УДК: 341.1/8

***Никитенко С.В.***

*Аспирант*

*Кафедра международного права*

*Санкт-Петербургский государственный университет*

**ОБЪЯСНИМОСТЬ КАК ОСНОВНОЕ ЮРИДИЧЕСКОЕ СВОЙСТВО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА**

**Аннотация:** Активное развитие ИИ-технологий позволяет автоматизировать процессы во многих сферах деятельности, повышая производительность и общее качество жизни. Однако, высокая автономность интеллектуальных алгоритмов, обеспечиваемая высокой сложностью архитектуры и производимых вычислений, неизбежно приводит к возникновению проблемы объяснимости действий и решений такого ИИ (проблема «чёрного ящика»). Низкая интерпретируемость интеллектуальной системы существенно усложняет применение традиционных правовых институтов и механизмов, обеспечивающих права человека, среди которых: деликтная ответственность, получение согласия на обработку персональных данных и медицинское вмешательство. Анализ применимого международного регулирования и национального законодательства позволяет рассматривать объяснимость как одно из ключевых юридических свойств систем искусственного интеллекта.

**Ключевые слова:** искусственный интеллект, интеллектуальная система, алгоритм, объяснимость, чёрный ящик, деликт

**EXPLAINABILITY AS A FUNDAMENTAL LEGAL FEATURE OF ARTIFICIAL INTELLIGENCE**

**Abstract:** The rapid development of AI technologies leads to the automation of processes in many fields of life, increasing productivity and overall quality of life. However, the high autonomy of intelligent algorithms, ensured by the high complexity of the architecture and the calculations produced, inevitably leads to the problem of explainability of acts and decisions of such AI (the "black box" problem). The low interpretability of the intellectual system significantly challenges the application of traditional legal instruments and procedures ensuring human rights, among them: tort liability, consent to personal data processing and medical interventions. The analysis of applicable international regulation and national legislation suggests explainability as one of the key legal features of artificial intelligence systems.

**Keywords:** artificial intelligence, intelligent system, algorithm, explainability, black box, tort

Искусственный интеллект (далее – ИИ), выступая во главе четвёртой промышленной революции [5] и являясь сквозной технологией, лежащей в основе цифровой экономики [20], способен позитивно повлиять практически на все сферы жизни общества, особенно – экономику за счёт высокой производительности и возможности автоматизации рутинных процессов. Однако, распространение ИИ также порождает трудноразрешимые проблемы социально-экономического, этического и юридического характера.

Обращаясь к существующему регулированию ИИ-технологии, отметим отсутствие какой-либо целостной нормативно-правовой базы, закрепляющей правовой режим искусственного интеллекта. При этом, до настоящего момента отсутствует единый подход по ряду принципиальных вопросов, которые касаются дефиниции ИИ, правового режима результатов его деятельности, а также ответственности за вред, причинённый действиями интеллектуальных систем.

Одним из основных свойств ИИ-алгоритмов, которые затрудняют решение обозначенных проблем, является его низкая интерпретируемость и необъяснимость принимаемых системой решений, вызванная сложностью архитектуры и принципов действия искусственного интеллекта. Как отмечается некоторыми исследователями, иногда даже сами разработчики затрудняются понять, каким образом системой был получен результат [15]. Оператор ИИ-алгоритма способен лишь загружать входные данные и задать ограниченные параметры – остальное заранее запрограммировано разработчиками [3]. В научной литературе данный вопрос получил название проблемы «черного ящика» [14], а также нашёл отражение на федеральном уровне – в Концепции развития регулирования отношений в сфере технологий искусственного интеллекта и робототехники до 2024 год, утв. Распоряжением Правительства РФ от 19.08.2020 № 2129-р.

Осложняющим фактором является наличие обратной зависимости между объяснимостью ИИ-системы и её эффективностью, что приводит нас вопросу о надлежащей оценке ценностей, особенно, если на кону стоит здоровье и жизнь человека. Так, предварительное изготовление разработчиками подробной «карты решений», к которым может прийти алгоритм, не применимо к ИИ-технологиям, которые способны самостоятельно, бесконтрольно обучаться и выстраивать собственные логические цепочки и выявлять закономерности [15]. Представленный способ решения проблемы интерпретируемости наиболее уместен для т.н. экспертных систем, которые являлись технологическими предшественниками искусственного интеллекта и функционировали исключительно в рамках чётко очерченного набора «сценариев» с заранее определёнными причинно-следственными цепочками (метод «древа решений»): примерами таких систем являются программы «DXplain» [10] и «MYCIN» [9].

Вместе с тем, системы, разработанные на основании указанного метода принятия решений, отличаются низкой точностью полученных результатов в сравнении с методами машинного обучения или глубокими нейронными сетями [19]. В связи с чем, нормативно закреплённое требование о необходимости разработки ИИ-систем исключительно на архитектуре «древа решений» (примат объяснимости над остальными факторами) также не является способом решения проблемы, поскольку это существенно снизит эффективность применения технических решений и может в целом привести к технологической рецессии.

Прямым следствием указанной проблемы является непредсказуемость действий ИИ, вследствие чего существенно затрудняется контроль качества работы систем и создаётся риск принятия решений, нарушающих права человека. Кроме того, возникают трудности в вопросе привлечения к ответственности за причинённый ИИ-алгоритмом вред (сложности в установлении причинно-следственной связи), а также при получении информированного согласия на обработку персональных данных или согласия на медицинское вмешательство.

Необъяснимость ИИ также снижает уровень доверия к технологии и возможностям её применения [2] (к примеру, врач способен объяснить свои действия и решения, а также несёт за них юридическую, моральную и профессиональную ответственность [13]), что способно привести к стагнации разработок в ИИ-сфере [24]. Наконец, отмечается такое следствие проблемы «чёрного ящика», как предвзятость результатов, которая вытекает из возможности алгоритмов на основе анализа больших данных находить невидимые человеку взаимосвязи [1]. Некоторые авторы отмечают, что ИИ может быть необъективен в связи с установлением причинно-следственной взаимосвязи между двумя абсолютно не связанными между собой факторами, основываясь исключительно на статистическом анализе явлений [13].

В совокупности, данные обстоятельства свидетельствуют об особом «статусе» объяснимости среди остальных свойств искусственного интеллекта, среди которых: автономность, обучаемость и др. [11]. В связи с чем, в настоящее время всё больший интерес вызывает такое направление исследований, как разработка моделей т.н. «объяснительного интеллекта» [19].

Отметим также, что практически в каждом нормативно-правовом акте национального или международного уровня, посвящённом ИИ-проблематике, мы можем обнаружить положения, посвящённые обеспечению объяснимости ИИ, в частности, через реализацию соответствующего принципа прозрачности и/или объяснимости.

Так, в Рекомендации ЮНЭСКО об этических аспектах ИИ [23] обеспечение прозрачности рассматривается в качестве неотъемлемого условия эффективной реализации механизма ответственности (п. 37). Отмечается, что принцип прозрачности должен быть реализован вместе с принципом объяснимости, который направлен на обеспечение понятности и представление разъяснений в отношении полученных с помощью ИИ-систем результатов, включая доступность для понимания поведения алгоритма, в том числе, исходных данных и их влияния на результат (п. 40). Аналогичным образом принципы прозрачности и объяснимости закреплены в п. 1.3 и пп. «b» п. 1.4 Рекомендации Совета ОЭСР по искусственному интеллекту, принятые 22.05.2019 г. [8] Таким образом, необходимость имплементации данных принципов в регулирование ИИ-отношений закреплена на международном уровне, что свидетельствует об их ключевом для отрасли значении.

Кроме того, в Регламенте Европейского союза «Об унифицированных правилах регулирования искусственного интеллекта и внесении изменений в некоторые законодательные акты Союза» [7], направленного на установление общего правового режима ИИ-технологий, мы можем обнаружить положения о стандартизации требований по обеспечению прозрачности и доступности сведений об ИИ-системе.

Объяснимость и корреспондирующий ей принцип прозрачности имеют ключевое значение для обеспечения конфиденциальности персональных данных. Непосредственным образом закрепляет принцип прозрачности только Регламент № 2016/679 Европейского союза «О защите физических лиц при обработке персональных данных <…> (Общий Регламент о защите персональных данных)» [6], однако, практически всеми региональными и внутригосударственными актами предусмотрен механизм получения информированного согласия на обработку персональных данных, который является «сердцевиной» обозначенного принципа, в частности, Конвенцией Совета Европы о защите частных лиц в отношении автоматизированной обработки данных личного характера 1981 г., ETS № 108 [17] (ратифицирована Россией 15.05.2013) и ст. 9 Федерального закона РФ «О персональных данных» от 27.07.2006 № 152-ФЗ [21].

Наиболее подробно принцип прозрачности раскрыт в упомянутом Регламенте ЕС № 2016/679. Суть принципа состоит в предоставлении субъекту данных как перед обработкой, так в процессе, причём, перечень таких сведений заранее определён. Помимо стандартной информации о целях обработки, категории обрабатываемых данных и т.д., субъекту данных также должны быть предоставлены сведения о наличии автоматизированного процесса принятия решений, включая достоверные сведения о соответствующей логической схеме алгоритма (пп. «f» п. 2 ст. 13 Регламента). Как отмечается европейским регулятором, такие сведения должны в отчетливо видимой, понятной и разборчивой форме дать субъекту общее представление о предполагаемой обработке.

В части охраны здоровья, механизм получения информированного согласия пациента на медицинское вмешательство с использованием ИИ-технологий позволяет принять взвешенное и осознанное решение: пациент должен понимать, каким образом будет осуществлено вмешательство, поскольку от этого зависит его здоровье и жизнь [15]. Необходимость получения информированного согласия является общепризнанным условием для осуществления медицинского вмешательства (см., к примеру, ст. 20 Федеральный закон от 21.11.2011 № 323-ФЗ «Об основах охраны здоровья граждан в Российской Федерации» [22], ст. 5 Конвенции Совета Европы о защите прав и достоинства человека в связи с применением достижений биологии и медицины 1997 г. [16] и ст. 6 Всеобщей декларации ЮНЕСКО о биоэтике и правах человека 2005 г. [12]).

Вместе с тем, указанные акты не раскрывают механизм получения такого информированного согласия, в связи с чем, остаются нерешёнными проблемы, вызванные низкой интерпретируемостью интеллектуальных систем и следующей из этого сложностью в предоставлении адекватной, полной информации в доступной форме о предстоящем вмешательстве. Вновь обозначим, что решение данной проблемы выходит за рамки работы, и ключевое значение в контексте рассматриваемой проблематики имеет тот факт, что получение любой информации об алгоритме для получения согласия на вмешательство непосредственным образом взаимосвязано с интерпретируемостью соответствующего ИИ.

Анализ существующих доктринальных позиций и относимого регулирования (как национального, так и международного) позволяет нам рассматривать объяснимость в качестве основного элемента правового режима интеллектуальных технологий, а корреспондирующий принцип прозрачности – одним из ключевых принципов ИИ-отрасли в целом.

Вместе с тем, законодателю и международному регулятору ещё предстоит разработать комплекс мер по разрешению проблемы «черного ящика», направленных на обеспечение доступности информационных процессов, обрабатываемых ИИ-системами, а также определения формы и объёма предоставления сведений заинтересованным лицам, что в свете обозначенных выводов должно иметь первостепенное значение.

**Список использованной литературы**

1. Antoniades Ch., Asselbergs F.W., Vardas P. Год в сердечно-сосудистой медицине, 2020 г.: цифровое здравоохранение и инновации // Российский кардиологический журнал. – 2021. № 26 (3). С. 114-124.
2. Char D.S., Shah N.H., Magnus D. Implementing Machine Learning in Health Care – Addressing Ethical Challenges // The New England Journal of Medicine. – 2018. – № 378(11). С. 981-983. – DOI: 10.1056/NEJMp1714229.
3. Ching T., Himmelstein D.S., Beaulieu-Jones B.K., Kalinin A.A., Do B.T., Way G.P., et al. Opportunities and obstacles for deep learning in biology and medicine [Электронный ресурс] // Journal of the Royal Society Interface. – 2018. – № 15(141). – 47 с. – DOI: 10.1098/rsif.2017.0387. – URL: <https://royalsocietypublishing.org/doi/pdf/10.1098/rsif.2017.0387>.
4. Kamensky S. Artificial Intelligence and Technology in Health Care: Overview and Possible Legal Implications // DePaul journal of health care law. – 2020. – Т. 21. – № 3. – С. 1-18.
5. Kemp R. Legal Aspects of Artificial Intelligence (v. 2.0) [Электронный ресурс] // Kemp IT Law. – 2016. – URL: <https://www.kempitlaw.com/wp-content/uploads/2016/11/Legal-Aspects-of-AI-Kemp-IT-Law-v2.0-Nov-2016-.pdf>.
6. On the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation): Регламент (ЕС) Европейского Парламента и Совета Европейского союза № 2016/679 от 27.04.2016 [Электронный ресурс] // Официальный сайт законодательства Европейского союза и других официальных документов Европейского союза «EUR-Lex». – Режим доступа: <https://eur-lex.europa.eu/>.
7. Proposal for a 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: Регламент (ЕС) Европейского Парламента и Совета Европейского союза № COM/2021/206 от 21.04.2021 [Электронный ресурс] // Официальный сайт законодательства Европейского союза и других официальных документов Европейского союза «EUR-Lex». – Режим доступа: <https://eur-lex.europa.eu/>.
8. Recommendation on Artificial Intelligence № OECD/LEGAL/0449: принята Советом Организации экономического сотрудничества и развития (ОЭСР) 22.05.2019 [Электронный ресурс] // LEGALINSTRUMENTS.OECD.ORG – официальный сайт ОЭСР. – URL: <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>.
9. Shortliffe E.H. et al. Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system // Computers and biomedical research. – 1975. – Т. 8. – №. 4. – С. 303-320.
10. Using decision support to help explain clinical manifestations of disease // MGHLCS.ORG – официальный сайт Лаборатории компьютерных наук Массачусетской больницы общего профиля. – URL: <http://www.mghlcs.org/projects/dxplain>.
11. Васильев А.А., Шпоппер Д., Матаева М.Х. Термин «искусственный интеллект» в российском праве: доктринальный анализ // Юрислингвистика. – 2018. – № 7-8. – С. 35-44.
12. Всеобщая декларация о биоэтике и правах человека: принята резолюцией Генеральной конференции ЮНЕСКО по докладу Комиссии III на 18-м пленарном заседании 19.10.2005 [Электронный ресурс] // Официальный сайт ООН. – URL: <https://www.un.org/ru/documents/decl_conv/declarations/bioethics_and_hr.shtml>.
13. Иванова А.П. Правовые проблемы использования искусственного интеллекта в сфере здравоохранения // Социальные и гуманитарные науки. Отечественная и зарубежная литература. Серия 4: государство и право. Реферативный журнал. – 2021. – № 1. – С. 151-159.
14. Ившин А.А., Гусев А.В., Новицкий Р.Э. Искусственный интеллект: предиктивная аналитика перинатального риска // Вопросы гинекологии, акушерства и перинатологии. – 2020. – № 19 (6). – С. 133-144.
15. Киселева А.Ю. Применение искусственного интеллекта в здравоохранении: аспекты медицинского права // Медицинское право: теория и практика. – 2020. – Т. 6. – № 2 (12). – С. 29-38.
16. Конвенция о защите прав человека и человеческого достоинства в связи с применением достижений биологии и медицины («Конвенция о правах человека и биомедицине»): Совет Европы, ETS № 164, заключена в г. Овьедо 04.04.1997 [Электронный ресурс] // Справочная правовая система «Консультант-Плюс». – Режим доступа: <http://www.consultant.ru/>.
17. Конвенция о защите частных лиц в отношении автоматизированной обработки данных личного характера: Конвенция Совета Европы, ETS № 108, открыта для ратификации 28.01.1981, вступила в силу 01.10.1985 [Электронный ресурс] // Официальный сайт Совета Европы. – URL: <https://www.coe.int/en/web/conventions/full-list?module=treaty-detail&treatynum=108>.
18. Малышкин А.В. Интегрирование искусственного интеллекта в общественную жизнь: некоторые этические и правовые проблемы // Вестник СПбГУ. Право. – 2019. – Т. 10. – № 3. – С. 444-460.
19. Мелдо А.А., Уткин Л.В., Трофимова Т.Н., 2020 г. Искусственный интеллект в медицине: современное состояние и основные направления развития интеллектуальной диагностики // Лучевая диагностика и терапия. – 2020. – Т. 11. – № 1. – С. 9-17.
20. Национальная стратегия развития искусственного интеллекта на период до 2030 года: Указ Президента РФ от 10.10.2019 № 490 [Электронный ресурс] // Справочная правовая система «Консультант-Плюс». – Режим доступа: <http://www.consultant.ru/>.
21. О персональных данных: Федеральный закон РФ от 27.07.2006 № 152-ФЗ, ред. от 02.07.2021 [Электронный ресурс] // Справочно-правовая система «КонсультантПлюс». – Режим доступа: <http://www.consultant.ru/>.
22. Об основах охраны здоровья граждан в Российской Федерации: Федеральный закон от 21.11.2011 № 323-ФЗ, ред. от 02.07.2021 [Электронный ресурс] // Справочная правовая система «Консультант-Плюс». – Режим доступа: <http://www.consultant.ru/>.
23. Рекомендация об этических аспектах искусственного интеллекта: принята ЮНЭСКО в ходе 41-й сессии, проходившей в Париже с 9 по 24 ноября 2021 г. [Электронный ресурс] // UNESDOC.UNESCO.ORG – Цифровая библиотека ЮНЭСКО. – URL: [https://unesdoc.unesco.org/ark:/48223/pf0000380455\_rus](https://unesdoc.unesco.org/ark%3A/48223/pf0000380455_rus).
24. Черных Е.Е. Цифровая медицина: риски правореализации инноваций в сфере здравоохранения // Юридическая наука и практика: Вестник нижегородской академии МВД России. – 2020. – № 4(52). – С. 84-94.