Право. Журнал Высшей школы экономики. 2025. Том 18. № 3.
Law. Journal of the Higher School of Economics. 2025. Vol. 18, no 3.

Российское право: состояние, перспективы, комментарии

Научная статья
УДК: 342.5, 342.7
JEL: K23, K38
DOI:10.17323/2072-8166.2025.3.28.55

Прозрачность государственного управления в условиях автоматизированного принятия решений

Павел Петрович Кабытов[1], Никита Алексеевич Назаров[2]

[1, 2] Институт законодательства и сравнительного правоведения при Правительстве Российской Федерации, Россия 117218, Москва, Большая Черемушкинская ул., д. 34,

[1] kapavel.v@yandex.ru, https://orcid.org/0000-0001-8656-5317

[2] naznikitaal@gmail.com, https://orcid.org/0000-0002-3997-0886

Аннотация

В условиях активного внедрения автоматизированных систем принятия решений и систем искусственного интеллекта в деятельность органов публичной власти актуализируется проблема поддержания надлежащего уровня прозрачности государственного управления, имеющая критическое значение для соблюдения принципов верховенства права и защиты фундаментальных прав граждан. Настоящая работа ставит целью провести комплексную систематизацию и критический анализ сложившихся в российском и зарубежном праве, а также в правовой теории подходов к решению указанной проблемы. Методологическую основу исследования составили общенаучные (анализ, синтез, системный подход) и частные (сравнительно-правовой, формально-юридический) методы изучения. В статье рассматриваются концептуальные основания и практические вызовы реализации требований прозрачности и объяснимости автоматизированного принятия решений и систем искусственного интеллекта, включая их роль в повышении доверия, обеспечении подотчетности, предотвращении дискриминации и укреплении легитимности публичного управления. Основное внимание уделено критическому

анализу широкого спектра механизмов прозрачности (классифицируемых, в частности, по направленности на систему в целом или на отдельное решение, а также по моменту сообщения информации — *ex ante* или *ex post*): раскрытие порядка или логики принятия решений, «право на объяснение», контрфактологические объяснения, раскрытие данных и программного кода/модели, аудит и общественный контроль, информирование о применении, а также использование объяснимых/интерпретируемых моделей и иных технических решений. По каждому механизму выявлены преимущества, недостатки и трудности реализации — конфликт с защитой интеллектуальной собственности, техническая затрудненность имплементации и интерпретации, «эффект Расёмона» и фундаментальная проблема «черного ящика» систем искусственного интеллекта. Обосновывается вывод о недостаточности применения отдельных инструментов и необходимости разработки гибкого, риск-ориентированного и контекстуально-зависимого комплексного подхода. Подчеркивается актуальность имплементации адаптированных и системно увязанных механизмов в российское законодательство для поддержания надлежащего уровня прозрачности в условиях автоматизации государственного управления в России.

Ключевые слова

автоматизированное решение; система искусственного интеллекта; прозрачность государственного управления; право на объяснение; контрфактологические объяснения; алгоритмическая подотчетность.

Благодарности: Исследование выполнено за счет гранта Российского научного фонда № 23-78-01254, https://rscf.ru/project/23-78-01254/.

Для цитирования: Кабытов П.П., Назаров Н.А. Прозрачность государственного управления в условиях автоматизированного принятия решений // Право. Журнал Высшей школы экономики. 2025. Том 18. № 3. С. 28–55. DOI:10.17323/2072-8166.2025.3.28.55

Russian Law: Conditions, Perspectives, Comments

Research article

Transparency of Public Administration in Context of Automated Decision-Making

Pavel P. Kabytov[1], Nikita A. Nazarov[2]

[1, 2] Institute of Legislation and Comparative Legal Studies under Government of the Russian Federation, 34 Bolshaya Cheremushkinskaya Str., Moscow 117218, Russia,

[1] kapavel.v @yandex.ru, https://orcid.org/0000-0001-8656-5317

[2] naznikitaal@ gmail.com.https: //orcid.org/0000-0002-3997-0886

Abstract

In context of active implementation of automated systems decision-making and artificial intelligence systems into activities of public authorities, a problem of maintaining an adequate level of transparency in public administration is becoming increasingly relevant. The issue is critical for upholding principles of rule of law and protecting fundamental rights of citizens. The work aims to conduct a comprehensive systematization and critical analysis of current approaches to solving the problem in Russian and foreign law, as well as in legal theory. The methodological basis of the research includes general research methods (analysis, synthesis, systematic approach) and specific scholar methods (comparative legal, formal legal). The article consistently examines the conceptual foundations and practical challenges of implementing transparency and explainability requirements for automated systems decision-making and artificial intelligence systems, including their role in increasing trust, maintaining accountability, preventing discrimination, and strengthening legitimacy of public administration. The main attention is paid to a detailed and critical analysis of a wide range of transparency mechanisms (classified, in particular, according to their focus on the system as a whole or on a specific decision, as well as by the timing of information provision — *ex ante* or *ex post*): disclosure of the procedure or logic of decision-making, the «right to explanation», counterfactual explanations, disclosure of data and program code/models, audit and public control, information about application, as well as use of explainable/interpretable models and other technical solutions. For each mechanism, advantages, disadvantages, and difficulties of practical implementation are identified like conflicts with intellectual property protection, technical complexity of implementation and interpretation, and the fundamental «black box» problem of artificial intelligence systems. The conclusion substantiates the insufficiency of applying individual tools and the necessity of developing a flexible, risk-oriented, and context-dependent comprehensive approach.

Keywords

automated decision; artificial intelligence system; transparency of public administration; right to explanation; counterfactual explanations; algorithmic accountability.

Acknowledgments: The study was carried out with a grant from the Russian Scientific Foundation 23-78-01254, https://rscf.ru/project/23-78-01254/.

For citation: Kabytov P.P., Nazarov N.A. (2025) Transparency of Public Administration in Context of Automated Decision-Making. *Law. Journal of the Higher School of Economics,* vol. 18, no. 3, pp. 28–55 (in Russ.) DOI:10.17323/2072-8166.2025.3.28.55

Введение

Последнее десятилетие во всем мире характеризуется одной из ключевых тенденций в развитии государственного управления — количественным и качественным расширением применения «алгоритмов», систем «искусственного интеллекта» при принятии органами публичной власти юридически значимых или оказывающих

иное существенное влияние решений. Точкой приложения перечисленных технологий в деятельности органов власти выступают правотворческая и правоприменительная деятельность, а также отправление правосудия.

Для описания новых явлений правовой действительности в юридический словарь вошло понятие «автоматизированное принятие решений». Использование таких систем — принципиально новая тенденция в деятельности органов публичной власти. Хотя автоматизация государственного управления, использование различных аналитических систем и систем поддержки принятия решений начались еще в середине XX века, их первоначальное влияние на конечный результат управленческих решений, способы и характер взаимодействия органов публичной власти и адресатов их решений оставалось незначительным. Они были нацелены на информационное обслуживание органов власти, фактически формировали первичную информационную базу последующего принятия решения госслужащими. Сохранялись классическая схема взаимодействия органов публичной власти и адресатов их решений, высокий уровень подконтрольности и проверки информационных процессов. Уже тогда в доктрине ставились принципиальные вопросы о разграничении ответственности при использовании автоматизированных систем в государственном управлении, о правовых последствиях ошибок таких систем [Венгеров А.Б., 1979: 88, 245], которые в последние два десятилетия вновь привлекли внимание ученых-юристов.

Повышенный интерес вызван как очередным этапом развития технологий, именуемых искусственным интеллектом (далее — ИИ), так и принципиальным увеличением их влияния на ландшафт социального взаимодействия, закономерностей возникновения, изменения и прекращения правоотношений, их сущность и содержание. Применительно к сфере государственного управления такие изменения наблюдаются, во-первых, в части повсеместной имплементации в деятельность органов публичной власти автоматизированных систем принятия решений, предназначенных для поддержки принятия решений, формирования рекомендаций или автоматического принятия решений на основе имеющихся в распоряжении органов публичной власти данных. Полномочия принятия правотворческих, правоприменительных и судебных решений, ранее осуществляемые исключительно уполномоченными должностными лицами от лица органа публичной власти и судьями, все чаще частично или полностью «делегируется» информационным системам.

В судебных системах США и КНР тестируются и применяются рекомендательные и прогностические системы, призванные не только выработать предварительное решение по делу, но и оценить фактические обстоятельства дела, вероятность рецидива[1]. Органы исполнительной власти Канады[2], Австралии[3], стран Европейского союза[4], России[5] активно используют и продолжают имплементировать автоматизированные системы принятия решений.

Во-вторых, в части расширения использования в автоматизированных системах принятия решений технологий, именуемых ИИ, характеризующихся чрезвычайной технической сложностью и непредсказуемостью функционирования. Обработка информации, аналитика данных такими системами и принятые по их итогам автоматизированных решений характеризуются невоспроизводимостью, необъяснимостью, ограниченной подконтрольностью и проверяемостью для рядовых служащих и должностных лиц.

В результате с одной стороны сокращаются затраты временных, трудовых и иных ресурсов для реализации функций органов публичной власти, ускоряется и упрощается получение государственных услуг, снижаются административная нагрузка на субъектов предпринимательства и коррупционные риски при автоматизации административных процедур. С другой стороны, принципиально изменяется схема взаимодействия органов публичной власти и адресатов их решений (между органом публичной власти и адресатом решения появляется «посредник» в виде информационной системы), процедуры принятия решений, размываются базовые принципы их взаимодействия. На фоне ускоренного внедрения информационных технологий государственного управления право и его доктрина в силу

[1] Available at: URL: https://web.archive.org/web/20241215031419/https://www.chinacourt.org/article/detail/2024/11/id/8215844.shtml (дата обращения: 07.05.2025); Humanizing Justice: The transformational impact of AI in courts, from filing to sentencing. 2024. Available at: URL: https://www.thomsonreuters.com/en-us/posts/ai-in-courts/humanizing-justice/ (дата обращения: 24.05.2025)

[2] Directive on Automated Decision-Making. Available at: URL: https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592 (дата обращения: 11.12.2023)

[3] Automated decision-making better practice guide. Available at: URL: https://apo.org.au/node/306481 (дата обращения: 06.04.2024)

[4] Gesetz über die Möglichkeit des Einsatzes von datengetriebenen Informationstechnologien bei öffentlich-rechtlicher Verwaltungstätigkeit vom 16. März 2022. Available at: URL: https://www.gesetze-rechtsprechung.sh.juris.de/bssh/document/jlr-ITEGSHpP1 (дата обращения: 10.12.2023)

[5] Федеральный закон от 02.10.2007 № 229-ФЗ (ред. от 23.11.2024) «Об исполнительном производстве» // СЗ РФ. 2007. № 41. Ст. 4849.

своего традиционного отставания от развития общественных отношений не сформулировали адекватного ответа на все многообразие рисков, порождаемых внедрением автоматизированных систем принятия решений в государственное управление.

Исходной точкой преодоления всего многообразия последствий внедрения автоматизированных систем в деятельность органов власти, включая их использование при принятии юридически значимых решений, выступает разрешение проблемы прозрачности и объяснимости. Речь идет об объяснимости как функционирования и результатов работы самих автоматизированных систем, так и принятых с их помощью юридически значимых решений, что имеет значение как для адресатов этих решений, так и для государственных служащих и должностных лиц. При этом самой природе многих систем ИИ присущи непрозрачность и необъясненность (проблема «черного ящика» [6]), что вступает в диалектическое противоречие с принципом открытости деятельности органов публичной власти — одним из ее базовых принципов деятельности, прямо вытекающим из конституционных установлений.

Именно это противоречие предопределило повышенную активность в доктринальном освоении проблематики прозрачности государственного управления в условиях автоматизированного принятия решений и использования систем ИИ, а также усилия теории и практики по идентификации оптимальных методов и средств поддержания надлежащего уровня прозрачности в деятельности органов публичной власти. Некоторые из таких предложенных наукой методов и средств уже получили закрепление в законодательстве отдельных зарубежных правопорядков (Канада, Австралия, Германия).

Между тем российская правовая доктрина только приступает к освоению проблематики объяснимости и прозрачности государственного управления: ставится вопрос о нормативном закреплении принципа прозрачности ИИ [Талапина Э.В., 2024: 36–39]. Однако полноценная систематизация знаний о методах и средствах поддержания надлежащего уровня прозрачности в деятельности органов публичной власти в условиях широкого внедрения автоматизированных процедур принятия решений и использования систем ИИ пока не осуществлена. При этом фактическое использование органами публичной власти автоматизированных систем принятия ре-

[6] См.: AI's mysterious 'black box' problem, explained | University of Michigan-Dearborn. Available at: URL: https://umdearborn.edu/news/ais-mysterious-black-box-problem-explained (дата обращения: 04.04.2025)

шений, обеспечивающих принятие юридически значимых решений в полностью автоматизированном режиме, предопределяет обязательную имплементацию в позитивное право механизмов поддержания надлежащего уровня прозрачности. Такое положение дел обусловливает целесообразность и актуальность систематизации уже сформулированных в научной доктрине воззрений, концепций и подходов к проблематике обеспечения прозрачности в деятельности органов публичной власти при использовании указанных технологий.

1. Объяснимость и прозрачность как основополагающие требования к процедуре автоматизированного принятия решений

В иностранных правопорядках сложилась однозначная позиция: требования объяснимости и прозрачности процедуры автоматизированного принятия решений в государственном управлении закреплены в качестве основополагающих. Сначала эти требования нашли отражение в более чем 73-х этических руководствах для ИИ; в дальнейшем одни иностранные правопорядки включили автоматизированные решения в закон об административных процедурах и распространили на них все обязательные требования, в том числе требования прозрачности и объяснимости, а другие правопорядки приняли специальные нормативные акты, в которых также закрепили эти требования.

При этом, как отмечают исследователи, в юридический лексикон входит понятие демократической транспарентности, претендующее на статус фундаментальной характеристики современной демократии. Помимо семантических нюансов данного понятия, транспарентность может быть определена как информационная политика государства, построенная на принципах открытости и доступности гражданам информации о деятельности органов публичной власти и лиц, их представляющих [Пилипенко А.Н., 2019: 204].

Именно виду отсутствия прозрачности и объяснимости голландский суд запретил использование технологии «SyRI» для борьбы с мошенничеством в области социального обеспечения и незаконных схем, связанных с доходами, налогами и отчислениями на социальное страхование, а также в сфере трудового законодательства[7]. Суд в обоснование решения указал, что в отсутствие достаточной

[7] ECLI:NL: RBDHA: 2020:865. Available at: URL: https://deeplink.rechtspraak.nl/uitspraak?id=ECLI:NL:RBDHA:2020:865 (дата обращения: 07.04.2025)

и прозрачной защиты права на уважение частной жизни возникает «сдерживающий эффект», когда без уверенности в надежной защите конфиденциальности граждане с меньшей вероятностью будут поставлять данные или это будет иметь меньшую поддержку. При этом важность прозрачности с точки зрения того насколько она поддается проверке, также велика, поскольку использование модели риска и анализа, проводимого в этом контексте, несет риск непреднамеренных дискриминационных последствий.

Между тем во многих исследованиях и правовых актах нет четкого различия между прозрачностью и объяснимостью в автоматизированных системах принятия решений. Кроме того, не сложился единый подход к пониманию их содержания как самостоятельных категорий. Так, в некоторых исследованиях под объяснением понимают разъяснение набора правил, в других — «дерево решений» (decision tree), а в третьих — прототип (особенно в контексте изображений); схожая неоднозначность в трактовках существует и в отношении содержания понятия «прозрачность» [Guidotti R., Monreale A. et al., 2019: 36]. Хотя вопрос о точном соотношении понятий «прозрачность» и «объяснимость» остается дискуссионным, в настоящем исследовании ввиду их схожей функциональной направленности как требований к процедуре автоматизированного принятия решений они используются преимущественно как взаимозаменяемые.

2. Правовые механизмы обеспечения объяснимости и прозрачности

Требования объяснимости и прозрачности автоматизированных решений получили закрепление в позитивном праве в целом ряде юрисдикций. Однако в нормативных актах правовые механизмы их обеспечения зачастую не закреплены, либо закреплены, но в преломлении правоприменительной практики определено, что они не обеспечивают надлежащего уровня охраны прав и законных интересов человека и гражданина. Во многом это свидетельствует о противоречиях между прозрачностью и объяснимостью как нормативным идеалом и его воплощении.

Между тем механизмы и подходы к объяснению и прозрачности различаются в зависимости от целевой направленности и способа реализации [Wachter S., Mittelstadt B., Floridi L., 2017: 78]. Так, требование прозрачности и объяснимости может быть направлено на:

функциональную часть системы, т.е. на логику, значимость, предполагаемые последствия и общую функциональность автоматизи-

рованной системы принятия решений, например, спецификация требований к системе, «деревья решений», заранее определенные модели, критерии и структуры классификации;

отдельные решения, т.е. обоснование, причины и индивидуальные обстоятельства каждого автоматизированного решения, например, взвешивание функций, определяемые машиной правила принятия решений для конкретного случая, информация о ссылочных или профильных группах.

По способу объяснимости и прозрачности существуют как: *ex ante* — объяснение, которое происходит ранее автоматического принятия решения; *ex post* — объяснение, сообщаемое после принятия автоматизированное решения. Тем самым объяснение *ex ante* направлено на объяснение функциональной части системы, а объяснение *ex post* позволяет понять как причины принятия решения, так и соответствующие аспекты функционирования системы.

Анализ зарубежной литературы и позитивного права позволил выявить следующие основные механизмы объяснения и прозрачности: a) раскрытие порядка или логики принятия решений; b) право на объяснение; c) контрфактологические объяснения; d) раскрытие данных, на основе которых принимаются автоматизированные решения; e) раскрытие программного кода и (или) модели ИИ; f) аудит и общественный контроль; g) информирование (раскрытие) о применении автоматизированной системы принятия решений; h) различные технические методы, в том числе объяснимые или интерпретируемые модели.

Ниже указанные подходы рассмотрены подробнее. Для удобства анализа они сгруппированы на преимущественно правовые (включая организационно-правовые) и технические, хотя многие из них носят комплексный, междисциплинарный характер, а их реализация зачастую требует сочетания нормативных предписаний, организационных мер и технологических решений.

2.1. Раскрытие информации о логике или порядке принятия решений в рамках законодательства о персональных данных

Большинство автоматизированных систем принятия решений функционируют на основе обработки персональных данных. Эта обработка по общему правилу должна соответствовать законодательству о персональных данных. Одна из первых попыток законодательного регулирования автоматизированных решений была предпринята в Законе Франции 1978 года о защите данных. Так, в своей первоначальной форме ст. 2 Закона запрещала судебные, административные

или личные решения, связанные с оценкой человеческого поведения, поскольку они основывались исключительно на автоматической обработке данных, которая определяла профиль или личность соответствующего лица.

Впоследствии в Директиве 95/46/ЕС Европейского парламента и Совета от 24.10.1995 «О защите физических лиц в связи с обработкой личных данных и о свободном перемещении таких данных»[8] и в Регламенте № 2016/679 Европейского парламента и Совета ЕС «О защите физических лиц при обработке персональных данных и о свободном обращении таких данных, а также об отмене Директивы 95/46/ЕС (Общий Регламент о защите персональных данных)» (далее — GDPR) была закреплена аналогичная правовая норма. Однако в течение всего срока их действия они редко применялись [Kuner C., Bygrave L., Docksey C., 2019: 528, 529]. В дальнейшем вследствие «брюссельского эффекта» схожее регулирование распространилось на значительную часть стран мира. В связи с этим большинство научных исследований проводится на основе GDPR.

Как в российском, так и в европейском регулировании содержится запрет на принятие решений на основании исключительно автоматизированной обработки персональных данных, кроме специальных случаев, указанных в законе. При этом общий подход к регулированию персональных данных заключается в том, что все способы их обработки действуют запретительно для всех лиц (только в случае согласия), но только в случае автоматизированных решений законодатель прописал явный запрет.

Можно предположить, что закрепление строгого регулирования обусловлено в том числе: 1) европейским скептицизмом по отношению к предвзятости и потенциально ложным решениям, которые могут быть приняты автоматическими средствами, если они не проверяются человеком; 2) необходимостью наделения субъекта данных дополнительными гарантиями защиты его прав и законных интересов.

Сфера применения

Закрепленная в законодательстве о персональных данных сфера применения автоматизированных решений имеет особенности, и не каждая автоматизированная система принятия решений подпадает под нее, так как в нее попадает только исключительно авто-

[8] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Available at: URL: http://data.europa.eu/eli/dir/1995/46/oj/eng (дата обращения: 07.04.2025)

матизированная обработка персональных данных. Так, если автоматизированная обработка использовалась только для подготовки доказательств, в то время как фактическое решение было принято человеком, то (согласно европейскому подходу) исключительно автоматизированная обработка персональных данных не осуществлялась [Wachter S., Mittelstadt B., Floridi L., 2017: 88]. Кроме того, в мире существует тенденция, согласно которой в некоторых сферах правоприменения запрещены исключительно автоматизированные решения, что не позволяет применять данную норму[9].

Механизм прозрачности

Если автоматизированная система принятия решений входит в сферу применения закона о персональных данных, оператор должен соблюсти несколько требований, в том числе обязанность обеспечения прозрачности принятия решения. В российском законодательстве эта обязанность находит отражение в ст. 16 Федерального закона от 27.07.2006 № 152-ФЗ «О персональных данных» — обязанность разъяснить порядок принятия решения[10]. В отличие от этого подхода в европейском законодательстве существует обязанность разъяснить содержательную информация о «примененной логике» принятия решения (ст. 13(2)(f), 14(2)(g) и 15(1)(h) GDPR). В доктрине считается, что этот подход предполагает больше конкретики по сравнению с российским, так как потенциально не дает оператору ограничиваться абстрактной информацией о системе[11]. При этом европейский регулятор подчеркнул, что оператор должен сообщить содержательную информацию о задействованной логике, а не обязательно сложное объяснение используемых алгоритмов или раскрытие полного алгоритма, однако информация должна быть настолько полной, чтобы субъект данных мог понять причины принятия решения[12].

---

[9] См.: Directive on Automated Decision-Making. Available at: URL: https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592 (дата обращения: 07.04.2025); Gesetz über die Möglichkeit des Einsatzes von datengetriebenen Informationstechnologien bei öffentlich-rechtlicher Verwaltungstätigkeit (IT-Einsatz-Gesetz — ITEG) vom 16. März 2022. Available at: URL: https://www.gesetze-rechtsprechung.sh.juris.de/bssh/document/jlr-ITEGSHpP1 (дата обращения: 07.04.2025)

[10] Федеральный закон от 27.07.2006 № 152-ФЗ (ред. от 08.08.2024) «О персональных данных» // СЗ РФ. 2006. № 31. Ст. 3451.

[11] Савельев А.И. Научно-практический постатейный комментарий к Федеральному закону «О персональных данных». М., 2021 // СПС Консультант Плюс.

[12] Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679. Available at: URL: https://ec.europa.eu/newsroom/article29/items/612053 (дата обращения: 07.04.2025)

Кроме того, в науке есть суждение: значимая информация должна интерпретироваться применительно к субъекту данных. То есть информация о логике должна быть значимой для него, особенно для человека и, предположительно, без технических знаний, а также проверка значимости информации должна быть функциональной, привязанной к некоторым действиям, которые объяснение дает субъекту данных, например, к праву оспорить решение [Selbst A.D., Powles J., 2017: 236].

Таким образом, в рамках закона о персональных данных реализовано объяснение функциональности системы путем *ex ante*. При этом указанный механизм является ограниченным по нескольким основаниям:

1. Интеллектуальная собственность и коммерческая тайна. Теория и практика исходят из того, что субъектам персональных данных не сообщают полной и точной информации о логике принятия решений, поскольку лежащие в их основе модели и компьютерный код защищены как коммерческой тайной, так и интеллектуальной собственностью[13]. При этом оператор в каждом случае должен стремиться к соблюдению разумного баланса между своими законными интересами (например, защитой коммерческой тайны и интеллектуальной собственности на используемые модели и алгоритмы) и правами и законными интересами субъектов персональных данных, в частности, их правом на получение информации о логике автоматизированного принятия решений. Однако на практике достижение указанного баланса сопряжено со значительными трудностями, и зачастую оператор отдает приоритет собственной защите коммерческих интересов в ущерб полной реализации прав субъектов данных.

2. Формальная открытость при сохранении общественно значимой информации в секрете. Нынешний подход к объяснению порядка или логики автоматизированного принятия решений также позволяет технологическим компаниям использовать нарративы, создающие впечатление открытости перед пользователями, при фактическом сохранении высокой степени конфиденциальности ключевых элементов таких систем. Например, обзор политики конфиденциальности Google демонстрирует сочетание обилия весьма подробной информации о типах собираемых данных, частично собранной в удобной пользователю форме, с крайне расплывчатыми, общими (и

[13] См.: ECLI: EU: C: 2023: 957. Available at: URL: https://curia.europa.eu/juris/document/document.jsf;jsessionid=59BF6046EEA31E86D50793AFC0115814?text=&docid=280426&pageIndex=0&doclang=EN&mode=req&dir=&occ=first&part=1&cid=601767 (дата обращения: 07.04.2025)

практически не имеющими содержательного наполнения) положениями о целях обработки данных, декларируемых как улучшение пользовательского опыта [Felzmann H., Villaronga E.F. et al., 2019: 8].

3. Отказ в разъяснении порядка или логики принятия решения, так как автоматизированная система принятия решений не входит в сферу применения нормы. Операторы не признают или делают все возможное чтобы доказать, что их автоматизированные системы не принимают юридических или значимых решений, или же это не исключительно автоматизированные решения, включая в процесс принятия решения «номинального» исполнителя[14].

4. *Незнание и неактивность субъектов персональных данных*. Большинству лиц, не имеющих компетенций в области программирования и разработки информационных технологий, раскрытие логики работы алгоритма не даст полезной информации, при этом если пользователь алгоритма, например, ИИ, разберется в его логике и найдет техническую ошибку, то внесение изменений в алгоритм будет невозможно, ибо для этого потребуется пересмотреть весь заложенный в алгоритм математический инструментарий. Одновременно социологические исследования показывают, что если автоматизированная система принятия решений нарушила права субъектов, то 13% респондентов не знали, что имеют право запрашивать информацию о способе обработки их персональных данных, а 77% никогда не чувствовали необходимости делать такой запрос [Wulf A.J., Seizov O., 2024: 239].

5. Затрудненность объяснения, как от определенных данных будет меняться итоговое решение. Будущее использование данных трудно предсказать, поэтому право на предварительное объяснение того, как данные могут быть использованы в будущем, вряд ли неосуществимо. Во многих алгоритмах машинного обучения отсутствует линейность. Так, в сквозных моделях нейронных сетей (например, глубоком обучении) связи между входными данными и решением сложным образом зависят от всех входных данных. Поэтому все больше исследователей пытается разработать объяснимые или интерпретируемые системы ИИ[15].

---

[14] EU Law Analysis: The Ola & Uber judgments: for the first time a court recognizes a GDPR right to an explanation for algorithmic decision-making. EU Law Analysis. Available at: URL: https://eulawanalysis.blogspot.com/2021/04/the-ola-uber-judgments-for-first-time.html (дата обращения: 07.04.2025)

[15] См.: Explainable Artificial Intelligence. Available at: URL: https://nsarchive.gwu.edu/sites/default/files/documents/5794867/National-Security-Archive-David-Gunning-DARPA.pdf (дата обращения: 06.04.2025)

Таким образом, предлагаемые законодательством о защите персональных данных механизмы не обеспечивают надлежащего уровня прозрачности и объяснимости, что негативно влияет на уровень защищенности прав и законных интересов человека и гражданина, так как, с одной стороны, доступ ко многим типам данных на практике обеспечивается редко, а с другой стороны, существует множество противоречий в самой норме для решений на основании исключительно автоматизированной обработки персональных данных. При этом механизм объяснимости и прозрачности в рамках законодательства о персональных данных направлен на прозрачность именно работы системы, а не итогового решения.

2.2. Концепция «права на объяснение» индивидуального решения

Ввиду ограниченности действующих в GDPR механизмов прозрачности и объяснимости многие исследователи указывают на необходимость закрепления нового права субъекта данных под условным названием «права на объяснение». Содержание предлагаемого права связывалось с возможностью изображения работы алгоритма в понятной человеку форме, позволяющей как минимум проследить связь между входными данными и результатом (прогнозом) [Goodman B., Flaxman S., 2017: 50–57]. Теоретические основания для такого права усматривались в п. 71 преамбулы GDPR, согласно которому лицу, подвергшемуся автоматизированному принятию решений «должны принадлежать соответствующие гарантии, которые должны включать информацию для субъекта данных и право на вмешательство человека, на выражение своей точки зрения, *на получение объяснения решения, принятого после такой оценки*, и на оспаривание решения» [курсив наш. — П.К., Н.Н.].

Однако в дальнейшем в доктрине было отмечено, что нахождение упоминания о праве на объяснение в преамбуле делает его юридически не обязательным для исполнения. Утверждалось, что в рамках GDPR специальное «право на объяснение» отсутствует, а есть лишь ограниченное «право субъекта данных на получение информации» о логике принятия решений, а именно, субъекту данных по его запросу должна быть сообщена информация, как автоматизированная система работает в целом, для каких целей и с каким прогнозируемым воздействием, прежде чем автоматизированные решения будут приняты (согласно ст. 13(2)(f), 14(2)(g) и 15(1)(h) GDPR) [Wachter S., Mittelstadt B., Floridi L., 2017: 77]. Вместе с этим в этом же исследова-

нии подчеркивалось, что право на оспаривание решения может быть затруднено или лишено смысла, если субъект данных не может понять, как было принято оспариваемое решение. Поэтому во многих исследованиях отмечалась необходимость закрепления полноценного права на объяснение.

Отчасти реагируя на дискуссию и признавая специфику функционирования систем ИИ, законодатель принял Регламент Европейского парламента и Совета ЕС 2024/1689 от 13.06.2024 «Установление согласованных правил в области ИИ (Закон об искусственном интеллекте) и внесение изменений в некоторые законодательные акты Союза» (далее — Закон ЕС об ИИ)[16]. В акте закреплена ст. 86 («Право на объяснение индивидуального решения»), устанавливающая, по существу, что лицо, чьи права и интересы существенно затронуты решением, принятым с использованием определенных систем ИИ, регулируемых данным Законом, имеет право получить от организации, применившей систему, понятные разъяснения о роли этой системы в принятии данного решения.

Общие требования к прозрачности (ст. 13) и человеческому надзору (ст. 14) для систем ИИ высокого риска, заложенные еще в первоначальном проекте Закона ЕС об ИИ (COM/2021/206 final) и сохраненные в итоговой редакции, подчеркивали важность понимания работы таких систем[17]. Однако явное право именно затронутого лица на получение понятных и содержательных объяснений (согласно формулировке ст. 86 и п. 171 преамбулы) стало результатом развития законодательной мысли в ходе обсуждения Регламента, возможно, как раз под влиянием продолжавшейся научной и общественной дискуссии о необходимости упрочения гарантий прав граждан.

2.3. Контрфактологические объяснения

Многие ученые, разочаровавшись в теперешнем механизме прозрачности и объяснимости автоматизированных решений в сфере персональных данных, стали предлагать иные механизмы. Так,

[16] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonized rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).

[17] Proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) 2021. Available at: URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021PC0206 (дата обращения: 04.04.2025)

С. Вахтер, Б. Миттельштадт и К. Рассел предложили контрфакты как средство объяснения индивидуальных решений [Wachter S., Mittelstadt B., Russell C., 2017: 841–887]. Этот подход гарантирует субъектам данных содержательное объяснение для понимания итогового решения, основания его оспаривания и советы: как субъект данных может изменить свое поведение или ситуацию, чтобы, возможно, получить нужное ему решение (например, одобрение кредита).

Однако единого определения данного механизма не сформировано. Вероятно, оптимально следующее: «Контрфактологические объяснения представляют собой класс постфактумных объяснений (*post hoc*) интерпретируемости, которые сообщают человеку, подвергшемуся решению моделью машинного обучения, понятную информацию о результатах моделирования и стратегию достижения альтернативного (будущего) решения» [Ferrario A., Loi M., 2022: 82736]. Примерами контрфактного объяснения являются следующие: «Вам было отказано в кредите, поскольку ваш годовой доход составлял 30 000 фунтов стерлингов»[18]. Схожий пример был предложен в другой статье: «Если бы доход был бы на 1000 $ выше текущего, и если бы клиент полностью погасил текущие долги перед другими банками, то кредит был бы принят». Исходя из данных примеров, допустимо утверждать, что контрфактологическое объяснение позволяет клиенту понять, что нужно ему изменить для получения кредита. При этом в некоторых случаях путем этого объяснения можно увидеть предвзятость или дискриминацию, содержащиеся в автоматизированной системе принятия решений.

Кроме того, контрфактологическое объяснение является объяснением «что, если» (what if) для автоматизированных решений. Это во многом аналогично тому, как дети учатся на контрфактических примерах, и позволяют автоматически исследовать желаемые сценарии «что, если». Тем самым контрфактические утверждения находятся на самом высоком уровне шкалы интерпретируемости Перла, поскольку они отвечают, почему было принято решение, подчеркивая, какие изменения во входных данных могли бы привести к

[18] При этом основное содержание контрфактологического объяснения должно иметь следующую форму: Если бы *q* было ложным, *S* не поверил бы *p*. Мы утверждаем, что в этом случае q служит объяснением веры S в p, поскольку S придерживается убеждения p только тогда, когда q истинно, и что изменение q также приведет к изменению веры S. Ключевым моментом является то, что такие утверждения описывают только убеждения S, которые не обязательно отражают реальность. Таким образом, эти утверждения могут быть сделаны без знания какой-либо причинной связи между q и p.

другому результату. Эти объяснения означают объяснение причинно-следственной связи. Согласно литературе по когнитивной психологии, контрфакты помогают рассуждать на основе объяснений, которые выявляют причинно-следственные связи. «Причина» — это значения признаков входного экземпляра, «вызвавшие» прогноз, а «следствие» — это прогнозируемый результат. В предыдущем примере претендент на кредит может обнаружить, что кредит был бы принят, если бы его доход был на 1000 долл. выше текущего и если бы он полностью возвратил долги [Guidotti R., 2022: 55].

Поэтому в отличие от общепринятого подхода к регулированию персональных данных контрфактологические объяснения являются продолжением развития концепции «право на объяснение», согласно которому обеспечение объяснимости и прозрачности происходит путем *ex post* итогового решения. Эти объяснения также помогают находить предвзятость и дискриминацию в автоматизированных решениях. При этом контрфактологические объяснения близки к языку рассуждений, т.е. к формальной логике, и таким образом пользователь может понять элементарные логические правила. На основе логических правил легче построить повествование, понятное пользователям с разным опытом.

Помимо этого некоторые специалисты предлагают закрепить контрфактологическое объяснение как общий критерий, тем самым выйдя за рамки ограничений GDPR, и использовать контрфакты в качестве безусловных объяснений. Эти объяснения должны даваться всякий раз, когда это требуется, независимо от результата (положительное или отрицательное решение), было ли решение основано исключительно на автоматизированных процессах и их (юридических или аналогичных значимых) последствиях [Wachter S., Mittelstadt B., Russell C., 2017: 841–887].

Однако исследователи отмечают противоречия в контрфактологических объяснениях. Путем анализа сделан вывод, что эти объяснения не способны обрабатывать и объяснять недостающие атрибуты; например, если входные данные имеют отсутствующее значение для теперешнего набора атрибутов, то метод объяснения (или система генерирования объяснений) не может быть применен. Тем самым система принятия решений может основывать решения на наборе скрытых от пользователя функций, так как соответствующие контрфактологические данные могут быть либо неполными (поскольку они не учитывают «основные функции»), либо недействительными, поскольку «основные функции» неизвестны [Guidotti R., 2022: 53].

Например, в вышеизложенном случае в заявке на кредит на передний план выводится запрошенная клиентом сумма. Вместе с тем в автоматизированной системе в качестве фоновых признаков могут также рассматриваться другие атрибуты, например, долги родственников или друзей заявителя.

2.4. Раскрытие данных, на основе которых принимаются автоматизированные решения

Данные являются важной составляющей автоматизированной системы принятия решений, на основе которой она обучается, функционирует и принимает автоматизированные решения. Поэтому неудивительно, что многие исследователи предлагают обеспечить прозрачность и объяснимость на основе данных. Этот подход заключается в раскрытии данных неограниченному кругу лиц, включая их источник; метод сбора, оценки их качества и пробелов; методы, используемые для очистки и стандартизации данных.

Распространение этих данных позволяет раскрывать потенциальное влияние системы на интересы личности и может служить отправной точкой исследования влияния этой автоматизированной системы на соответствующие социальные группы, а также позволяет заявителям выяснить, использовать ли эту автоматизированную систему принятия решений [Mittelstadt B., 2016: 4991–5002]. При этом путем раскрытия данных ученые, работающие с машинным обучением, могут проводить различные исследования, в том числе оценивая эпистемологическую обоснованность, надежность и ограничения модели ИИ[19].

Кроме того, органы, планирующие использовать автоматизированные системы принятия решений, должны тщательно продумывать вопросы происхождения данных и доступа к ним. Это особенно необходимо, когда органы исполнительной власти заключают соглашения с внешними третьими сторонами, поскольку это может повлиять на будущую способность обеспечивать предусмотренную законом прозрачность принятия административных решений. При отсутствии должного внимания к вопросам происхождения данных и доступа к ним обозначается (или усугубляется) асимметрия информации между разработчиками систем ИИ и субъектами, их внедряющими. Стоит добавить, что создается двойственная асимметрия — как для органа, осуществляющего государственные функ-

[19] См: Internet Research: Ethical Guidelines 3.0 Association of Internet Researchers. Available at: URL: https://aoir.org/reports/ethics3.pdf (дата обращения: 06.04.2025)

ции, так и для граждан, которые не знают, на каких данных обучалась автоматизированная система принятия решений и принимает ли она автоматизированные решения.

Некоторые исследователи также предлагают реконструировать регулирование персональных данных так, чтобы гарантировать, что субъект данных имеет возможность передавать все данные, касающиеся его или ее, включая «предполагаемые»; может оценить точность, полноту, актуальность этих данных (насколько это позволяют Регламент и цели обработки); если эти данные составляют основу (автоматизированного) решения, которое затрагивает (или может затронуть) его или ее законные интересы, субъект данных должен иметь доступ, по крайней мере, ко всей информации, необходимой для оценки и, возможно, поддержания оспаривания этого решения (путем обращения к собеседнику-человеку или залу суда) [Troisi E., 2022: 197]. Однако этот подход вызывает опасение, что оператор персональных данных будет собирать большой массив достоверных данных о субъекте. Это во многом не сочетается с принципами законодательства о персональных данных, в частности, с принципом минимизации данных.

2.5. Раскрытие программного кода и (или) модели искусственного интеллекта

Раскрытие программного кода и модели ИИ означает размещение этих объектов в открытом доступе. Механизм делает осуществимым общественный контроль позволяющий выявлять неисправности в автоматизированной системе принятия решений. Однако имеются законодательные и политические ограничения, которые не позволяют в полной мере применять этот механизм. По общему правилу программный код и модель ИИ охраняются в рамках интеллектуальной собственности или коммерческой тайны, и их раскрытие может нанести вред правообладателям.

Иностранные правопорядки неодинаково подходят к такому раскрытию программного кода и модели ИИ. Так, в Великобритании не было зарегистрировано ни одного случая, когда суд выносил постановление о раскрытии исходного кода системы поддержки принятия решений сторонам в судебном процессе, даже в связи с удивительно большим количеством споров, касающихся систем государственного сектора, где вопросы авторского права могли бы считаться менее значимыми [Edwards L., Veale M., 2018: 12].

В Швеции в муниципалитете Треллеборга с 2017 года применяются полностью автоматизированные решения по заявлениям на

социальные пособия. Вскоре это вызвало множество споров, общественная критика варьировалась от незаконного делегирования решений алгоритмическим системам, которые не поддерживаются правовыми нормами для муниципалитетов, до вопросов прозрачности, а также будущего работы и статуса государственных служащих в целом. В итоге гражданин подал иск в суд, утверждая, что исходный код используемого программного обеспечения подпадает под действие шведского принципа публичного доступа к официальным записям (Offentlighetsprincipen)[20]. Суд постановил, что исходный код должен быть доступен общественности и полностью подпадает под принцип публичного доступа.

2.6. Аудит и общественный контроль

Аудит и общественный контроль позволяют провести внешнюю или внутреннюю проверку автоматизированных систем принятия решений. Так, государственные или контролируемые государством механизмы наблюдения за алгоритмами можно рассматривать как меры внешнего контроля. При этом неинструментальный подход (который манифестируется исследователем как более современный), не ставя под сомнение законности и «правозащитности» решения, ориентирует процедуру на принятие согласованного, рационального решения, открытого и прозрачного, в том числе для всех форм контроля.

Нередко выдвигается идея «алгоритма TÜV»[21]. Такая система могла бы быть системой независимой проверки алгоритмов на предмет правовых нарушений или дискриминации на основе баз данных с последующей сертификацией соответствия. Чтобы наилучшим образом понять препятствия, связанные с потенциальными конкурирующими правовыми интересами в защите коммерческой тайны и интеллектуальной собственности, в доктрине предлагается разработать процедуру административного контроля, которая использовала бы камеральный механизм, гарантирующий, что исходный код и

[20] Känsliga uppgifter spreds via kod till biståndsrobot. Available at: URL: https://www.dagenssamhalle.se/samhalle-och-valfard/socialtjanst/kansliga-uppgifter-spreds-via-kod-till-bistandsrobot/ (дата обращения: 06.04.2025)

[21] TÜV — аббревиатура, которая в переводе примерно означает «общество технического надзора». TÜV — частные предприятия в Германии, которые оказывают услуги сертификации безопасности и инспектирования. См: Available at: URL: https://en.wikipedia.org/wiki/Technischer_%C3%9Cberwachungsverein (дата обращения: 06.04.2025)

особенности рассматриваемого алгоритма доступны только уполномоченным инспекторам и что соблюдение этими лицами обязанности конфиденциальности обеспечивалось посредством уголовных наказаний [Santosuosso A., Pinotti G., 2020: 56].

Кроме того, в рамках действующего регулирования аудит и общественный контроль уже закреплены. Так, ст. 42 GDPR предусматривает, что государства-члены ЕС, надзорные органы, Европейский совет защиты данных и Комиссия обязаны поощрять создание механизмов сертификации в области защиты персональных данных, а также знаков и маркировок (или печатей и знаков), позволяющих операторам и обработчикам демонстрировать соответствие их обработки требованиям Регламента. В Великобритании ICO уже объявила тендер на получение сертификационного органа для использования печати конфиденциальности Великобритании, хотя процесс был прерван выходом страны из Европейского союза [Edwards L., Veale M., 2018: 46–54].

В качестве еще одного инструмента саморегулирования предполагается разработка Кодекса поведения программистов самообучающихся систем [Santosuosso A., Pinotti G., 2020: 56]. Это обязало бы программистов соблюдать этические и юридические стандарты. Такая система добровольного обязательства соблюдать свод правил может сочетаться с мерами сертификации (которые уже установлены в GDPR). В качестве вознаграждения предложено совместить членство в схеме со снижением гражданской ответственности за нарушения защиты данных.

Некоторые ученые предлагают создать специальный регулирующий орган надзора над поставщиками услуг [Tutt A., 2017: 83–123]. Этот орган будет вправе «классифицировать алгоритмы по типам на основе их предсказуемости, объяснимости и общего интеллекта», чтобы определить необходимость создания регулирования. Некоторые виды машинного обучения могут, например, быть запрещены или строго ограничены для систем персонализации контента, чтобы предотвратить предвзятость или дискриминацию уязвимых социальных групп, включая группы, выделяемые по признаку политических убеждений, с помощью использования опосредованных признаков (прокси-функций), которые косвенно указывают на принадлежность к таким группам [Mittelstadt B., 2016: 4991–5002]. Регулирующий орган или другая доверенная третья сторона будет вправе также потребовать раскрытия компаниями качественной информации, которая «обеспечивает значимое уведомление о том, как функционирует алгоритм, насколько он полезен и какие ошибки он,

скорее всего, совершает», не раскрывая при этом деталей конструкции, являющихся собственностью компании [Tutt A., 2017: 83–123].

2.7. Информирование о применении автоматизированной системы принятия решений

Как один из возможных механизмов обеспечения прозрачности функционирования автоматизированной системы принятия решений в государственном управлении рассматривается обязанность уполномоченных органов информировать субъектов (граждан, организации) о том, что решение, затрагивающее их права и интересы, было принято (полностью или частично) с использованием таких систем, включая системы ИИ. Суть этого механизма не только в уведомлении о факте применения технологии, но и в создании предпосылок реализации других прав и гарантий.

Допустимо выделить следующие цели введения этого механизма:

реализация права знать, права видеть: данная обязанность напрямую связана с фундаментальным правом на информацию и принципами открытости государственного управления, так как граждане должны понимать, как принимаются затрагивающие их решения, в том числе с участием технологий или без [Талапина Э.В., 2024: 36–39];

обеспечение процессуальной справедливости: знание о применении автоматизированной системы принятия решений является необходимым условием для того, чтобы субъект мог адекватно оценить ситуацию и воспользоваться другими правами: запросить дополнительную информацию, потребовать человеческого вмешательства или пересмотра решения, эффективно его обжаловать;

формирование доверия: открытое информирование о применении систем ИИ и автоматизированных систем принятия решений (даже если сама система сложна) может способствовать повышению доверия граждан к этим технологиям в государственном управлении, демонстрируя готовность органов власти к подотчетности.

Между тем несмотря на важность этого механизма, простое уведомление о факте применения систем ИИ и автоматизированных систем принятия решений порождает непонимание и часто рассматривается в науке как необходимое, но недостаточное условие полноценной (или «осмысленной») прозрачности. В частности, остаются открытыми вопросы о: 1) содержании уведомления: достаточно ли указать только факт использования ИИ, или нужно уточнять тип системы, ее роль в процессе принятия решения? Насколько детальным должно быть это информирование? [Edwards L., Veale M., 2017:

65]; 2) своевременности: когда должно происходить уведомление — до начала процедуры, в процессе или только при сообщении итогового решения? 3) полезности: обеспечивает ли знание о применении этих технологий возможность понять логику решения или его оспорить, особенно при отсутствии доступа к значимым объяснениям (explanations)? [Selbst A.D., Powles J., 2017: 233-242].

3. Технические механизмы обеспечения объяснимости и прозрачности

3.1. Объяснимые или интерпретируемые модели

Самой популярной сферой исследований в технической науке является создание объяснимых моделей, позволяющих обеспечить прозрачность и объяснимость автоматизированного решения «по-умолчанию» *(*by default*)*. При этом в российской науке вместе с объяснимыми моделями выделяют также интерпретируемые модели. Основное их отличие в том, что интерпретируемые модели способны описывать внутреннюю структуру системы понятным способом с использованием ясных правил и метрик объяснения. Можно сказать, что эти технологии объясняют «понятно как, но непонятно почему» принимается автоматизированное решение. В отличие от этого объяснимая модель — возможность кратко описать, почему модель работает (не вдаваясь в подробности). Объяснимый искусственный интеллект (далее–ОИИ) характеризуют так: «непонятно как, но понятно почему», т.е. причина принятия того или иного решения понятна и может быть даже вполне обоснована, но алгоритм, описывающий переход от причины к явлению, остается неявным.

Концепция ОИИ является техническим решением способа объяснения итоговых автоматизированных решений. В доктрине подчеркивается, что ОИИ и связанные с ним технологии кажутся разумными: пользователь алгоритма (или кто-то, чьи правовые или экономические интересы могут быть затронуты решением, принятым алгоритмом) должен иметь доступ к понятному (даже упрощенному) описание функционирования алгоритма. Кроме того, специалисты считают, что не следует отвергать различных подходов к объяснениям, не испытывая их в различных правовых контекстах, поскольку потребности и риски различаются в разных секторах. То же самое относится и к разработке правовых норм, связанных с объяснимостью и прозрачностью, — законодатели должны формировать закон на основе проблем, а не подгонять проблемы к закону.

3.2. Другие технические подходы

В технических дисциплинах есть и другие попытки создания механизмов прозрачности. Авторы одной из научных работ подчеркнули, что способность объяснять модели машинного обучения становится все более важной[22]. Для объяснения моделей с черным ящиком предложена модель экстракции. Этот подход заменяет сложную модель более объяснимой. Л. Дайвер и Б. Шафер предложили использовать технику визуализации процесса, известную как сеть Петри (Petri net) для достижения целей конфиденциальности по дизайну [Diver L., Schafer B., 2017: 68–90]. Этот подход содержит интуитивно-понятные визуальные понятия о состоянии системы и потоке информации внутри правовых и технических моделей. Между ними могут воплощать цели законодательства с самого начала разработки программного обеспечения, в то время как юристы могут получить понятие о внутренней работе программного обеспечения без необходимости понимать его код. В другом исследовании предложен подход скользящей шкалы (sliding scale system); в этом подходе происходит адаптация причинно-следственной связи [Bathaee Y., 2017: 1–50].

В другой работе ученые на основе исследований в области аудита алгоритмов получили понятия о технических характеристиках систем, работающих на ИИ, и на этой основе выдвинули точку зрения, что одним из важных направлений, нацеленных на повышение прозрачности алгоритмов машинного обучения, является What-If Tool — веб-приложение TensorBoard с открытым исходным кодом, которое позволяет пользователям анализировать модели машинного обучения [Felzmann H. et al., 2019: 1–14].

GDPR ввел ряд новых положений, которые довольно радикально не столько наделяют индивидуальными правами субъекта персональных данных, сколько пытаются создать среду, в которой в будущем будут создаваться менее «токсичные» автоматизированные системы. Эти идеи — результат долгой эволюции технологии «конфиденциальность по замыслу» как способа создания систем, обеспечивающих конфиденциальность или дружественных к конфиденциальности, как правило, на добровольной, а не обязательной основе [Edwards L., Veale M., 2018: 46–54].

Кроме того, автоматизированный процесс принятия решений показал много преимуществ для бизнеса и общества, но у этого также

[22] Bastani O., Kim C., Bastani H. Interpretability via Model Extraction. Available at: URL: http://arxiv.org/abs/1706.09773 (дата обращения: 06.04.2025)

есть цена. Давно известно, что высокий уровень автоматизации принятия решений часто приводит к различным недостаткам, таким как предвзятость в решениях и деквалификация работников. На основе анализа этих двух недостатков ученые разработали новую систему поддержки принятия решений, а именно — интеллектуальную помощь в принятии решений (Intelligent Decision Assistance). Эта система дополняет процесс принятия решений человеком с помощью объяснимого ИИ. В самом решении не содержится явных рекомендаций.

Заключение

Несмотря на признание требований прозрачности и объяснимости основополагающими для доверия, подотчетности и законности деятельности органов публичной власти, ни один из нынешних правовых или технических механизмов не является самодостаточным для обеспечения прозрачности этой деятельности. Каждый из них имеет существенные ограничения, связанные с защитой интеллектуальной собственности, технической сложностью, возможностью манипуляций и фундаментальной проблемой «черного ящика» для сложных систем ИИ. Только сочетание разных механизмов позволяет повысить уровень прозрачности автоматизированного принятия решений.

В условиях отсутствия в России комплексного регулирования автоматизированного принятия решений в государственном управлении систематизация подходов и механизмов, а также анализ их недостатков формируют основу дальнейшей дискуссии и разработки нормативных решений. Прозрачность автоматизированного принятии решений требует комплексного подхода, сочетающего применение различных инструментов (правовых, технически, организационных). Так, целесообразно выработать необходимый универсальный набор механизмов к обеспечению прозрачности в условиях внедрения автоматизированных систем принятия решений и систем ИИ в государственном управлении, который будет применяться во всех сферах государственного управления, а в каждой отдельной сфере путем риск-ориентированного подхода будет происходить добавление иных механизмов. Только такой гибкий, риск-ориентированный и контекстуально-зависимый комплексный подход позволит соблюсти необходимый баланс между потребностями государственного управления, защитой прав участников правоотношений и стимулированием безопасного технологического развития в условиях развития автоматизации государственного управления.

Список источников

1. Венгеров А.Б. Правовые основы автоматизации управления народным хозяйством СССР. М.: Высшая школа, 1979. 245 с.

2. Пилипенко А.Н. Франция: к цифровой демократии // Право. Журнал Высшей школы экономики. 2019. Том 13. №4. С. 185-207. doi: https://doi.org/10.17323/2072-8166.2019.4.185.207.

3. Талапина Э.В. Принцип прозрачности использования искусственного интеллекта // Государственная власть и местное самоуправление. 2024. № 7. С. 36–39.

4. Bathaee Y. The artificial intelligence black box and the failure of intent and causation. Harvard Journal of Law & Technology, 2017, no 2, pp. 889–938.

5. Diver L., Schafer B. Opening the black box: Petri nets and Privacy by Design. International Review of Law, Computers & Technology, 2017, no 1, pp. 1–39.

6. Edwards L., Veale M. Enslaving the Algorithm: From a «Right to an Explanation» to a «Right to Better Decisions»? IEEE Security & Privacy, 2018, no 3, pp. 46–54. doi: https://doi.org/10.1109/MSP.2018.2701152.

7. Felzmann H., Villaronga E.F. et al. Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. Big Data & Society, 2019, no. 1, pp. 1–14. doi: https://doi.org/10.1177/2053951719860542.

8. Ferrario A., Loi M. The Robustness of Counterfactual Explanations over Time. IEEE Access, 2022, no. 10, pp. 82736–82750. doi: https://doi.org/10.1109/ACCESS.2022.3196917.

9. Goodman B., Flaxman S. European Union Regulations on Algorithmic Decision Making and a «Right to Explanation». AI Magazine, 2017, no. 3, pp. 50–57. doi: https://doi.org/10.1609/aimag.v38i3.2741.

10. Guidotti R. Counterfactual explanations and how to find them: literature review and benchmarking. Data Mining and Knowledge Discovery, 2022, vol. 38, pp. 2770–2824. doi: https://doi.org/10.1007/s10618-022-00831-6.

11. Guidotti R., Monreale A. et al. A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys. 2019. №5. pp. 1–42. doi: https://doi.org/10.1145/3236009.

12. Kuner C., Bygrave L.A., Docksey C. The EU General Data Protection Regulation (GDPR): a commentary. Oxford: University Press, 2019, 1393 p.

13. Mittelstadt B. Auditing for Transparency in Content Personalization Systems. International Journal of Communication, 2016, vol. 10, pp. 4991–5002.

14. Santosuosso A., Pinotti G. Bottleneck or Crossroad? Problems of Legal Sources Annotation and Some Theoretical Thoughts. Stats, 2020, vol. 3, no. 3, pp. 376–395. doi: https://doi.org/10.3390/stats3030024.

15. Selbst A.D., Powles J. Meaningful information and the right to explanation. International Data Privacy Law. 2017, vol. 7, no. 4, pp. 233–242. doi: https://doi.org/10.1093/idpl/ipx022.

16. Troisi E. Automated Decision Making and right to explanation. The right of access as ex post information. European Journal of Privacy Law & Technologies. 2022, no. 1, pp. 182–202.

17. Tutt A. An Fda for Algorithms. Administrative Law Review, 2017, no. 1, pp. 83–123.

18. Wachter S., Mittelstadt B., Floridi L. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation.

International Data Privacy Law, 2017, vol. 7, no. 2, pp. 76–99. doi: https://doi.org/10.1093/idpl/ipx005.

19. Wulf A.J., Seizov O. «Please understand we cannot provide further information»: evaluating content and transparency of GDPR-mandated AI disclosures. AI & SOCIETY, 2024, vol. 39, no. 1, pp. 235–256. doi: https://doi.org/10.1007/s00146-022-01424-z.

References

1. Bathaee Y. (2017) The Artificial Intelligence Black Box and the Failure of Intent and Causation. *Harvard Journal of Law & Technology,* no. 2, pp. 889–938.

2. Diver L., Schafer B. (2017) Opening the Black Box: Petri Nets and Privacy by Design. *International Review of Law, Computers & Technology*, no. 1, pp. 1–39.

3. Edwards L., Veale M. (2018) Enslaving the Algorithm: From a Right to an Explanation to a Right to Better Decisions? *IEEE Security & Privacy*, no. 3, pp. 46–54. doi: https://doi.org/10.1109/MSP.2018.2701152.

4. Felzmann H., Villaronga E.F. et al. (2019) Transparency You Can Trust: Transparency Requirements for Artificial Intelligence between Legal Norms and Contextual Concerns. *Big Data & Society,* no. 1, pp. 1–14. doi: https://doi.org/10.1177/2053951719860542.

5. Ferrario A., Loi M. (2022) The Robustness of Counterfactual Explanations over Time. *IEEE Access,* no. 10, pp. 82736–82750. doi: https://doi.org/10.1109/ACCESS.2022.3196917.

6. Goodman B., Flaxman S. (2017) European Union Regulations on Algorithmic Decision Making and a Right to Explanation. *AI Magazine,* no. 3, pp. 50–57. doi: https://doi.org/10.1609/aimag.v38i3.2741.

7. Guidotti R., Monreale A. et al. (2019) A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, no. 5, pp. 1–42. doi: https://doi.org/10.1145/3236009.

8. Guidotti R. (2022) Counterfactual Explanations and How to Find them: Literature Review and Benchmarking. *Data Mining and Knowledge Discovery,* vol. 38, pp. 2770–2824. doi: https://doi.org/10.1007/s10618-022-00831-6.

9. Kuner C., Bygrave L.A., Docksey C. (2019) The EU General Data Protection Regulation (GDPR): a commentary. Oxford: University Press, 1393 p.

10. Mittelstadt B. (2016) Auditing for Transparency in Content Personalization Systems. *International Journal of Communication.* no. 10, pp. 4991–5002.

11. Pilipenko A.N. (2019) France: towards Digital Democracy. *Pravo. Journal Vysshey shkoly ekonomiki*=Law. Journal of the Higher School of Economics, vol. 12, no. 4, pp. 185–207. doi: https://doi.org/10.17323/2072-8166.2019.4.185.207. (in Russ.)

12. Santosuosso A., Pinotti G. (2020) Bottleneck or Crossroad? Problems of Legal Sources Annotation and some Theoretical Thoughts *Stats*, vol. 3, no. 3, pp. 376–395. doi: https://doi.org/10.3390/stats3030024.

13. Selbst A.D., Powles J. (2017) Meaningful Information and the Right to Explanation. *International Data Privacy Law*, vol. 7, no. 4, pp. 233–242. doi: https://doi.org/10.1093/idpl/ipx022.

14. Talapina E.V. (2024) Principle of Transparency in the use of Artificial Intelligence. *Gosudarstvennaya vlast i mestnoe samoupravlenie*=State Power and Local Self-Government, no. 7, pp. 36–39 (in Russ.)

15. Troisi E. (2022) Automated Decision Making and Right to Explanation. The Right of Access as *ex post* Information. *European Journal of Privacy Law & Technologies,* no. 1, pp. 182–202.

16. Tutt A. (2017) An Fda for Algorithms. *Administrative Law Review*, no. 1, pp. 83–123.

17. Vengerov A.B. (1979) *Legal Bases of Management Automation in the National Economy of the USSR*. Moscow: Vysshaya shkola, 245 p. (in Russ.)

18. Wachter S., Mittelstadt B., Floridi L. (2017) Why a Right to Explanation of Automated Decision-Making does not Exist in the General Data Protection Regulation. *International Data Privacy Law, v*ol. 7, no. 2, pp. 76–99. doi: https://doi.org/10.1093/idpl/ipx005.

19. Wulf A.J., Seizov O. (2024) Please Understand We Cannot Provide Further Information: Evaluating Content and Transparency of GDPR-Mandated AI Disclosures. *AI & SOCIETY, v*ol. 39, no. 1, pp. 235–256. doi: https://doi.org/10.1007/s00146-022-01424-z.

Информация об авторах:

П.П. Кабытов — кандидат юридических наук, ведущий научный сотрудник.

Н.А. Назаров — младший научный сотрудник.

Information about the authors:

P.P. Kabytov — Candidate of Sciences (Law), Leading Researcher.

N.A. Nazarov — Junior Researcher.

Вклад авторов:

П.П. Кабытов — введение, ч. 3, заключение;

Н.А. Назаров — введение, чч. 1, 2, 3, заключение.

Contribution of the authors:

P. P. Kabytov — introduction, part 3, conclusion;

N.A. Nazarov — introduction, parts 1, 2, 3, conclusion.

Статья поступила в редакцию 14.04.2025; одобрена после рецензирования 26.05.2025; принята к публикации 23.06.2025.

The article was submitted to editorial office 14.04.2025; approved after reviewing 26.05.2025; accepted for publication 23.06.2025.