УДК 004.056:004.8:006(470+100)

РЕГУЛИРОВАНИЕ ВОПРОСОВ БЕЗОПАСНОСТИ, ДОВЕРИЯ И УСТОЙЧИВОСТИ ПРИ ИСПОЛЬЗОВАНИИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА: РОССИЙСКИЕ И МЕЖДУНАРОДНЫЕ СТАНДАРТЫ

REGULATION OF SAFETY, TRUST, AND RESILIENCE IN ARTIFICIAL INTELLIGENCE: RUSSIAN AND INTERNATIONAL STANDARDS

Иконников А.В., студент кафедры «Интел-лектуальные системы информационной безопасности», ФГБОУ ВО «МИРЭА – Российский технологический университет», г. Москва

Спирин А.А., канд. техн. наук, доцент кафедры «Интел-лектуальные системы информационной безопасности», ФГБОУ ВО «МИРЭА – Российский технологический университет», г. Москва

Ikonnikov A.V., student of the Department of "Intelligent Information Security Systems", MIREA – Russian Technological University, Moscow

Spirin A.A., Ph.D., docent of the Department of "Intelligent Information Security Systems", MIREA – Russian Technological University, Moscow

На фоне интенсивного внедрения искусственного интеллекта в различные сектора экономики и сферы социальной жизни возникает острое противоречие между скоростью технологического прогресса и отсутствием зрелых механизмов управления сопутствующими рисками. Данная статья посвящена анализу этого разрыва. Цель работы – систематизировать и сопоставить формирующуюся «архитектуру» российских и международных стандартов, регулирующих безопасность, доверие, робастность (устойчивость) интеллектуальных систем. Резюмировано, что глобальная и национальная регуляторика переходит от декларативных принципов к созданию конкретных, технически верифицируемых требований. Авторский вклад состоит в том, что разрозненные стандарты (касающиеся терминологии, оценки рисковых факторов, формальных методов верификации нейронных сетей) представлены как элементы единой, взаимосвязанной системы – «архитектуры доверия».

Ключевые слова: доверие, искусственный интеллект, международные стандарты, национальные стандарты, робастность, стандартизация.

Для цитирования: Иконников А.В., Спирин А.А. Регулирование вопросов безопасности, доверия и устойчивости при использовании искусственного интеллекта: российские и международные стандарты // Информационно-экономические аспекты стандартизации и технического регулирования. 2026. № 1(88). С. 11–15.

Amid the intensive deployment of artificial intelligence across various sectors of the economy and spheres of social life, an acute contradiction emerges between the pace of technological progress and the lack of mature mechanisms for managing the associated risks. This article addresses the analysis of this gap. The purpose of the study is to systematize and compare the emerging "architecture" of Russian and international standards governing the safety, trustworthiness, and robustness (resilience) of intelligent systems. It is concluded that both global and national regulatory frameworks are shifting from declarative principles toward the development of concrete, technically verifiable requirements. The author's contribution lies in conceptualizing fragmented standards, relating to terminology, risk factor assessment, and formal methods for neural network verification, as elements of a unified, interconnected system referred to as an "architecture of trust".

Keywords: trust, artificial intelligence, international standards, national standards, robustness, standardization.

For citation: Ikonnikov A.V., Spirin A.A. Regulation Of Safety, Trust, And Resilience In Artificial Intelligence: Russian And International Standards. Information and Economic Aspects of Standardization and Technical Regulation. 2026; 1(88): 11–15. (In Russ.).

ВВЕДЕНИЕ

Актуальность темы обусловлена стремительным внедрением технологий искусственного интеллекта в значимые сферы экономики и государственного управления. Это усиливает риски, которые сопряжены с безопасностью, надежностью, общественным доверием к алгоритмическим решениям. Отсутствие унифицированных и согласованных подходов к регулированию ИИ, асимметрия между национальными и международными стандартами порождают правовую неопределенность, затрудняют трансграничное использование технологий. В складывающихся условиях сопоставительный анализ российских и международных стандартов приобретает особую ценность, поскольку помогает выявить точки конвергенции и расхождения в подходах к обеспечению безопасности, устойчивости, доверия, а также определить ориентиры гармонизации регулирования на фоне ускоренной цифровизации.

Стремительное проникновение технологий искусственного интеллекта (ИИ) в самые разные сферы социально-экономической жизни знаменует собой новую технологическую революцию и ставит перед обществом, государством, бизнесом комплексные вызовы, связанные с управлением рисками. Вопросы безопасности, этики, интерпретируемости, устойчивости систем искусственного интеллекта (СИИ) перестают быть исключительно академическими, они переходят в разряд ключевых факторов, которые определяют возможность их практического внедрения в критически важных областях (медицина, транспорт, финансы, государственное управление и т.д.). Отсутствие прозрачных и общепринятых механизмов контроля способно привести к непредсказуемым сбоям, дискриминации, нарушению конфиденциальности данных и, как следствие, к подрыву общественного доверия к технологии в целом. В рассматриваемом контексте формирование адекватной регуляторной среды становится не барьером для инноваций, а весьма значимым условием их устойчивого, безопасного развития. Глобальная дискуссия о регулировании ИИ постепенно смещается от декларативных принципов к созданию конкретных нормативно-технических инструментов, среди которых центральную роль играет стандартизация. Именно стандарты призваны заложить «архитектуру доверия» – речь идет о едином понятийном аппарате и измеримых критериях, помогающих оценивать и верифицировать такие сложные свойства ИИ, как надежность и робастность.

Анализ современной литературы по обсуждаемой теме показывает, что исследователи по-разному акцентируют внимание на аспектах регулирования искусственного интеллекта с точки зрения безопасности, доверия и устойчивости. Часть авторов сосредоточена на международных стандартах ISO и их роли в формировании глобальной нормативной базы. Так, Л. Н. Варламова рассматривает систему стандартов ISO/IEC, обеспечивающих безопасность и прозрачность алгоритмов [1], К. Локетт и В. Скиданова описывают подходы к формализации требований к ИИ на раннем этапе стандартизации [2]. Другие исследователи делают акцент на разработке национальных документов: С.В. Гарбук и А.П. Шалаев предлагают концептуальную структуру российского комплекса норм, ориентированного на интеграцию с международными практиками [3, 4]; М.В. Екатеринин, А.А. Колючкин с коллегами раскрывают проблематику предотвращения технологических рисков ИИ [5, 6]. Отдельный блок публикаций посвящен этическим и гуманитарным аспектам. Так, Н.А. Лисицина и А.В. Линкина исследуют соотношение стандартов этики ИИ и защиты персональных данных [7], С.П. Прохоров рассматривает эволюцию этических принципов в международной практике [8].

Невзирая на схожую направленность работ, в источниках имеют место расхождения относительно механизмов имплементации в национальное законодательство. Недостаточно изученными остаются вопросы относительно оценки эффективности стандартов ИИ в обеспечении доверия пользователей и устойчивости инфраструктур.

Целью настоящего исследования является систематизация формируемой «архитектуры» российских и международных стандартов, регулирующих безопасность, доверие, робастность (устойчивость) интеллектуальных систем.

В ходе написания статьи применялись методы сравнительно-правового анализа, контент-оценки нормативных актов, системного подхода, обобщения, интерпретации научных источников.

АНАЛИЗ НАЦИОНАЛЬНЫХ СТАНДАРТОВ В ОБЛАСТИ СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Формирование доверия к системам искусственного интеллекта – это процесс институционального, технического и социального обеспечения предсказуемого, объяснимого, контролируемого функционирования СИИ, основанный на соблюдении установленных стандартов безопасности, устойчивости, ответственности, а также на возможности их верификации, аудита, подотчетности в рамках правового и этического регулирования. Речь идет о многоплановой задаче, требующей согласованного подхода к терминологии, методологии оценки, управлению рисками на всех этапах жизненного цикла. Российская система стандартизации в описываемой области активно развивается, закладывая нормативный «фундамент», во многом гармонизированный с международными разработками [3, 7].

Так, основополагающим документом, с помощью которого создается единое понятийное поле, является ГОСТ Р 71476–2024 «Искусственный интеллект. Концепции и тер-

минология»[1]. Этот стандарт, соответствующий международному ISO/IEC 22989:2022, систематизирует ключевые термины, определяет взаимосвязи между ними. Посредством него вводятся формальные определения для таких понятий, как «система искусственного интеллекта», «машинное обучение», «нейронная сеть», а также для характеристик СИИ, в том числе «прозрачность», «интерпретируемость», «предвзятость». Установление единой терминологии служит критически важным первым шагом, поскольку устраняется неоднозначность трактовок и создается базис для разработки более специализированных технических регламентов и методик оценивания.

Центральное место в регулировании занимает концепция «доверия», которая рассматривается не как субъективное ощущение, а как измеримое свойство системы, подтверждаемое объективными данными [1, 5, 8], включая их качество [9]. ГОСТ Р 59276–2020[2] является ключевым документом в данной области. В нем представлена рамочная структура, определяющая свойства, формирующие доверие к СИИ, включая надежность, безопасность, прозрачность, управляемость и ответственность.

Заложенные в ГОСТ Р 59276–2020 принципы могут служить основой для риск-ориентированного подхода на протяжении всего жизненного цикла системы – от проектирования и сбора данных до эксплуатации и вывода из использования. Такой подход предполагает выявление потенциальных угроз и факторов, способных снизить уровень доверия к системе, их оценку и принятие мер по обеспечению требуемого уровня надежности и безопасности [10]. Для каждой конкретной системы это может включать идентификацию потенциальных рисков, оценку их вероятности и тяжести последствий, а также разработку соответствующих мер смягчения.

Если упомянутый выше стандарт задает общую концептуальную основу, то последующие документы детализируют технические аспекты обеспечения надежности и устойчивости СИИ.

Так, робастность (устойчивость) – это способность нейронной сети сохранять корректность и стабильность функционирования, а также приемлемый уровень качества результатов при воздействии искажений, шумов, непредвиденных вариаций во входных данных, при изменении условий эксплуатации, не представленных явно в обучающей выборке. Она опирается на устойчивость модели к случайным ошибкам, целенаправленным возмущениям, частичным нарушениям структуры данных, что помогает смягчить риски деградации решений в сочетании с повышением надежности применения систем искусственного интеллекта в реальных и критически значимых сценариях [2]. Описываемый аспект очень важен, поскольку в реальном мире данные редко бывают идеальными [6]. Ниже представлена серия из двух стандартов, посвященных этому вопросу.

ГОСТ Р 70462.1–2022[3] вводит классификацию угроз (например, состязательные атаки, случайные шумы, сдвиг данных) и содержит общие подходы к тестированию. Помогает разработчикам и заказчикам понять, от каких именно вызовов должна быть защищена система.

ГОСТ Р ИСО/МЭК 24029-2–2024[4] переходит от обзора к конкретным методикам. Формальные методы – это математически строгие подходы, которые позволяют как протестировать систему на конечном наборе примеров, так и доказать (или опровергнуть) ее устойчивость к целому классу возможных возмущений. Благодаря этому обеспечивается гораздо более высокий уровень гарантий надежности по сравнению с эмпирическим тестированием.

Еще одно весьма значимое направление – интеграция ИИ в системы, где отказ может привести к ущербу для жизни и здоровья людей. Проект национального стандарта ПНСТ 836–2023[5] (на основе ISO/IEC DTR 5469) решает именно эту задачу. Анализируется, как традиционные подходы к функциональной безопасности (например, стандарты серии IEC 61508) реально адаптировать для систем с компонентами на основе машинного обучения. Ключевая проблема здесь проявляется в том, что СИИ по своей природе являются недетерминированными, и доказать их стопроцентную корректность традиционными методами невозможно. С опорой на рассматриваемый стандарт предлагаются новые подходы к анализу опасностей и оценке рисков, где учитывается специфика ИИ (вероятностный характер ошибок, сложность верификации).

Российская система стандартизации ИИ развивается в тесной связи с международными усилиями (таблица), что является стратегически верным решением, обеспечивающим совместимость технологий и регуляторных практик. Принятие серии ГОСТ Р ИСО/МЭК служит прямым свидетельством курса на гармонизацию. Это означает, что разработчики и регуляторы в Российской Федерации работают в едином поле с мировым сообществом, применяя общие методологии и метрики. Между тем существуют и

[1] ГОСТ Р 71476–2024 (ИСО/МЭК 22989:2022). Искусственный интеллект. Концепции и терминология искусственного интеллекта. (Введ. 2025-10-01).

[2] ГОСТ Р 59276–2020. Системы искусственного интеллекта. Способы обеспечения доверия. Общие положения. – Введ. 2021-03-01.

[3] ГОСТ Р 70462.1–2022 / ISO/IEC TR 24029-1–2021. Информационные технологии. Интеллект искусственный. Оценка робастности нейронных сетей. Часть 1. Обзор. (Введ. 2023-01-01).

[4] ГОСТ Р ИСО/МЭК 24029-2–2024. Искусственный интеллект. Оценка робастности нейронных сетей. Часть 2. Методология использования формальных методов. (Введ. 2025-01-01).

[5] ПНСТ 836–2023 (ISO/IEC DTR 5469). Искусственный интеллект. Функциональная безопасность и системы искусственного интеллекта. (Срок действия: с 2024-01-01 до 2027-01-01).

национальные акценты. Например, ГОСТ Р 59276–2020, хотя и опирается на международные концепции, является самостоятельной разработкой, агрегирующей различные аспекты доверия в единую структуру, которая адаптирована для российских реалий.

Таким образом, российская нормативная база не изолирована – она является частью глобального процесса по созданию универсальных «правил игры» для разработчиков и пользователей ИИ.

Анализ текущего состояния стандартизации в характеризуемой сфере позволяет сформулировать ряд рекомендаций по последующему развитию регуляторной базы.

Так, существующие документы носят общий, рамочный характер. Для их практического применения целесообразно создание отраслевых профилей (к примеру, для беспилотного автотранспорта, кредитного скоринга), которые конкретизировали бы требования к безопасности, робастности и прозрачности с учетом специфических рисков каждого домена. Например, уже существует целая серия стандартов, разработанных специально для медицинской диагностики, начиная с ГОСТ Р 59921.1–2022[6]. Создание стандартов для других областей решит проблему чрезмерной абстрактности базовых стандартов и весомо облегчит их внедрение.

Для того чтобы стандарты работали, необходим механизм подтверждения соответствия. Видится уместным проработать вопрос о создании аккредитованных лабораторий и центров сертификации, которые способны проводить независимую оценку СИИ (особенно в критически значимых сферах) на соответствие требованиям ГОСТ по робастности, предвзятости и безопасности. Новизна предлагаемого шага проявляется в переходе от добровольного применения стандартов к обязательному подтверждению соответствия для систем с высоким уровнем риска.

[6] ГОСТ Р 59921.1–2022. Системы искусственного интеллекта в клинической медицине. Часть 1. Клиническая оценка. (Введ. 2022-09-01).

ЗАКЛЮЧЕНИЕ

Посредством проведенного анализа показывается, что в России и на международном уровне активно формируется многоуровневая система технического регулирования, призванная обеспечить надежность и устойчивость систем искусственного интеллекта. Исследуемый процесс движется от общих принципов к конкретным, измеримым, верифицируемым требованиям, заложенным в национальных и международных стандартах. Ключевой парадигмой этого регулирования становится концепция «доверия», понимаемая как комплексное свойство системы. Российская Федерация, с одной стороны, деятельно гармонизирует свою нормативную базу с международной, принимая стандарты ISO/IEC в качестве национальных ГОСТов (как в случае с оценкой робастности), что обеспечивает включенность в глобальный технологический контекст, а с другой – разрабатывается и собственные комплексные документы, где систематизируются подходы к укреплению доверия. Отмечено, что его «архитектура» строится на следующих принципах: унифицированной терминологии, принципах риск-ориентированного анализа на протяжении жизненного цикла, конкретных методологиях оценки технических характеристик.

Практическое значение проведенного исследования заключается в том, что оно предоставляет разработчикам, регуляторам, заказчикам СИИ структурированное видение той нормативной «рамки», которая будет определять их деятельность в ближайшие годы.

Дальнейшие исследования в данной области представляется логичным сосредоточить на развитии технических и методологических показателей, критически значимых для

Сравнительный анализ стандартов и международный контекст (составлено автором на основе анализа рассматриваемых стандартов)

АСПЕКТ СТАНДАРТИЗАЦИИ	ПОДХОДЫ		СООТНОШЕНИЕ
	Российский	Международный	
Терминология	ГОСТ Р 71476–2024	ISO/IEC 22989:2022	Высокая степень согласованности. Российский стандарт является идентичным международному
Управление рисками и доверие	ГОСТ Р 59276–2020	ISO/IEC TR 24028 (Trustworthiness), ISO/IEC 23894 (Risk Management)	Концептуальная общность. ГОСТ Р 59276–2020 представляет собой комплексную структуру, ISO/IEC предлагает серию более гранулярных документов
Робастность нейронных сетей	ГОСТ Р ИСО/МЭК 24029 (части 1 и 2)	ISO/IEC 24029 (parts 1 & 2)	Прямое принятие международного стандарта, полная гармонизация
Функциональная безопасность	ПНСТ 836–2023	ISO/IEC DTR 5469	Прямое использование проекта международного технического отчета в качестве базиса для национального стандарта

безопасного и ответственного применения ИИ, включая объяснимость и интерпретируемость моделей, устойчивость к ошибкам, внешним воздействиям, воспроизводимость результатов, корректность формальной верификации алгоритмов. Пристальное внимание следует уделить анализу правовых и этических последствий стандартизации и сертификации автономных систем, поскольку именно через формирование измеримых и проверяемых технических требований стандартизация обеспечивает доверие, безопасность, зрелость рынка технологий искусственного интеллекта.

Список литературы

1. Варламова Л.Н. Международные стандарты ИСО, регламентирующие вопросы искусственного интеллекта (ИИ) // Делопроизводство. 2024. № 2. С. 13–19.
2. Локетт К., Скиданова В. Стандарты для искусственного интеллекта // Управление качеством. 2020. № 2. С. 74–76.
3. Гарбук С.В., Шалаев А.П. Перспективная структура национальных стандартов в области искусственного интеллекта // Стандарты и качество. 2021. № 10. С. 26–33.
4. Гарбук С.В. Метод оценки влияния параметров стандартизации на эффективность создания и применения систем искусственного интеллекта // Информационно-экономические аспекты стандартизации и технического регулирования. 2022. № 3(67). С. 4–14.
5. Екатеринин М.В. Искусственный интеллект: предотвращение рисков с помощью стандартов // Методы менеджмента качества. 2022. № 7. С. 44–47.
6. Колючкин А.А., Кудрявцев Д.П., Ковалеров Я.Д. Технологические аспекты развития искусственного интеллекта, стандарты и безопасность // Студенческий научный форум 2025. Сборник статей XVII Международной научно-практической конференции. – Пенза: Наука и Просвещение, 2025. – С. 60–62.
7. Лисицина Н.А., Линкина А.В. Этические принципы и стандарты в сфере искусственного интеллекта и обработки персональных данных // Молодежь в науке: экономика, технологии и инновации. Материалы Международной научно-практической конференции. – Воронеж: ИПЦ «Научная книга», 2023. – С. 385–389.
8. Прохоров С.П. Международные стандарты по этике искусственного интеллекта: история и развитие // Материалы II Международной конференции Российского национального комитета по истории и философии науки и техники РАН, посвященной 300-летию Российской академии наук. – М.: ИИЕТ им. С.И. Вавилова РАН, 2024. – С. 67–70.
9. Бурый А.С., Погодин И.М. Оценка качества больших данных часть 1. Основные понятия и метрики // Информационно-экономические аспекты стандартизации и технического регулирования. 2024. № 3(78). С. 49–58.
10. Glikson E., Woolley A.W. Human trust in artificial intelligence: Review of empirical research // Academy of management annals. 2020. Т. 14. № 2. С. 627–660.

References

1. Varlamova L.N. International ISO Standards Governing Artificial Intelligence (AI) Issues. Office Management. 2024, no. 2, pp. 13–19. (In Russian).
2. Locke K., Skidanova V. Standards for Artificial Intelligence. Quality Management, 2020, no. 2, pp. 74–76. (In Russian).
3. Garbuk S.V., Shalaev A. P. Promising Structure of National Standards in the Field of Artificial Intelligence Standards and Quality, 2021, no. 10, pp. 26–33. (In Russian).
4. Garbuk S.V. A method for assessing the impact of standardization parameters on the effectiveness of creating and applying artificial intelligence systems. Information and economic aspects of standardization and technical regulation, 2022, no. 3(67), pp. 4–14. (In Russian).
5. Ekaterinin M.V. Artificial Intelligence: Risk Prevention Using Standards. Quality Management Methods, 2022, no. 7, pp. 44–47. (In Russian).
6. Koluchkin A.A., Kudryavtsev D.P., Kovalerov Y.D. Technological aspects of the development of artificial intelligence, standards and security. Student Scientific Forum 2025. Collection of articles of the XVII International Scientific and Practical Conference. Penza: 2025, pp. 60–62. (In Russian).
7. Lisitsina N.A., Linkina A.V. Ethical principles and standards in the field of artificial intelligence and personal data processing. Youth in Science: Economics, Technologies and Innovations. Materials of the International Scientific and Practical Conference. Voronezh: 2023, pp. 385–389. (In Russian).
8. Prokhorov S.P. International Standards on the Ethics of Artificial Intelligence: History and Development // Materials of the II International Conference of the Russian National Committee on the History and Philosophy of Science and Technology of the Russian Academy of Sciences, dedicated to the 300th Anniversary of the Russian Academy of Sciences. Moscow: 2024, pp. 67–70. (In Russian).
9. Buryi A.S., Pogodin I.M. Ocenka kachestva bol'shih dannyh chast' 1. Osnovnye ponyatiya i metriki // Information and economic aspects of standardization and technical regulation. 2024, no. 3(78), pp. 49–58. (In Russian).
10. Glikson E., Woolley A.W. Human trust in artificial intelligence: Review of empirical research. Academy of management annals. 2020, no. 14(2), pp. 627–660.