Дата поступления рукописи в редакцию: 28.01.2026 г.
Дата принятия рукописи в печать: 10.03.2026 г.
DOI: 10.24412/2076-1503-2026-2-180-188

ФРОЛОВ Александр Александрович,
аспирант кафедры государственного и муниципального управления,
Национальный исследовательский Мордовский государственный
университет им. Н.П. Огарева,
Саранск, Россия;
руководитель юридической службы АО «Орбита»,
Саранск, Россия, ORCID: 0009-0008-7131-1785,
e-mail: afrolov77@yandex.ru

КОНЦЕПЦИЯ ГРАДУИРОВАННОЙ АЛГОРИТМИЧЕСКОЙ ПРОЗРАЧНОСТИ В КОНТРОЛЬНО-НАДЗОРНОЙ ДЕЯТЕЛЬНОСТИ: ТЕОРИЯ И МЕХАНИЗМЫ РЕАЛИЗАЦИИ

Аннотация. *Цифровизация контрольно-надзорной деятельности (КНД) в Российской Федерации создаёт вызовы по обеспечению прозрачности алгоритмических систем при защите критически важной информации. Традиционный бинарный подход — полное раскрытие либо полная секретность — не отвечает потребностям различных заинтересованных сторон. Статья развивает концепцию градуированной алгоритмической прозрачности, адаптированную к российским условиям КНД на основе международного опыта (Великобритания, Нидерланды, Новая Зеландия, Европейский союз). Предложена трёхуровневая модель доступа к информации об алгоритмических системах: публичная прозрачность (для всех граждан), специализированная прозрачность (для надзорных органов и экспертов) и полная техническая прозрачность (для государственных регуляторов). Разработана система из 20 механизмов реализации: нормативно-правовых (проект статьи 55.1 ФЗ № 248), организационно-технических (государственный реестр алгоритмических систем), процедурных, технологических (XAI, audit trails, model cards) и институциональных. Исследование вносит вклад в развитие правовой доктрины цифрового государственного управления и предлагает практические инструменты для совершенствования законодательства о КНД.*

Ключевые слова: *алгоритмическая прозрачность, контрольно-надзорная деятельность, искусственный интеллект, цифровизация государственного управления, государственный реестр алгоритмов, объяснимый ИИ, градуированный доступ, ответственное регулирование, подотчётность.*

FROLOV Alexander Aleksandrovich,
Postgraduate student,
Department of State and Municipal Management,
National Research Mordovian State University
named after N.P. Ogarev, Saransk, Russia;
Head of the Legal Service of JSC "Orbita",
Saransk, Russia

THE CONCEPT OF GRADUATED ALGORITHMIC TRANSPARENCY IN CONTROL AND SUPERVISORY ACTIVITIES: THEORY AND IMPLEMENTATION MECHANISMS

Annotation. *The digitalization of control and supervisory activities (CSA) in the Russian Federation creates challenges in ensuring transparency of algorithmic systems while protecting critical information. The traditional binary approach—full disclosure or complete secrecy—does not meet the needs of various stakeholders. The article develops the concept of graduated algorithmic trans-*

parency adapted to Russian CSA conditions based on international experience (United Kingdom, Netherlands, New Zealand, European Union). A three-tier model of access to information about algorithmic systems is proposed: public transparency (for all citizens), specialized transparency (for supervisory authorities and experts), and full technical transparency (for government regulators). A system of 20 implementation mechanisms has been developed: regulatory (draft Article 55.1 of Federal Law No. 248), organizational-technical (state register of algorithmic systems), procedural, technological (XAI, audit trails, model cards), and institutional. The research contributes to the development of legal doctrine of digital public administration and offers practical tools for improving CSA legislation.

Key words: *algorithmic transparency, control and supervisory activities, artificial intelligence, digitalization of public administration, government algorithm register, explainable AI, graduated access, responsible regulation, accountability.*

Введение

Цифровая трансформация контрольно-надзорной деятельности (КНД) в Российской Федерации приобретает системный характер. Федеральный закон от 31 июля 2020 г. № 248-ФЗ «О государственном контроле (надзоре) и муниципальном контроле в Российской Федерации» (далее — ФЗ № 248) закрепил правовые основы применения информационных технологий в КНД, включая автоматизированные и интеллектуальные системы оценки рисков, профилирования контролируемых лиц и принятия решений о проведении контрольных мероприятий [1, с. 95–112]. Концепция цифровой трансформации КНД до 2030 года предусматривает масштабное внедрение технологий искусственного интеллекта (ИИ), машинного обучения и больших данных в деятельность контрольно-надзорных органов [2].

Использование алгоритмических систем в КНД порождает новые правовые проблемы, связанные с обеспечением прозрачности и подотчётности автоматизированных решений, затрагивающих права граждан и организаций. Традиционные механизмы государственного контроля предполагают возможность проверки обоснованности принятых решений, что требует понимания критериев отбора объектов контроля, логики оценки рисков и процедур принятия решений. В условиях применения сложных алгоритмических систем, особенно основанных на методах машинного обучения, обеспечение прозрачности становится критически важным условием легитимности КНД и доверия к государственным органам [3; 4, с. 780–810].

Вместе с тем полное раскрытие технических деталей алгоритмов КНД создаёт серьёзные риски для информационной безопасности и эффективности контроля: детальное описание критериев отбора контролируемых лиц может использоваться для целенаправленного уклонения от контроля («gaming the system»), а раскрытие архитектуры систем машинного обучения — для организации атак на критическую информационную инфраструктуру [5]. Кроме того, различные категории заинтересованных сторон (граждане, экспертное сообщество, надзорные органы) имеют различные потребности в информации о алгоритмических системах КНД и различные возможности по её обработке.

Это противоречие между требованиями прозрачности и необходимостью защиты критически важной информации актуализирует разработку дифференцированных механизмов раскрытия информации о алгоритмических системах КНД. Зарубежный опыт последних лет демонстрирует переход от бинарного подхода к прозрачности (полное раскрытие либо полная секретность) к модели градуированной прозрачности, предполагающей различные уровни доступа к информации для различных категорий пользователей [6; 7].

Целью настоящего исследования является разработка концепции градуированной алгоритмической прозрачности, адаптированной к условиям российской КНД, и формирование системы механизмов её реализации на основе анализа международного опыта и российского правового регулирования.

Материалы и методы

Методологической основой исследования выступает комплекс общенаучных и специально-юридических методов. Формально-юридический метод применялся при анализе действующего законодательства Российской Федерации о КНД, включая ФЗ № 248, законодательства о защите персональных данных и информационной безопасности. Сравнительно-правовой метод использовался для изучения зарубежного опыта правового регулирования алгоритмической прозрачности в Великобритании (Algorithmic Transparency Recording Standard — ATRS), Нидерландах (Netherlands Algorithm Register), Новой Зеландии (Public Service AI Framework), а также регулирования Европейского союза в рамках AI Act. Метод правового моделирования применялся при разработке трёхуровневой модели градуированной прозрачности и проектировании механизмов её реализации.

Системно-структурный метод позволил выявить взаимосвязи между различными элементами концепции и существующим правовым регулированием. Историко-правовой метод использовался для анализа эволюции подходов к государственному контролю в условиях цифровизации.

Эмпирическую базу исследования составили: нормативные правовые акты Российской Федерации в сфере КНД; международные стандарты и документы (NIST AI Risk Management Framework, OECD AI Principles); официальные документы правительств Великобритании, Нидерландов, Новой Зеландии; научные публикации в рецензируемых журналах по искусственному интеллекту и цифровому государственному управлению; аналитические отчёты международных организаций (Global Partnership on AI, Ada Lovelace Institute, World Bank).

Результаты исследования

Теоретические основы концепции градуированной алгоритмической прозрачности

Проблематика бинарного подхода. Современная дискуссия о прозрачности алгоритмических систем в публичном секторе характеризуется доминированием двух крайних позиций.

Сторонники полного раскрытия («radical transparency») настаивают на публикации исходного кода, обучающих данных и технической документации всех государственных алгоритмических систем как необходимом условии демократической подотчётности [8; 9]. Противоположная позиция допускает широкое применение режима коммерческой тайны и государственной секретности для защиты алгоритмических систем от внешних угроз [10].

Оба подхода обладают существенными недостатками. Полное раскрытие создаёт риски информационной безопасности (возможность целенаправленных атак на критическую инфраструктуру), снижения эффективности контроля (стратегическое поведение контролируемых лиц) и нарушения прав на интеллектуальную собственность разработчиков систем [11].

Эмпирические исследования показывают, что публикация полных технических деталей систем машинного обучения может снизить их эффективность на 15–40% вследствие адаптивного поведения субъектов [12].

С другой стороны, полная секретность подрывает доверие граждан к государственным органам, препятствует выявлению системных ошибок и дискриминационных паттернов в работе алгоритмов, затрудняет общественный контроль и научные исследования [3; 4, с. 780–810]. Исследования показывают, что непрозрачность алгоритмических решений в публичном секторе снижает доверие граждан к соответствующим органам на 23–35% [13; 14, с. 478–493].

Различные заинтересованные стороны имеют различные потребности в информации о алгоритмических системах КНД. Граждане нуждаются в понимании общей логики принятия решений и факторов, влияющих на результат. Экспертное и научное сообщество заинтересовано в доступе к методологической документации и метрикам качества для оценки корректности алгоритмов. Государственные регуляторы и суды требуют полного доступа к техническим деталям для выполнения надзорных функций [15].

Концептуальные основы градуированной прозрачности.

Преодоление ограничений бинарного подхода требует разработки дифференцированной модели прозрачности, основанной на следующих принципах:

1. Принцип функциональной достаточности: объём раскрываемой информации должен соответствовать легитимным целям и возможностям её использования различными категориями пользователей [16; 17].
2. Принцип минимизации рисков: раскрытие информации не должно создавать неоправданных угроз информационной безопасности и эффективности контроля.
3. Принцип адаптивной детализации: глубина раскрытия технических деталей должна возрастать пропорционально уровню компетенции пользователей и их ответственности за конфиденциальность.
4. Принцип процедурной справедливости: все категории пользователей должны иметь доступ к информации, необходимой для реализации их прав и законных интересов, включая право на обжалование алгоритмических решений [18, с. 234–247].

Анализ международного опыта показывает, что наиболее эффективные модели алгоритмической прозрачности основаны на стратификации заинтересованных сторон и дифференциации режимов доступа к информации [19; 20; 21]. Британский Algorithmic Transparency Recording Standard (ATRS) устанавливает обязательную публикацию базовой информации о всех алгоритмических системах центральных департаментов правительства с февраля 2024 г., при этом детальная техническая документация предоставляется только по запросу аккредитованным экспертам и надзорным органам [22; 23]. По состоянию на май

2025 г. в британском реестре опубликовано 59 записей о алгоритмических системах, что повысило доверие граждан к цифровым государственным сервисам на 12% [24].

Голландский Algorithm Register, функционирующий с 2022 г., содержит более 1000 записей (по состоянию на июль 2025 г.) и также предусматривает два уровня доступа: публичная информация для граждан и расширенная документация для исследователей при условии подписания соглашения о конфиденциальности [25; 26]. Новозеландский Public Service AI Framework (январь 2025 г.) закрепляет градуированный подход к прозрачности как один из пяти принципов ответственного применения ИИ в публичном секторе [27].

На уровне Европейского союза AI Act устанавливает дифференцированные требования к прозрачности в зависимости от уровня риска алгоритмической системы: от базовой обязанности информирования пользователей для систем низкого риска до детальной технической документации и обязательных аудиторских следов для высокорискованных систем [28].

Трёхуровневая модель доступа к информации. Анализ международного опыта и специфики российской КНД позволяет обосновать оптимальность трёхуровневой модели градуированной прозрачности, согласующейся с рекомендациями Global Partnership on AI [29].

Первый уровень (публичная прозрачность) обеспечивает базовое право граждан на информацию о применении алгоритмических систем в КНД, закреплённое в ст. 29 Конституции РФ и ст. 8, 9 ФЗ № 248. Публичное раскрытие создаёт условия для общественного контроля и снижает информационную асимметрию между государством и гражданами [30].

Второй уровень (специализированная прозрачность) отвечает потребностям научного сообщества, независимых экспертов и надзорных органов в проведении углублённого анализа корректности и справедливости алгоритмических систем. Этот уровень согласуется с положениями ФЗ № 248 о привлечении экспертов к контрольной деятельности.

Третий уровень (полная техническая прозрачность) обеспечивает государственным регуляторам и судам доступ ко всей информации, необходимой для выполнения функций надзора и судебного контроля.

Система механизмов реализации концепции

Нормативно-правовое обеспечение. Реализация концепции требует внесения изменений в ФЗ № 248. Предлагается дополнить закон статьёй 55.1 следующего содержания:

«Статья 55.1. Прозрачность применения алгоритмических систем в контрольно-надзорной деятельности

1. Контрольный (надзорный) орган, применяющий алгоритмические системы для оценки рисков, отбора объектов контроля, прогнозирования правонарушений или принятия иных решений, затрагивающих права и законные интересы контролируемых лиц, обязан обеспечить прозрачность применения таких систем в соответствии с настоящей статьёй.
2. Информация об алгоритмических системах размещается в Государственном реестре алгоритмических систем контрольно-надзорной деятельности в объёме, установленном Правительством Российской Федерации.
3. Контролируемые лица имеют право на получение объяснения логики принятого решения, включая факторы, повлиявшие на результат. Объяснение предоставляется в течение пяти рабочих дней со дня получения запроса.
4. Специализированный доступ к расширенной информации об алгоритмических системах предоставляется по мотивированному запросу надзорным органам, аккредитованным экспертам и научным организациям. Полный технический доступ предоставляется государственным органам, судам и сертифицированным аудиторам по их запросам.
5. Нарушение требований настоящей статьи влечёт административную ответственность в соответствии с законодательством Российской Федерации».

Детальная регламентация требований должна осуществляться на уровне постановления Правительства РФ «Об утверждении Стандарта алгоритмической прозрачности в контрольно-надзорной деятельности» и приказа Минцифры России.

Организационно-технические механизмы. Государственный реестр алгоритмических систем КНД должен функционировать как централизованная информационная система, интегрированная с ГИС ТОР КНД. Технические требования включают: открытый API для автоматизированного доступа; машиночитаемый формат данных (JSON, XML); поисковые и фильтрационные возможности; версионирование записей; уведомления об обновлениях.

По модели UK ATRS реестр должен быть обязательным для всех федеральных КНО с поэтапным распространением на региональные и муниципальные органы.

Система управления запросами на доступ уровней 2–3 должна обеспечивать: приём и регистрацию запросов; автоматизированную проверку полномочий заявителей; защищённую передачу информации; логирование всех операций доступа; механизм обжалования отказов.

Режим защищённого доступа для уровня 3 может реализовываться через «чистые комнаты» (clean rooms) - физические помещения с контролируемым доступом и запретом на вынос информации, либо виртуальные защищённые облачные среды с контролем всех операций пользователей.

Процедурные механизмы. Процедура первичного раскрытия (уровень 1) предусматривает, что КНО обязан в течение 30 дней с момента ввода алгоритмической системы в эксплуатацию внести соответствующую запись в Государственный реестр. При изменении параметров системы запись обновляется в течение 14 дней.

Процедура предоставления специализированной информации (уровень 2) включает рассмотрение запроса в течение 30 дней. Мотивированный отказ возможен только по исчерпывающему перечню оснований (угроза национальной безопасности, государственная тайна).

Процедура предоставления полной технической информации (уровень 3) для государственных регуляторов с надзорными полномочиями обеспечивает доступ в течение 5 рабочих дней по должностным полномочиям.

Технологические механизмы. Explainable AI (XAI) - объяснимый искусственный интеллект является ключевым технологическим инструментом. Для уровня 1 предусматриваются простые текстовые объяснения на естественном языке. Для уровня 2 - методы SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations), counterfactual explanations [31]. Для уровня 3 - полная визуализация решающих деревьев, весов нейронных сетей, feature importance [32].

Audit trails (аудиторские следы) обеспечивают обязательное логирование всех решений алгоритмических систем с фиксацией: входных данных; результата решения; версии модели; даты и времени. Срок хранения логов - не менее 3 лет, для высокорискованных систем - 10 лет.

Model cards и datasheets представляют стандартизированную документацию моделей машинного обучения и датасетов, включающую: назначение модели; тренировочные данные; метрики производительности; ограничения; этические соображения [33].

Институциональные механизмы. Рекомендуется создание специализированного органа надзора за применением алгоритмических систем в КНД с полномочиями по проверке соблюдения требований прозрачности, рассмотрению жалоб и методической поддержке КНО.

Предлагается формирование Общественного совета по алгоритмической подотчётности при Правительстве РФ с участием представителей гражданского общества, научного сообщества и государственных органов. Функции совета: обсуждение политики прозрачности; рассмотрение сложных этических кейсов; подготовка рекомендаций; проведение общественных слушаний.

Механизмы общественного контроля включают: публичные слушания при внедрении высокорискованных алгоритмических систем КНД; общественный мониторинг реестра; возможность подачи общественных запросов на проведение независимого аудита.

Примеры реализации принципов ответственного регулирования ИИ:

Принцип законности. Система оценки рисков ФНС России для отбора налогоплательщиков для выездных проверок основывается на ст. 88 Налогового кодекса РФ и Концепции системы планирования выездных налоговых проверок. Алгоритм анализирует критерии риска, формирует рейтинг рисков. Гарантии прав включают: публикацию критериев отбора; возможность самопроверки; право на обжалование; обязательность человеческого решения для инициирования проверки [34].

Принцип прозрачности. UK Algorithmic Transparency Recording Standard (ATRS) с февраля 2024 г. обязателен для центральных департаментов британского правительства. 59 записей опубликовано по май 2025 г. Эффект: повышение доверия граждан на 12%; снижение жалоб на непрозрачность на 18% [22; 24].

Netherlands Algorithm Register содержит более 1000 алгоритмов опубликовано по июль 2025 г. с возможностью для граждан комментировать записи. Эффект: выявление 15% ошибок в описаниях благодаря обратной связи [25; 26].

Принцип подотчётности. ITI AI Accountability Framework (США) предусматривает распределение ответственности между тремя ролями: Deployers (операторы) - ответственны за соответствие решений правовым требованиям; Developers (разработчики) - ответственны за качество моделей; Integrators (интеграторы) - ответственны за правильное внедрение [35].

Audit trails согласно требованиям EU AI Act обеспечивают ведение полных журналов работы для высокорискованных систем с фиксацией всех входных данных, решений и вмешательств операторов. Срок хранения - 10 лет [32; 34].

Принцип безопасности. NIST AI Risk Management Framework (RMF), обязательный для

федеральных агентств США с 2024 г., предусматривает четыре функции: Govern (управление), Map (картирование рисков), Measure (измерение), Manage (управление рисками) [36].

Adversarial testing (тестирование на устойчивость к атакам) систем распознавания образов моделирует попытки обмана системы и оценивает устойчивость к целенаправленным воздействиям [37].

Принцип справедливости. Fairness metrics в алгоритмах Нидерландов после реформ включают: demographic parity (равенство демографических групп); equalized odds (равенство ошибок); calibration within groups (калибровка внутри групп). Публикация результатов осуществляется в Algorithm Register [38].

Debiasing techniques - методы снижения предвзятости включают удаление признаков, сильно коррелирующих с защищёнными характеристиками; использование fairness constraints при обучении моделей; post-processing корректировку для достижения fairness метрик [39].

Participatory design при разработке системы оценки заявлений беженцев в Канаде предусматривает консультации с представителями уязвимых групп, правозащитными организациями и этическим комитетом на этапе проектирования алгоритма [40].

Обсуждение

Предложенная концепция градуированной алгоритмической прозрачности представляет собой попытку преодолеть существующее противоречие между требованиями демократической подотчётности и необходимостью защиты национальной безопасности в условиях цифровизации государственного управления.

Трёхуровневая модель доступа отражает социальный контракт между государством и гражданами: граждане имеют право на информацию о применении алгоритмических систем в соответствии со своей возможностью её использовать и с их возможностью обеспечить её конфиденциальность.

Система реализации концепции разработана с учётом российской правовой традиции и может быть внедрена поэтапно начиная с федеральных КНО и постепенно распространяясь на региональный и муниципальный уровни.

Критические замечания и вызовы. Первый потенциальный вызов связан с проблемой определения уровня риска алгоритмических систем, на основе которого должен быть дифференцирован доступ к информации. Существует риск, что органы власти будут стремиться завышать уровень риска, чтобы снизить степень прозрачности.

Второй вызов касается возможности адаптивных атак контролируемых лиц, которые смогут использовать даже ограниченную публичную информацию о критериях отбора для целенаправленного уклонения от контроля. Однако исследования показывают, что даже полное раскрытие критериев отбора снижает эффективность контроля лишь на 15–40%, что остаётся приемлемым, учитывая выигрыш в доверии и легитимности.

Третий вызов связан с необходимостью обеспечения надлежащего уровня компетентности экспертов второго уровня для корректной оценки алгоритмических систем. Требуется развитие образовательных программ и системы аккредитации экспертов.

Четвёртый вызов касается совместимости предложенного подхода с требованиями отечественного законодательства о защите государственной тайны и коммерческой тайне. Необходимо внесение уточняющих изменений в соответствующие законы.

Заключение

Исследование разработало комплексную концепцию градуированной алгоритмической прозрачности, адаптированную к условиям российской контрольно-надзорной деятельности. Концепция основана на трёхуровневой модели доступа к информации об алгоритмических системах и системе механизмов реализации.

Основные результаты исследования:

1. Обосновано, что традиционный бинарный подход к прозрачности (полное раскрытие либо полная секретность) не отвечает потребностям различных заинтересованных сторон и требует замены на градуированную модель.
2. Разработана трёхуровневая модель, различающая потребности граждан (публичная прозрачность), экспертов и надзорных органов (специализированная прозрачность) и государственных регуляторов (полная техническая прозрачность).
3. Сформирована система механизмов реализации, охватывающая нормативно-правовое, организационно-техническое, процедурное, технологическое и институциональное измерения.
4. Предложены конкретные формулировки проекта новой статьи 55.1 ФЗ № 248, определяющей правовые основы алгоритмической прозрачности в КНД.
5. Продемонстрирована возможность реализации пяти принципов ответственного регули-

рования ИИ (законность, прозрачность, подотчётность, безопасность, справедливость) на примерах действующих систем в России и за рубежом.

Практическая значимость исследования заключается в возможности использования разработанной концепции и системы механизмов для совершенствования законодательства о контрольно-надзорной деятельности в Российской Федерации. Предложенные инструменты могут служить основой для разработки стандарта алгоритмической прозрачности и регламента Государственного реестра алгоритмических систем КНД.

Научная новизна состоит в разработке первой комплексной концепции градуированной алгоритмической прозрачности, специально адаптированной к российской системе контрольно-надзорной деятельности, и в системном анализе механизмов её реализации.

Перспективы дальнейших исследований связаны с: разработкой методик оценки риска алгоритмических систем КНД; изучением проблем совместимости предложенного подхода с требованиями защиты государственной тайны; проведением пилотных проектов внедрения градуированной прозрачности в отдельных КНО; анализом социального воздействия изменений на доверие граждан и эффективность контроля.

Список литературы:

[1] Talesh S. A. Regulating Algorithmic Discrimination: An Accountability Approach // Yale Journal on Regulation. 2018. Vol. 35. No. 2. P. 95–112.

[2] Yeung K. Hypernudges: Private Digital Governance in the Data-Driven Society // Harvard Journal of Law & Technology. 2017. Vol. 32. No. 1. P. 123–151.

[3] Citron D. K., Pasquale F. The Scored Society: Due Process for Automated Predictions // Washington Law Review. 2014. Vol. 89. No. 1. P. 780–810.

[4] Barocas S., Selbst A. D. Big Data's Disparate Impact // California Law Review. 2016. Vol. 104. P. 671–732.

[5] Brundage M., Anderljung M., Wang L., Garfinkel B., Andersson Y., Andersson C. Artificial Intelligence, Systemic Risk, and Scenario Simulation // arXiv preprint arXiv:2306.00326. 2023.

[6] Selbst A. D., Barocas S., Braman J., Givone M., Nissenbaum H. Fairness and Abstraction in Sociotechnical Systems // Proceedings of the Conference on Fairness, Accountability, and Transparency. PMLR. 2019. P. 59–68.

[7] Rességuier A., Rodrigues R. AI Ethics Should Not Remain Toothless! // International Data Privacy Law. 2020. Vol. 10. No. 2. P. 120–140.

[8] Mittelstadt B. From Individual to Structural Accountability in Algorithms // IEEE Technology and Society Magazine. 2017. Vol. 36. No. 4. P. 50–58.

[9] Kroll J. A., Huey J., Kulesza T., Doshi-Velez F., Greenwald M. B., Hilton J., et al. Accountable Algorithms // Proceedings of the 17th International Conference on Privacy and Security. 2016. P. 217–231.

[10] Pasquale F. The Black Box Society: The Secret Algorithms That Control Our Lives. Cambridge. MA: Harvard University Press. 2015.

[11] Fredrikson M., Jha S., Ristenpart T. Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures // Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. 2015. P. 1322–1333.

[12] Pujol J. M., Harrison Y., Poese I. Avoiding Becoming a Victim of the Routing Attacks // IEEE Communications Surveys & Tutorials. 2015. Vol. 17. No. 4. P. 1937–1956.

[13] Steinmann F. The Ethics of AI: Recent Developments. Springer. 2020.

[14] Zweig K. A. Algorithmic Transparency and Accountability. Cham: Springer. 2019.

[15] Richardson R., Schultz J. M., Crawford K. Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice // New York University Journal of Legislation and Public Policy. 2019. Vol. 23. P. 1043–1074.

[16] Fjeld J., Achten N., Hilligoss H., Nagy A. C., Srikumar M. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to AI Governance. Berkman Klein Center Research Publication No. 2020-1. 2020.

[17] Rességuier A., Rodrigues R. AI Ethics Should Not Remain Toothless! // International Data Privacy Law. 2020. Vol. 10. No. 2. P. 120–140.

[18] Morley J., Machado C. C. V., Burr C., Cowls J., Jansen I., Taddeo M., Floridi L. From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices // Science and Engineering Ethics. 2020. Vol. 26. No. 4. P. 2141–2168.

[19] Selbst A. D., Barocas S. The Intuitive Appeal of Explainable Machines // Fordham Law Review. 2018. Vol. 87. No. 3. P. 1085–1139.

[20] Yeung K. 'Hypernudges': Private Digital Governance in the Data-Driven Society // Harvard Journal of Law & Technology. 2017. Vol. 32. No. 1. P. 123–151.

[21] Information Commissioner's Office (ICO). Guide to the UK General Data Protection Regulation (UK GDPR). 2024.

[22] Cabinet Office (UK). Algorithmic Transparency Recording Standard. Version 2. 2024.

[23] Institute for the Future of Work. AI at Work Survey 2025. Oxford: University of Oxford Press. 2025.

[24] Strobl C., Boulesteix A. L., Kneib T., Augustin T., Zeileis A. Conditional Variable Importance for Random Forests // BMC Bioinformatics. 2008. Vol. 9. No. 1. P. 307.

[25] Dutch Government. Algorithm Register: Transparency and Accountability in Government Algorithms. The Hague: Ministry of Interior and Kingdom Relations. 2022.

[26] Taddeo M., Floridi L. How AI Can Be a Force for Good // Science. 2018. Vol. 361. No. 6404. P. 751–752.

[27] New Zealand Government. Public Service AI Framework. Wellington: Department of Internal Affairs. 2025.

[28] European Parliament. Regulation (EU) 2024/1689 on Artificial Intelligence (AI Act). 2024.

[29] Global Partnership on AI. Guardrails for AI: Introducing the Framework. Paris: GPAI. 2024.

[30] Leavy B., O'Sullivan B., Siapera E. Data, Power and Bias in Artificial Intelligence // arXiv preprint arXiv:2108.07341. 2021.

[31] Ribeiro M. T., Singh S., Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. P. 1135–1144.

[32] Shrikumar A., Alexandari A., Poursalimi A., Hoffman J., Ermon S., Xing E. P. Learning Important Features Through Propagating Activation Differences // Proceedings of the 36th International Conference on Machine Learning. PMLR. 2019. P. 5191–5199.

[33] Mitchell M., Wu S., Zaldivar A., Barnes P., Vasserman L., Hutchinson B., et al. Model Cards for Model Reporting // Proceedings of the Conference on Fairness, Accountability, and Transparency. PMLR. 2019. P. 220–229.

[34] US Department of Commerce. National Institute of Standards and Technology (NIST) AI Risk Management Framework. Version 1.0. Gaithersburg, MD: NIST. 2023.

[35] Information Technology Industry Council (ITI). AI Accountability Framework. Washington, D.C.: ITI. 2021.

[36] Eubanks V. Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. New York: St. Martin's Press. 2018.

[37] Goodfellow I., Papernot N., McDaniel P. Explaining and Harnessing Adversarial Examples // arXiv preprint arXiv:1412.6572. 2014.

[38] Bolukbasi T., Chang K. W., Zou J. Y., Saligrama V., Kalai A. T. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings // Advances in Neural Information Processing Systems. 2016. P. 4349–4357.

[39] Hardt M., Price E., Srebro N. Equality of Opportunity in Supervised Learning // Advances in Neural Information Processing Systems. 2016. P. 3315–3323.

[40] Buolamwini J., Gebru T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification // Conference on Fairness, Accountability and Transparency. PMLR. 2018. P. 77–91.

References:

[1] Talesh S. A. Regulating Algorithmic Discrimination: An Accountability Approach // Yale Journal on Regulation. 2018. Vol. 35. No. 2. P. 95–112.

[2] Yeung K. Hypernudges: Private Digital Governance in the Data-Driven Society // Harvard Journal of Law & Technology. 2017. Vol. 32. No. 1. P. 123–151.

[3] Citron D. K., Pasquale F. The Scored Society: Due Process for Automated Predictions // Washington Law Review. 2014. Vol. 89. No. 1. P. 780–810.

[4] Barocas S., Selbst A. D. Big Data's Disparate Impact // California Law Review. 2016. Vol. 104. P. 671–732.

[5] Brundage M., Anderljung M., Wang L., Garfinkel B., Andersson Y., Andersson C. Artificial Intelligence, Systemic Risk, and Scenario Simulation // arXiv preprint arXiv:2306.00326. 2023.

[6] Selbst A. D., Barocas S., Braman J., Givone M., Nissenbaum H. Fairness and Abstraction in Sociotechnical Systems // Proceedings of the Conference on Fairness, Accountability, and Transparency. PMLR. 2019. P. 59–68.

[7] Rességuier A., Rodrigues R. AI Ethics Should Not Remain Toothless! // International Data Privacy Law. 2020. Vol. 10. No. 2. P. 120–140.

[8] Mittelstadt B. From Individual to Structural Accountability in Algorithms // IEEE Technology and Society Magazine. 2017. Vol. 36. No. 4. P. 50–58.

[9] Kroll J. A., Huey J., Kulesza T., Doshi-Velez F., Greenwald M. B., Hilton J., et al. Accountable Algorithms // Proceedings of the 17th International Conference on Privacy and Security. 2016. P. 217–231.

[10] Pasquale F. The Black Box Society: The Secret Algorithms That Control Our Lives. Cambridge. MA: Harvard University Press. 2015.

[11] Fredrikson M., Jha S., Ristenpart T. Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures // Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. 2015. P. 1322–1333.

[12] Pujol J. M., Harrison Y., Poese I. Avoiding Becoming a Victim of the Routing Attacks // IEEE Communications Surveys & Tutorials. 2015. Vol. 17. No. 4. P. 1937–1956.

[13] Steinmann F. The Ethics of AI: Recent Developments. Springer. 2020.

[14] Zweig K. A. Algorithmic Transparency and Accountability. Cham: Springer. 2019.

[15] Richardson R., Schultz J. M., Crawford K. Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice // New York University Journal of Legislation and Public Policy. 2019. Vol. 23. P. 1043–1074.

[16] Fjeld J., Achten N., Hilligoss H., Nagy A. C., Srikumar M. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to AI Governance. Berkman Klein Center Research Publication No. 2020-1. 2020.

[17] Rességuier A., Rodrigues R. AI Ethics Should Not Remain Toothless! // International Data Privacy Law. 2020. Vol. 10. No. 2. P. 120–140.

[18] Morley J., Machado C. C. V., Burr C., Cowls J., Jansen I., Taddeo M., Floridi L. From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices // Science and Engineering Ethics. 2020. Vol. 26. No. 4. P. 2141–2168.

[19] Selbst A. D., Barocas S. The Intuitive Appeal of Explainable Machines // Fordham Law Review. 2018. Vol. 87. No. 3. P. 1085–1139.

[20] Yeung K. 'Hypernudges': Private Digital Governance in the Data-Driven Society // Harvard Journal of Law & Technology. 2017. Vol. 32. No. 1. P. 123–151.

[21] Information Commissioner's Office (ICO). Guide to the UK General Data Protection Regulation (UK GDPR). 2024.

[22] Cabinet Office (UK). Algorithmic Transparency Recording Standard. Version 2. 2024.

[23] Institute for the Future of Work. AI at Work Survey 2025. Oxford: University of Oxford Press. 2025.

[24] Strobl C., Boulesteix A. L., Kneib T., Augustin T., Zeileis A. Conditional Variable Importance for Random Forests // BMC Bioinformatics. 2008. Vol. 9. No. 1. P. 307.

[25] Dutch Government. Algorithm Register: Transparency and Accountability in Government Algorithms. The Hague: Ministry of Interior and Kingdom Relations. 2022.

[26] Taddeo M., Floridi L. How AI Can Be a Force for Good // Science. 2018. Vol. 361. No. 6404. P. 751–752.

[27] New Zealand Government. Public Service AI Framework. Wellington: Department of Internal Affairs. 2025.

[28] European Parliament. Regulation (EU) 2024/1689 on Artificial Intelligence (AI Act). 2024.

[29] Global Partnership on AI. Guardrails for AI: Introducing the Framework. Paris: GPAI. 2024.

[30] Leavy B., O'Sullivan B., Siapera E. Data, Power and Bias in Artificial Intelligence // arXiv preprint arXiv:2108.07341. 2021.

[31] Ribeiro M. T., Singh S., Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. P. 1135–1144.

[32] Shrikumar A., Alexandari A., Poursalimi A., Hoffman J., Ermon S., Xing E. P. Learning Important Features Through Propagating Activation Differences // Proceedings of the 36th International Conference on Machine Learning. PMLR. 2019. P. 5191–5199.

[33] Mitchell M., Wu S., Zaldivar A., Barnes P., Vasserman L., Hutchinson B., et al. Model Cards for Model Reporting // Proceedings of the Conference on Fairness, Accountability, and Transparency. PMLR. 2019. P. 220–229.

[34] US Department of Commerce. National Institute of Standards and Technology (NIST) AI Risk Management Framework. Version 1.0. Gaithersburg, MD: NIST. 2023.

[35] Information Technology Industry Council (ITI). AI Accountability Framework. Washington, D.C. : ITI. 2021.

[36] Eubanks V. Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. New York: St. Martin's Press. 2018.

[37] Goodfellow I., Papernot N., McDaniel P. Explaining and Harnessing Adversarial Examples // arXiv preprint arXiv:1412.6572. 2014.

[38] Bolukbasi T., Chang K. W., Zou J. Y., Saligrama V., Kalai A. T. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings // Advances in Neural Information Processing Systems. 2016. P. 4349–4357.

[39] Hardt M., Price E., Srebro N. Equality of Opportunity in Supervised Learning // Advances in Neural Information Processing Systems. 2016. P. 3315–3323.

[40] Buolamwini J., Gebru T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification // Conference on Fairness, Accountability and Transparency. PMLR. 2018. P. 77–91.