включая государства, частный сектор, научное сообщество и гражданское общество. Авторы осторожно признают, что ИИ может способствовать достижению устойчивого

развития и решению глобальных проблем, таких как изменение климата, бедность и неравенство. Однако для этого необходимо обеспечить этичное и ответственное использование технологии.

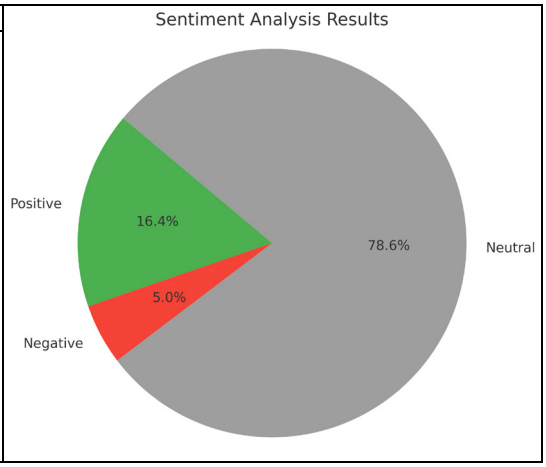
Из общей тональности текста можно заключить: это попытка дать объективную оценку соотношения рисков и положительных перспектив использования технологий ИИ во всех сферах жизни человечества. При этом оценка рисков носит не риторический характер. Способ постановки вопросов обнаруживает серьезную обеспокоенность последствиями неконтролируемой «весны искусственного интеллекта». Вместе с тем с учетом общей сегодняшней геополитической нестабильности и наличия в текущем историческом периоде конфликта цивилизационных платформ и ценностных систем необходимо сделать важное замечание: оценка этических рисков развития технологий ИИ основывается на глобальных ценностных установках без учета региональных интересов. Представленный обзор указывает на этические проблемы ИИ для «человечества вообще» в оптике глобальных ценностей глобального мира [5]. Рассмотрение локальных или региональных ценностных аспектов ИИ требует отдельного исследования.

1.2. Эмоциональные оценки перспектив: между оптимизмом и настороженностью

Из-за колоссальных объемов научной литературы по этическим аспектам ИИ, которая появилась за последние пять лет, не представляется возможным сделать всеобъемлющий обзор. Рассмотрение тематических областей применения правил регулирования в большинстве видов человеческой деятельности (образование, наука, культура, обмен информацией и управление данными, здравоохранение, сельское хозяйство, экология, экономика) обращает внимание на так называемые сентимент-маркеры. Результаты проведенного сентимент-анализа рекомендаций ЮНЕСКО [2] красноречиво свидетельствуют о нейтральной оценке (с элементами осторожного оптимизма) перспектив и значимости применения ИИ в общественно-экономической сфере; вместе с тем очевидно присутствие негативных оценок (табл. 1).

Таблица 1. **Результаты сентимент-анализа текста «Рекомендаций по этике ИИ» ЮНЕСКО [2]**

Позитивно	Негативно	Нейтрально
91	28	436



В документе однозначно указывается на то, что внедрение ИИ в социальные и производственно-технологические процессы неизбежно сопровождается рядом рисков и угроз, требующих постоянной экспертизы. Алгоритмические системы, опирающиеся на диахронические наборы данных, склонны воспроизводить и закреплять предвзятые точки зрения, что приводит к дискриминации маргинализированных групп и ставит под сомнение справедливость принимаемых решений. Одновременно с этим использование ИИ в обработке персональной информации порождает новые вызовы конфиденциальности, требующие разработки этических и правовых механизмов регулирования принципов обмена и хранения данных. К тому же уязвимость ИИ к внешним атакам и сбоям поднимает вопросы безопасности и надежности, обостряя проблему ответственности за техногенные риски. Процессы автоматизации рутинных задач, базирующиеся на ИИ, трансформируют рынок труда, вытесняя человеческую деятельность и поднимая социально-антропологические вопросы о роли человека в технологическом обществе. Отметим, что принятие решений на основе алгоритмов неизбежно сталкивается с ограничениями машинной рациональности, что вызывает сложные этические дилеммы, связанные с правомерностью и обоснованностью действий человеко-компьютерных систем. Также важно, что использование ИИ для генерации дезинформации и манипуляции общественным мнением угрожает основам социального порядка и девальвирует эпистемическую ценность категории истины в эпоху цифровых технологий [6].

2. Методы выборки и анализа литературы по областям

Существует множество работ, в которых применены новейшие методы контент-анализа при обзоре литературы по проблемам этики ИИ. Особое внимание обращает на себя работа [7], в которой авторы последовательно раскрывают негативные факторы, мешающие сообществам принять единые правила этического дизайна систем ИИ. Можно согласиться с утверждением, что дизайн ИИ-систем должен включать в себя не только архитектурные компоненты решений разработчиков, но и встроенные культурные, этические и юридические принципы регулирования, из которых ключевыми (всего в результате анализа было выявлено 22) являются прозрачность, приватность, ответственность и честность. Общим итогом работы служит следующий тезис:

Вызовы этики ИИ связаны с проблемами в формулировке, интерпретации и применении этических принципов, а также с недостаточной прозрачностью и несовершенством механизмов регулирования, технологическими и институциональными ограничениями.

В нашем систематическом обзоре в качестве ключевых параметров выборки и анализа литературы будет принят список затрудняющих факторов из указанной работы А.А. Хана с соавт. (табл. 2).

Таблица 2. **Факторы, затрудняющие внедрение этики ИИ [7]**

Индекс фактора	Наименование затрудняющего фактора	Англоязычный эквивалент
Ф-01	Недостаток этических знаний	Lack of ethical knowledge
Ф-02	Размытые принципы	Vague principles
Ф-03	Чрезмерно общие принципы	Overly broad principles
Ф-04	Конфликт в практическом применении	Practical application conflicts
Ф-05	Различная интерпретация принципов	Divergent principle interpretations
Ф-06	Недостаточное техническое понимание	Insufficient technical understanding
Ф-07	Дополнительные ограничения	Additional constraints
Ф-08	Отсутствие аудита и мониторинга	Lack of audit and monitoring
Ф-09	Отсутствие юридических рамок	Absence of legal frameworks
Ф-10	Бизнес-интересы	Business interests
Ф-11	Плюрализм этических методов	Pluralism of ethical approaches
Ф-12	Этические дилеммы	Ethical dilemmas
Ф-13	Искажения в работе алгоритмов	Algorithmic biases/distortions
Ф-14	Недостаток руководств и стратегий	Lack of guidelines and strategies
Ф-15	Отсутствие межкультурного сотрудничества	Lack of cross-cultural collaboration

Основное отличие состоит в том, что, во-первых, в наш обзор попали только работы 2022–2024 гг. (ситуация на рынке ИИ кардинально изменилась за последний год), во-вторых, этот список применен не только к этическим аспектам, но и к нескольким технологическим доменам. Ключевой вопрос, ответ на который формулируется в ходе исследования, звучит так:

Какие ключевые факторы затрудняют этическое внедрение ИИ в разных профессиональных доменах (образование, наука, здравоохранение, управление данными)?

Основной замысел состоит в том, чтобы отсортировать наиболее значимые статьи по поисковому терму и провести с помощью инструментов ИИ анализ полнотекстовых версий статей. Значимость определяется не по общему числу цитирований, а по максимальному количеству цитирований в год. Такой метод сортировки позволяет попасть в выборку не только ранним популярным публикациям, но и новым, оцененным и процитированным в последнее время.

До того как проводить детальную проработку наиболее репрезентативных текстов, необходимо провести анализ данных из выборки глобальных библиографических баз. В качестве источника данных используется платформа «Publish or Perrish» – это программное обеспечение, которое извлекает и анализирует академические источники из глобальных библиографических баз метаданных [8]. Программа анализирует публикации по наукометрическим параметрам и предоставляет ряд метрик для проведения сравнительного анализа, включая количество статей, даты публикации, общее количество цитат и индекс Хирша.

При извлечении данных использовался поисковый терм «AI Ethics Challenges». В качестве источников данных послужили такие системы, как Crossref и Google Scholar. Всего проанализированы метаданные более 1 000 источников за 2022–2025 гг.

В качестве данных для анализа использовалось поле с заголовками статей. То есть на вход подается текстовый массив, который сформировался из заголовков статей, извлеченных по поисковому терму «AI Ethics Challenges». Процесс анализа данных в данном исследовании включает несколько ключе-

вых этапов. Сначала данные после сортировки в «Publish or Perrish» загружаются и просматриваются для понимания их структуры. Затем текстовые данные очищаются: приводятся к нижнему регистру, удаляются знаки препинания и служебные слова. После этого проводится частотный анализ слов, чтобы выявить наиболее распространенные термины. Далее тексты преобразуются в числовые векторы, что позволяет проанализировать содержание заголовков. Используется метод поиска скрытых тем, который группирует слова таким образом, чтобы выявить общие концепции. Все статьи делятся на группы (кластеры) по их схожести. Для каждой группы находятся характерные слова, описывающие ее содержание. Результаты визуализируются с помощью графиков, где каждая точка представляет статью, а цвет указывает на принадлежность к определенной группе. Также проводится анализ корреляций между различными характеристиками статей, такими как количество цитирований и возраст публикации. Это помогает понять, какие факторы влияют на популярность работы. В итоге составляется список самых цитируемых статей для каждой группы, что позволяет выделить наиболее значимые исследования в каждой области[1].

В результате проведенного анализа было выявлено пять наиболее репрезентативных тематических кластеров:

0: Взаимодействие человека и ИИ: агентность, справедливость.

1: Юридические и медицинские аспекты ИИ: проблема интерпретируемости.

2: Моральные дилеммы в этике ИИ.

3: Управление и регулирование ИИ, влияние на общество.

4: Этика ИИ в образовании, педагогические практики.

Распределение кластеров в двумерном пространстве показано на рис. 1.

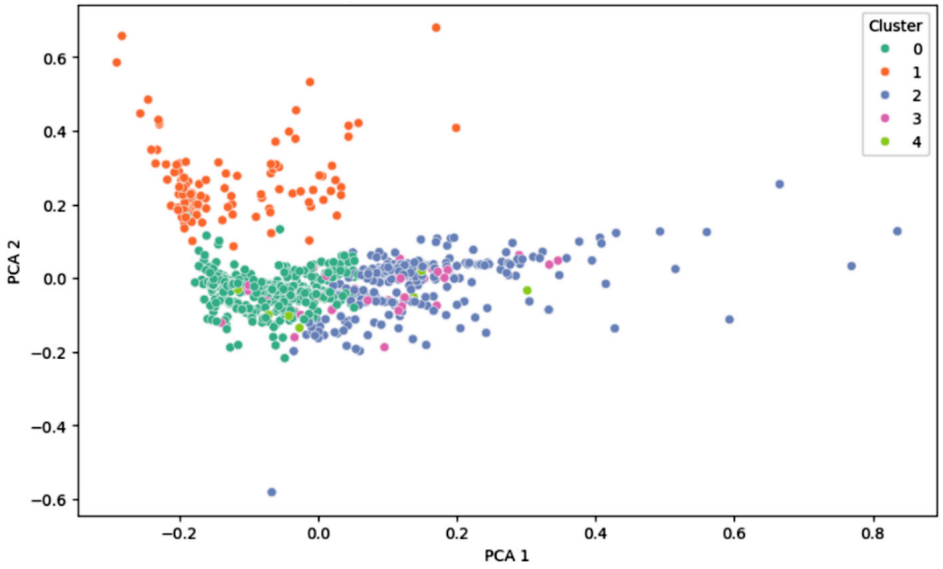


Рис. 1. Распределение тематических кластеров статей по этическим проблемам ИИ

[1] Ознакомиться с исходным кодом можно перейдя по ссылке: https://colab.research.google.com/drive/13fTKxNHxq74I4bOtNz37W5JlkYQPNtw-?usp=sharing

Наиболее плотные группы составляют кластеры 0 (Взаимодействие человека и ИИ: агентность, справедливость) и 2 (Моральные дилеммы в этике ИИ). Из каждой тематической области мы выбрали топ-10 самых цитируемых статей в год. Дальнейшие рассуждения представляют собой авторское обобщение извлеченной информации[1]. Ниже мы рассмотрим наиболее проблемные области из некоторых кластеров и их пересечений, дадим характеристику и общую оценку этическим вызовам.

3. Этические вызовы современного этапа развития ИИ

3.0. Взаимодействие человека и ИИ: агентность, справедливость

Одной из наиболее обсуждаемых проблем сегодня является возможность разработки алгоритмического дизайна этических регулятивов в сфере ИИ. Такой «инженерный» подход к этике вполне объясним, так как бизнес (главный бенефициар развития ИИ) ставит задачи перед разработчиками, а не перед гуманитарным экспертным сообществом. В работе [9] мы видим детальный разбор понятий, этических ограничений (фреймворков) и механизмов их реализации на разных уровнях применения технологии ИИ.

На раннем этапе становления технологий ИИ угрозы, вызовы и этические проблемы были, скорее, умозрительными. Обсуждения этих тем перешли в практическую плоскость вместе с широким распространением интеллектуальных систем на всех уровнях производственных и управленческих процессов. С конца 80-х гг. XX в. происходит интенсификация обсуждений этического компонента технологий ИИ. В 20-е гг. XXI в. мы видим переход ко «второй фазе этики ИИ» [10], в ходе которого происходит смещение от вопросов типа «что такое этические ограничения?» к вопросам типа «как воплотить в жизнь этические ограничения?». Стоит отметить, что многие принципы восходят к классическим формулировкам, применявшимся к растущему рынку ИТ в конце прошлого века: «Этика информационной эпохи строится на четырех ключевых понятиях: конфиденциальность, точность, право собственности и доступность (privacy, accuracy, property, and accessibility)» или «Должны быть гарантии, что информационные технологии и информация, которую они обрабатывают, используются для укрепления достоинства человечества» [11. P. 655].

Сегодня список этических вызовов разросся и вышел далеко за пределы перечисленных понятий. Уровни детализации и спецификации каждой проблемной области усложнились. Например, Э. Прем разработал структурированный список подходов к созданию этичного ИИ, выявив широкий спектр инструментов и методик – от алгоритмов до общих фреймворков и инструментов, которые могут стать частью инфраструктуры этики ИИ (наборы данных, сообщества или лицензионные модели) [9]. Свои наборы решений предлагает целый ряд авторов [12, 13].

Анализ наглядно выявляет, что подходы могут варьировать по уровню конкретности: от программного кода до концептуальных рассуждений. Кон-

[1] Не все статьи, выбранные таким методом, попадут в список литературы, поэтому прилагается ссылка на таблицу с полным перечнем ключевых источников: https://github.com/kagort/databases2/blob/master/PoPCites.csv

кретные решения, такие как алгоритмы, обычно решают лишь одну этическую проблему, тогда как более абстрактные подходы часто реагируют на комплексные этические вызовы общественного уровня. Мы видим, что технические подходы сосредоточены на ограниченном круге вопросов, таких как объяснимость и справедливость ИИ, которые трансформировались в целые области машинного обучения. Хотя множество исследований выявило сопоставимые наборы принципов для этичного ИИ, техническая реализация этих принципов демонстрирует такое богатое и зачастую противоречивое разнообразие подходов, что сегодня нет возможности говорить об универсальных решениях.

В вопросах о справедливой ИИ нужно признать, что инструментальный подход порождает много проблем взаимосвязи этики и дизайна ИИ. Для успешной реализации этичных систем необходимо изучать их применение в конкретных контекстах, аналогично медицинской этике. Это может включать разработку стандартных ситуаций, публикацию лучших практик и их обсуждение. Предполагается, что такие подходы можно сочетать со стандартными аудитами, сертификатами, декларациями и регулированием [14].

Относительно проблемы агентности вопросы стоят еще острее. Современный спектр этических проблем, связанных с искусственными автономными агентами, охватывает широкие сферы: беспилотный транспорт, роботы в сфере здравоохранения, автономные оружейные комплексы и многое другое. Тем не менее сегодня пока недостаточно теоретических, методологических или эмпирических исследований, в которых раскрывался бы универсальный «рецепт» для понимания, например, проблемы распределения моральной ответственности при использовании ИИ-роботов.

Для преодоления подобных затруднений в некоторых работах [15, 16] разрабатывается концептуальная рамка, которая интерпретирует этические последствия применения ИИ-роботов на основе дескриптивной и нормативной этической теории. З. Тот с коллегами в результате скрупулезного анализа пришли к выводу, что при росте агентности ИИ возрастают и моральная ответственность, и механизмы ее распределения между множеством акторов и институтов. Этика искусственных интеллектуальных агентов может раскрываться в трех аспектах: пользовательский, производственный и институциональный. То есть этические компоненты должны реализовываться на трех уровнях:

1) в пользовательском опыте взаимодействия с ИИ-агентами;

2) в инженерных схемах на этапе разработки;

3) на институциональном уровне за счет механизмов государственного или муниципального управления.

Нормативная база при этом должна обладать способами распределения моральной ответственности между всеми уровнями. Для разработки такой нормативной базы необходимо определение «границ ответственности».

Еще более острым выглядит вопрос о моральной ответственности в сфере военного применения систем так называемой ситуативной осведомленности (Situational Awareness) ИИ. В работах, посвященных внедрению интеллектуальных комплексов в военную сферу [16–18], рассматриваются различные технологии: разведывательные методы сбора данных, сети датчиков, программно-определяемые радиоустройства, беспилотные летательные

аппараты, автономные транспортные средства, системы определения местоположения выстрелов, интернет вещей, туманная вычислительная мощность, слияние информации, автоматическое распознавание событий в потоковых видео и системы расширенного зрения. Эти технологии могут быть применены в различных областях – там, где ситуативная осведомленность играет ключевую роль: поле боя, городские боевые действия, «серая зона» войны, государственная безопасность, защита критической инфраструктуры. Тем не менее в этих работах этический аспект применения ИИ-технологий вообще не упоминается. То есть парадоксальным образом вопросы этики здесь стоят особенно остро именно потому, что они полностью обходятся. Скорее исключением из правил представляются аналитические обзоры по этическим проблемам автономных военных летальных комплексов, сделанные организациями типа Международного комитета Красного Креста: «Принимая на себя большую когнитивную нагрузку в военных операциях, системы поддержки на основе ИИ могут ослабить человеческий элемент морального и этического принятия решений. Переход от суждений, руководимых человеком, к суждениям с помощью ИИ может подорвать способность командиров полностью брать на себя моральную ответственность за свои решения, в конечном итоге ставя под угрозу военные добродетели» [19].

В итоге взаимодействие систем ИИ и человека остается проблемной областью, требующей межинституциональной работы. Обзор источников подтвердил, что чем более прикладной характер носят исследования по взаимодействию искусственных интеллектуальных систем и человека, тем менее остро формулируются (а иногда и вовсе опускаются) этические вопросы. Ситуативная осведомленность как результат такого взаимодействия позволяет преодолеть операциональные осложнения, которые возникают при обработке больших объемов разнородных данных, постоянном их обновлении и при необходимости эффективной визуализации комплексной информации для принятия решений. Поиск инструментальных подходов к этике ИИ приводит к такому разнообразию решений еще на этапе разработки потому, что проблема моральной ответственности перед другими акторами (индивидуальный пользователь или государственные регуляторы) не встает.

3.1. Юридические и медицинские аспекты ИИ, проблема интерпретируемости

Проблема интерпретируемости ИИ в областях человеческой деятельности, где целеполагание ориентировано на содержание этических категорий (юриспруденция, медицина, финансы, беспилотные транспортные системы), сформировала такое направление, как «Объяснимый ИИ» (Explainable AI, XAI). Explainable AI характеризуется рядом ключевых свойств: прозрачность, интерпретируемость, доверие. Одно из основных требований – прозрачность. Логика работы модели должна быть легко интерпретируемой. Это может достигаться, например, за счет использования менее сложных алгоритмов, таких как деревья решений, вместо глубоких нейронных сетей. Интерпретируемость – результаты и выводы системы должны быть представлены в доступной форме для широкого круга пользователей, включая неспециалистов. Система обязана предоставлять подробные объяснения относительно того, какие входные данные повлияли на итоговое решение и почему был

выбран конкретный вариант действия. Доверие системе обеспечивается прозрачностью и интерпретируемостью, что особенно критично в регулируемых отраслях, где важно не только получить верный результат, но и объяснить его этиологические и телеологические свойства [11].

Стоит отметить, что запрос со стороны системы здравоохранения на прозрачность и объяснимость принципов работы экспертных систем и систем принятия решений на базе ИИ был сформулирован задолго до повсеместного внедрения LLM. В 1970-х гг. разработаны новые способы представления структурированных экспертных баз знаний о клинических и биомедицинских проблемах с использованием причинных, таксономических, ассоциативных и фреймовых моделей [20]. В последнее время остро стоит вопрос о прозрачности и распределении моральной ответственности при принятии решений с опорой на данные систем ИИ.

Результаты обобщения обзоров [21] по ключевым проблемам, связанным с технологиями ИИ в здравоохранении, представлены в табл. 3.

Таблица 3. **Корреляция вызовов с ИИ-технологиями**

ИИ-технологии	Вызовы
Прогностические алгоритмы	Черный ящик: сложность интерпретации результатов. Отсутствие учета человеческой автономии
Системы поддержки принятия решений	Необходимость информированного согласия. Возможная дискриминация из-за предвзятости в данных. Проблемы объяснимости
Технологии на основе личных данных	Риск нарушения конфиденциальности. Сложности в управлении данными
Модели прогнозирования госпитализаций/смертности	Предвзятость на основе типа страхования (по полу, расе, медицинским показаниям)
Встроенный (embedded) ИИ	Угроза нарушения автономии пациента. Возможные технические сбои или ошибки в работе роботов
ИИ для общественного блага	Сложности в обеспечении справедливого доступа. Необходимость защиты конфиденциальности при использовании данных
Алгоритмы машинного обучения в диагностике	Непрозрачность алгоритмов («черный ящик»). Отсутствие полной интерпретируемости результатов
Все виды технологий на основе ИИ	Дискриминация из-за предвзятых алгоритмов. Проблемы с объяснением работы систем
ИИ в первичной медико-санитарной помощи	Отсутствие совместного принятия решений с пациентами. Зависимость от решений ИИ может быть источником опасности
Общие проблемы всех технологий	Потенциальная потеря конфиденциальности данных. Баланс между пользой технологии и рисками для человека. Сложности в обеспечении справедливости и с отсутствием предвзятости

Среди основных принципов особенно значимыми являются:

• *уважение к человеческой автономии*, что предполагает обеспечение прав пациента на принятие решений и необходимость получения информированного согласия;

• *предотвращение вреда* через обеспечение безопасности технологий ИИ и минимизацию рисков ошибок или негативных последствий их применения;

• *справедливость* – реализуется через устранение предвзятости и дискриминации, а также обеспечение равного доступа к медицинским услугам для всех групп населения;

• *объяснимость* алгоритмов, направленная на преодоление проблемы «черного ящика» и создание понятных для пользователей моделей;

• *защита конфиденциальности пациентов*, включая соблюдение норм законодательства, таких как GDPR (Общий регламент о защите данных).

Однако реализация этих принципов сталкивается с рядом серьезных затруднений. Существует риск снижения роли человека в процессе принятия решений, что может быть обусловлено недостаточной ориентацией технологий на потребности пациента. Вместе с тем технические сбои или ошибки в работе ИИ представляют собой потенциальную угрозу безопасности, особенно в условиях чрезмерной зависимости от технологий. Также необъективные данные, формирующие обучающие датасеты, могут усиливать социальную дискриминацию. При этом сложность интерпретации работы моделей глубокого обучения создает препятствия для достижения прозрачности и объяснимости решений. Одной из самых острых проблем остается проблема защиты данных, так как использование ИИ повышает риск утечки личной информации пациентов, требуя постоянного балансирования между безопасностью данных и полезностью технологий. Если обобщать вышесказанное, то напрашивается тривиальный вывод: такого рода вызовы подчеркивают необходимость разработки комплексных этических подходов для успешного внедрения искусственных интеллектуальных систем в здравоохранение и сопровождающую юридическую практику. Ниже будет представлен и обоснован пессимистичный взгляд на саму возможность подобной разработки.

3.2. Моральные дилеммы и этика ИИ

Особое место в обзорах по этическим проблемам ИИ занимают исследования, посвященные моральным дилеммам в тех областях, где делегирование моральной ответственности машинам, с одной стороны, повышает скорость, точность и эффективность принимаемых решений, с другой – оставляет сам акт делегирования вне области моральных оценок. Последнее при фатальных ошибках приводит к неразрешимым этическим и юридическим ситуациям. Стоит отметить, что сегодня хорошо исследованы психологические аспекты доверия искусственным интеллектуальным системам. В этических дилеммах наблюдается противоречие: люди предпочитают исключать ИИ из решений, связанных с моральной ответственностью (при вынесении медицинских решений, например), но эта тенденция ослабевает, если система напрямую влияет на жизнь оператора. При возникновении моральных ошибок люди чаще обвиняют ИИ, считая его неспособным к моральным суждениям [22].

Пользователи компьютерных систем еще на этапе самых ранних разработок ИИ были склонны к своего рода пользовательскому анимизму (эффект бессознательного одушевления технических устройств). Этот эффект сохраняется и усиливается при работе с современными системами ИИ и оказывает сильное влияние на оценку выбора в моральных дилеммах. Ситуация осложняется тем, что, обладая высокой компетентностью и низкой эмоциональностью, ИИ в моральных дилеммах избегает деонтологических решений, предпочитая утилитарные (последние лучше поддаются компьютерному инжинирингу). При этом такие решения чаще вызывают негативную реакцию со стороны людей. Подобное положение вещей связано с рядом факторов, которые сегодня исследуются в прикладном ключе целыми исследовательскими группами [23]. Указанные факторы можно представить следующим списком:

1. Предписывание ИИ-агентам человеческих норм [24].

2. Запрет на бездействие (люди принимают бездействие как реакцию на дилемму со стороны человека и не допускают такое поведение ИИ-агента).

3. Высокая степень влияния ситуативного контекста при оценивании моральных действий ИИ-агента.

Особый резонанс (если судить по количеству цитирований) вызывают обзоры этических проблем с резко очерченной пессимистичной точкой зрения авторов [25. P. 61; 26]. Например, Л. Мунн полагает, что повсеместная активная разработка этических принципов – это социальный ответ на стимул, идущий от возрастающей роли технологий ИИ в общественной и частной сферах. Но при этом данные принципы изолированы, бессвязны, «беззубы» и применяются в бизнес-сферах (IT-индустрия, образование), ориентированных в первую очередь на рентабельность, а не на этику. В итоге образуется непреодолимый разрыв между постулируемыми принципами и реальными технологическими практиками. Это связано с тем, что социальные концепты, закладываемые в этические директивы, не выразимы языком технических инструкций. Также важно, что неэтичный ИИ является продуктом неэтичной индустрии, в связи с чем наблюдается парадоксальная ситуация: технологический прорыв в сфере ИИ влияет на все сферы общественно-экономической жизни, но при этом существуют силы, которые прямо сопротивляются регулированию. В работе Л. Мунна обращают на себя внимание элементы «революционной» риторики, указывающей на «ИИ-капитализм», «деколонизацию сферы ИИ», «манифест ИИ-справедливости» и т.п. [26].

На наш взгляд, «проблема перевода» концептуальных систем с языка бизнес-утилитаризма на языки этических ограничений недооценена и требует всесторонней проработки со стороны гуманитарной экспертизы.

Выводы

В данном исследовании был представлен систематический анализ современных исследований по этике искусственного интеллекта, проведенный в несколько этапов, начиная с контент-анализа научной литературы 2022–2025 гг., с последующей структуризацией извлеченных данных методами кластеризации и тематического моделирования. В результате было выявлено несколько основных кластеров проблематизации этических аспектов ИИ, включая взаимодействие человека и ИИ, юридические и медицинские аспекты, моральные дилеммы, управление и регулирование. Основная проблема в сфере этики ИИ сегодня – разрыв между теоретическими положениями и их практической реализацией, преобладание инструментального подхода над гуманитарным анализом и недостаточная проработка вопросов распределения моральной ответственности. Этический ракурс новых технологий не вписывается в стратегические проекты институтов власти и бизнеса, инвестирующих в разработку технологических инноваций в сфере ИИ.

К нерешенным сложным проблемам относятся разработка универсальных механизмов распределения моральной ответственности при взаимодействии человека и ИИ, обеспечение прозрачности и объяснимости работы алгоритмов в медицинской и юридической сферах, решение моральных дилемм при делегировании решений ИИ, преодоление рисков дискриминации и предвзятости в работе алгоритмов, а также защита персональных данных при

использовании ИИ-систем в сфере образования и здравоохранения. Особую остроту приобретают вопросы военного применения ИИ, где этический аспект часто игнорируется в угоду технологической и боевой эффективности.

Таким образом, можно сделать итоговое заключение: несмотря на значительные усилия в осмыслении этических вызовов, в ходе нового витка развития технологий искусственного интеллекта сохраняется комплекс нерешенных проблем, требующих разработки практических механизмов реализации этических принципов в условиях непредсказуемо меняющегося технологического ландшафта.

Список источников

1. *Stryker C., Kavlakoglu E.* What is AI? URL: https://www.ibm.com/think/topics/artificial-intelligence

2. *UNESCO.* Recommendation on the Ethics of Artificial Intelligence. 2022.

3. *Zhuravlev D.V., Smolin V.S.* The neural network revolution of artificial intelligence and it's development options. 2023. P. 223–244.

4. *Ambartsoumean V.M., Yampolskiy R.V.* AI Risk Skepticism, A Comprehensive Survey. 2023. URL: https://arxiv.org/abs/2303.03885

5. *Сергейчик Е.М.* Глобальные ценности глобального мира // Вестник Санкт-Петербургского университета. Философия и конфликтология. 2021. Т. 3, № 37. С. 532–543.

6. *Leyra-Curiá S., Soler J.P.* Lying in the age of artificial intelligence: A call to moral and legal responsibility // Church, Communication and Culture. 2023. Vol. 8, № 2. P. 135–153.

7. *Khan A.A. et al.* Ethics of AI: A Systematic Literature Review of Principles and Challenges. 2021. URL: https://arxiv.org/abs/2109.07906

8. *Harzing A.-W.* Using the Publish or Perish software: Crafting your career in academia. London : Tarma Software Research Ltd. UK, 2023. 375 p.

9. *Prem E.* From ethical AI frameworks to tools: a review of approaches // AI Ethics. 2023. Vol. 3, № 3. P. 699–716.

10. *Morley J. et al.* From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices // Sci Eng Ethics. 2020. Vol. 26, № 4. P. 2141–2168.

11. *Hulsen T.* Explainable Artificial Intelligence (XAI): Concepts and Challenges in Healthcare // AI. 2023. Vol. 4, № 3. P. 652–666.

12. *Konstantis K.* The Main Challenges of AI Ethics: Historical Contextualization, Black-Boxing, Social Biases, Labor Invisibility // AIES. 2025. Vol. 7, № 2. P. 23–25.

13. *Schuett J., Reuel A.-K., Carlier A.* How to design an AI ethics board // AI Ethics. 2024. https://doi.org/10.1007/s43681-023-00409-y

14. *Danks D.* Digital Ethics as Translational Ethics // Advances in Human and Social Aspects of Technology / eds. I. Vasiliu-Feltes, J. Thomason. IGI Global, 2021. P. 1–15.

15. *Tóth Z. et al.* The Dawn of the AI Robots: Towards a New Framework of AI Robot Accountability // J. Bus Ethics. 2022. Vol. 178, № 4. P. 895–916.

16. *Munir A., Aved A., Blasch E.* Situational Awareness: Techniques, Challenges, and Prospects // AI. 2022. Т. 3, № 1. P. 55–77.

17. *He L. et al.* Artificial intelligence technology in battlefield situation awareness // E3S Web Conf. 2022. Vol. 360. 01063. URL: https://www.e3s-conferences.org/10.1051/e3sconf/202236001063

18. *Lee C.-E. et al.* Deep AI military staff: cooperative battlefield situation awareness for commander's decision making // J. Supercomput. 2023. Vol. 79, № 6. P. 6040–6069.

19. *Klaus M.* Transcending weapon systems: the ethical challenges of AI in military decision support systems // Humanitarian Law & Policy. 2024. URL:https://blogs.icrc.org/law-and-policy/2024/09/24/transcending-weapon-systems-the-ethical-challenges-of-ai-in-military-decision-support-systems/

20. *Kulikowski C.A.* An Opening Chapter of the First Generation of Artificial Intelligence in Medicine: The First Rutgers AIM Workshop, June 1975 // Yearbook of Medical Informatics. 2015. Vol. 24, № 01. P. 227–233.

21. *Karimian G., Petelos E., Evers S.M.A.A.* The ethical issues of the application of artificial intelligence in healthcare: a systematic scoping review // AI Ethics. 2022. Vol. 2. № 4. P. 539–551.

22. *Zhang Z., Chen Z., Xu L.* Artificial intelligence and moral dilemmas: Perception of ethical decision-making in AI // Journal of Experimental Social Psychology. 2022. Vol. 101. P. 104327. URL: https://linkinghub.elsevier.com/retrieve/pii/S0022103122000464

23. *Malle B.F. et al.* People's judgments of humans and robots in a classic moral dilemma // Cognition. 2025. Vol. 254. C. 105958. https://linkinghub.elsevier.com/retrieve/pii/S0010027724002440

24. *Chu Y., Liu P.* Machines and humans in sacrificial moral dilemmas: Required similarly but judged differently? // Cognition. 2023. Vol. 239. P. 105575. https://linkinghub.elsevier.com/retrieve/pii/S0010027723002093

25. *Heilinger J.-C.* The Ethics of AI Ethics. A Constructive Critique // Philos. Technol. 2022. Vol. 35, № 3.

26. *Munn L.* The uselessness of AI ethics // AI Ethics. 2023. Vol. 3, № 3. P. 869–877.

References

1. Stryker, C. & Kavlakoglu, E. (n.d.) *What is AI?* [Online] Available from: https://www.ibm.com/think/topics/artificial-intelligence

2. UNESCO. (2022) *Recommendation on the Ethics of Artificial Intelligence.* [Online] Available from: https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence

3. Zhuravlev, D.V. & Smolin, V.S. (2023) The neural network revolution of artificial intelligence and its development options. *Futurity designing. Digital reality problems.* Proc. of the 6th International Conference. February 2–3, 2023. Moscow. pp. 223–244.

4. Ambartsoumean, V.M. & Yampolskiy, R.V. (2023) *AI Risk Skepticism, A Comprehensive Survey*. [Online] Available from: https://arxiv.org/abs/2303.03885

5. Sergeyichik, E.M. (2021) Global'nye tsennosti global'nogo mira [Global values of a global world]. *Vestnik Sankt-Peterburgskogo universiteta. Filosofiya i konfliktologiya*. 3(37). pp. 532–543.

6. Leyra-Curiá, S. & Soler, J.P. (2023) Lying in the age of artificial intelligence: A call to moral and legal responsibility. *Church, Communication and Culture*. 8(2). pp. 135–153.

7. Khan, A.A. et al. (2021) *Ethics of AI: A Systematic Literature Review of Principles and Challenges*. [Online] Available from: https://arxiv.org/abs/2109.07906

8. Harzing, A.-W. (2023) *Using the Publish or Perish Software: Crafting your Career in Academia*. London: Tarma Software Research Ltd.

9. Prem, E. (2023) From ethical AI frameworks to tools: A review of approaches. *AI Ethics*. 3(3). pp. 699–716.

10. Morley, J. et al. (2020) From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science and Engineering Ethics*. 26(4). pp. 2141–2168.

11. Hulsen, T. (2023) Explainable Artificial Intelligence (XAI): Concepts and Challenges in Healthcare. *AI*. 4(3). pp. 652–666.

12. Konstantis, K. (2025) The Main Challenges of AI Ethics: Historical Contextualization, Black-Boxing, Social Biases, Labor Invisibility. *AIES*. 7(2). pp. 23–25.

13. Schuett, J., Reuel, A.-K. & Carlier, A. (2024) How to design an AI ethics board. *AI Ethics*. 5. pp. 863–881. DOI: 10.1007/s43681-023-00409-y

14. Danks, D. (2021) Digital Ethics as Translational Ethics. In: Vasiliu-Feltes, I. & Thomason, J. (eds) *Advances in Human and Social Aspects of Technology*. IGI Global. pp. 1–15.

15. Tóth, Z. et al. (2022) The Dawn of the AI Robots: Towards a New Framework of AI Robot Accountability. *Journal of Business Ethics*. 178(4). pp. 895–916.

16. Munir, A., Aved, A. & Blasch, E. (2022) Situational Awareness: Techniques, Challenges, and Prospects. *AI*. 3(1). pp. 55–77.

17. He, L. et al. (2022) Artificial intelligence technology in battlefield situation awareness. *E3S Web Conf*. 360. 01063. [Online] Available from: https://www.e3s-conferences.org/10.1051/e3sconf/202236001063

18. Lee, C.-E. et al. (2023) Deep AI military staff: Cooperative battlefield situation awareness for commander's decision making. *Journal of Supercomputing*. 79(6). pp. 6040–6069.

19. Klaus, M. (2024) *Transcending weapon systems: The ethical challenges of AI in military decision support systems*. [Online] Available from: https://blogs.icrc.org/law-and-policy/2024/09/24/transcending-weapon-systems-the-ethical-challenges-of-ai-in-military-decision-support-systems/

20. Kulikowski, C.A. (2015) An Opening Chapter of the First Generation of Artificial Intelligence in Medicine: The First Rutgers AIM Workshop, June 1975. *Yearbook of Medical Informatics*. 24(1). pp. 227–233.

21. Karimian, G., Petelos, E. & Evers, S.M.A.A. (2022) The ethical issues of the application of artificial intelligence in healthcare: A systematic scoping review. *AI Ethics*. 2(4). pp. 539–551.

22. Zhang, Z., Chen, Z. & Xu, L. (2022) Artificial intelligence and moral dilemmas: Perception of ethical decision-making in AI. *Journal of Experimental Social Psychology*. 101. pp. 104327. [Online] Available from: https://linkinghub.elsevier.com/retrieve/pii/S0022103122000464

23. Malle, B.F. et al. (2025) People's judgments of humans and robots in a classic moral dilemma. *Cognition*. 254. pp. 105958. [Online] Available from: https://linkinghub.elsevier.com/retrieve/pii/S0010027724002440

24. Chu, Y. & Liu, P. (2023) Machines and humans in sacrificial moral dilemmas: Required similarly but judged differently? *Cognition*. 239. pp. 105575. [Online] Available from: https://linkinghub.elsevier.com/retrieve/pii/S0010027723002093

25. Heilinger, J.-C. (2022) The Ethics of AI Ethics. A Constructive Critique. *Philosophy and Technology*. 35(3). p. 61.

26. Munn, L. (2023) The uselessness of AI ethics. *AI Ethics*. 3(3). pp. 869–877.

Сведения об авторе:
Барышников П.Н. – доктор философских наук, доцент, профессор кафедры исторических, социально-философских дисциплин, востоковедения и теологии Пятигорского государственного университета, Министерство науки и высшего образования (Пятигорск, Россия). E-mail: pnbaryshnikov@pgu.ru

Автор заявляет об отсутствии конфликта интересов.

Information about the author:
Baryshnikov P.N. – Dr. Sci. (Philosophy), docent, professor at the Department of Historical, Social and Philosophical Disciplines, Oriental Studies and Theology, Pyatigorsk State University, Ministry of Science and Higher Education (Pyatigorsk, Russian Federation). E-mail: pnbaryshnikov@pgu.ru

The author declares no conflicts of interests.

Статья поступила в редакцию 04.04.2025;
одобрена после рецензирования 21.05.2025; принята к публикации 30.06.2025
The article was submitted 04.04.2025;
approved after reviewing 21.05.2025; accepted for publication 30.06.2025