



Дата поступления рукописи в редакцию: 15.05.2026 г.
Дата принятия рукописи в печать: 29.06.2026 г.

DOI: 10.24412/2076-1503-2026-6-911-916

ЛЫСАК Владислав Витальевич,
ООО «Юридическое бюро «МАУП», юрист,
e-mail: vlad_word@mail.ru

Научный руководитель:
ГЕЛЬДИБАЕВ Мовлад Хасиевич,
профессор, доктор юридических наук,
e-mail: mail@law-books.ru

ОБУЧЕНИЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА: BIG DATA VS DATASET

Аннотация. Активное развитие искусственного интеллекта и его интеграция в гражданский оборот обусловили необходимость формирования единого понятийно-категориального аппарата, характеризующего объекты, задействованные при создании интеллектуальных систем. В доктрине и нормативных правовых актах прослеживается смешение терминов «большие данные», «датасет», «массив данных», «набор данных», что препятствует последующему определению правового режима каждого из обозначенных объектов. В статье анализируются содержание указанных терминов, их соотношение друг с другом и взаимосвязь с процессом обучения искусственного интеллекта. Использование формально-юридического, сравнительно-правового и логико-семантического методов позволило показать, что большие данные и датасеты не являются тождественными категориями. Большие данные характеризуются не только совокупностью информационных объектов, но и системой средств их сбора, хранения и обработки. Датасет, напротив, представляет собой идентифицируемую, структурированную и подготовленную под определенную задачу совокупность данных. Автор приходит к выводу о том, что обучение искусственного интеллекта непосредственно осуществляется не на больших данных в их естественном состоянии, а на датасетах, интегрированных в российское правовое поле в формах «набор данных» и «массив данных».

Ключевые слова: обучение искусственного интеллекта; правовой режим; большие данные; массив данных; машинное обучение; датасет; набор данных; искусственный интеллект.

LYSAK Vladislav Vitalyevich,
Legal Bureau MAiP LLC, lawyer

Scientific supervisor:
GELDIBAЕV Movlad Khasievich,
Professor, Doctor of Law

TRAINING ARTIFICIAL INTELLIGENCE: BIG DATA VS DATASET

Annotation. The active development of artificial intelligence and its integration into civil circulation have necessitated the formation of a unified conceptual and categorical framework describing the objects involved in the creation of intelligent systems. Both legal doctrine and regulatory legal acts reveal a conflation of the terms «big data», «dataset», «data array» and «data set», which complicates the subsequent determination of the legal regime of each of these objects. The article analyzes the content of these terms, their correlation with one another and their connection with the process of artificial intelligence training. The use of formal legal, comparative legal and logical-semantic methods makes it possible to show that big data and datasets are not identical categories. Big data is characterized not only by a set of informational objects, but also by a system of means for

their collection, storage and processing. A dataset, by contrast, is an identifiable, structured and task-oriented set of data. The author concludes that artificial intelligence is trained directly not on big data in its natural state, but on datasets integrated into the Russian legal framework under the terms «data set» and «data array».

Key words: *artificial intelligence training; legal regime; big data; data array; machine learning; dataset; data set; artificial intelligence.*

Введение

В последнее десятилетие сфера искусственного интеллекта получила широкое распространение. Технологии, основанные на машинном обучении, активно разрабатываются и интегрируются практически во все сферы жизнедеятельности человека. США и КНР задают темп научных исследований в обозначенном направлении, вынуждая другие страны направлять существенные средства на разработку собственных аналогов для сохранения суверенитета в эпоху информационных технологий. В законодательстве многих стран появляются нормы, регламентирующие новые поведенческие акты, связанные с разработкой и эксплуатацией искусственного интеллекта (далее - ИИ).

Между тем успешная интеграция разрабатываемых ИИ-технологий в гражданский оборот невозможна без специализированного правового регулирования. Особо остро вопрос введения соответствующих норм касается основополагающих категорий, связанных с правовым режимом объектов, задействованных при создании систем ИИ.

Одним из существенных процессов, реализуемых разработчиками, выступает обучение интеллектуальной системы. Уже сейчас выделяется проблема раздробленности понятийно-категориального аппарата, используемого как в доктрине, так и на практике при характеристике объектов, применяемых для обучения ИИ. Во многих исследованиях и актах государственных органов происходит смешение понятий «большие данные» (big data), «датасет» (dataset), «массив информации», «набор данных» и иных словесных конструкций. Но оправдано ли отождествление указанных выражений? Не имеет ли каждое из них самостоятельное содержание, предполагающее отличный от иных правовой режим?

Проблему формирования единообразного понятийного аппарата в информационной сфере затрагивали и другие исследователи. В. Ю. Карпычев обращал внимание на неоднородность употребления термина «большие данные», который в документах федерального уровня принимал форму «больших объемов данных» [1]. Э. И. Лескина и М. М. Панарина указывают на синонимизацию обозначенного понятия с «массивом данных», а также на различные интерпретации «совокупности данных», принимающей формы

«массива данных», «наборов данных», «датасетов» [2].

Отмеченное иллюстрирует потребность в формировании единой структуры терминов для будущих норм специализированного федерального закона, регламентирующего общественные отношения в сфере создания и использования ИИ.

Таким образом, актуальность настоящего исследования заключается в необходимости стабилизации понятийного аппарата, характеризующего объекты, предназначенные для обучения ИИ.

Характеристика больших данных (big data)

Одним из базовых процессов, выступающих этапом создания ИИ, является его обучение. Для реализации данной деятельности необходим колоссальный объем данных, на основании которого алгоритмы выявляют паттерны и закономерности, строят вероятностную иерархию. При этом процесс обучения ИИ часто соотносят с big data. Так, в некоторых работах авторами указывается, что именно большие данные используются для обучения ИИ [3].

Понятие «большие данные» начало применяться примерно с 2005 года и по мере увеличения объемов информационных объектов стало развиваться и интегрироваться в различные виды деятельности.

Содержание указанной категории толкуется в доктрине неоднозначно. Ряд авторов интерпретирует big data как некоторый объем информационных объектов. С. С. Кривушева рассматривает большие данные через призму информационной совокупности [4]. Э. И. Лескина использует близкий подход: big data раскрывается ею как значительный по объему комплекс сведений [5]. Н. В. Володина конкретизирует этот признак и относит к большим данным масштабные потоки структурированных и неструктурированных цифровых данных, обработка которых невозможна с применением традиционных программных средств [6].

Между тем отмеченные позиции сужают сущностное содержание больших данных, фактически сводя их понимание к обычной совокупности объектов. В свою очередь, характеристика и структура big data не ограничиваются лишь объемом однотипных благ, включенных в общую кате-

горию «информация» или «сведения», а представляют собой более обширное понятие, в которое среди прочего входят механизмы обработки данных. Big data представляет собой не только средство хранения, но и технологическую среду, в рамках которой происходит преобразование информационных объектов.

Так, например, Д. А. Петров рассматривает большие данные не как простую совокупность сведений, а в виде объемного и разнородного информационного массива, пригодного к программной обработке, что обуславливает возможность задействовать их в качестве самостоятельных объектов правоотношений в гражданском обороте [4].

Л. Ю. Василевская и Е. Б. Подузова, анализируя спектр элементов, охватываемых обозначенным понятием, рассматривают большие данные как информационную систему, указывая на то, что они не могут существовать без симбиоза материального носителя и программных средств обработки [7; 8].

В свою очередь, Е. В. Ульянова рассматривает big data как технологию, охватывающую не отдельный программный продукт и не самостоятельное вычислительное средство, а взаимосвязанную систему программ и утилит, обеспечивающих сбор, обработку, хранение и анализ сведений в условиях, когда традиционные инструменты не справляются с возложенной задачей [9].

На наш взгляд, учитывая специфику составляющих элементов больших данных, наиболее приемлемой и эксплицирующей внутреннее содержание представляется характеристика big data в качестве информационной системы, которая в силу своей экономической ценности в гражданском обороте должна иметь выражение в качестве объекта гражданских прав.

Big data не имеет единообразного состава и может объединять данные с различным правовым режимом, например, объекты, охраняемые авторским правом, и персональные данные. Подобная разнородность также указывает на невозможность сведения больших данных к простому набору сведений.

Ввиду усиливающейся важности big data раскрытие указанного термина нашло отражение в официальных нормативных актах в качестве обозначения структурированных и неструктурированных данных огромных объемов и значительного многообразия, эффективно обрабатываемых горизонтально масштабируемыми программными инструментами (абзац 8 раздела 1 Концепции создания цифровой аналитической платформы предоставления статистических данных, утвержденной распоряжением Правительства Российской Федерации от 17.12.2019 № 3074-р).

В государственных стандартах были отражены существенные признаки, которые характеризуют big data: объем, скорость обработки, разнообразие и вариативность (пункт 3.1.2 ГОСТ Р ИСО/МЭК 20546-2021 «Информационные технологии. Большие данные. Обзор и словарь»).

При этом процесс формирования отмеченных признаков происходил эволюционным путем, что также указывает на недопустимость сведения рассматриваемой категории к простой совокупности. Ранее свойства big data рассматривались с позиции трех составляющих: объема, разнообразия и скорости, однако в последующем приведенный список был дополнен изменчивостью и достоверностью, что обусловило закрепление перечня, приведенного в вышеизложенном ГОСТе [10].

Таким образом, большие данные начали свое формирование еще до активной фазы разработок ИИ. В существовании big data отсутствует направленность на исключительное обеспечение нужд интеллектуальных систем. Большие данные имеют масштабную экономическую ценность и активно применяются для расширенного анализа, позволяющего выявить незаметные ранее корреляции и тенденции, осуществить прогнозирование и выполнять иные действия. Так, например, А. Н. Гулемин полагает, что основной целью сбора и обработки больших данных выступает создание новой информации [11]. Отмеченное позволяет заключить, что big data представляет собой сложную структуру, характеризующуюся не только хранением, но и обработкой данных.

Анализ содержания понятия датасет (dataset)

Понятие «датасет» (dataset / data set) появилось и получило свое распространение за пределами Российской Федерации, однако ввиду важности характеризующего им компонента оно постепенно стало интегрироваться и в российскую действительность.

Сравнительно-правовой анализ зарубежных правовых порядков, проведенный О. В. Костиной, М. А. Слизовской и М. А. Гапоненко, демонстрирует разнородность подходов к определению рассматриваемого термина. В законодательстве США dataset прямо не определен, однако закреплено схожее понятие «информационный ресурс» (data asset) - совокупность элементов данных или наборов данных, которые могут быть сгруппированы вместе. В сингапурском регулировании отсутствует закрепление понятия «dataset», однако, устанавливая требования к данным, используемым для разработки моделей ИИ, законодательство обозначенной страны фактически признает существование датасета. Аналогично рассматриваемое понятие отсутствует среди нормативных

правовых актов Республики Корея, содержащих, однако, закрепление терминов «публичные данные» и «информация». Юрисдикция Европейского союза раскрывает содержание обозначенного понятия в актах, регулирующих узкие сферы, например, в миграционном направлении под датасетом понимается набор информации, записанный в базу данных отпечатков пальцев Европейского союза, соответствующий одному набору отпечатков пальцев субъекта данных и состоящий из биометрических данных, буквенно-цифровых данных и, при наличии, отсканированной цветной копии удостоверения личности или проездного документа [12]. Закон Республики Казахстан от 17.11.2025 № 230 VIII «Об искусственном интеллекте» не использует термин «датасет». Вместо него в статье 1 закреплена категория «библиотека данных» - «совокупность структурированных и (или) сгруппированных данных, пригодных для обучения моделей искусственного интеллекта». Отдельно назван «изготовитель библиотеки данных» - лицо, организовавшее сбор, обработку и структурирование таких данных. Следовательно, казахстанский законодатель связывает обучающие данные не с массивом информации в естественном состоянии, а с предварительно организованной, обработанной и структурированной совокупностью, пригодной для обучения модели ИИ. По содержанию эта конструкция близка к датасету, хотя сам термин «датасет» в законе прямо не закреплён.

Сопоставление отмеченных подходов выявляет общую закономерность: ни одна развитая юрисдикция нормативно не отождествляет обучение систем ИИ с big data в их естественном состоянии. Каждый из приведенных актов оперирует структурированной выборкой, по содержанию совпадающей с категорией датасета.

В российской правовой системе понятие dataset отражено и раскрыто в государственных стандартах. Пункты 3.3 ГОСТ Р 59925-2021 «Информационные технологии. Большие данные. Техническое задание. Требования к содержанию и оформлению» и 3.1.11 ГОСТ Р ИСО/МЭК 20546-2021 «Информационные технологии. Большие данные. Обзор и словарь» используют единый подход: датасет раскрывается как «идентифицируемая совокупность данных, доступная для получения или скачивания в одном либо нескольких форматах».

Близкая дефиниция содержится в пункте 3.2.5 ГОСТ Р 71476-2024 (ИСО/МЭК 22989:2022) «Искусственный интеллект. Концепции и терминология искусственного интеллекта». В указанном стандарте датасет определяется через два признака: идентифицируемость совокупности данных и представление таких данных в общем формате.

При этом обозначенные стандарты фактически синхронизируют три термина - «набор данных», «массив данных» и «датасет». Их содержание раскрывается как тождественное, что указывает на отсутствие самостоятельного технического различия между указанными категориями в терминологии стандартизации, то есть все три термина имеют одно значение, являясь одним понятием.

Иной уровень детализации используется в Национальной стратегии развития искусственного интеллекта на период до 2030 года, утвержденной Указом Президента Российской Федерации от 10.10.2019 № 490. Подпункт «д» пункта 5 Стратегии определяет набор данных через состав структурированных или сгруппированных по определенным признакам данных, которые соответствуют требованиям законодательства Российской Федерации и необходимы для разработки программ для ЭВМ на основе ИИ.

Категория «набор данных» применяется и за пределами цифровой сферы. Так, ГОСТ ISO 10993-1-2021 «Изделия медицинские. Оценка биологического действия медицинских изделий. Часть 1. Оценка и исследования в процессе менеджмента риска» связывает набор данных с информацией, полученной из различных источников и используемой для характеристики биологического ответа на медицинское изделие. В состав такой информации стандарт включает физические и химические характеристики, сведения о токсичности и иные данные.

Подобное использование подтверждает универсальность термина: вне зависимости от отрасли он обозначает структурированный массив сведений, подготовленный для решения определенной задачи.

Таким образом, датасет представляет собой совокупность структурированных для определенных целей данных. Между тем в российской правовой действительности понятие «dataset», сохраняя свое значение, обретает форму «набора данных» или «массива данных» - совокупности, предназначенной для обучения ИИ. Как следует из Национальной стратегии развития искусственного интеллекта на период до 2030 года, преимущество отдается первому из отмеченных терминов. Аналогичная логика прослеживается в проекте Федерального закона «Об основах государственного регулирования сфер применения технологий искусственного интеллекта в Российской Федерации» с ID 166424, подготовленного Минцифры России. В частях 5 и 7 статьи 2 законопроекта именно «набор данных» интерпретируется как основа для обучения ИИ.

Заключение

В естественном виде большие данные сами по себе не пригодны для непосредственной эксплуатации в целях обучения ИИ. Разработчики интеллектуальных систем производят предварительную систематизацию, отбор и приведение к единому формату совокупности материалов, входящих в big data. Именно результатом такой обработки является dataset. Соотношение между большими данными и датасетом строится по принципу «сырьевой источник - готовый продукт». Big data в контексте соотнесения их со вторым явлением представляют собой неупорядоченный массив сведений, тогда как датасет - подготовленный к машинной обработке набор, ориентированный на конкретную задачу.

Примером отмеченного является стадия аннотирования данных. Э. И. Лескина обращает внимание на связь качества наборов данных с их очисткой, классификацией, комментированием и иной обработкой, необходимой для последующего машинного обучения. Отсюда следует, что датасет формируется не самим фактом накопления информации, а через придание ей признаков, позволяющих алгоритму выявлять целевые зависимости. В этом смысле data set ближе к технологически подготовленной выборке, чем к массиву сведений, существующему до обработки [13].

Разграничение отмеченных понятий прослеживается и в государственных стандартах. Так, ГОСТ Р ИСО/МЭК 20546-2021 и ГОСТ Р 59925-2021 в разделе «Термины и определения» dataset и big data употребляются в качестве двух самостоятельных категорий.

Проведенный анализ позволяет сформулировать несколько итоговых положений:

- понятия «большие данные» и «датасет» имеют различное содержание и не подлежат отождествлению;
- big data представляет собой самостоятельное явление, объединяющее совокупность данных и систему средств их сбора, хранения и обработки. Обучение ИИ непосредственно осуществляется не на больших данных в их естественном состоянии, а на датасетах - структурированных, размеченных и подготовленных под конкретную задачу выборках, формируемых путем переработки материалов, входящих в big data. Сосуществование указанных категорий в нормативных актах требует их четкого разграничения, поскольку смешение терминов способно породить пробелы в режиме каждого из объектов и затруднить распределение прав участников оборота.

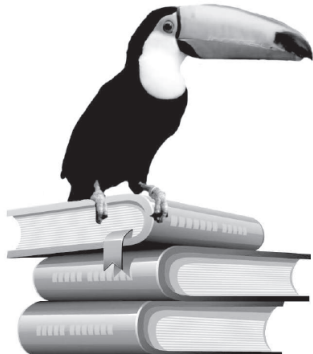
Список литературы:

- [1] Карпычев В. Ю. Правовое регулирование больших данных: пусть будет // Юрист. 2022. № 4. С. 68-73.
- [2] Лескина Э. И., Панарина М. М. Правовой режим данных в эпоху больших данных: правовые дилеммы и способы их разрешения // Закон. 2025. № 9. С. 91-100.
- [3] Андреасян А. А. Анализ подходов к определению больших данных в законодательстве Российской Федерации // Вестник Университета имени О. Е. Кутафина. 2024. № 10 (122). С. 168-175.
- [4] Право цифровой среды: монография / под ред. Т. П. Подшивалова, Е. В. Титовой, Е. А. Громовой. Москва, 2022. 896 с.
- [5] Лескина Э. И. Big Data как актив: задачи правового обеспечения оборота данных средствами публичного права // Законодательство. 2026. № 2. С. 22-29.
- [6] Володина Н. В. Дальнейшие перспективы развития цифрового права в современном обществе // Современное право. 2025. № 8. С. 5-10.
- [7] Василевская Л. Ю. Big data в механизме формирования основных направлений национального проекта «Экономика данных»: взгляд цивилиста на проблему // Lex Russica. 2024. № 1. С. 9-21.
- [8] Подузова Е. Б. Информационные системы: проблемы определения и толкования в свете нотариальной практики // Наследственное право. 2024. № 2. С. 16-19.
- [9] Ульянова Е. В. Big Data: технология, принципы и архитектура // Журнал Суда по интеллектуальным правам. 2020. № 30. С. 32-41.
- [10] Корбат Л. А. Большие данные (big data) как объект правового регулирования // Право и государство: теория и практика. 2024. № 2 (230). С. 159-162.
- [11] Гулемин А. Н. Пределы обработки больших объемов данных для целей получения информации о человеке: правовой аспект // Электронное приложение к Российскому юридическому журналу. 2022. № 6. С. 52-57.
- [12] Костина О. В., Слизовская М. А., Гапоненко М. А. Перспективы правового регулирования датасетов в Российской Федерации. Часть I // ИС. Промышленная собственность. 2026. № 2. С. 79-85.
- [13] Лескина Э. И. Правовое регулирование индустрии аннотации данных как способ обеспечения качества данных // Вестник ВГУ. Серия: Право. 2025. №3 (62). С. 64-70.

References:

- [1] Karpychev V. Yu. Legal regulation of big data: let it be//Lawyer. 2022. № 4. S. 68-73.
- [2] Leskina E. I., Panarina M. M. The legal regime of data in the era of big data: legal dilemmas and ways to resolve them//Law. 2025. № 9. S. 91-100.
- [3] Andreyanov A. A. Analysis of approaches to the definition of big data in the legislation of the Russian Federation//Bulletin of the O. E. Kutafin University. 2024. № 10 (122). S. 168-175.
- [4] The right of the digital environment: monograph/ed. T. P. Podshivalova, E. V. Titova, E. A. Gromova. Moscow, 2022. 896 p.
- [5] Leskina E. I. Big Data as an asset: tasks of legal support of data circulation by means of public law//Legislation. 2026. № 2. S. 22-29.
- [6] Volodina N.V. Further prospects for the development of digital law in modern society//Modern law. 2025. № 8. S. 5-10.
- [7] Vasilevskaya L. Yu. Big data in the mechanism for forming the main directions of the national project "Data Economics": the civilist's view of the problem//Lex Russica. 2024. № 1. S. 9-21.
- [8] Poduzova E. B. Information systems: problems of definition and interpretation in the light of notarial practice//Hereditary law. 2024. № 2. S. 16-19.
- [9] Ulyanova E.V. Big Data: technology, principles and architecture//Journal of the Court for Intellectual Rights. 2020. № 30. S. 32-41.
- [10] Corbat L. A. Big data as an object of legal regulation//Law and state: theory and practice. 2024. № 2 (230). S. 159-162.
- [11] Gulemin A. N. Limits of processing large amounts of data for the purpose of obtaining information about a person: legal aspect//Electronic Appendix to the Russian Legal Journal. 2022. № 6. S. 52-57.
- [12] Kostina O. V., Slizovskaya M. A., Gaponenko M. A. Prospects for the legal regulation of datasets in the Russian Federation. Part I//IE. Industrial property. 2026. № 2. S. 79-85.
- [13] Leskina E. I. Legal regulation of the data annotation industry as a way to ensure data quality//Vestnik VSU. Series: Right. 2025. №3 (62). S. 64-70.





ЮРКОМПАНИ

www.law-books.ru

Юридическое издательство
«ЮРКОМПАНИ»

Издание учебников,
учебных и методических
пособий, монографий,
научных статей.

Профессионально.

В максимально
короткие сроки.

Размещаем
в РИНЦ, E-Library.