УДК 349

ПРАВОВОЕ РЕГУЛИРОВАНИЕ ОТНОШЕНИЙ В СФЕРЕ РАЗРАБОТКИ И ПРИМЕНЕНИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

А.И. Шевандо

преподаватель кафедры права интеллектуальной собственности юридического факультета, Белорусский государственный университет, ShavandaAI@bsu.by

Аннотация. *В ходе исследования было рассмотрено понятие «искусственный интеллект» и выделены его ключевые признаки. Также были проанализированы такие риски, как алгоритмическая непрозрачность и дискриминация в автоматизированных решениях, возникающие при использовании искусственного интеллекта. По итогам исследования были сделаны предложения по совершенствованию национального законодательства.*

Ключевые слова: *искусственный интеллект, алгоритмическая прозрачность, дискриминация и предвзятость, автономность, обучающие данные.*

Abstract. *The study examines the concept of "artificial intelligence" and identifies its key characteristics. It also analyzes risks such as algorithmic opacity and discrimination in automated decision-making that arise from the use of artificial intelligence. Based on the results of the research, proposals were made to improve national legislation.*

Keywords: *artificial intelligence, algorithmic transparency, discrimination and bias, autonomy, training data.*

Бурное развитие искусственного интеллекта (далее – ИИ) стало одним из ключевых технологических трендов XXI века, радикально меняющим экономику, управление, науку и повседневную жизнь. ИИ уже сегодня используется в различных сферах: от медицины до образования, финансов и правоприменительной деятельности. Вместе с тем столь стремительное распространение ИИ сопровождается появлением новых юридических вызовов, связанных с вопросами безопасности, ответственности, защиты прав человека, приватности, интеллектуальной собственности и недопущения дискриминации. Как показывают данные *AI Incident Database*, только в 2024 году было зафиксировано свыше 200 случаев вреда, причиненного системами ИИ [1], что ставит вопрос о необходимости правового регулирования отношений в сфере разработки и применения ИИ.

В настоящей статье будут проанализированы некоторые виды рисков, возникающих при разработке и внедрении систем ИИ. Поскольку ИИ от-

личается от традиционного программного обеспечения, необходимо также определить ключевые признаки понятия «искусственный интеллект». Понимание того, что отличает ИИ от иных программных решений, является необходимой предпосылкой для последующего анализа конкретных угроз и выработки эффективных механизмов их правового регулирования.

Некоторые исследователи определяют ИИ как «компьютерный алгоритм, предназначенный для выполнения логических задач, свойственных человеческому мышлению» [2, с. 35]. Такое определения является скорее описательным и не позволяет выделить отличительные признаки ИИ. Отсутствие четкого законодательного определения создает правовые риски и препятствует нормальному развитию технологий, так как затрудняет квалификацию моделей ИИ и размывает границы ответственности лиц, разрабатывающих и выпускающих такие системы в гражданский оборот.

Обратимся к Закону об искусственном интеллекте (англ. *Artificial Intelligence Act*, далее – Закон об ИИ), принятому в 2024 году в Европейском союзе и являющемуся первым в мире нормативным правовым актом, регулирующим отношения в сфере применения и разработки ИИ. Согласно ст. 3 Закона об ИИ под системой искусственного интеллекта понимается машинная система, способная функционировать с варьируемыми уровнями автономности, проявлять адаптивность после развертывания и которая, действуя для достижения явных или подразумеваемых целей, способна делать выводы на основе входных данных путем создания выходных результатов, таких как прогнозы, контент, рекомендации или решения, способных оказывать влияние на физическую или виртуальную среду.

Определение системы ИИ сформулировано достаточно широко и может включать в себя системы, основанные не только на машинном обучении, но и на логических методах и подходах, использующих базы знаний. Ключевыми признаками таких систем являются способность делать выводы (англ. *capability to infer*) и варьируемый уровень автономности (англ. *varying levels of autonomy*) при функционировании. Согласно п.12 Преамбулы Закона об ИИ способность делать выводы выходит за рамки простой обработки данных, поскольку включает в себя обучение, моделирование и рассуждение. Система ИИ должна быть способна воспринимать входные данные (в том числе из окружающей среды), обрабатывать их с помощью заданного набора алгоритмов и производить выходную информацию в виде прогнозов, рекомендаций или решений, которые могут повлиять на физическую или виртуальную среду. Автономность, которую система ИИ должна проявлять при

функционировании, заключается в наличии определенной степени независимости ее действий от участия человека и способности функционировать без вмешательства человека. Иными словами, простая автоматизация, статистическое программное обеспечение или полностью детерминированные алгоритмы по типу «если X, то Y» не подпадают под определение систем ИИ, поскольку они не обладают признаками автономности [3].

В определении также указывается на способность системы ИИ проявлять адаптивность, а также то, что такая система должна являться машинной. Адаптивность означает способность системы ИИ к самообучению после ввода в эксплуатацию, позволяющую ей изменяться и совершенствовать свое поведение в процессе использования. Термин «машинная» указывает на то, что системы ИИ разрабатываются с помощью машин и функционируют на их основе. При этом под машиной понимается как аппаратное, так и программное обеспечение, поддерживающее работу системы ИИ [4].

В Республике Беларусь легальная дефиниция ИИ закреплена в Постановлении Совета Министров Республики Беларусь от 21.04.2023 № 280 «О мерах по реализации Указа Президента Республики Беларусь от 07.04.2022 № 136». Под ИИ понимается комплекс технологических решений, позволяющий имитировать когнитивные функции человека (в том числе самообучение и поиск решений без заранее заданного алгоритма) и получать при выполнении конкретных задач результаты, сопоставимые с результатами интеллектуальной деятельности человека, и включающий в себя информационно-коммуникационную инфраструктуру, программное обеспечение, процессы и сервисы по обработке данных и поиску решений.

Такое определение является не очень удачным по ряду причин. Во-первых, признак «имитация когнитивных функций человека» представляется избыточно широким и будет размывать понятие ИИ. Любой алгоритм, формализующий задачу, может «имитировать» элементарные когнитивные действия человека. Во-вторых, упор на внешнюю схожесть результатов ИИ с результатами интеллектуальной деятельности человека игнорирует тот факт, что системы ИИ действуют по иным правилам и решают проблемы «нечеловеческими» способами. Представляется, что законодатель должен отойти от концепции «антропоцентричности» систем ИИ и при определении их правового статуса сосредоточиться на сущностных характеристиках, таких как автономность, способность самообучаться и принимать решения с определенной степенью самостоятельности [5].

Именно автономность, обучаемость, способность выявлять закономерности в данных и делать выводы отличают системы ИИ от других техноло-

гий. Эти характеристики обеспечивают гибкость ИИ, позволяют решать разноплановые задачи и, соответственно, оптимизировать процессы практически в любой сфере: от экономики до медицины. Однако те же особенности порождают ряд рисков, среди которых выделяют низкую прозрачность алгоритмов и принятие решений, способных быть дискриминационными и предвзятыми.

Одним из наиболее серьезных вызовов при внедрении современных систем ИИ является их недостаточная прозрачность. Под прозрачностью обычно подразумевают доступность и наглядность внутренних механизмов, обеспечивающих вычислительные результаты, генерируемые системами ИИ [7]. Внутренние механизмы обработки данных и принятия решений в ИИ часто функционируют как «черный ящик» (англ. *black box*), что затрудняет проверку корректности работы, воспроизведение результатов и выявление системных ошибок. Такая непрозрачность ограничивает возможности аудита и оценки алгоритмов пользователями, регуляторами и разработчиками, повышает риски скрытой дискриминации и подрывает доверие к технологиям ИИ, особенно в критически важных сферах.

Примером алгоритмической непрозрачности может служить инцидент, произошедший в рамках судебного разбирательства *Mata v. Avianca* (Нью-Йорк, 2023 г.). Адвокат истца представил суду пояснения, в которых обосновывал позицию своего клиента со ссылками на несуществующие прецеденты, сгенерированные *ChatGPT*. Данное дело наглядно иллюстрирует проблему «галлюцинаций» ИИ – генерации ложной информации, выдаваемой за достоверную. Подобные искажения возникают вследствие неинтерпретируемости сложных моделей (особенно нейросетей), где ключевые механизмы обработки данных скрыты внутри сложных слоев и недоступны для анализа.

Для преодоления подобных рисков активно развивается направление объяснимого искусственного интеллекта (англ. *eXplainable AI, XAI*). Его цель – повышение прозрачности моделей ИИ и укрепление доверия пользователей к результатам, формируемым такими системами [7]. Для выполнения этих задач используют такие методы, как **LIME** (англ. *Local Interpretable Model-agnostic Explanations*) и **SHAP** (англ. *SHapley Additive exPlanations*). SHAP – это метод, основанный на теории игр. Его цель состоит в том, чтобы объяснить любую модель ИИ, рассматривая каждый признак как игрока и результат модели как вознаграждение [7]. Иными словами, использование такого метода означает оценку влияния каждого фактора на итоговый вывод ИИ. Метод *LIME* работает по принципу локальной аппроксимации: чтобы объяснить конкретное предсказание модели, *LIME* создает

множество слегка измененных версий исходного входа (например, текста или изображения), пропускает их через модель ИИ и на основе полученных ответов строит простую интерпретируемую модель, которая хорошо описывает поведение сложной модели в непосредственной близости от исходного случая [8, с. 830]. Это позволяет выяснить, какие признаки (например, слова или пиксели) сыграли наибольшую роль в принятии конкретного решения.

С проблемой низкой прозрачности алгоритмов тесно связаны риски предвзятости и дискриминации в решениях ИИ. Примером такой проблемы является довольной хрестоматийный случай, когда в 2015 году компания Amazon начала использовать для найма сотрудников систему ИИ, которая была обучена на резюме кандидатов за предыдущие 10 лет. Из-за того, что в технической отрасли большинство работников являются мужчинами, рекрутинговая модель занижала оценки резюме, которые содержали в себе слова «women», отдавай предпочтение мужским резюме.

Этот пример показывает, что ИИ, подобно человеку, который перенимает чужие стереотипы, может «усвоить» скрытые предвзятости в ходе обучения. В результате даже без явного указания чувствительных признаков (пол, раса, возраст) модель может выдавать решения, неблагоприятные для отдельных людей или групп, если в обучающей выборке присутствовали соответствующие смещения [9].

В литературе выделяют несколько причин, по которым результаты ИИ могут приводить к дискриминации:

1) алгоритмическая предвзятость при моделировании ИИ, когда в результате оптимизации внутренних параметров модель искусственно «сглаживает» влияние редких признаков, недооценивает их вклад и фактически игнорирует характеристики меньшинств;

2) предвзятость в обучающих данных, возникающая из-за использования наборов с неравномерным распределением признаков (например, при разработке инструмента для найма преобладает информация, взятая из резюме молодых кандидатов);

3) предвзятость при использовании ИИ, возникающая, когда модель применяется за пределами сферы или контекста, для которых она была обучена [10].

Решение проблемы с дискриминацией в решениях ИИ может быть достигнуто за счет разнообразия обучающих данных и использования метрик справедливости (англ. *fairness metrics*), позволяющих выявить неравномерное отношение ИИ к определенным группам.

Алгоритмическая непрозрачность и предвзятость в решениях моделей ИИ могут стать причиной вреда, нанесенного правам и законным интересам человека. Роль правового регулирования заключается в минимизации таких рисков и возложении на разработчиков и производителей систем ИИ обязанностей по обеспечению прозрачности алгоритмических механизмов и предотвращению дискриминации при автоматизированном принятии решений. Это может быть обеспечено за счет введения законодательных требований к раскрытию ключевой информации о системах ИИ, включая их архитектуру, принципы функционирования, а также характеристику и происхождение данных, использованных на этапе обучения. В социально чувствительных сферах, таких как здравоохранение, кредитование, трудоустройство и др., предполагается установление норм, обязывающих обеспечивать эффективный человеческий надзор, позволяющий контролировать и, при необходимости, корректировать решения, принимаемые ИИ.

Подобные механизмы уже находят отражение в законодательстве Европейского союза. Так, Закон об ИИ возлагает на поставщиков высокорисковых систем следующие обязанности:

1) использовать для обучения моделей высококачественные, релевантные и репрезентативные данные (ст. 10);

2) обеспечивать фиксацию всех событий на протяжении всего жизненного цикла системы ИИ (ст. 12);

3) сопровождать свои системы четкими инструкциями, включающими информацию о функциональных возможностях и ограничениях модели, а также о потенциальных рисках ее применения (ст. 13);

4) разрабатывать ИИ таким образом, чтобы был обеспечен эффективный человеческий контроль (ст. 14).

Подводя итоги проведенного исследования, следует отметить, что построение надежных и безопасных систем ИИ возможно лишь при наличии эффективного и комплексного правового регулирования. Оно служит не только средством обеспечения прозрачности и контроля над функционированием ИИ, но и важным механизмом защиты фундаментальных прав человека в условиях цифровой трансформации.

Одним из ключевых элементов такого регулирования является формулирование точного и содержательного определения искусственного интеллекта. Представляется, что существующее в законодательстве Республики Беларусь понятие не отражает такие сущностные признаки ИИ, как способность к обучению, автономность и адаптивность.

Наряду с формированием точного определения ИИ, приоритетом должно стать правовое регулирование рисков, связанных с функционированием систем ИИ. В статье рассмотрены некоторые виды таких рисков, включая алгоритмическую непрозрачность и дискриминацию в решениях ИИ. Для решения этих проблем необходимо законодательно закрепить за разработчиками моделей обязанности по обеспечению качества данных, используемых при обучении ИИ, прозрачности алгоритмов и человеческого контроля за решениями, принимаемыми ИИ.

Список использованных источников

1. AI Incident Database. – URL: https://incidentdatabase.ai/ (дата обращения 31.05.2025).
2. Artificial Intelligence and Intellectual Property / ed. Lee J. A., Hilty R. M., and Liu K. C. – Oxford : Oxford University Press, 2021. – 449 p.
3. Silva N. S. E. The Artificial Intelligence Act: critical overview // arXiv. – 2024. – URL: https://arxiv.org/abs/2409.00264 (дата обращения 31.05.2025).
4. Commission Guidelines on the Application of the Definition of an AI System and the Prohibited AI Practices Established in the AI Act // European Law Institute, 2024. – URL: https://www.europeanlawinstitute.eu/fileadmin/user_upload/p_eli/Publications/ELI_Response_on_the_definition_of_an_AI_System.pdf (дата обращения 31.05.2025).
5. Ядревский, О. О. Гражданско-правовые аспекты функционирования искусственного интеллекта // ИПС «ЭТАЛОН-ONLINE» (дата обращения 31.05.2025).
6. Almado M. Technical AI Transparency: A Legal View of the Black Box // SSRN. – 2025. – URL: https://ssrn.com/abstract=5096913 (дата обращения 31.05.2025).
7. Salih A., Raisi-Estabragh Z., Ilaria B.G. et al. A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME // arXiv. – 2023. – URL: https://arxiv.org/abs/2305.02012 (дата обращения 31.05.2025).
8. Dey S., Chowdhury T. R. A Comparative Survey of SHAP and LIME: Explaining Machine Learning Models for Transparent AI // International Journal of Innovative Research in Education. – 2024. – Vol. 11, № 6. – P. 827-835.
9. Daniel V., Suárez J.L. Discrimination, Bias, Fairness, and Trustworthy AI // Applied Sciences. – 2022. – № 12(22). – URL: https://doi.org/10.3390/app12125826 (дата обращения 31.05.2025).
10. Ferrer X., Nuenen v. T., Jose M. S. et al. Bias and Discrimination in AI: a cross-disciplinary perspective // arXiv. – 2020. – URL: https://arxiv.org/abs/2008.07309 (дата обращения 31.05.2025).