УДК 342.5 ББК 67.400.6

К. А. КАРАМАНОВ
Российская государственная академия интеллектуальной собственности (РГАИС)

K. A. KARAMANOV
Russian State Academy of Intellectual Property (RGAIS)

АЛГОРИТМИЧЕСКАЯ ПРОЗРАЧНОСТЬ КАК ПРИНЦИП ПРАВОВОГО РЕГУЛИРОВАНИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

ALGORITHMIC TRANSPARENCY AS A PRINCIPLE OF LEGAL REGULATION OF ARTIFICIAL INTELLIGENCE

Аннотация. В статье рассматривается проблема алгоритмической прозрачности как ключевого принципа регулирования искусственного интеллекта в условиях цифровой трансформации публичного управления. Обосновывается необходимость перехода от формального раскрытия логики ИИ-систем к созданию механизмов, обеспечивающих реальную интерпретируемость и подотчётность принимаемых решений. Указывается, что в обеспечении прозрачности ИИ заинтересованы как государственные органы, так и граждане, для которых доступ к объяснению решений является основой правовой защищённости и доверия к цифровым сервисам. Особое внимание уделяется концепции контрфактуальных объяснений как способу достижения баланса между технической реализуемостью и юридической верифицируемостью алгоритмических решений. Анализируются этические, правовые и социальные риски, связанные с «чёрным ящиком» ИИ, включая угрозу алгоритмической предвзятости и дискриминации. Рассматриваются положения

Abstract. The article examines the issue of algorithmic transparency as a key principle in regulating artificial intelligence in the context of digital transformation of public administration. It substantiates the need to move from formal disclosure of AI system logic to the creation of mechanisms that ensure real interpretability and accountability of decisions made. It is noted that both public authorities and citizens are interested in ensuring AI transparency, as access to decision explanations forms the basis for legal protection and trust in digital services. Special attention is given to the concept of counterfactual explanations as a means of achieving a balance between technical feasibility and legal verifiability of algorithmic decisions. The article analyzes ethical, legal, and social risks associated with the «black box» nature of AI, including the threat of algorithmic bias and discrimination.

стратегических документов РФ, касающихся объяснимого искусственного интеллекта, а также зарубежные подходы к интерпретируемости. Целью исследования является обоснование алгоритмической прозрачности как фундаментального условия правового регулирования ИИ и предложения направлений по его институциональному закреплению. Методологическую основу исследования составляют общенаучные методы (анализ, синтез, индукция, дедукция), а также специальные юридические методы: формально-юридический анализ нормативных актов и сравнительно-правовой анализ зарубежных подходов к объяснимому ИИ. Сделан вывод о необходимости междисциплинарного подхода, включающего технические, правовые и этические элементы, для обеспечения подотчётности ИИ и защиты прав субъектов в цифровом правовом пространстве.

Ключевые слова: алгоритмическая прозрачность, объяснимый ИИ, цифровое правовое регулирование, подотчётность, предвзятость, правовая интерпретируемость,контрфактуальные объяснения.

It also reviews strategic documents of the Russian Federation related to explainable AI, as well as foreign approaches to interpretability. The purpose of the study is to justify algorithmic transparency as a fundamental condition for the legal regulation of AI and to propose directions for its institutional implementation. The methodological basis of the study includes general scientific methods (analysis, synthesis, induction, deduction), as well as specialized legal methods: formal legal analysis of regulatory acts and comparative legal analysis of international approaches to explainable AI. The study concludes that a multidisciplinary approach, encompassing technical, legal, and ethical elements, is necessary to ensure AI accountability and to protect the rights of individuals in the digital legal space.

Keywords: algorithmic transparency, explainable AI, digital legal regulation, accountability, bias, legal interpretability, counterfactual observation.

ВВЕДЕНИЕ

Внедрение систем искусственного интеллекта (ИИ) в публичные сферы — правоприменение, администрирование, оказание услуг населению — стало неотъемлемой частью цифровой трансформации современного общества. По мере усложнения технологий всё острее встаёт вопрос о прозрачности алгоритмов, лежащих в основе принимаемых решений, и, как следствие, о пределах и формах их правового регулирования.

Многочисленные положения, регламентирующие требования к прозрачности ИИ, разрабатываются и принимаются законодательными органами различных государств, отражая растущий интерес к обеспечению подотчётности и объяснимости цифровых систем, особенно в случаях, когда их действия затрагивают фундаментальные права человека[1].

Особую обеспокоенность вызывает то, что для обычного пользователя информационные технологии становятся всё менее контролируемыми и предсказуемыми. Люди часто не понимают, какие именно данные обрабатываются, каким образом происходит анализ и на основании каких критериев система принимает то или иное решение. В связи с этим формируется общественный и научный запрос на то, чтобы логика работы алгоритмов была открыта и доступна для понимания. Эти вызовы

[1] Arbeitsgruppe «Digitaler Neustart». Zwischenbericht der Arbeitsgruppe «Digitaler Neustart» zur Frühjahrskonferenz der Justizministerinnen und Justizminister am 6. und 7. Juni 2018 in Eisenach. // https://www.justiz.nrw/JM/justizpol_themen/digitaler_neustart (дата обращения: 25.03.2025).

способствовали формированию термина «алгоритмическая прозрачность» — сначала в области информатики, а затем и в юридической сфере, где он используется для описания принципов регулирования ИИ и других цифровых технологий, влияющих на обработку данных.

На уровне российского законодательства идея алгоритмической прозрачности получила закрепление в Национальной стратегии развития искусственного интеллекта до 2030 года, утверждённой Указом Президента РФ № 490 от 10 октября 2019 года[2]. В числе заявленных принципов стратегии присутствуют как объяснимость результатов работы ИИ, так и недискриминационный доступ к информации о механизмах функционирования таких систем. Прозрачность рассматривается как основа доверия пользователей к ИИ, а также как средство обеспечения прав и свобод граждан в условиях растущей цифровизации. Не случайно аналогичные требования к раскрытию логики алгоритмов предусмотрены и в международных актах, таких как Общий регламент Европейского союза по защите данных (GDPR)[3].

Современные системы, особенно в контексте взаимосвязанных устройств и датчиков, объединённых в единую цифровую сеть для сбора и обмена данными (интернет вещей), обрабатывают колоссальные массивы информации, включая персональные данные различных лиц. Это делает прозрачность не просто желательной, а критически необходимой характеристикой информационных технологий. Стремление обеспечить контроль за алгоритмами приводит к законодательным инициативам, направленным на раскрытие механизмов принятия решений и перечня обрабатываемых данных. Всё это требует не только правового осмысления, но и технологической и этической проработки.

Однако несмотря на присутствие термина «алгоритмическая прозрачность» в нормативных документах, его содержание остаётся размытым. В науке не существует единой позиции о границах применимости этого понятия: одни исследователи трактуют его крайне узко — применительно только к ИИ, другие рассматривают в более широком контексте цифровых платформ и автоматизированных решений. В результате появляется необходимость в комплексном исследовании сущности алгоритмической прозрачности: от технологических особенностей до нормативных рамок и этических ограничений.

Основными источниками, раскрывающими теоретико-правовые и технические аспекты алгоритмической прозрачности, стали работы Буррелл Дж., Ананни М., Кроуфорд К., Вахтера С., Миттельштадта Б., Флориди Л., в которых исследуются понятия «чёрного ящика», прозрачности и объяснимости алгоритмических решений. Технические ограничения и методы интерпретации ИИ-моделей подробно анализируются в трудах Самека В., Монтевона Г., Балдуззи Д. и др. Российские авторы — Мартынов А. В., Харитонова Ю. С., Воробьёва И. Б. — акцентируют внимание на правовых и этических аспектах применения ИИ в публичной сфере и правоприменительной практике. Исследователи, в основном, фокусируются либо на нормативном обосновании прозрачности как элемента справедливости и подотчётности, либо на технических аспек-

[2] Указ Президента Российской Федерации № 490 от 10.11.2019 (вместе с Национальной стратегией развития искусственного интеллекта на период до 2030 года). (2019). Собрание законодательства Российской Федерации, 41, ст. 5700 // http://www.kremlin.ru/acts/bank/44731 (дата обращения:25.03.2025)

[3] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 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). // https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679 (дата обращения: 30.03.2025)

тах интерпретируемости моделей машинного обучения. Однако значительная часть работ остаётся в рамках декларативного подхода, ограничиваясь анализом правовых норм и этических принципов без глубокого учёта специфики функционирования современных ИИ-систем.

Новизна настоящей работы состоит в комплексном теоретико-правовом осмыслении алгоритмической прозрачности с учётом реальных технических ограничений современных ИИ-моделей. В исследовании предпринимается попытка объединить подходы технических наук и юриспруденции для уточнения содержания данного принципа, обоснования его места в системе правового регулирования и выявления границ его применимости в правоприменительной деятельности. Это позволяет не только конкретизировать содержание понятия «алгоритмическая прозрачность», но и выработать предложения по его нормативному закреплению с учётом технологических реалий.

Таким образом, возникает необходимость в теоретико-правовом осмыслении алгоритмической прозрачности как правового принципа, подлежащего закреплению в системе регулирования ИИ. Важно определить, каким образом данный принцип может быть интегрирован в правовую систему и какова его роль в обеспечении справедливости, предсказуемости и законности функционирования интеллектуальных систем.

МАТЕРИАЛЫ И МЕТОДЫ ИССЛЕДОВАНИЯ

В исследовании применялись как универсальные, так и специальные юридические методы, обеспечившие комплексный подход к анализу алгоритмической прозрачности как принципа правового регулирования в сфере искусственного интеллекта.

Анализ использовался для выделения структурных элементов алгоритмической прозрачности и её правовой сущности. В ходе анализа были выявлены четыре ключевых аспекта, характерных для правового принципа: объяснимость, подотчётность, верифицируемость и недискриминационность. Это позволило обосновать прозрачность не только как техническое требование, но и как нормативную категорию.

Синтез способствовал объединению этических, технических и юридических позиций, что дало возможность сформировать целостную концепцию алгоритмической прозрачности как принципа публичного управления в условиях цифровизации, обладающего нормативной силой и способного влиять на регулирование.

Индукция применялась при обобщении эмпирического материала — положений стратегических и нормативных документов Российской Федерации, в которых алгоритмическая прозрачность фактически признаётся ценностью и ориентиром регулирования. Это подтверждает её эволюцию в сторону правового принципа.

Дедукция использовалась для выведения алгоритмической прозрачности из базовых правовых принципов, таких как законность, доступность правосудия и подотчётность власти, что позволило раскрыть её внутреннюю нормативную логику и показать её соответствие системе юридических принципов.

Формально-юридический метод был использован для системного анализа действующего законодательства и стратегических документов Российской Федерации в сфере цифрового регулирования и искусственного интеллекта, включая Национальную стратегию развития искусственного интеллекта. Анализ выявил отсутствие нормативной определённости ключевых понятий.

Сравнительно-правовой метод позволил провести правовой анализ зарубежного нормативного материала, демонстрирующего высокую степень проработанности вопросов прозрачности алгоритмических решений. В частности, в рамках анали-

за положений Регламента (ЕС) 2016/679 (GDPR) были исследованы механизмы обеспечения права на получение «значимой информации о логике, применяемой в автоматизированной обработке данных», что непосредственно соотносится с концептом алгоритмической объяснимости и прозрачности. Особое внимание уделено также докладу рабочей группы «Digitaler Neustart» (Германия, 2018), в котором обоснована необходимость внедрения процедур объяснения решений, принимаемых с участием ИИ, в правовую систему, включая механизм контрфактуальных объяснений.

Таким образом, применённые методы позволили рассматривать алгоритмическую прозрачность не просто как характеристику ИИ-систем, а как принцип современного правового регулирования, обладающий нормативной значимостью и способствующий обеспечению законности, предсказуемости и справедливости в цифровом управлении.

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Результаты проведённого анализа подтверждают наличие системной проблемы, связанной с непрозрачностью современных моделей искусственного интеллекта, в частности нейросетевых архитектур. На протяжении нескольких десятилетий системы ИИ базировались на чётко определённых правилах и логике, которую можно было проследить и объяснить. Такие подходы, как экспертные системы и деревья решений, обеспечивали интерпретируемость и понимание как для специалистов, так и для конечных пользователей.

Однако с переходом к моделям глубокого обучения, использующим многослойные нейронные сети и большие объёмы данных, акцент сместился в сторону эффективности и точности за счёт потери прозрачности [1]. Эти архитектуры функционируют по принципу скрытых параметров и нелинейных зависимостей, что затрудняет прямое объяснение причин, по которым принимается то или иное решение. Это явление получило наименование «проблема чёрного ящика» — высокая результативность моделей сочетается с невозможностью верифицировать или интуитивно понять их выводы.

В ходе анализа выявлены несколько ключевых факторов, способствующих снижению алгоритмической прозрачности:

1) Нелинейная и многомасштабная природа внутренних представлений нейросетей [2].

Поскольку нейросеть не просто складывает и умножает числа, а делает это по сложным, «изогнутым» законам, которые нельзя выразить простой формулой, один и тот же вход может привести к разным результатам в зависимости от контекста. Отдельные нейроны могут реагировать на локальные признаки, другие — на глобальные зависимости, что делает интерпретацию вклада отдельных компонентов модели затруднительной. Также Современные алгоритмы машинного обучения, особенно в случае нейросетевых архитектур, представляют собой не единый подход, а совокупность разнообразных методов — от простых статистических моделей до многоуровневых глубинных сетей. Из-за этого становится практически невозможно построить однозначную и интерпретируемую цепочку логических рассуждений между вводимыми данными и полученным результатом. В отличие от традиционных формального вывода или даже простых эвристик, в ИИ-моделях отсутствует возможность «логически» объяснить, почему на основе таких-то данных система выдала именно такое решение. Это особенно критично в юридическом контексте, где требуется точная реконструкция процесса принятия решений — как для целей апелляции, так и для установления законности или нарушений прав. Поскольку такие системы зачастую функционируют как «чёрные ящики», в которых вход и выход связаны нелинейной и неочевидной трансформацией, проверить

корректность или обоснованность результата постфактум затруднительно. Это ограничивает правоприменительную проверку, ставит под угрозу принцип подотчётности и делает невозможной ретроспективную реконструкцию решений по историческим логам, особенно в случае использования различных техник и моделей в рамках одной системы[4].

2) Глубина архитектуры и эффект «разбитого градиента» (shattered gradient problem) [3].

В нейросетях, особенно в глубоких (deep learning) моделях, глубина архитектуры означает количество слоёв, через которые проходит информация в процессе её обработки. Чем больше слоёв в сети, тем сложнее становится процесс обучения и интерпретации влияния каждого компонента на результат. Когда сеть имеет множество слоёв, градиент, который вычисляется в обратном направлении, может ослабевать на каждом шаге, проходя через слои, и в конечном итоге стать очень маленьким (или «разбитым»), когда достигает начальных слоёв. Это приводит к тому, что в конечных слоях нейросети практически не происходит обновления весов, потому что градиенты слишком малы для того, чтобы существенно изменить параметры. Этот процесс называется «проблема исчезающих градиентов» (vanishing gradient problem). В других случаях, наоборот, градиенты могут расти слишком быстро и становиться слишком большими, что приводит к нестабильности обучения. Это называется «взрывные градиенты» (exploding gradients).

3) Проблема «состязательных-примеров» (adversarial examples) — малозаметные для человека изменения во входных данных, способные полностью изменить выход модели [3]. В рамках «состязательной-атаки» можно модифицировать текстовые данные таким образом, чтобы система неправильно интерпретировала ситуацию, несмотря на то что для человека эти изменения будут практически незаметны. Например, система может ошибочно классифицировать человека как более склонного к правонарушению. Это подчёркивает уязвимость современных ИИ и ставит под сомнение их возможность их массового применения в публичной сфере.

С юридической точки зрения отсутствие алгоритмической прозрачности порождает ряд значимых правовых проблем, препятствующих эффективному нормативному регулированию и защите прав граждан:

1) Нарушение принципа юридической определённости и прозрачности: невозможность восстановления причинно-следственной связи между входными параметрами и финальным решением ИИ-системы затрудняет как проведение независимого аудита, так и последующую юридическую оценку действий таких систем. Это прямо затрагивает возможность реализации принципов правовой предсказуемости и подотчётности.
2) Отсутствие верифицируемости — нет возможности проверить, были ли соблюдены правила, стандарты или логика, принятая в законодательстве или внутренней политике организации.
3) Угроза дискриминации и предвзятости: использование скрытых закономерностей в обучающих выборках, недоступных для анализа и коррекции, может привести к воспроизводству и усилению социально-неприемлемых различий, противоречащих принципам равенства и недискриминации, закреплённым как в международных, так и в национальных правовых актах.

[4] Bundesanstalt für Finanzdienstleistungsaufsicht. Big Data trifft auf künstliche Intelligenz. Herausforderungen und Implikationen für Aufsicht und Regulierung von Finanzdienstleistungen. 2018. URL: https://www.bafin.de/SharedDocs/Downloads/DE/dl_bdai_studie.pdf;jsessionid=FBED7288D92E4337384780CF96C47F27.internet012?__blob=publicationFile&v=3 (дата обращения: 21.04.2025).

4) Отсутствие мотивировочной части решений: аналогично необходимости обоснования судебного акта, применение ИИ в сферах, затрагивающих права граждан, требует предоставления интерпретируемых и воспроизводимых объяснений. Отказ от такой мотивации делает невозможным правовую верификацию решений и нарушает принцип справедливости.
5) Правовая неопределённость в вопросе распределения ответственности: существующая модель правовой ответственности не приспособлена к ситуации, в которой решение принимается ИИ-системой, действующей автономно. Это порождает сложность в установлении субъекта ответственности — будь то разработчик, оператор или организация — и препятствует защите прав пострадавших лиц.

Стоит также отметить проблему конфликта между интерпретируемостью решений и уровнем подготовки пользователя. Сложность алгоритмов ИИ часто противоречит уровню подготовки конечного пользователя, особенно если речь идёт о гражданине, не обладающем специальными техническими или юридическими знаниями. Даже если предоставить подробное объяснение логики, лежащей в основе того или иного решения, оно может остаться непонятым или интерпретированным неверно. Возникает необходимость в таком формате объяснения, который был бы одновременно и корректным с точки зрения внутренней логики алгоритма, и доступным для восприятия неспециалиста.

Таким образом, алгоритмическая прозрачность может рассматриваться как новый правовой принцип, отражающий необходимость адаптации традиционных юридических механизмов к условиям цифровой среды. Как и иные принципы (законность, справедливость, правовая определённость), прозрачность выполняет функцию ориентира для нормотворчества, правоприменения и оценки правомерности деятельности цифровых систем.

Алгоритмическая прозрачность обладает рядом признаков правового принципа: она носит объективный характер, является инструментом оценки поведения субъектов, и направлена на обеспечение прав и свобод личности. Её реализация требует нормативного закрепления как в общих актах цифрового регулирования, так и в отраслевом законодательстве, например, в сфере административных процедур, судебного контроля, обработки персональных данных.

Следует подчеркнуть, что алгоритмическая прозрачность не сводится к технологической интерпретируемости моделей ИИ — в правовом смысле она означает возможность воспроизводимого и правомерного объяснения принятого решения, достаточного для последующего контроля, обжалования и установления ответственности. Таким образом, прозрачность должна выступать не только как технический, но и как юридически значимый стандарт, обеспечивающий соблюдение базовых принципов правового государства.

Таким образом, алгоритмическая прозрачность предстает не только как техническая характеристика, но и как самостоятельный правовой принцип, требующий системного внедрения в правовую практику и нормативную базу. Подтверждением этого служит подход, предложенный Мартыновым А. В. и Бундиным М. В., которые включают прозрачность в перечень ключевых принципов применения искусственного интеллекта в органах власти наряду с законностью, целесообразностью и безопасностью. Это свидетельствует о признании прозрачности как важного правового ориентира, необходимого для адаптации традиционных государственных процедур к вызовам цифровой эпохи и алгоритмического управления [4].

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ ИССЛЕДОВАНИЯ

Среди ученых указанных стран вопрос прозрачности ИИ порождает волну дебатов относительно технической осуществимости обеспечения прозрачности принятых ИИ решений. Роман Ямпольский указывает, что «если все, что у нас есть, — это «черный ящик», то невозможно понять причины сбоев и повысить безопасность системы, кроме того, если мы привыкнем принимать ответы искусственного интеллекта без объяснения причин, мы не сможем определять, если он начнет давать неправильные или манипулятивные ответы» [5].

Вопрос об обеспечении баланса между прозрачностью решений ИИ и их понятностью для широкого круга пользователей приобретает особую актуальность, особенно в случаях, когда получатель решения не обладает технической или юридической подготовкой. На данном фоне возникает дополнительная проблема, касающаяся доступности интерпретации решений ИИ для неспециалистов. Даже если такие решения сопровождаются техническим обоснованием, пользователь, не обладающий профильными знаниями, вряд ли сможет осмыслить, каким именно образом система пришла к тому или иному выводу. Это делает прозрачность скорее формальной, чем реальной [6,7].

Среди концептуальных решений данной проблемы в научной литературе выделяется идея контрфактуального объяснения. Согласно подходу, предложенному Вахтером, Миттельштадтом и Флориди, пользователю следует раскрывать лишь те факторы, которые повлияли бы на иной результат при изменении входных условий — таким образом, разъясняется не весь алгоритмический путь, а лишь ключевые моменты, существенные для понимания возможной альтернативы [8].

Такой способ представления информации обладает сразу несколькими достоинствами. Во-первых, он позволяет сфокусировать внимание пользователя на реально значимых аспектах, которые могли бы быть изменены для получения иного исхода. Во-вторых, контрфактуальные объяснения упрощают оспаривание решений в правовом порядке: поскольку указаны конкретные условия, влияющие на результат, они могут быть предметом анализа в судебном разбирательстве. В-третьих, такой подход снижает избыточную нагрузку на государственные органы: отпадает необходимость в раскрытии полной алгоритмической цепочки, особенно если в ней содержатся элементы коммерческой или государственной тайны.

Таким образом, применение модели контрфактуальных объяснений может стать одним из действенных инструментов в решении задачи повышения подотчетности и интерпретируемости ИИ-решений, особенно в правоприменительной практике и в рамках взаимодействия граждан с государственными цифровыми сервисами. Это особенно важно в условиях, когда пользователи зачастую не обладают достаточными знаниями для анализа технической стороны алгоритмов, но нуждаются в чётком понимании, каким образом можно повлиять на результат в будущем или оспорить текущий.

Говоря о вопросе предвзятости ИИ, важно отметить, что его использование в расследовательской деятельности требует строгого соблюдения этических норм. Как подчеркивает И. Б. Воробьева, применение ИИ в органах, таких как Следственный комитет и МВД России, должно сопровождаться разработкой специальных норм, направленных на ограничение алгоритмической предвзятости [9].

Ю. С. Харитонова и В. С. Савина также отмечают необходимость минимизации предвзятости алгоритмов путем конструирования оборота данных таким образом, чтобы исключить явные групповые признаки [10]. Однако, как они подчеркивают, техническая составляющая машинного обуче-

ния и его целевая функция могут вступать в определенный конфликт при реализации данного подхода.

Таким образом, обсуждаемые подходы — включая модель контрфактуальных объяснений — представляют собой не только технический механизм повышения интерпретируемости, но и потенциальную форму реализации принципа алгоритмической прозрачности в правовой практике. С позиции теории права, алгоритмическая прозрачность должна рассматриваться как новый нормативный ориентир в условиях цифровизации государственного управления. Она выполняет функцию обеспечения подотчётности алгоритмических систем, гарантий правового контроля и реализации прав личности на доступ к информации о значимых для него решениях.

В условиях отсутствия прямой нормативной регламентации алгоритмического принятия решений в ряде правовых систем, включение принципа прозрачности в рамки правового регулирования становится необходимым шагом. Это позволит перейти от этико-технического регулирования к формированию юридических стандартов объяснимости, включая обязательность контрфактуального разъяснения при использовании ИИ в публичной сфере. Принцип прозрачности в данном контексте выступает как гарантия законности, предсказуемости последствий и возможности юридической защиты.

ЗАКЛЮЧЕНИЕ

Решение вопроса «черного ящика» является одной из приоритетных задач для органов власти, для решения которой необходимы систематическая работа в указанной сфере, совершенствование как понятийного аппарата, так и учет философско-этических аспектов взаимодействия ИИ и человека, а также преодоление вопросов технической грамотности и оснащенности в области ИИ. Когда ИИ используется в контекстах, связанных с принятием важных решений, общество заинтересовано в том, чтобы понимать, как работает система, подвергать сомнению её результаты и требовать обоснования решений, принятых с помощью ИИ [8].

Алгоритмическая прозрачность становится одной из ключевых ценностей в развитии ИИ. Она определяется не только как доступ к внутренней логике модели, но и как способность представлять выводы в форме, понятной для специалистов и неспециалистов. Это понятие объединяет в себе:

Техническую объяснимость — способность ИИ-систем предоставлять понятные объяснения своих решений,

Этическую подотчётность — обеспечение ответственности разработчиков и пользователей ИИ за последствия его применения, включая возможность отслеживания и исправления ошибок и предвзятостей в алгоритмах.

Социальную справедливость - гарантия того, что ИИ-системы не увековечивают существующие социальные предвзятости и дискриминацию, способствуя равному обращению с людьми.

И правовую верифицируемость решений ИИ-систем — возможность юридической проверки решений, принимаемых ИИ, что требует прозрачности и документирования процессов, лежащих в основе этих решений.

На данном этапе развития машинного обучения можно выделить как его плюсы, так и минусы. Среди недостатков особенно стоит отметить риск появления поведенческой адаптации или искажённой интерпретации данных, возникающий в результате возможного вмешательства злоумышленников в открытые базы данных.

Анализ действующих нормативных документов, включая Национальную стратегию развития искусственного интеллекта до 2030 года, а также концепции цифрового регулирования в Российской Федерации, демонстрирует, что проблема обеспечения прозрачности алгоритмических систем при-

знана на уровне государственной политики как одна из ключевых. В документах подчёркивается необходимость разработки комплексных подходов к созданию объяснимого искусственного интеллекта (explainable artificial intelligence, XAI), внедрения унифицированных стандартов интерпретируемости, а также эффективных механизмов пользовательского контроля и верификации принимаемых ИИ решений[5].

Решение проблемы «чёрного ящика» в использовании искусственного интеллекта в публичной сфере требует не отказа от технологий, а их осознанной интеграции с учётом юридических, этических и технических аспектов. Как отмечал Герберт Саймон [11], даже человеческие решения подвержены когнитивным и мотивационным предубеждениям, что делает применение ИИ потенциально более объективным инструментом при условии его правильной настройки и контроля, а это требует активного участия законодателей и юристов в формировании технологий, обеспечивающих понятные и обоснованные выводы ИИ, доступные как специалистам, так и широкой общественности.

Таким образом решение проблемы «чёрного ящика» в использовании искусственного интеллекта представляет собой одну из приоритетных задач для органов публичной власти, требующую системной работы в нормативной, технической и этико-правовой плоскости. В условиях цифровизации сферы государственного управления и правоприменения становится очевидным, что алгоритмическая прозрачность должна рассматриваться не только как технологическая характеристика, но и как самостоятельный правовой принцип, необходимый для обеспечения законности, подотчётности и соблюдения прав личности.

Алгоритмическая прозрачность — это не факультативная цель технического прогресса, а нормативно значимая категория, подлежащая закреплению в законодательстве и применению в правоприменительной практике. Она должна быть понята и реализована как универсальный принцип публичного управления в цифровую эпоху, обеспечивающий правомерность, предсказуемость и воспроизводимость решений, принимаемых с участием ИИ.

Анализ действующих стратегических документов, включая Национальную стратегию развития ИИ и концепции цифрового регулирования в РФ, подтверждает, что государство признаёт необходимость разработки правовых механизмов объяснимого ИИ (XAI), внедрения стандартов интерпретируемости и формирования механизмов юридической ответственности.

Следовательно, достижение алгоритмической прозрачности требует не отказа от технологии, а её правового осмысления и институционализации. Прозрачность как правовой принцип — основа устойчивого и легитимного внедрения ИИ в социотехнические системы. Её реализация невозможна без активного участия юристов, законодателей и правоприменителей в формировании регулирующих стандартов, обеспечивающих как защиту прав личности, так и доверие к цифровым государственным институтам.

[5] Gunning, D. (2017). Explainable artificial intelligence (XAI). Defense advanced research projects agency (DARPA) // https://nsarchive.gwu.edu/sites/default/files/documents/5794867/National-Security-Archive-David-Gunning-DARPA.pdf (дата обращения:1.04.2025)

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Tutt A. An FDA for algorithms // Administrative Law Review.* 2017, Vol. 69, P. 83. — DOI: 10.2139/ssrn.2747994.
2. *Samek W., Montavon G., Lapuschkin S., Anders C. J., Müller K. R.* Explaining deep neural networks and beyond: A review of methods and applications // Proceedings of the IEEE. 2021. Vol. 109, No. 3. P. 247–278. — DOI: 10.1109/jproc.2021.3060483.
3. *Balduzzi D., Frean M., Leary L., Lewis J. P., Ma K. W. D., McWilliams B.* The shattered gradients problem: If resnets are the answer, then what is the question? // International Conference on Machine Learning. — 2017, P. 342–350. — DOI: 10.48550/arXiv.1702.08591.
4. *Мартынов А. В.*, Бундин М. В. О правовых принципах применения искусственного интеллекта при осуществлении органами исполнительной власти контрольно-надзорной деятельности // Журнал российского права. — 2020. — № 10 — С. 59–75. — DOI: 10.12737/jrl.2020.120.
5. *Yampolskiy R. V.* Unexplainability and incomprehensibility of artificial intelligence. 2019. — DOI: 10.48550/arXiv.1907.03869.
6. *Burrell J.* How the machine «thinks": Understanding opacity in machine learning algorithms // Big Data & Society. 2016, Vol. 3, No. 1, P. 1–12. — DOI: 10.1177/2053951715622512.
7. *Ananny M., Crawford K.* Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability // New Media & Society. 2018. Vol. 20, No. 3, P. 973–989. DOI: 10.1177/1461444816676645.
8. *Wachter S., Mittelstadt B., Floridi L.* Counterfactual explanations without opening the black box: Automated decisions and the GDPR // Harvard Journal of Law & Technology. — 2017, Vol. 31, No. 2. — DOI: 10.2139/ssrn.3063289.
9. *Воробьёва И. Б.* Этические аспекты использования систем искусственного интеллекта при расследовании преступлений // Вестник Саратовской государственной юридической академии. 2022. — № 4 (147). — С. 162–172. — DOI: 10.24412/2227-7315-2022-4-162-172.
10. *Харитонова Ю. С., Савина В. С., Паньини Ф.* Предвзятость алгоритмов искусственного интеллекта: вопросы этики и права // Вестник Пермского университета. Юридические науки. — 2021. — № 53. — С. 488–515. — DOI: 10.17072/1995-4190-2021-53-488-515.
11. *Simon H. A.* Administrative Behavior. New York: Simon and Schuster, — 2016. — 384 p. — DOI 10.1093/oxfordhb/9780199646135.013.22.

REFERENCES

1. *Tutt A.* An FDA for Algorithms. // Administrative Law Review. 2017. Vol. 69. P. 83. DOI: 10.2139/ssrn.2747994.
2. *Samek W., Montavon G., Lapuschkin S., Anders C. J., Müller K. R.* Explaining deep neural networks and beyond: A review of methods and applications. // Proceedings of the IEEE. 2021. Vol. 109, No. 3. P. 247–278. DOI: 10.1109/jproc.2021.3060483.
3. *Balduzzi D., Frean M., Leary L., Lewis J. P., Ma K. W. D., McWilliams B.* The shattered gradients problem: If ResNets are the answer, then what is the question? // International Conference on Machine Learning. 2017. P. 342–350. DOI: 10.48550/arXiv.1702.08591.
4. *Martynov A. V., Bundin M. V.* O pravovykh printsipakh primeneniya iskusstvennogo intellekta pri osushchestvlenii organami ispolnitel'noy vlasti kontrol'no-nadzornoy deyatel'nosti [On legal principles of using artificial intelligence in supervisory activities of executive authorities] // Journal of Russian Law– 2020. – No. 10. – P. 59–75. DOI: 10.12737/jrl.2020.120.
5. *Yampolskiy R. V.* Unexplainability and incomprehensibility of artificial intelligence. — 2019. — DOI: 10.48550/arXiv.1907.03869.
6. *Burrell J.* How the machine «thinks": Understanding opacity in machine learning algorithms. // Big Data & Society. — 2016, Vol. 3, No. 1, P. 1–12. — DOI: 10.1177/2053951715622512.
7. *Ananny M., Crawford K.* Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. // New Media & Society. — 2018, Vol. 20, No. 3, P. 973–989. DOI: 10.1177/1461444816676645.

8. *Wachter S., Mittelstadt B., Floridi L.* Counterfactual explanations without opening the black box: Automated decisions and the GDPR. // Harvard Journal of Law & Technology. — 2017. Vol. 31, No. 2. — DOI: 10.2139/ssrn.3063289.
9. *Vorob'yova I. B.* Ehticheskie aspekty ispol'zovaniya sistem iskusstvennogo intellekta pri rassledovanii prestupleniy [Ethical aspects of using artificial intelligence systems in criminal investigations] // Bulletin of the Saratov State Law Academy — 2022, No. 4 (147), P. 162–172. — DOI: 10.24412/2227-7315-2022-4-162-172.
10. *Kharitonova Yu. S., Savina V. S., Panini F.* Predvzyatost' algoritmov iskusstvennogo intellekta: voprosy ehtiki i prava [Bias of artificial intelligence algorithms: issues of ethics and law] // Perm University Herald. Juridical Sciences — 2021, No. 53, P. 488–515. — DOI: 10.17072/1995-4190-2021-53-488-515.
11. *Simon H. A.* Administrative Behavior. New York: Simon and Schuster, 2016. 384 p. DOI 10.1093/oxfordhb/9780199646135.013.22

Информация об авторе

КАРАМАНОВ КОНСТАНТИН АНДРЕЕВИЧ
Аспирант Юридического факультета, Российская государственная академия интеллектуальной собственности (РГАИС), Москва, Российская Федерация
e-mail: KonstantKar@yandex.ru

Научный руководитель
АЛЬБОВ АЛЕКСЕЙ ПАВЛОВИЧ
доктор юридических наук, профессор, профессор кафедры гражданского и предпринимательского права, Российская государственная академия интеллектуальной собственности (РГАИС)

About the author

KONSTANTIN A. KARAMANOV
Postgraduate student of the Faculty of Law, Russian State Academy of Intellectual Property (RGAIS), Moscow, Russian Federation
ORCID:0009-0006-6121-1380

Scientific supervisor:
ALEKSEY PAVLOVICH ALBOV
Doctor of Law, Professor of the Department of Civil and Business Law, Russian State Academy of Intellectual Property (RGAIS)

Статья поступила в редакцию 18.05.2025; одобрена после рецензирования 08.06.2025; принята к публикации 21.06.2025.

Автор прочитал и одобрил окончательный вариант рукописи.

The article was submitted 18.05.2025; approved after reviewing 08.06.2025; accepted for publication 21.06.2025.

The author read and approved the final version of the manuscript.