Хмелевской Илья Романович[1]**,**
Ассистент кафедры теоретических и публичных дисциплин
Тюменского государственного университета,
i.r.khmelevskoi@utmn.ru

ЭТИКО-ПРАВОВЫЕ ОСНОВАНИЯ ОБЪЯСНИМОСТИ РЕШЕНИЙ ИИ В ВЫСОКОРИСКОВЫХ СФЕРАХ[2]

Аннотация. В условиях стремительного внедрения систем искусственного интеллекта (ИИ) в социально значимые сферы возрастает необходимость обеспечения их объяснимости как этико-правового императива. Особенно актуальной эта задача становится при использовании ИИ в высокорисковых областях — здравоохранении, правосудии, социальной политике и безопасности, где последствия ошибочных или непрозрачных решений могут затронуть фундаментальные права человека. В настоящей статье анализируются минимальные критерии объяснимости, позволяющие достичь разумного баланса между эффективностью и автономностью ИИ-систем, с одной стороны, и необходимостью соблюдения принципов справедливости, прозрачности и ответственности, защиты интересов личности — с другой. Внимание сосредоточено на вопросах нормативного закрепления права на объяснение в контексте современных инициатив и разработок, а также на раскрытии оснований, определяющих рамки допустимости применения ИИ в социально чувствительных областях.

Ключевые слова: искусственный интеллект, объяснимость, высокорисковые сферы.

Khmelevskoi Ilia Romanovich,
Assistant of the Department of Theoretical and Public Law Disciplines of
University of Tyumen,
i.r.khmelevskoi@utmn.ru

[1] Научный руководитель: Зенин Сергей Сергеевич, канд. юр. наук, доцент, директор Института государства и права, проректор Тюменского университета

[2] Исследование выполнено при поддержке Министерства науки и высшего образования Российской Федерации в рамках проекта "Фундаментальные проблемы методики разработки и связанного с ней правового и этического регулирования в сфере применения систем и моделей искусственного интеллекта" (FEWZ-2024-0052)

ETHICAL AND LEGAL FOUNDATIONS OF AI DECISION EXPLAINABILITY IN HIGH-RISK SPHERES

Abstract. In the context of the rapid integration of artificial intelligence (AI) systems into socially significant domains, the need to ensure their explainability is becoming an ethical and legal imperative. This issue is particularly relevant in high-risk areas such as healthcare, justice, social policy, and security, where the consequences of erroneous or opaque decisions may affect fundamental human rights. This article examines the minimum criteria for AI explainability that enable a reasonable balance between the efficiency and autonomy of AI systems on the one hand, and the need to uphold the principles of fairness, transparency, responsibility, and the protection of individual interests on the other. Particular attention is given to the legal recognition of the right to explanation in the context of contemporary initiatives and regulatory developments, as well as to the ethical foundations that define the boundaries of permissible AI use in socially sensitive contexts.

Keywords: artificial intelligence, explainability, high-risk spheres.

Развитие искусственного интеллекта (ИИ) и его внедрение в высокорисковые сферы жизни общества (здравоохранение, транспорт, финансы, правосудие, государственное управление) требует не только технической надежности, но и прозрачности, которая обеспечивается объяснимостью решений ИИ. Однако этико-правовые аспекты объяснимости остаются предметом активных дискуссий, поскольку между необходимостью защиты прав человека и технологическими возможностями часто возникает противоречие[1].

К числу сфер применения искусственного интеллекта с высоким уровнем риска относятся области, в которых автоматизированные решения могут оказывать существенное влияние на права и свободы граждан, их здоровье, безопасность или благополучие. Наиболее уязвимыми в этом контексте считаются следующие сферы[2]: (1) здравоохранение (например, диагностика заболеваний,

[1] Jobin A., Ienca M., Vayena E. The global landscape of AI ethics guidelines // Nature Machine Intelligence. 2019. Vol. 1, No. 9. P. 396.
[2] Wei M., Zhou Z. AI Ethics Issues in Real World: Evidence from AI Incident // preprint. 2022. P. 5-7.

управление медицинскими данными)[1]; (2) финансовый сектор (например, кредитный скоринг, оценка страховых рисков); (3) государственное управление; (4) транспорт (например, автономные транспортные средства)[2]; (5) образование и трудоустройство (например, автоматизированные системы отбора персонала).

Ошибки, предвзятость или недостаточная объяснимость решений ИИ в указанных областях могут привести к необратимым последствиям: нарушению прав человека, усилению социальной несправедливости, потере доверия к институтам и даже угрозе жизни и здоровью отдельных лиц.

Объяснимость (explainability) искусственного интеллекта — это способность системы предоставлять ясные, доступные для восприятия и интерпретации обоснования своих решений, направленные как на конечного пользователя, так и на иные заинтересованные стороны, включая специалистов по контролю, разработчиков, аудиторов и лиц, подвергающихся воздействию этих решений. Объяснимость служит ключевым звеном в обеспечении доверия к ИИ, особенно в контексте высокорисковых применений[3].

К минимальным критериям объяснимости могут относиться следующие компоненты: (1) прозрачность — предоставление информации о принципах работы алгоритма, архитектуре модели, типе используемых данных и предположениях, лежащих в основе её функционирования. Прозрачность является базовым условием для понимания внутренней логики системы и последующего контроля[4]; (2) интерпретируемость — способность человека понять, каким образом и почему система пришла к конкретному решению; (3) аудируемость (auditability) — возможность внешней или внутренней проверки решений, при-

[1] Corfmat M., Martineau J.T., Régis C. High-reward, high-risk technologies? An ethical and legal account of AI development in healthcare // BMC Medical Ethics. 2025. Vol. 26. No. 4. p.19.
[2] Ivanova, L., Kalashnikov, N. The Liability Limits of Self-Driving Cars / // Legal Issues Journal. 2022. Vol. 9, No. 1. P.5-9
[3] Robbins M.D. AI Explainability: Regulations and Responsibilities [Электронный ресурс]. ResearchGate, 2019. P. 1–3. URL: https://www.researchgate.net/publication/336197243_AI_ Explainability_Regulations_and_Responsibilities (дата обращения: 14.04.2025).
[4] Lipton Z.C. The Mythos of Model Interpretability // Communications of the ACM. 2016. P. 96–98.

нимаемых системой, на предмет их корректности, беспристрастности и соответствия нормативным требованиям[1]; (4) трассируемость (traceability) — способность установить, какие исходные данные и этапы обработки повлияли на конечное решение[2].

Таким образом, объяснимость ИИ представляет собой многоаспектную характеристику, охватывающую технические, правовые и этические измерения. Её реализация требует учёта как особенностей модели, так и контекста применения. В сфере высокорисковых областей эти критерии должны быть реализованы таким образом, чтобы обеспечить баланс между эффективностью ИИ и защитой интересов государства, общества и человека.

Одной из ключевых задач объяснимости является выявление возможных предвзятостей и дискриминационных решений, что особенно важно для соблюдения прав человека[3]. Прозрачность алгоритмов позволяет выявлять и устранять такие эффекты, что поддерживается рядом международных стандартов (например, OECD AI Principles)[4].

Этико-правовые аспекты объяснимости искусственного интеллекта представляют собой комплекс требований, находящихся на пересечении моральных норм, правовых принципов и общественных ожиданий. Они обеспечивают баланс между технологическим развитием и защитой фундаментальных прав и свобод человека в условиях цифровизации.

Так, выделяется право на объяснение, которое предполагает, что лицо, на которое повлияло автоматизированное решение, должно иметь возможность получить понятную информацию о сути, логике и причинах принятого решения. С этической точки зрения, это право способствует признанию автономии личности

[1] Fernsel L., Kalff Y., Simbeck K. Assessing the Auditability of AI-integrating Systems: A Framework and Learning Analytics Case Study // HTW University of Applied Sciences, Berlin. 2024. 28 p.
[2] High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI [Электронный ресурс]. Brussels: European Commission, 2019. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (дата обращения: 14.04.2025).
[3] Barocas S., Selbst A.D. Big Data's Disparate Impact // California Law Review. 2016. Vol. 104. P. 671–675.
[4] Organisation for Economic Co-operation and Development (OECD). OECD Principles on Artificial Intelligence [Электронный ресурс]. Paris, 2019. URL: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 (дата обращения: 14.04.2025).

и её способности принимать осознанные решения, особенно в контексте цифрового взаимодействия. Наряду с этим, важным направлением является обеспечение алгоритмической справедливости и недопущение дискриминации[1].

Вопрос ответственности также приобретает особое значение. В условиях использования автономных систем важно определить, кто именно несёт правовую и моральную ответственность за их действия — разработчик, оператор или пользователь. ИИ-системы должны быть не только технологически надёжными, но и прозрачными, чтобы можно было восстановить логику принятого решения и оценить его допустимость. Показательным примером формализации требований к ответственности за использование искусственного интеллекта выступает законодательство Европейского союза. Регламент ЕС об искусственном интеллекте (AI Act)[2] вводит классификацию ИИ-систем по уровням риска и устанавливает требования к высокорисковым технологиям, включая обязательную документацию, аудит, обеспечение объяснимости и соблюдение технических и этических стандартов. Особое внимание в Регламенте ЕС уделено распределению ответственности между участниками рынка. В зависимости от характера нарушения меры могут применяться как к пользователям, так и к поставщикам — физическим или юридическим лицам, разрабатывающим, продающим или вводящим в эксплуатацию ИИ-системы на территории ЕС[3].

Таким образом, объяснимость решений, принимаемых системами искусственного интеллекта, представляет собой не только техническое требование, но и важнейший этико-правовой инструмент обеспечения прозрачности, справедливости и защиты прав личности. Совокупная реализация критериев объяснимости позволяет формировать доверие к ИИ-системам, снижать риски дискриминации

[1] Талапина Э. В. Обработка данных при помощи искусственного интеллекта и риски дискриминации // Право. Журнал Высшей школы экономики. 2022. Т. 15. № 1. С. 4-27.
[2] Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts. Электронный ресурс. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj (дата обращения: 13.04.2025).
[3] Хмелевской, И.Р. Правовое регулирование юридической ответственности за действия искусственного интеллекта: опыт Европейского союза // Право в эпоху искусственного интеллекта: перспективные вызовы и современные задачи: сборник научных статей по материалам Международного научно-практического форума, Тюмень, 17–19 октября 2024 года. Тюмень: ТюмГУ-Press, 2024. С. 291-295.

и обеспечивать условия для правовой подотчётности. В заключение следует подчеркнуть, что в рамках статьи предложены подходы к формированию минимального набора требований к объяснимости и их реализация возможна только при условии комплексного, междисциплинарного подхода.

Список литературы

1. Barocas S., Selbst A.D. Big Data's Disparate Impact // California Law Review. – 2016. – Vol. 104. – P. 671–732.

2. Corfmat M., Martineau J.T., Régis C. High-reward, high-risk technologies? An ethical and legal account of AI development in healthcare // BMC Medical Ethics. — 2025. — Vol. 26. — Article No. 4. — P. 19.

3. Fernsel L., Kalff Y., Simbeck K. Assessing the Auditability of AI-integrating Systems: A Framework and Learning Analytics Case Study // HTW University of Applied Sciences. – Berlin. – 2024. – 28 p.

4. Ivanova L., Kalashnikov N. The Liability Limits of Self-Driving Cars // Legal Issues Journal. – 2022. – Vol. 9, No. 1. – P. 1–14.

5. Jobin A., Ienca M., Vayena E. The global landscape of AI ethics guidelines // Nature Machine Intelligence. — 2019. — Vol. 1, No. 9. — P. 389–399.

6. Lipton Z.C. The Mythos of Model Interpretability // Communications of the ACM. – 2016. – P. 96–100.

7. Robbins M.D. AI Explainability: Regulations and Responsibilities // ResearchGate. – 2019. – 15 p.

8. Wei M., Zhou Z. AI Ethics Issues in Real World: Evidence from AI Incident Database // Preprint. – 2022. – 11 p.

9. Талапина Э. В. Обработка данных при помощи искусственного интеллекта и риски дискриминации // Право. Журнал Высшей школы экономики. – 2022. – Т. 15. – № 1. – С. 4–27.

10. Хмелевской И.Р. Правовое регулирование юридической ответственности за действия искусственного интеллекта: опыт Европейского союза // Право в эпоху искусственного интеллекта: перспективные вызовы и современные задачи: сб. науч. ст. по материалам Междунар. науч.-практ. форума. Тюмень: ТюмГУ-Press, 2024. – С. 291–295.