Научная статья
УДК 342.511:004.8
DOI: 10.14529/law250410

СОВРЕМЕННОЕ СОСТОЯНИЕ И ПРОБЛЕМЫ ПРАВОВОГО РЕГУЛИРОВАНИЯ ПРИМЕНЕНИЯ НАБОРОВ ДАННЫХ ДЛЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

А. В. Минбалеев, *alexmin@bk.ru, https://orcid.org/0000-0001-5995-1802*
Д. П. Осипов, *dpo.osipov@yandex.ru, https://orcid.org/0009-0004-6812-9692*
Московский государственный юридический университет имени О. Е. Кутафина (МГЮА), г. Москва, Россия

Аннотация. В статье исследуются современное состояние и проблемы правового регулирования применения наборов данных для искусственного интеллекта, анализируются нормы Модельного закона «О технологиях искусственного интеллекта», принятого Межпарламентской ассамблеей Содружества Независимых Государств, в части регулирования набора данных для технологий искусственного интеллекта.

Установлено, что активное внедрение технологий искусственного интеллекта с каждым днем приводит к стремительному расширению сфер применения наборов данных. В связи с этим анализируется применение наборов данных в строительной сфере, в здравоохранении, а также в юридической сфере, в том числе в судебной.

Авторы приходят к выводу о том, что решать комплексные вопросы регулирования формирования, использования и доступа к наборам данных возможно за счет принятия следующих мер: создание открытых платформ проверки лицензионных соглашений и характеристик наборов данных; регистрация наборов данных, обученных на охраняемых результатах интеллектуальной деятельности, в уполномоченных федеральных органах исполнительной власти; совершенствование корпоративных политик по формированию наборов данных; расширение взаимодействия с экспертами-юристами; организационные и правовые процедуры контроля происхождения наборов данных – лицензирования, формирования, обучения и использования. Все эти меры поэтапно должны быть закреплены в законодательстве Российской Федерации.

Ключевые слова: данные, искусственный интеллект, международное регулирование, наборы данных, правовое регулирование, СНГ, стратегическое планирование, технологии искусственного интеллекта, цифровые технологии.

Благодарности. Статья подготовлена в рамках программы стратегического академического лидерства «Приоритет–2030».

Для цитирования: Минбалеев А. В., Осипов Д. П. Современное состояние и проблемы правового регулирования применения наборов данных для искусственного интеллекта // Вестник ЮУрГУ. Серия «Право». 2025. Т. 25. № 4. С. 72–77. DOI: 10.14529/law250410.

Original article
DOI: 10.14529/law250410

THE CURRENT STATE AND PROBLEMS OF LEGAL REGULATION OF THE USE OF DATA SETS FOR ARTIFICIAL INTELLIGENCE

A. V. Minvaleev, *alexmin@bk.ru, https://orcid.org/0000-0001-5995-1802*
D. P. Osipov, *dpo.osipov@yandex.ru, https://orcid.org/0009-0004-6812-9692*
Kutafin Moscow State Law University (MSAL), Moscow, Russia

Abstract. This article examines the current state and challenges of legal regulation of the use of datasets for artificial intelligence, analyzing the provisions of the Model Law "On Artificial Intelligence Technologies," adopted by the Interparliamentary Assembly of the Commonwealth of Independent States, as they relate to the regulation of datasets for artificial intelligence technologies.

It has been established that the active implementation of artificial intelligence technologies is rapidly expanding the scope of dataset application. In this regard, the use of datasets in the construction, healthcare, and legal spheres, including the judiciary, is analyzed.

The authors conclude that complex issues of regulating the formation, use, and access to datasets can be addressed by adopting the following measures: creating open platforms for verifying license agreements and dataset characteristics; registering datasets trained on protected intellectual property with authorized federal executive bodies; improving corporate policies for the formation of datasets; expanding interaction with legal experts; organizational and legal procedures for monitoring the origin of datasets – licensing, formation, training, and use. All these measures should be gradually enshrined in Russian Federation legislation.

Keywords: data, artificial intelligence, international regulation, datasets, legal regulation, CIS, strategic planning, artificial intelligence technologies, digital technologies.

Acknowledgments. The article was prepared within the framework of the strategic academic leadership program "Priority 2030".

For citation: Minvaleev A. V., Osipov D. P. The current state and problems of legal regulation of the use of data sets for artificial intelligence. *Bulletin of the South Ural State University. Series "Law".* 2025. vol. 25. no. 4. pp. 72–77. (in Russ.) DOI: 10.14529/ law250410.

В условиях активного развития технологий искусственного интеллекта одним из наиболее острых вопросов является закрепление правового режима и регулирование вопросов формирования и использования наборов данных для искусственного интеллекта. Хотя сфера использования наборов данных не ограничивается применением для обучения и работы с искусственным интеллектом, но именно для этих целей данный объект все чаще становится объектом пристального государственного внимания.

Согласно ст. 2 Модельного закона 18 апреля 2025 г. № 58-8 «О технологиях искусственного интеллекта», принятого Межпарламентской ассамблеей Содружества Независимых Государств (далее – Модельный закон об ИИ), под наборами данных применительно к технологиям искусственного интеллекта понимается «совокупность данных, прошедших предварительную подготовку (обработку) в соответствии с требованиями законодательства и необходимых для разработки и функционирования технологий искусственного интеллекта и (или) систем с использованием технологий искусственного интеллекта». Определение представляет собой переработку из первой редакции Национальной стратегии развития искусственного интеллекта на период до 2030 года, утвержденной Указом Президента Российской Федерации от 10 октября 2019 г. № 490 «О развитии искусственного интеллекта в Российской Федерации» (далее – Национальная стратегия ИИ), а также ряда технических стандартов Российской Федерации.

В гл. 6 Модельного закона об ИИ освещены вопросы использования, обработки, хранения и защиты данных. В соответствии с положениями Модельного закона об ИИ процессы формирования, подготовки, разметки использования наборов данных регламентируются на национальном уровне. В качестве признаков данных выделяются: достоверность, достаточность и целесообразность, соответствие принципам управления данными; процессы сбора, обеспечения полноты, репрезентативности, хранения, реализации политики доступа, этапы подготовки, оценки, аудита должны соответствовать требованиям оператора. В качестве требований предусмотрены обязательная предварительная обработка и подготовка для машинного обучения. Этапы такой подготовки включают очистку данных, масштабирование и нормализацию, кодирование категориальных данных, разделение на тренировочный и тестовый наборы.

Предусматривается также и ряд требований для наборов данных: соответствие стандартам, принятие мер для недопущения дискриминации, обеспечение защиты конфиденциальности и персональных данных.

Активное внедрение технологий искусственного интеллекта с каждым днем приводит к стремительному расширению сфер применения наборов данных [1–3; 6]. Так, грамотное формирование наборов данных для целей искусственного интеллекта в сфере строи-

тельства позволяет осуществлять мониторинг регионов с целью анализа протекающих строительных процессов и уровня урбанизации региона, динамическое ценообразование и бенчмаркинг на основе искусственного интеллекта, мастер-планирование территорий жилых комплексов, внедрять платформу интерактивной отчетности, продукты автономного строительства, автоматизированное размещение объектов на территории и т.д. Примерами таких данных могут быть космические снимки со спутников, данные по инфраструктуре, застройке, данные по объемам производства и прозрачности закупок материалов, технические паспорта многоквартирных домов, цифровые модели объектов капитального строительства, фото-, видео-, аудиоматериалы со строительных объектов [4].

В сфере здравоохранения искусственный интеллект может применяться с целью повышения качества диагностики клинически значимых изменений посредством использования интеллектуального ассистента, который в проспективном режиме будет передавать врачу-человеку данные о потенциальных патологиях, повышения эффективности управления значимыми для медицинской организации ресурсами с учетом сезонности заболеваний, персонализации предложений лечения для разных типов заболеваний и т.д. Для этого могут быть использованы данные, например, рентгенографических исследований, статистика заболеваемости в зависимости от пола, возраста, региона проживания, статистика по уровню удовлетворенности граждан медицинскими услугами, релевантные базы органов публичной власти [4].

Процессы формирования наборов данных приобретают актуальность и для совершенствования юридической деятельности. Так, одним из предметов исследований становится процесс извлечения и классификации юридически значимых понятий из различных документов (legal Information extraction, LIE), необходимый для поиска судебной практики по аналогичным делам, прогнозирования итоговых судебных решений, развития ситуационного поиска (в вопросно-ответной форме). Данная работа осложняется многими барьерами: во-первых, языковыми, поскольку существует объективная потребность в размеченных данных на языке государства использования, во-вторых, контекстуальными, не позволяющими оперировать четкими юридическими категориями, имеющими собственную специфику, их межотраслевыми связями и др. [5].

Более того, опыт внедрения интеллектуальных решений в судебную систему показал, что провалы автоматизации процессов по обезличиванию судебных решений связаны с недооценкой ручной разметки данных, позволяющей искусственному интеллекту эффективно выполнять поставленные перед ним задачи, а также юридической квалификации специалистов (аннотаторов), размечающих данные. Так, при автоматизации процессов обезличивания данных в базе судебных решений «Judilibre» в качестве персональных данных были обезличены имена лошадей. Исследователи, проанализировав полный цикл работы искусственного интеллекта и показав неспособность ИИ работать автономно без грамотной разметки и непрерывного мониторинга, пришли к выводу, что объективность результатов использования заключается во множестве субъективных решений, принятых на подготовительном этапе работы с данными [8].

Сегодня большая часть данных аккумулируется в федеральных органах исполнительной власти. Разработчики и отраслевые пользователи ИИ-решений часто не всегда представляют, какие конкретно данные будут полезны для целей применения ИИ и в каких конкретно видах деятельности возможен быстро достижимый социальный эффект. Для эффективного обеспечения этих процессов возрастает потребность в стабильном взаимодействии уполномоченных государственных органов с организациями, методической поддержке, активной деятельности по привлечению организаций-потребителей данных, техническом обеспечении хранения наборов данных. Следовательно, управление в условиях формирования экономики данных недостаточно характеризовать с позиций целенаправленного воздействия для достижения необходимого результата, государство становится участником диалога с организациями и вырабатывает эффективную организационно-правовую модель, направленную на развитие использования наборов данных.

В этой связи видится верным решением нормативно предусмотреть обязанность профильных органов государственной власти предоставлять организациям открытые данные для формирования востребованных набо-

ров данных. При этом для организаций важно определить критерии участия, условия, основания предоставления, цели обработки данных, отдельно закрепить вопросы защиты интеллектуальных прав на результаты таких разработок. В случаях, когда разработчиком набора данных выступает сам уполномоченный орган государственной власти, за ним должны быть закреплены соответствующие полномочия.

Открытые данные на периодической основе публикуются на официальных сайтах федеральных органов исполнительной власти в информационно-телекоммуникационной сети «Интернет». В июле 2025 года Министерством экономического развития Российской Федерации запущен Портал открытых данных (https://data.gov.ru/), на котором размещены наборы данных по следующим категориям: безопасность, государство, досуг и отдых, здоровье, картография, культура, метеоданные, образование, спорт, строительство, торговля, транспорт, туризм, экология, экономика, электроника. Поставщиком таких данных являются органы публичной власти. Данные представляются в форматах «CSV», «JSON», «XML» и доступны начиная с 2000 года [11]. В условиях отсутствия порталов открытых данных, в частности данных о преступности, и одновременно высокого уровня урбанизации в Китае формирование востребованных дата-сетов осложняется. Между тем исследователи анализируют возможные источники для сбора данных (например, платформа судебных решений «China Judgments Online» и др.), предлагают осуществлять геокодирование территорий для точности определения мест преступлений, а также использовать время, место, типичные ситуации, информацию о жертвах и обвиняемых для формирования дата-сета.

В Российской Федерации и ранее активно развивались подобные порталы, преимущественно в сфере медицины.

Развитие и функционирование таких платформ снижает проблему предвзятости источников предоставляемых данных, что в перспективе должно повысить качество самих наборов данных, обеспечение которых является ресурсозатратным для заинтересованных в разработке субъектов и может повлечь негативные социальные эффекты.

Исследователи проблем эффективного формирования наборов данных также предлагают изучать деятельность организаций-потребителей данных путем сбора анкет (дата-шитов). Собираемая информация может быть структурирована на следующие разделы: общая информация (изучается цель и конкретная задача сбора данных, период сбора данных, от кого должны быть получены наборы данных, каков объем данных и т.д.), информация о сборе данных (то есть к какому типу относятся входные данные (признаковое описание объектов, матрица расстояний между объектами, временной ряд или сигнал, изображение или видеоряд), какие организации участвовали в процессе сбора данных и как они финансировались, проводился ли анализ потенциального воздействия набора данных и его использования на субъекты данных (например, анализ воздействия на защиту данных) и т.д.), информация о предварительной обработке данных, информация о публикации и использовании данных, информация о доступе к данным (какие предусмотрены уровни доступа, какова процедура такого доступа), информация о распространении и дальнейшем техническом обслуживании [7; 9; 10].

Актуальным направлением для научного осмысления и практики являются данные, обладателем которых является частный сектор. В ряде случаев для кооперации по вопросам развития ИИ требуется объединение таких данных. Узловыми вопросами являются готовность частного сектора на такой обмен, содержание и объем данных, а также возможное определение уполномоченного государственного органа, объединяющего и предоставляющего доступ на законных и справедливых основаниях. Безусловно, такая организационная модель небесспорна и существенно расширяет функции государства как участника диалога, о котором мы ранее указывали.

Аналогично разрабатывать наборы данных могут научно-исследовательские институты, данная работа финансируется за счет бюджетных средств. Такие разработки могут быть полезны для выстраивания бизнес-процессов на основе искусственного интеллекта и разработки новых решений для последующей коммерциализации.

Отдельного внимания заслуживают данные, на которые распространяются специальные правовые режимы (персональные данные, медицинская тайна, налоговая тайна и т.д.). Существуют также данные, отдельный правовой режим которых находится сегодня на эта-

пе разработки в России. Например, промышленные данные. Они существенным образом затрудняют процесс формирования наборов данных, межведомственную аналитику, а также установление экспериментальных правовых режимов с их использованием.

Представляется, что решать комплексные вопросы регулирования формирования, использования и доступа к наборам данных возможно за счет принятия следующих мер: создание открытых платформ проверки лицензионных соглашений и характеристик наборов данных; регистрация наборов данных, обученных на охраняемых результатах интеллектуальной деятельности, в уполномоченных федеральных органах исполнительной власти; совершенствование корпоративных политик по формированию наборов данных; расширение взаимодействия с экспертами-юристами; организационные и правовые процедуры контроля происхождения наборов данных – лицензирования, формирования, обучения и использования. Все эти меры поэтапно должны быть закреплены в законодательстве.

Полагаем, что параллельно с этим государство должно минимизировать административные барьеры в межведомственной аналитике, что позволит оперативно создавать универсальные цифровые профили, которые будут беспрепятственно объединять данные многих органов публичной власти, смогут охватить все отраслевые домены, функционально дадут возможность вновь использовать и развивать цифровые профили. В конечном счете, это во многом даст возможность моделировать меры поддержки, прогнозировать поведение населения, точно оценивать управленческие риски, персонализировать государственное управление для граждан и организаций, централизованно контролировать разработку и использование ИИ-решений.

Нормативная правовая гарантированность государственных задач в сфере формирования и развития наборов данных обеспечивается пока преимущественно посредством актов стратегических планирования и выражена фрагментарно. Законодателю предстоит принять еще ряд норм в данном направлении.

Список источников

1. Искусственный интеллект и право: от фундаментальных проблем к прикладным задачам / Д. Л. Кутейников, О. А. Ижаев, С. С. Зенин, В. А. Лебедев. М.: ООО «Проспект», 2025. 104 с.

2. Комплексное исследование правовых и этических аспектов, связанных с разработкой и применением систем искусственного интеллекта и робототехники: монография / В. В. Архипов, Г. Г. Камалова, В. Б. Наумов, А. В. Незнамов. Санкт-Петербург: НП-Принт, 2022. 336 с.

3. Механизмы и модели регулирования цифровых технологий: монография / А. В. Минбалеев, А. В. Мартынов, Г. Г. Камалова. М.: ООО «Проспект», 2023. 264 с.

4. Рабочая группа «Искусственный интеллект». URL: https://d-economy.ru/wg/ai/.

5. Cao Y., Sun Y., Xu C., Li Ch., Du J., Lin H. CAILIE1.0: A dataset for Challenge of AI in Law Information Extraction V1.0. AI Open 3. 2022. P. 208–212.

6. Definition of artificial intelligence in the context of the Russian legal system: a critical approach / V. V. Arkhipov, A. V. Gracheva, V. B. Naumov [et al.]. State and Law. 2022. no. 1. P. 168–178.

7. Gebru T., Morgenstern J., Vecchione B., Wortman J., Wallach H., Daumé III H., Crawford K. Datasheets for Datasets. Communications of the ACM. 2021, vol. 64, iss. 12, pp. 86–92. Available at: doi.org/10.1145/345872.

8. Girard-Chanudet C. Ground-truth is law: The invisible conceptual work behind AI. Big Data & Society. 2025. № 12 (2). Available at: doi.org/10.1177/20539517251352823.

9. Pushkarna M., Zaldivar A. Data Cards: Purposeful and Transparent Documentation for Responsible AI. In 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea. ACM, New York, NY, USA 51 Pages. Available at: doi.org/10.1145/3531146.3533231.

10. Tsay J., Braz A., Hirzel M., Shinnar A. and Mummert T. AIMMX: Artificial Intelligence Model Metadata Extractor," 2020 IEEE/ACM 17th International Conference on Mining Software Repositories (MSR), Seoul, Korea, Republic of, 2020, pp. 81–92. Available at: doi: 10.1145/3379597.3387448.

11. Zhang, Y., Kwan, MP. & Fang, L. An LLM driven dataset on the spatiotemporal distributions of street and neighborhood crime in China. Sci Data. 2025. no. 12 (467). Available at: doi.org/10.1038/s41597-025-04757-8.

References

1. Kuteynikov D. L., Izhaev O. A., Zenin S. S., Lebedev V. A. *Iskusstvennyy intellekt i pravo: ot fundamental'nykh problem k prikladnym zadacham* [Artificial intelligence and law: from fundamental problems to applied problems]. Moscow, 2025, 104 p.

2. Arkhipov V. V., Kamalova G. G., Naumov V. B., Neznamov A. V. Kompleksnoe issledovanie pravovykh i eticheskikh aspektov, svyazannykh s razrabotkoy i primeneniem sistem iskusstvennogo intellekta i robototekhniki [Comprehensive study of legal and ethical aspects related to the development and application of artificial intelligence and robotics systems]. St. Petersburg, 2022, 336 p.

3. Minbaleev A. V., Martynov A. V., Kamalova G. G. *Mekhanizmy i modeli regulirovaniya tsifrovykh tekhnologiy* [Mechanisms and models of digital technology regulation]. Moscow, 2023, 264 p.

4. *Rabochaya gruppa «Iskusstvennyy intellekt»* [Artificial Intelligence Working Group]. Available at: d-economy.ru/wg/ai/.

Информация об авторах

Минбалеев Алексей Владимирович, доктор юридических наук, профессор, заведующий кафедрой информационного права и цифровых технологий, Московский государственный юридический университет имени О. Е. Кутафина (МГЮА), г. Москва, Россия.

Осипов Даниил Павлович, стажер-исследователь Института информационной и медиабезопасности, Московский государственный юридический университет имени О. Е. Кутафина (МГЮА), г. Москва, Россия.

Information about the authors

Aleksey V. Minbaleev, Doctor of Sciences (Law), Professor, head Department of Information Law and Digital Technologies, Kutafin Moscow State Law University (MSAL), Moscow, Russia.

Daniil P. Osipov, Intern researcher at the Institute of Information and Media Security, Kutafin Moscow State Law University (MSAL), Moscow, Russia.

Статья поступила в редакцию 25 августа 2025 г.
Received August 25, 2025.