DOI 10.37882/2223-2966.2026.03.45

РИСК-ОРИЕНТИРОВАННЫЕ МОДЕЛИ УПРАВЛЕНИЯ ЖИЗНЕННЫМ ЦИКЛОМ ДОВЕРЕННЫХ СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В КРИТИЧЕСКИ ЗНАЧИМЫХ ОТРАСЛЯХ

RISK-BASED LIFECYCLE MANAGEMENT MODELS FOR TRUSTED ARTIFICIAL INTELLIGENCE SYSTEMS IN CRITICAL SECTORS

S. Tyryshkin
E. Ignatova
P. Parfenov

Summary. In critical sectors, artificial intelligence increasingly becomes part of operational and safety control loops, yet its behavior is shaped not only by software code but also by data, training procedures and the runtime environment. This makes trust assurance non-trivial and calls for a shift from fragmented checks to risk-based lifecycle management. The paper proposes a lifecycle model for trusted AI systems that integrates ISO/IEC 42001-aligned governance (GOST R ISO/IEC 42001-2024), AI system life cycle processes (GOST R 71539-2024), AI risk management guidance (ISO/IEC 23894:2023), the NIST AI RMF 1.0, and AI security practices (MITRE ATLAS, OWASP LLM Top 10). The suggested «evidence-based trust loop» includes criticality classification, stage-specific risk registers, data quality controls, independent validation, drift monitoring and change management. Practical examples from healthcare, energy and financial compliance illustrate how the model reduces hidden failures, model tampering and compliance risks through standardized assurance gates and auditable artifacts.

Keywords: trusted AI, risk management, life cycle, critical infrastructure, audit, verification and validation, data quality, model drift, MLOps, cybersecurity.

Тырышкин Сергей Юрьевич
кандидат технических наук, доцент,
Алтайский государственный технический университет им. И.И. Ползунова
service.vip-spe@yandex.ru
Игнатова Елена Ивановна
кандидат технических наук, доцент,
Санкт-петербургский государственный морской технический университет
ignatova384756@gmail.com
Парфенов Павел Дмитриевич
Аспирант, Северный (Арктический) федеральный университет имени М.В. Ломоносова
i@pparfenov.ru

Аннотация. В критически значимых отраслях искусственный интеллект становится частью контуров управления и безопасности, однако его поведение определяется не только программным кодом, но и данными, обучением и эксплуатационной средой. Это усложняет доказательство доверия и требует перехода от разрозненных проверок к риск-ориентированному управлению жизненным циклом. Цель статьи — обосновать и формализовать модель управления жизненным циклом доверенных ИИ-систем, интегрирующую требования стандартов ГОСТ Р ИСО/МЭК 42001-2024 и ГОСТ Р 71539-2024, рекомендации ISO/IEC 23894:2023 и подход NIST AI RMF 1.0, а также практики обеспечения безопасности ИИ (MITRE ATLAS, OWASP LLM Top 10). Предложена архитектура «контур доказательного доверия», включающая классификацию критичности, риск-регистры по стадиям, контроль качества данных, независимую валидацию, мониторинг дрейфа и управление изменениями. На примерах из здравоохранения, энергетики и финансового комплаенса показано, как модель снижает вероятность скрытых отказов, подмены моделей и регуляторных нарушений за счёт единых «ворот» допуска и измеримых артефактов аудита.

Ключевые слова: доверенный искусственный интеллект, управление рисками, жизненный цикл, критическая инфраструктура, аудит, верификация и валидация, качество данных, дрейф модели, MLOps, информационная безопасность.

Системы искусственного интеллекта (ИИ), используемые в здравоохранении, энергетике, транспорте, финансовом секторе, связи и промышленной автоматизации, всё чаще оказываются встроенными в критически значимые технологические контуры и информационные системы. Для таких контуров характерны повышенные требования к отказоустойчивости, предсказуемости поведения, защите данных и доказуемости соответствия регуляторным требованиям. При этом ИИ-компоненты принципиально отличаются от традиционного программного обеспечения тем, что функциональность определяется не только исходным кодом, но и качеством данных, процедурой обучения, настройками, а также изменчивостью эксплуатационной среды.

Датацентричность и адаптивность повышают эффективность, но одновременно усиливают неопределённость и порождают новые классы рисков: дрейф модели, состязательные воздействия, утечки через параметры, зависимость от внешних наборов данных и библиотек, ошибки в разметке, «скрытые» режимы деградации качества.

В 2023–2026 гг. наблюдается ускорение стандартизации и регулирования доверенного ИИ. На международном уровне развиваются рамки риск-менеджмента и доказательства доверия (NIST AI RMF 1.0) [1], конкретизируются методики анализа угроз для AI-enabled систем (MITRE ATLAS) [2], уточняются риски приложений на основе больших языковых моделей (OWASP LLM Top 10) [3]. В Европейском союзе принят риск-ориентированный правовой режим (Регламент (EU) 2024/1689, AI Act) [4]. В Российской Федерации в 2024–2025 гг. введены национальные стандарты по системе менеджмента ИИ (ГОСТ Р ИСО/МЭК 42001-2024) [5], процессам жизненного цикла ИИ-систем (ГОСТ Р 71539-2024) [6] и структуре жизненного цикла данных (ГОСТ Р 70889-2023) [7]. Сочетание этих документов задаёт нормативную базу, но практический вопрос остаётся открытым: каким образом интегрировать требования менеджмента, инженерных процессов и кибербезопасности в единую модель управления жизненным циклом доверенных ИИ-систем.

Цель статьи — предложить риск-ориентированную модель управления жизненным циклом доверенных ИИ-систем для критически значимых отраслей, объединяющую: 1) процессное описание жизненного цикла; 2) системный контур управления рисками; 3) набор измеримых артефактов, достаточных для аудита; 4) механизм непрерывного контроля эффективности мер в эксплуатации. Научная новизна заключается в формализации «контура доказательного доверия» как интеграции стандартов и практик безопасности ИИ в виде последовательности «ворот» допуска, применимых на всех стадиях — от постановки требований до вывода из эксплуатации.

Обзор публикаций за последние три года позволяет выделить четыре устойчивых направления, определяющих современную повестку доверенного ИИ в критически значимых отраслях.

Первое направление — институционализация риск-ориентированных подходов к ИИ-управлению. В NIST AI RMF риск трактуется как совокупность технических, организационных и социально-правовых последствий, а доверенность связывается с измеримыми характеристиками (надёжность, безопасность, устойчивость, объяснимость, управляемость) и процессами управления на протяжении всего жизненного цикла [1]. Риск-ориентированная логика закрепляется и в правовом поле: в AI Act ключевым механизмом выступает классификация систем по уровню риска, приоритетно регулируя «высокорисковые» применения в критических областях [4].

Второе направление — развитие стандартов, фиксирующих процессы жизненного цикла и управленческую рамку. ISO/IEC 5338:2023 описывает процессы жизненного цикла ИИ-систем [9], а ISO/IEC 23894:2023 предлагает рекомендации по управлению рисками ИИ и интеграции риск-менеджмента в AI-деятельность [10]. Российская адаптация процессного подхода закреплена в ГОСТ Р 71539-2024 [6], а управленческая надстройка — в ГОСТ Р ИСО/МЭК 42001-2024 [5]. Отдельно формализована структура жизненного цикла данных (ГОСТ Р 70889-2023) [7], что принципиально важно для датацентричных систем.

Третье направление — исследования угроз и уязвимостей ИИ-систем. На материалах профильных изданий показано, что атаки и уязвимости затрагивают все уровни: данные, модель, инфраструктуру и процессы эксплуатации [11; 12]. MITRE ATLAS описывает угрозы в терминах тактик и техник, что делает базу пригодной для моделирования угроз и построения сценариев тестирования [2]. OWASP фиксирует типовые риски для приложений на основе LLM по всему жизненному циклу, включая обучение, развёртывание и эксплуатацию [3].

Четвёртое направление — практики сертификации, аудита и формирования доверенных контуров разработки. В отечественной дискуссии развивается концепт «доверенного искусственного интеллекта», связываемый с методиками аудита, качеством разработки и формированием доверенных репозиториев компонентов [8; 17; 21; 22]. Конвергенция управленческих стандартов и инженерных практик безопасности требует прикладной модели, которая связывает эти контуры в единый механизм управления жизненным циклом доверенных ИИ-систем именно для критически значимых отраслей. Эту задачу и решает предлагаемая в статье модель.

Критически значимые отрасли предъявляют к ИИ-системам требования выходящие за рамки точности прогнозов или полноты классификации. В условиях КЗО ключевым становится обеспечение безопасного поведения системы при неопределённости и ошибках. Анализ научных работ и профессиональных рекомендаций позволяет сгруппировать проблемные зоны в три блока.

1) Технический блок: а) уязвимость к состязательным воздействиям на данные и модель; б) деградация качества вследствие дрейфа данных/концепции; в) непрозрачность сложных моделей и трудности проверки корректности; г) риски цепочки поставок (наборы данных, библиотеки, базовые модели) [2; 3; 11; 12; 18; 19].

2) Организационный блок: а) отсутствие единого владельца риска и «сквозной» ответственности за ИИ-компонент; б) разрывы между разработкой и эксплуатацией (DevOps/MLOps), приводящие к неконтролируемым изменениям; в) фрагментарность документации и невозможность воспроизвести обучение и результаты валидации.
3) Регуляторно-правовой блок: а) необходимость классификации системы по уровню риска и соответствия требованиям к высокорисковым применениям [4; 13]; б) выполнение требований к защите персональных данных и к безопасности критической информационной инфраструктуры; в) обеспечение отчётности и готовности к аудиту в рамках отраслевого надзора.

С учётом этих блоков предлагается рассматривать доверие к ИИ-системе как проверяемое свойство, поддерживаемое набором артефактов и процедур. На уровне операционализации доверие удобно описывать через три группы критериев.

1) Свойства безопасности и устойчивости: наличие безопасных режимов, ограничения на автономность, предсказуемые стратегии деградации.
2) Свойства защищённости: устойчивость к атакующим воздействиям, защищённая цепочка поставок, контроль целостности моделей и данных.
3) Свойства управляемости и доказуемости: трассируемость требований и данных, воспроизводимость обучения, независимая валидация, прозрачность и проверяемость решений аудитом.

Эти критерии соотносятся с процессными требованиями ГОСТ Р 71539-2024 [6] и управленческими требованиями ГОСТ Р ИСО/МЭК 42001-2024 [5], а также с практиками NIST AI RMF [1]. На рисунке 1 представлена матрица трассируемости свойств доверия по ключевым стадиям жизненного цикла.

Предлагаемая модель исходит из того, что управлять необходимо не «моделью машинного обучения» как отдельным артефактом, а системой ИИ как совокупностью данных, моделей, программных компонентов, инфраструктуры, организационных процедур и внешних ограничений. В качестве базовой рамки берётся процессное описание жизненного цикла (ГОСТ Р 71539-2024) [6], а в качестве надстройки — система менеджмента ИИ (ГОСТ Р ИСО/МЭК 42001-2024) [5]. Контур управления рисками строится с опорой на ISO/IEC 23894:2023 [10] и NIST AI RMF 1.0 [1] и дополняется практиками обеспечения безопасности ИИ (MITRE ATLAS) [2].

Ключевым элементом модели выступают «ворота допуска» — формализованные контрольные точки, в которых принимается решение о переходе на следующую стадию жизненного цикла на основании доказательных артефактов. Ворота отражают принцип: в критически значимых отраслях недостаточно достичь целевых метрик качества; необходимо доказать приемлемость остаточного риска.

На рисунке 2 показана общая схема риск-ориентированного управления жизненным циклом, включающая восемь стадий и восемь ворот допуска (G0–G7). На каждой стадии формируется набор артефактов, подлежащих аудиту и повторяемой проверке.

Для единообразия оценки предлагается использовать комбинированную метрику приоритета риска, адаптированную к инженерным задачам ИИ-систем:

$$R = P \times I \times D, \quad (1)$$

где P — вероятность реализации сценария (учитывая статистику инцидентов и уязвимости контура), I — тяжесть последствий для безопасности, непрерывности и прав (включая отраслевые требования), D — обнаруживаемость (вероятность выявить деградацию до наступления ущерба). Компонент D особенно важен для ИИ-

Матрица трассируемости свойств доверия по стадиям жизненного цикла

Свойство доверия	Требования	Данные	Модель	ВиВ	Ввод	Эксплуатация
Безопасность (Safety)	✓	✓	✓	✓	✓	✓
Информационная безопасность	✓	✓	✓	✓	✓	✓
Надёжность и устойчивость	✓	✓	✓	✓	✓	✓
Объяснимость/ аудитируемость	✓	✓	✓	✓	✓	✓
Приватность и ПДн	✓	✓	✓	✓	✓	✓
Недискриминация и качество данных	✓	✓	✓	✓		✓
Управляемость и отказоустойчивость	✓	✓	✓	✓	✓	✓

Рис. 1. Матрица трассируемости свойств доверия по стадиям жизненного цикла (составлено автором на основе [1; 5–7; 10])

Рис. 2. Риск-ориентированное управление жизненным циклом доверенной ИИ-системы: стадии и ворота допуска (составлено автором на основе [1–3; 5–7; 9–10])

систем из-за латентных отказов и дрейфа, когда метрики ухудшаются постепенно и не фиксируются стандартным мониторингом.

В практической реализации модель предполагает:

1) Классификацию критичности и риск-категории применения (G1): определяются границы допустимого риска, режимы автономности, требования к человеку-в-контуре, а также обязательные артефакты аудита.
2) Управление данными (G2): формируется «паспорт данных» и карта потоков данных; вводятся контроль качества, репрезентативности, правомерности использования, процедуры разметки и контроля смещения. Требования синхронизируются с ГОСТ Р 70889-2023 [7].
3) Управление моделью (G3–G4): обеспечиваются воспроизводимость обучения, контроль версий, защита от подмены и несанкционированной модификации, протоколирование гиперпараметров и сред, независимая валидация и тестирование состязательной устойчивости [11; 12].
4) Управление вводом в эксплуатацию (G5): проводится оценка готовности к эксплуатации, определяется стратегия безопасного отказа, пороговые значения метрик и критерии остановки/отката.
5) Управление эксплуатацией (G6): вводятся мониторинг дрейфа, наблюдаемость, журналирование значимых решений, регулярные переоценки риска и учёт инцидентов; изменения рассматриваются как управляемые релизы с повторным прохождением ворот.
6) Управление выводом из эксплуатации (G7): обеспечивается сохранность доказательных артефактов, корректное прекращение обработки данных, перенос ответственности на резервные контуры и анализ уроков (таблица 1).

Ниже приведены три типовых сценария для критически значимых отраслей, демонстрирующие применение риск-ориентированных ворот допуска и артефактов аудита.

Таблица 1.

Минимальный набор доказательных артефактов по воротам допуска: составлено автором на основе [1–3; 5–7; 9–11]

Ворота	Ключевые артефакты	Опора на стандарты/ практики
G1	Паспорт ИИ-системы; классификация критичности; границы автономности; критерии приемлемости риска	ГОСТ Р ИСО/МЭК 42001-2024; ISO/IEC 23894; NIST AI RMF; AI Act
G2	Паспорт данных; карта потоков; протокол разметки; метрики качества и смещения; правовые основания	ГОСТ Р 70889-2023; ISO/IEC 23894; NIST AI RMF
G3	Репозиторий версий; SBOM/ML-SBOM; протокол обучения; контроль среды; защита целостности	ГОСТ Р 71539-2024; MITRE ATLAS
G4	Отчет ВиВ; тесты робастности; сценарии угроз; отчёт об объяснимости; протокол независимой проверки	NIST AI RMF; OWASP LLM Top 10; MITRE ATLAS
G5	План ввода; критерии отката; пороги метрик; план реагирования на инциденты; обучение персонала	ГОСТ Р ИСО/МЭК 42001-2024; ГОСТ Р 71539-2024
G6	Мониторинг дрейфа; журнал решений; отчеты инцидентов; периодическая переоценка риска; управление изменениями	ГОСТ Р 71539-2024; ISO/IEC 23894; NIST AI RMF
G7	План вывода; архив доказательств; прекращение обработки данных; итоговый отчет об уроках	ГОСТ Р ИСО/МЭК 42001-2024; ГОСТ Р 70889-2023

А) Клиническая поддержка принятия решений (лучевая диагностика). ИИ-модель используется для предварительного триажа исследований и подсказок врачу. Ключевой риск — ложное отрицание при редких патологиях, усиливаемое смещением данных (например, изменение протоколов съёмки) и дрейфом. На стадии G2 обязательна проверка репрезентативности и документирование источников; на G4 — независимая валидация

на отложенных выборках и стресс-тесты на артефакты изображений; на G6 — мониторинг дрейфа по распределениям признаков и по контролируемым подмножествам случаев.

Б) Прогнозирование аварийных режимов в энергосистеме. Модель прогнозирует вероятность перегрузки и рекомендует переключения. Риски носят системный характер: ошибка может привести к каскадному эффекту и нарушению непрерывности. Поэтому на G1 фиксируются границы автономности: рекомендации допускаются только при наличии подтверждения диспетчером, а в системе предусмотрены безопасные сценарии отказа. На G5 и G6 ключевыми становятся наблюдаемость, журналирование решений и процедуры быстрого отключения ИИ-компонента с переходом на регламентные алгоритмы.

В) Комплаенс-контур банка (скрининг санкций и противодействие отмыванию). Применение LLM для извлечения атрибутов и первичной классификации документов ускоряет обработку, но повышает риск галлюцинаций и некорректных выводов. В рекомендациях регулятора акцентируются принципы прозрачности, безопасности и ответственного управления рисками, а также необходимость регулярной проверки качества ИИ и соблюдения конфиденциальности персональных данных [14]. Отраслевые обзоры отмечают рост внедрений ИИ и необходимость более зрелого выбора сценариев и инфраструктуры, включая требования к комплаенсу и стоимости владения [15]. OWASP-классы рисков (внедрение вредоносных инструкций, небезопасная обработка выходных данных, цепочка поставок программного обеспечения) удобны как чек-лист для построения тестов на G4 и процедур мониторинга на G6 [3] (таблица 2).

Предложенная модель ориентирована на практическую применимость в организациях, эксплуатирующих ИИ-системы как часть критических контуров. Её внедрение требует решения трёх методических вопросов.

Во-первых, необходимо согласовать шкалы вероятности и последствий для разнородных рисков (технических, организационных, правовых). Стандарты системы менеджмента ИИ задают общую логику управления рисками, но не определяют отраслевые критерии; поэтому целесообразно формировать отраслевые профили риска и «каталоги контролей» по типам систем (медицинские решения, диспетчерские контуры, скоринговые модели).

Во-вторых, важно обеспечить измеримость и воспроизводимость доказательств. В критических отраслях аудит должен опираться на артефакты, которые можно проверить независимо: версии данных и моделей, протоколы обучения, результаты тестов, журналы решений, метрики дрейфа. Развитие практик threat-informed defense для ИИ и инженерная детализация сценариев атак становятся практической опорой для построения тестов на устойчивость и для расследования инцидентов [23].

В-третьих, требуется выстроить управление изменениями как управляемые релизы. Любая модификация данных, базовой модели, библиотек или параметров развёртывания должна трактоваться как потенциальное изменение уровня риска и, следовательно, как причина повторного прохождения соответствующих ворот допуска. Именно этот принцип переводит «доверие» из декларации в управляемое свойство.

В целом модель соответствует тенденции перехода от «проверки качества модели» к управлению рисками всей ИИ-системы и её окружения. Наличие в России стандартов ГОСТ Р ИСО/МЭК 42001-2024 и ГОСТ Р 71539-2024 создаёт базу для институционализации таких практик, а интеграция с MITRE ATLAS и OWASP LLM Top 10 позволяет учитывать современный ландшафт угроз.

Таблица 2.

Фрагмент риск-регистра по стадиям жизненного цикла (пример): составлено автором по материалам [1–3; 6–7; 11–12; 14–15]

Сценарий риска	Отрасль	Стадия	P (1–5)	I (1–5)	Ключевые меры/артефакты
Смещение данных из-за смены протокола диагностики ⟶ рост ложных отрицаний	Здравоохранение	G2/G6	3	5	Паспорт данных; независимая валидация; мониторинг дрейфа; пороги остановки
Подмена/модификация модели в контуре развёртывания	Энергетика	G3/G5	2	5	Контроль целостности; подписанные артефакты; ограничение прав; проверка в воротах
Внедрение вредоносных инструкций через внешние документы ⟶ искажение извлечённых атрибутов	Финансы	G4/G6	4	4	Тесты OWASP LLM; фильтрация входов/выходов; журналирование; человек-в-контуре
Деградация качества вследствие сезонности и изменения режима нагрузки	Транспорт	G6	3	4	Дрейф-метрики; переоценка риска; управляемые релизы; план отката

Риск-ориентированное управление жизненным циклом доверенных ИИ-систем является необходимым условием безопасного внедрения ИИ в критически значимые отрасли. В статье показано, что актуальные тенденции — стандартизация жизненного цикла, риск-ориентированное регулирование и развитие практик безопасности ИИ — могут быть объединены в единую модель «контура доказательного доверия».

Практическая ценность предложенной модели заключается в том, что модель:

1) формализует ворота допуска G0–G7 и минимальный набор аудируемых артефактов;
2) интегрирует управление данными, моделью и эксплуатацией в единый риск-контур;
3) снижает вероятность латентных отказов и атак за счёт независимой валидации и мониторинга дрейфа;
4) обеспечивает готовность к отраслевому и внутреннему аудиту благодаря воспроизводимости доказательств.

Дальнейшие исследования целесообразно направить на разработку отраслевых профилей рисков и эталонных каталогов контролей для конкретных классов высокорисковых ИИ-систем, а также на методики количественной оценки эффективности мер (включая экономическую оценку остаточного риска).

ЛИТЕРАТУРА

1. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. January 2023. — URL: https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf (дата обращения: 20.02.2026).
2. MITRE ATLAS™ (Adversarial Threat Landscape for Artificial-Intelligence Systems). — URL: https://atlas.mitre.org/ (дата обращения: 20.02.2026).
3. OWASP Top 10 for Large Language Model Applications (2023–2025). — URL: https://owasp.org/www-project-top-10-for-large-language-model-applications/ (дата обращения: 20.02.2026).
4. AI Act: Regulatory framework on artificial intelligence (Regulation (EU) 2024/1689). European Commission. — URL: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai (дата обращения: 20.02.2026).
5. ГОСТ Р ИСО/МЭК 42001-2024. Искусственный интеллект. Система менеджмента. — URL: https://protect.gost.ru/document1.aspx?control=31&baseC=6&id=263961 (дата обращения: 20.02.2026).
6. ГОСТ Р 71539-2024 (ИСО/МЭК 5338:2023). Искусственный интеллект. Процессы жизненного цикла системы искусственного интеллекта. — URL: https://protect.gost.ru/v.aspx?control=8&baseC=6&id=263966 (дата обращения: 20.02.2026).
7. ГОСТ Р 70889-2023 (ИСО/МЭК 8183:2023). Искусственный интеллект. Структура жизненного цикла данных. — URL: https://protect.gost.ru/v.aspx?control=8&baseC=6&id=244522 (дата обращения: 20.02.2026).
8. Аветисян А.И. Доверенный искусственный интеллект // Вестник Российской академии наук. 2024. Т. 94. № 3. С. 200–209. DOI: 10.31857/S0869587324030039.
9. ISO/IEC 5338:2023 Information technology — Artificial intelligence — AI system life cycle processes. — URL: https://www.iso.org/standard/81118.html (дата обращения: 20.02.2026).
10. ISO/IEC 23894:2023 Information technology — Artificial intelligence — Guidance on risk management. — URL: https://www.iso.org/standard/77304.html (дата обращения: 20.02.2026).
11. Котенко И.В., Саенко И.Б., Лаута О.С., Васильев Н.А., Садивников В.В. Атаки и методы защиты в системах машинного обучения: анализ современных исследований // Вопросы кибербезопасности. 2024. № 1(59). С. 24–37. DOI: 10.21681/2311–3456-2024-1-24-37.
12. Легашев Л.В., Жигалов А.Ю. Исследование состязательных атак на регрессионные модели машинного обучения в беспроводных сетях 5G // Вопросы кибербезопасности. 2024. № 3(61). С. 61–67. DOI: 10.21681/2311–3456-2024-3-61-67.
13. Карцхия А.А., Макаренко Г.И. Правовые горизонты технологий искусственного интеллекта: национальный и международный аспект // Вопросы кибербезопасности. 2024. № 1(59). С. 2–14. DOI: 10.21681/2311–3456-2024-1-2-14.
14. Пять принципов для искусственного интеллекта: рекомендации регулятора по применению ИИ. Банк России. 09.07.2025. — URL: https://www.cbr.ru/press/event/?id=25755 (дата обращения: 20.02.2026).
15. ИИ в финтехе — 2025: перспективные сценарии и лучшие практики. Альфа-Банк / T1. Ноябрь 2025. — URL: https://alfabank.st/marketing/d3/2b/marketing/II_v_fintekhe.pdf (дата обращения: 20.02.2026).
16. Росстандарт. Стандарты по направлению «Искусственный интеллект» (перечень действующих ГОСТ). — URL: https://www.rst.gov.ru/portal/gost/home/standarts/aistandarts (дата обращения: 20.02.2026).
17. Намиот Д.Е., Ильюшин Е.А. Доверенные платформы искусственного интеллекта: сертификация и аудит // CyberLeninka. 2024. — URL: https://cyberleninka.ru/article/n/doverennye-platformy-iskusstvennogo-intellekta-sertifikatsiya-i-audit (дата обращения: 20.02.2026).
18. Салов И.В. Угрозы информационной безопасности больших языковых моделей нейросетей // CyberLeninka. 2024. — URL: https://cyberleninka.ru/article/n/ugrozy-informatsionnoy-bezopasnosti-bolshih-yazykovyh-modeley-neyrosetey (дата обращения: 20.02.2026).
19. Костогрызов А.И. Анализ угроз злоумышленной модификации модели машинного обучения для систем с искусственным интеллектом // CyberLeninka. 2023. — URL: https://cyberleninka.ru/article/n/analiz-ugroz-zloumyshlennoy-modifikatsii-modeli-mashinnogo-obucheniya-dlya-sistem-s-iskusstvennym-intellektom (дата обращения: 20.02.2026).

20. Будко Д.В. Концепция риск-ориентированного подхода в условиях развития технологий искусственного интеллекта // CyberLeninka. 2024. — URL: https://cyberleninka.ru/article/n/kontseptsiya-risk-orientirovannogo-podhoda-v-usloviyah-razvitiya-tehnologiy-iskusstvennogo-intellekta (дата обращения: 20.02.2026).
21. Билятдинов К.З. Управление качеством разработки отечественных систем доверенного искусственного интеллекта // CyberLeninka. 2025. — URL: https://cyberleninka.ru/article/n/upravlenie-kachestvom-razrabotki-otechestvennyh-sistem-doverennogo-iskusstvennogo-intellekta (дата обращения: 20.02.2026).
22. Намиот Д.Е. Об оценке доверия к системам искусственного интеллекта // CyberLeninka. 2025. — URL: https://cyberleninka.ru/article/n/ob-otsenke-doveriya-k-sistemam-iskusstvennogo-intellekta (дата обращения: 20.02.2026).
23. Secure AI with Threat-Informed Defense (v2). MITRE CTID. 09.05.2025. — URL: https://ctid.mitre.org/blog/2025/05/09/secure-ai-v2/ (дата обращения: 20.02.2026).
24. Искусственный интеллект обрел Кодекс этики на финансовом рынке // ComNews. 11.07.2025. — URL: https://www.comnews.ru/content/240125/2025-07-11/2025-w28/1008/iskusstvennyy-intellekt-obrel-kodeks-etiki-finansovom-rynke (дата обращения: 20.02.2026).