doi: 10.51639/2713-0576_2026_6_2_60
Научная статья
УДК 004.912
ГРНТИ 20.23.25
ВАК 2.3.5

Особенности сертификации нейросетевых моделей в условиях формирования нормативной базы

Владислав Валерьевич Душкин
Краснодарское высшее военное училище имени генерала армии С.М. Штеменко, г. Краснодар, Россия
dushkin@list.ru

Аннотация

В статье рассматриваются актуальные подходы к сертификации нейросетевых моделей как особого вида продукции, содержащей искусственный интеллект. Анализируются формирующиеся нормативно-правовые требования в Российской Федерации и международные стандарты в данной области. Выявляются ключевые особенности оценки соответствия нейросетевых моделей, включая специфику подтверждаемых показателей, методологию тестирования и процедурные аспекты сертификации.

Предлагается классификация метрик качества и безопасности, а также рассматриваются проблемы обеспечения достоверности оценки. Делается вывод о необходимости комплексного подхода к сертификации, учитывающего как технические параметры, так и аспекты этики и безопасности.

Ключевые слова: нейросетевые модели, сертификация, искусственный интеллект, стандартизация, оценка соответствия, метрики качества, безопасность ИИ.

Введение

Стремительное внедрение технологий искусственного интеллекта (ИИ) в критически важные сферы – здравоохранение, транспорт, финансы, государственное управление – актуализирует проблему подтверждения качества и безопасности соответствующей продукции. Нейросетевые модели, как ядро современных систем ИИ, обладают рядом особенностей, отличающих их от традиционных программных продуктов: недетерминированность поведения, зависимость от обучающих данных, сложность интерпретации принимаемых решений, способность к обобщению и, одновременно, уязвимость к специфическим искажениям входной информации.

Данные обстоятельства обусловливают необходимость формирования специальных механизмов оценки соответствия. Как справедливо отмечается в проекте национального стандарта по сертификации продукции с использованием ИИ, создание системы требований «повышает уровень доверия к результату сертификации ... и способствует укреплению технологического суверенитета России».

Цель настоящей статьи – систематизировать особенности сертификации нейросетевых моделей, проанализировать формирующуюся нормативную базу и выявить ключевые проблемы в данной области.

В Российской Федерации активно формируется нормативная база, регламентирующая требования к системам искусственного интеллекта и процедуры их оценки. Технический комитет по стандартизации ТК 164 «Искусственный интеллект» выступает основным разработчиком документов в данной области. К числу действующих и разрабатываемых нормативных актов, имеющих значение для сертификации нейросетевых моделей, относятся:

– ПНСТ 835-2023 «Искусственный интеллект. Оценка эффективности моделей и алгоритмов машинного обучения в задаче классификации» (введен в действие с 01.01.2024 на период до 01.01.2027). Стандарт устанавливает методологию оценки эффективности моделей машинного обучения, включая требования к выбору показателей, подготовке наборов данных и интерпретации результатов [1];

– ГОСТ Р 70462.1-2022 – определяет методы оценки качества работы алгоритмов машинного обучения и способности нейронных сетей поддерживать заданный уровень производительности [2];

– ГОСТ Р 59276-2020 – посвящен способам обеспечения доверия к системам искусственного интеллекта, что является важнейшим аспектом сертификации [3];

– ГОСТ Р 71752-2024 – регламентирует структуру и содержание технического задания на разработку систем ИИ, что создает основу для последующей верификации [4].

Параллельно с российскими формируется и международная нормативная база:

– ISO/IEC 42006:2025 – документ определяет критерии компетентности, и надежности органов, проводящих аудит и сертификацию в области ИИ [5];

– ISO/IEC 24028 – посвящен вопросам обеспечения доверия к системам ИИ (transparency, robustness, safety) [6];

– ISO/IEC 23053 – определяет структуру описания систем ИИ и жизненного цикла их функционирования [7].

Таким образом, формируется двухуровневая система регулирования: на уровне требований к самой продукции (нейросетевым моделям) и на уровне требований к органам по сертификации и процедурам подтверждения соответствия.

Специфика нейросетевых моделей как объекта сертификации проявляется в многообразии показателей, подлежащих оценке. Традиционные подходы к сертификации программного обеспечения ориентированы преимущественно на проверку функциональных характеристик и стабильности работы. Для систем с ИИ этого недостаточно.

На основе анализа нормативных документов можно выделить следующие группы показателей, подлежащих подтверждению при сертификации нейросетевых моделей (таблица 1).

Как видно из таблицы, нейросетевые модели требуют многокритериальной оценки, причем многие показатели являются специфическими именно для систем ИИ.

В случае несбалансированных данных традиционные метрики могут завышать реальную эффективность модели, что требует применения дополнительных оценок – макроусредненных показателей, взвешенных метрик и других.

Сертификационные испытания нейросетевых моделей предполагают решение ряда методологических проблем:

– проблема репрезентативности тестовых данных. Для получения достоверных результатов тестовые наборы данных должны отражать реальное распределение входных воздействий в предполагаемых условиях эксплуатации. При этом необходимо исключить ситуацию, когда тестовые примеры попадали в обучающую выборку. В противном случае оценки эффективности будут завышенными и недостоверными;

Таблица 1 – Классификация показателей качества и безопасности нейросетевых моделей

Категория	Показатели	Нормативная основа	Методы оценки
Функциональная эффективность	Точность (accuracy), полнота (recall), F1-мера, AUC-ROC	ПНСТ 835-2023, ISO/IEC 4213 [8]	Тестирование на репрезентативных наборах данных, кросс-валидация
Робастность	Устойчивость к шумам, состязательным атакам (adversarial attacks), дрейфу данных	ISO/IEC 24028	Стресс-тестирование, тестирование на граничных значениях
Интерпретируемость	Объяснимость решений, возможность отслеживания логики вывода	ГОСТ Р 59276-2020, ISO/IEC 25023 [9]	Методы XAI (SHAP, LIME), анализ внимания
Безопасность	Отсутствие токсичных генераций, защита от утечек данных	SafetyBench [10] Perspective API [11]	Тестирование с провокационными запросами, анализ распределения выходов
Этическая приемлемость	Отсутствие дискриминационных предубеждений (bias), справедливость	–	Тестирование на стратифицированных выборках, анализ по защищенным атрибутам
Энергоэффективность	Вычислительная сложность, энергопотребление	–	Профилирование, бенчмаркинг

– проблема полноты тестирования. Генеративные нейросети работают в условиях высокой неопределенности и могут порождать бесконечное множество вариантов выходного контента. Это делает невозможным тестирование всех возможных сценариев. Решением становится применение методов выборочного тестирования с проверкой наиболее вероятных и наиболее критичных сценариев, а также оценка устойчивости модели к вариациям входных данных;

– проблема учета контекста. Одна и та же нейросетевая модель может демонстрировать различную эффективность в зависимости от конкретного применения;

– проблема динамического обновления. Нейросетевые модели в процессе эксплуатации могут дообучаться или адаптироваться к новым данным. Это создает сложность для сертификации по статической модели: требуется либо сертифицировать не конкретную модель, а процесс ее разработки и сопровождения (что соответствует подходу ISO/IEC 42006 к сертификации систем менеджмента), либо вводить процедуры периодической ресертификации [5].

Сертификация нейросетевых моделей как процедура подтверждения соответствия включает несколько этапов, представленных на рисунке 1.

Рисунок 1 – Этапы сертификации нейросетевых моделей

Особое значение приобретает этап формирования программы испытаний. В отличие от традиционного ПО, где тестовые сценарии могут быть достаточно полно определены на этапе разработки, для нейросетевых моделей требуется вероятностный подход к тестированию, учитывающий стохастическую природу их функционирования.

Современные исследования предлагают создание специализированных платформ валидации, интегрирующих различные модули оценки. Например, платформа AIVeritas, обеспечивает оценку качества данных, производительности модели, диагностику объяснимости и формирование сертификатов на основе картирования требований стандартов ISO/IEC.

Экспериментально подтверждено, что каждый модуль такой платформы является необходимым для обеспечения достоверности и полноты оценки.

Проведенный анализ позволяет выделить следующие ключевые проблемы в области сертификации нейросетевых моделей:

– недостаточная гармонизация требований. Существующие нормативные документы разрабатываются параллельно различными организациями, что создает риски дублирования и противоречивости требований;

– сложность верификации недетерминированного поведения. Нейросетевые модели, особенно большие языковые модели (LLM), демонстрируют вариативность ответов, что затрудняет применение классических подходов к тестированию;

– проблема полноты проверки безопасности. Потенциально опасные сценарии использования нейросетевых моделей (генерация вредоносного контента, раскрытие конфиденциальной информации, манипуляция пользователем) трудно полностью предусмотреть и проверить на этапе сертификации. Кроме того, модели могут демонстрировать латентные предубеждения, не проявляющиеся в стандартных тестах;

– отсутствие единых стандартов на данные обучения. Качество нейросетевой модели критически зависит от качества обучающих данных, однако требования к формированию датасетов остаются недостаточно формализованными;

– высокая стоимость сертификации. Учитывая необходимость многокритериальной оценки с привлечением специализированных лабораторий и вычислительных мощностей, сертификация нейросетевых моделей может быть экономически доступна только для крупных разработчиков.

Перспективы развития сертификации нейросетевых моделей связаны со следующими направлениями:

– разработка специализированных стандартов. Продолжение работы по созданию стандартов, учитывающих специфику различных типов нейросетевых моделей (генеративные, рекомендательные, классифицирующие) и областей их применения;

– внедрение механизмов «песочниц» (regulatory sandboxes). Создание экспериментальных правовых режимов, позволяющих тестировать инновационные модели ИИ в контролируемых условиях с участием регулятора;

– развитие методов непрерывной сертификации. Переход от разовой сертификации к мониторингу соответствия в процессе всего жизненного цикла модели с использованием автоматизированных средств контроля;

– стандартизация бенчмарков и тестовых наборов данных. Формирование официально признанных наборов тестовых данных для различных областей применения, что обеспечит сопоставимость результатов сертификации;

– интеграция требований к этичности и безопасности. Разработка методологии комплексной оценки, включающей не только технические показатели, но и соответствие этическим принципам, требованиям недискриминации и социальной ответственности.

Сертификация нейросетевых моделей представляет собой сложную, многоплановую задачу, решение которой требует учета как технических особенностей данных систем, так и формирующейся нормативно-правовой базы.

Заключение

Проведенный анализ позволяет сделать следующие выводы:

1. В Российской Федерации и на международном уровне активно развивается система стандартизации в области искусственного интеллекта, включающая как общие требования к системам ИИ, так и специальные требования к процедурам сертификации.
2. Специфика нейросетевых моделей как объекта сертификации обусловливает необходимость многокритериальной оценки, включающей показатели функциональной эффективности, робастности, интерпретируемости, безопасности, этической приемлемости и энергоэффективности.
3. Методология сертификационных испытаний должна учитывать вероятностную природу нейросетевых моделей, проблему репрезентативности тестовых данных и необходимость проверки устойчивости к состязательным атакам.
4. Процедуру сертификации целесообразно выстраивать как последовательность этапов: анализ документации, формирование программы испытаний, проведение тестирования, анализ результатов, выдача сертификата и инспекционный контроль.
5. Основными проблемами в рассматриваемой области являются недостаточная гармонизация требований, сложность верификации недетерминированного поведения, проблема полноты проверки безопасности и высокая стоимость сертификации.

Дальнейшее развитие института сертификации нейросетевых моделей будет способствовать повышению доверия к технологиям искусственного интеллекта, защите прав потребителей и укреплению технологического суверенитета.

Список источников

1. ПНСТ 835–2023. Искусственный интеллект. Оценка эффективности моделей и алгоритмов машинного обучения в задаче классификации: национальный стандарт Российской Федерации: издание официальное: утвержден и введен в действие Приказом Федерального агентства по техническому регулированию и метрологии от 29.12.2023 № 1649-пнст. М.: ФГБУ «РСТ», 2023. – 32 с.
2. ГОСТ Р 70462.1–2022 (ISO/IEC TR 24029–1:2021). Информационные технологии. Искусственный интеллект. Оценка робастности нейронных сетей. Часть 1. Обзор: издание официальное: утвержден и введен в действие Приказом Федерального агентства по техническому регулированию и метрологии от 29.11.2022 № 1355-ст. – М.: ФГБУ «РСТ», 2022. – 32 с.
3. ГОСТ Р 59276–2020. Системы искусственного интеллекта. Способы обеспечения доверия. Общие положения: национальный стандарт Российской Федерации: издание официальное: утвержден и введен в действие Приказом Федерального агентства по техническому регулированию и метрологии от 21.12.2020 № 1354-ст. – М.: Стандартинформ, 2020. – 16 с.
4. ГОСТ Р 71752–2024. Искусственный интеллект. Техническое задание. Требования к содержанию: национальный стандарт Российской Федерации: издание официальное: утвержден и введен в действие Приказом Федерального агентства по техническому регулированию и метрологии от 28.10.2024 № 1548-ст.– М.: ФГБУ «РСТ», 2024. – 20 с.
5. ISO/IEC 42006:2025. Информационные технологии. Искусственный интеллект. Требования к органам, проводящим аудит и сертификацию систем менеджмента искусственного интеллекта. – Женева: Международная организация по стандартизации, 2025. – VI, 34 с.
6. ISO/IEC TR 24028:2020. Информационные технологии. Искусственный интеллект. Обзор методов обеспечения доверия к искусственному интеллекту. – Женева: Международная организация по стандартизации, 2020. – VI, 42 с.
7. ISO/IEC 25023:2016. Системная и программная инженерия. Требования и оценка качества систем и программного обеспечения. Измерение качества системы и программного продукта. – Женева: Международная организация по стандартизации, 2016. – VIII, 44 с.
8. ISO/IEC TS 4213:2022. Информационные технологии. Искусственный интеллект. Оценка эффективности классификации с помощью машинного обучения. – Женева: Международная организация по стандартизации, 2022. – VI, 24 с.
9. ISO/IEC 25023:2016. Системная и программная инженерия. Требования и оценка качества систем и программного обеспечения. Измерение качества системы и программного продукта. – Женева: Международная организация по стандартизации, 2016. – VIII, 44 с.
10. SafetyBench: Оценка безопасности больших языковых моделей [Электронный ресурс] / Ceetal. – 2024. – Режим доступа: https://huggingface.co/ spaces/SafetyBench/ SafetyBench – (дата обращения: 05.03.2026).
11. Perspective API [Электронный ресурс] / Jigsaw, Google. – Режим доступа: https:// perspectiveapi.com/ (дата обращения: 05.03.2026).

Certification of Neural Network Models in the Context of the Emerging Regulatory Framework

Vladislav Valerievich Dushkin
Senior Researcher, Research Center, Krasnodar Higher School of Management, Krasnodar, Russia
dushkin@list.ru

Abstract

This article examines current approaches to the certification of neural network models as a special type of product containing artificial intelligence. Emerging regulatory requirements in the Russian Federation and international standards in this area are analyzed. Key features of conformity assessment for neural network models are identified, including the specifics of the indicators being verified, testing methodology, and procedural aspects of certification. A classification of quality and safety metrics is proposed, and challenges in ensuring the reliability of assessments are considered. A conclusion is drawn regarding the need for a comprehensive approach to certification that takes into account both technical parameters and ethical and safety aspects.

Keywords: neural network models, certification, artificial intelligence, standardization, conformity assessment, quality metrics, AI safety.