Научная статья
УДК 004.8

# ПРАВОВОЕ РЕГУЛИРОВАНИЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА: ОТ «ЧЕРНОГО ЯЩИКА» К ПРОЗРАЧНОСТИ АЛГОРИТМОВ

**Людмила Александровна Киселева[1], Александр Сергеевич Шайдуров[2]**
[1, 2] Уральский государственный лесотехнический университет, Екатеринбург, Россия
[1] kiselevala@m.usfeu.ru
[2] shaydurov-01@mail.ru

***Аннотация.*** Авторы исследуют проблему «черного ящика» в искусственном интеллекте с точки зрения российского права. Рассмотрены основные подходы к определению вины в российском законодательстве, проанализированы сложности их применения к искусственному интеллекту, а также предложены возможные пути их преодоления.

***Ключевые слова:*** искусственный интеллект, российское право, вина, ответственность, разработчик, юриспруденция, вред, «черный ящик»

***Для цитирования:*** Киселева Л. А., Шайдуров А. С. Правовое регулирование искусственного интеллекта: от «черного ящика» к прозрачности алгоритмов // Цивилизационные перемены в России. 2024. С. 89–96.

Original article

# LEGAL REGULATION OF ARTIFICIAL INTELLIGENCE: FROM THE “BLACK BOX” TO THE TRANSPARENCY OF ALGORITHMS

**Lyudmila A. Kiseleva[1], Alexander S. Shaidurov[2]**
[1, 2] Ural State Forest Engineering University, Yekaterinburg, Russia
[1] kiselevala@m.usfeu.ru
[2] shaydurov-01@mail.ru

***Abstract.*** The authors explore the problem of the “black box” in artificial intelligence from the point of view of issues of Russian law. The main approaches to determining guilt in Russian legislation are considered, the complexity of their application to artificial intelligence is analyzed, and possible solutions are proposed.

***Keywords:*** artificial intelligence, Russian law, guilt, responsibility, developer, jurisprudence, harm, “black box”

***For citation:*** Kiseleva L. A., Shaidurov A. S. Legal regulation of artificial intelligence: from the “black box” to the transparency of algorithms // Civilizational changes in Russia. 2024. P. 89–96.

В XXI в. искусственный интеллект (далее – ИИ) стремительно интегрируется во все сферы жизнедеятельности, в том числе в сферу правовых отношений. Законодательство РФ определяет искусственный интеллект как комплекс технологических решений, позволяющий имитировать когнитивные функции человека (включая самообучение и поиск решений) и получать при выполнении конкретных задач результаты, сопоставимые, как минимум, с результатами интеллектуальной деятельности человека [1].

Анализ большого объема данных, прогнозирование рисков и автоматизация рутинных задач – это лишь некоторые примеры применения ИИ, которые уже сегодня меняют юридическую практику. Стремительное развитие искусственного интеллекта порождает новые вызовы, в том числе правового характера.

Одной из проблем, стоящих перед современной юриспруденцией, является феномен «черного ящика» в отношении ИИ. Данная проблема отражена в работах многих исследователей, например Касперсена и Майкла [2], и связана с трудностями в установлении ответственности за последствия применения искусственного интеллекта.

В контексте юриспруденции «черный ящик» представляет собой алгоритм, используемый для анализа правовой информации или принятия решений, принципы работы которого не могут быть полностью проверены человеком. Как отмечают Паскуччи и Стайано, «черные ящики» в ИИ создают проблемы прозрачности и подотчетности, что особенно важно в юридическом контексте [3].

Можно увидеть результат работы алгоритма, но нельзя в полной мере проследить логику, которая привела к такому результату. Эта непрозрачность порождает серьезные вопросы о справедливости и объективности применения ИИ. Как в таких случаях определить виновных? Как доказать неправомерность действий ИИ, если мы не можем проследить логику его решений?

### *Проблема «черного ящика» в контексте права*

Российское право основано на принципе вины. Этот принцип предполагает, что ответственность наступает только за виновно совершенное общественно опасное деяние. Значит, для привлечения к ответственности за причиненный вред, – будь то вред имуществу (как материальному объекту – статья 1064 ГК РФ) или нематериальным благам (репутации, чести,

достоинству – статья 150 ГК РФ) [5], – необходимо установить наличие виновных действий или бездействия, а также причинно-следственную связь между ними и наступившим вредом.

В контексте использования искусственного интеллекта возникает ряд сложностей в связи с применением этого принципа.

Во-первых, непрозрачность алгоритмов ИИ затрудняет определение того, какие именно действия или бездействие привели к наступлению вреда. Если мы не можем проследить логику принятия решений ИИ, то и доказать наличие виновного поведения становится практически невозможно.

Во-вторых, возникает вопрос, можно ли вообще говорить о вине, когда речь идет об искусственном интеллекте. Ведь искусственный интеллект, в отличие от человека, не обладает сознанием и свободой воли и действует в соответствии с заданным алгоритмом. Так что пока не ясно, как на законодательном уровне давать правовую оценку действиям искусственного интеллекта: использовать уже существующие правовые нормы, регулирующие правоотношения людей, или разрабатывать новые, применительно к действиям ИИ.

В-третьих, «черный ящик» ИИ порождает сложный правовой вопрос о распределении ответственности за причиненный вред. Непонятно, кто должен нести ответственность: разработчик, оператор или сам ИИ как потенциальный «электронный субъект права».

В настоящее время в российском законодательстве нет однозначных ответов на эти вопросы, что делает проблему «черного ящика» еще более актуальной.

### *Зарубежный опыт решения проблемы «черного ящика» в ИИ*

Проблема «черного ящика» в искусственном интеллекте признается во всем мире, и многие страны уже предприняли шаги для ее решения. Рассмотрим опыт некоторых из них.

Евросоюз является одним из лидеров в области регулирования искусственного интеллекта. В 2021 г. Европейская комиссия представила проект Регламента об ИИ (Artificial Intelligence Act), который направлен на обеспечение надежности и безопасности систем ИИ [6].

Данный Регламент выдвигает следующие требования к искусственному интеллекту:

1. *Требования к прозрачности системы*. Разработчики искусственного интеллекта обязаны создать условия, при которых их системы прозрачны и понятны для человека, чтобы компетентные органы могли оценивать их соответствие требованиям Регламента.

2. *Обязательства по ведению документации*. Разработчики должны предоставлять подробную информацию о системе ИИ, включая ее целевое назначение, принципы работы, используемые данные и алгоритмы.

3. *Требования к контролю со стороны человека*. Для определенных категорий систем ИИ высокого риска предусмотрено обязательное участие человека в процессе принятия решений.

Хотя на федеральном уровне в США пока нет единого закона, регулирующего искусственный интеллект, тем не менее существуют законодательные инициативы на уровне отдельных штатов. Например, в 2019 г. штат Калифорния принял закон (California Bot Law) [7], который запрещает компаниям использовать боты для взаимодействия с потребителями без их согласия. Этот закон можно рассматривать как шаг в сторону повышения прозрачности в использовании ИИ.

Китай является одним из мировых лидеров в области развития и внедрения ИИ. В этой стране были приняты «Рекомендации по этике искусственного интеллекта» [8], которые призывают к разработке «надежного, контролируемого и упорядоченного» ИИ. В данном документе отмечается важность прозрачности и объяснимости алгоритмов ИИ, а также подчеркивается необходимость предотвращения дискриминации и нарушений прав и свобод человека и гражданина при использовании искусственного интеллекта.

Анализ зарубежного опыта показывает, что различные страны и организации предпринимают шаги по разработке правовых норм и стандартов, направленных на повышение прозрачности и подотчетности ИИ. Этот опыт может быть полезен и для России при формировании собственной нормативной базы в этой области.

## *Предложения экспертов в области искусственного интеллекта и права по решению проблемы «черного ящика» в ИИ*

Помимо инициатив на государственном уровне, существует ряд предложений от ведущих экспертов в области ИИ и права, которые могут способствовать решению проблемы «черного ящика».

Например, американский математик и специалист по анализу данных Кэти О'Нил, автор книги «Оружие математического поражения» [9], активно выступает за прозрачность алгоритмов, влияющих на жизнь людей. Он предлагает внедрить систему «алгоритмического аудита», которая обязывала бы организации, использующие ИИ в таких важных сферах, как образование, здравоохранение, трудоустройство, регулярно проводить независимую проверку своих алгоритмов на предмет предвзятости, дискриминации и других рисков.

Другой ведущий специалист в области ИИ, профессор калифорнийского университета Беркли Стюарт Рассел, предлагает сосредоточиться на создании «доказуемо полезного ИИ» [10].

Рассел считает, что разработчики ИИ должны уделять внимание созданию систем, цели которых полностью согласованы с целями человече-

ства. Такой подход, по его мнению, позволит снизить риски, связанные с неконтролируемым развитием ИИ, в том числе и с проблемой «черного ящика».

Таким образом, проблема «черного ящика» в ИИ представляет собой серьезный вызов, требующий комплексного подхода к его решению.

На данный момент в российском законодательстве отсутствуют специальные нормы, регулирующие отношения в сфере искусственного интеллекта.

Тем не менее можно выделить несколько *перспективных направлений,* которые могли бы способствовать решению данной проблемы:

1. Разработка интерпретируемых алгоритмов ИИ.

Одним из путей решения проблемы «черного ящика» является разработка алгоритмов ИИ, логика работы которых была бы более прозрачна и понятна человеку. Это позволило бы отследить цепочку принятия решений ИИ и определить причины, по которым он совершил то или иное действие.

В данном направлении уже ведутся активные исследования. Разрабатываются различные методы интерпретации моделей машинного обучения: от визуализации важности признаков до создания промежуточных моделей, имитирующих работу сложных алгоритмов в более понятной форме. Например, методы LIME [11] и SHAP [12] позволяют оценить вклад каждого входного параметра в конкретное решение модели.

Однако на сегодняшний день проблема интерпретируемости ИИ еще далека от своего окончательного решения. Сложные нейронные сети, демонстрирующие высокую эффективность во многих областях, зачастую остаются «черными ящиками» даже для своих создателей.

2. Создание «черных ящиков» с функцией аудита.

По аналогии с «черными ящиками» на самолетах, в сфере ИИ можно внедрить обязательную регистрацию всех ключевых действий и решений, принимаемых системой. Такие «лог-файлы» должны содержать достаточно информации, чтобы в случае инцидента специалисты смогли проанализировать работу ИИ и установить причины его некорректной работы.

Данный подход может быть особенно актуален для систем ИИ, используемых в критически важных областях, где ошибка может привести к серьезным последствиям. Однако необходимо тщательно продумать вопросы, связанные с хранением и защитой таких данных, чтобы не допустить нарушение конфиденциальности и других прав граждан.

3. Разработка новых правовых норм и стандартов для регулирования искусственного интеллекта.

В условиях стремительного развития ИИ необходима активная работа по усовершенствованию российского законодательства в этой сфере. Важно разработать четкие правовые нормы, регулирующие разработку, внедрение и использование ИИ, с учетом проблемы «черного ящика».

### *Пути решения проблемы «черного ящика» в ИИ на законодательном уровне*

Основываясь на анализе зарубежного опыта и мнении экспертов в области ИИ, мы предлагаем следующие пути решения:

1. Закрепить на законодательном уровне понятие «искусственный интеллект» и установить четкие критерии, позволяющие отнести системы к этой категории. Отсутствие единого общепризнанного определения ИИ не должно стать препятствием для такой законодательной инициативы. В качестве отправной точки можно использовать наработки ведущих ученых в этой сфере.

2. Обеспечить прозрачность и объяснимость алгоритмов ИИ, особенно в сферах, непосредственно влияющих на права и свободы граждан: здравоохранение, образование, отправление правосудия, финансовые услуги.

3. Разработать механизмы оценки и управления рисками, связанными с использованием ИИ, в том числе с проблемой «черного ящика».

4. Установить правовые процедуры, на основании которых будет распределяться ответственность между разработчиками, операторами и другими участниками этих отношений за вред, причиненный искусственным интеллектом.

5. В сферах, связанных с повышенным риском (например, медицина, транспорт, безопасность) целесообразно применять к разработчикам и операторам ИИ принцип «строгой ответственности». В этом случае они будут нести ответственность за причиненный ИИ вред, независимо от наличия вины с их стороны, если не докажут, что приняли все необходимые меры для предотвращения вреда.

При этом применение этого принципа должно быть взвешенным и обоснованным, чтобы не тормозить развитие инновационных технологий в сфере ИИ.

Подводя итоги, можно сказать, что проблема «черного ящика» в ИИ – это серьезный вызов, стоящий перед действующим законодательством: непрозрачность алгоритмов ИИ, особенно в условиях их стремительного развития и внедрения в различные сферы жизни, создает серьезные трудности с точки зрения определения вины, распределения ответственности и обеспечения справедливости.

Нами были предложены возможные пути решения данной проблемы, основанные на комплексном подходе с учетом зарубежного опыта и мнений экспертов. К ключевым направлениям можно отнести разработку более интерпретируемых алгоритмов ИИ, создание «черных ящиков» с функцией аудита, разработку новых правовых норм и стандартов для ИИ, а также адаптацию существующих правовых норм к реалиям развития ИИ.

Считаем, что решение проблемы «черного ящика» требует тесного взаимодействия специалистов в области права, социологии, этики и информационных технологий.

Создание нормативно-правовой базы, регулирующей ИИ, будет гарантировать ответственное и безопасное использование искусственного интеллекта в разных сферах жизнедеятельности.

*Список источников*

1. О развитии искусственного интеллекта в Российской Федерации : указ Президента РФ от 10 октября 2019 г. № 490. URL: https:// base.garant.ru/72838946/ (дата обращения: 18.10.2024).

2. Caspersen M. E., Michael, K. Artificial intelligence and the law: An ethical perspective // Ethics and Information Technology. 2017. № 19 (4). P. 265–278.

3. Pascucci P., Staiano J. Black-Box Society: Why We Should Be Concerned About the Rise of Artificial Intelligence in Legal Decision-Making // Law, Innovation and Technology. 2021. № 13 (1). P. 1–29.

4. Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks / J. Angwin, J. Larson, S. Mattu, L. Kirchner // ProPublica. 2016. URL: https://www.propublica.org/article/ machine-bias-risk-assessments-in-criminal-sentencing (дата обращения: 18.10.2024).

5. Гражданский кодекс Российской Федерации от 30.11.1994 № 51-ФЗ (ред. от 28.05.2022). URL: http://www.consultant.ru/document/cons_doc_ LAW_5142/ (дата обращения: 18.10.2024).

6. 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. // European Commission. 2021. URL: https://op.europa.eu/en/publication-detail//publication/ 604fd72c-e6dd-11ee-9ea8-01aa75ed71a1/language-en (дата обращения: 18.10.2024).

7. AB-375 Privacy: personal information: businesses // California Legislative Information. URL: https://leginfo.legislature.ca.gov/faces/billText Client.xhtml?bill_id=201720180AB375 (дата обращения: 18.10.2024).

8. O'Neil C. Weapons of math destruction: How big data increases inequality and threatens democracy. New York : Crown Publishers, 2016. 272 p.

9. Russell S. Human compatible: Artificial intelligence and the problem of control. Penguin, 2019. 348 p.

10. Ribeiro M. T., Singh, S., Guestrin C. "Why should i trust you?": Explaining the predictions of any classifier // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2016. P. 1135–1144.

11. Lundberg S. M., Lee S. I. A unified approach to interpreting model predictions // Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017. P. 4765–4774.

*References*

1. “On the development of artificial intelligence in the Russian Federation” : Decree of the President of the Russian Federation of October 10, 2019. № 490. URL: https://base.garant.ru/72838946 (accessed: 18.10.2024).

2. Caspersen M. E., Michael K. Artificial intelligence and the law: an ethical perspective // Ethics and information technology. 2017. 19 (4). P. 265–278.

3. Pascucci P., Staiano J. The Black Box Society: Why We should Be Concerned about the Increasing use of Artificial Intelligence in the Legal Decision-making Process // Law, InnovationandTechnology. 2021. № 13 (1). P. 1–29.

4. Machine bias: There’s software used across the country to predict future criminals. And it’s biased against blacks / J. Angwin, J. Larson, S. Mattu, L. Kirchner // ProPublica. 2016. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (дата обращения: 18.10.2024).

5. The Civil Code of the Russian Federation № 51 dated 11.30.1994 (as amended on 05.28.2022). URL: http://www.consultant.ru/document/cons_doc_LAW_5142/ (дата обращения: 18.10.2024).

6. 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. // European Commission. 2021. URL: https://op.europa.eu/en/publication-detail/-/publication/604fd72c-e6dd-11ee-9ea8-01aa75ed71a1/language-en (дата обращения: 18.10.2024).

7. AB-375 Privacy: personal information: businesses // California Legislative Information. URL: https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375 (дата обращения: 18.10.2024).

8. O'Neil C. Weapons of math destruction: How big data increases inequality and threatens democracy. New York : Crown Publishers, 2016. 272 p.

9. Russell S. Human compatible: Artificial intelligence and the problem of control. Penguin, 2019. 348 p.

10. Ribeiro M. T., Singh, S., Guestrin C. “Why should i trust you?”: Explaining the predictions of any classifier // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2016. P. 1135–1144.

11. Lundberg S. M., Lee S. I. A unified approach to interpreting model predictions // Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017. P. 4765–4774.