УДК 004.89

СТАНДАРТИЗАЦИЯ ТРЕБОВАНИЙ К ФРЕЙМВОРКАМ И БИБЛИОТЕКАМ ДОВЕРЕННОГО ОБУЧЕНИЯ

Куликова О.А., студент гр. 09-361, IV курс
Казанский (Приволжский) федеральный университет, г. Казань

По мере того, как искусственный интеллект всё глубже проникает в чувствительные сферы жизни – медицину, юриспруденцию, финансовые технологии, государственное управление – потребность в институционализированном доверии к машинному обучению становится критически важной. Уже недостаточно, чтобы модель была просто точной: она должна быть этически обоснованной, интерпретируемой, безопасной, соответствующей нормативам и технически воспроизводимой. В этом контексте особое значение приобретает стандартизация требований к фреймворкам и библиотекам доверенного обучения. Речь идёт не просто о технических спецификациях, а о формировании общего языка, инфраструктурных ориентиров и регулирующих принципов, которые определяют, каким должен быть фреймворк, чтобы считаться «доверенным». В данной статье проводится комплексный обзор концептуальных основ, задач и подходов к стандартизации фреймворков и библиотек в области доверенного машинного обучения [1].

Фреймворки машинного обучения – такие как PyTorch, TensorFlow, JAX, scikit-learn и их производные – стали де-факто стандартом разработки ИИ-систем. Однако большинство из них создавались с ориентацией на производительность, гибкость и исследовательскую открытость, а не на безопасность, нормативную совместимость или подотчётность. В результате даже широко используемые библиотеки могут содержать архитектурные уязвимости, не гарантировать отслеживаемость данных, не обеспечивать контроль версий или соответствие принципам недискриминации. Всё это создаёт необходимость в пересмотре парадигмы разработки ML-фреймворков и переходе к их проектированию как компонентов инфраструктуры доверия.

Цель стандартизации в данной области – выработка унифицированных требований, которым должны соответствовать фреймворки, предназначенные для использования в регулируемых и критически значимых областях. Эти требования могут быть сгруппированы в несколько ключевых блоков:

1. Требования к прозрачности и отслеживаемости (traceability). Фреймворк должен обеспечивать полную трассировку данных и метаданных

на всех этапах – от загрузки и предобработки до инференса и логирования. Это включает в себя версии данных, источники, применённые преобразования, используемые параметры, случайные сиды и архитектурные конфигурации. Такие функции необходимы для соблюдения принципов воспроизводимости и правовой подотчётности [2].

2. Требования к интерпретируемости и объяснимости. Современные модели часто ведут себя как «чёрные ящики». Стандартизированный фреймворк должен включать встроенные средства интерпретации – методы визуализации внимания, оценки вклада признаков, построения контрафактов, анализа локальных решений и глобальных зависимостей. Причём эти средства должны быть не факультативными, а базовой частью API, доступной в каждой модели.

3. Требования к устойчивости и надёжности. Фреймворки должны обеспечивать тестирование моделей на устойчивость к шуму, атакующим вмешательствам, ошибкам в данных и изменяющимся распределениям. Стандартизированные средства оценки устойчивости (robustness benchmarks) и инструментальные проверки (например, adversarial testing, stress testing) должны быть частью минимального набора функциональности.

4. Требования к безопасности и приватности. Библиотеки должны поддерживать реализацию алгоритмов с защитой конфиденциальности (например, дифференциальная приватность, федеративное обучение, homomorphic encryption), а также средства контроля доступа к моделям, данным и логам. Стандарты должны описывать, какие механизмы обязательны для включения в продуктивную сборку, и как проверяется их корректность [3].

5. Требования к юридической и этической совместимости. Все фреймворки, претендующие на доверенность, должны сопровождаться лицензией, описанием модели ответственности, возможностями формирования отчётов соответствия нормативам (compliance reporting) и поддержкой аннотаций о целях использования (например, «не для медицинской диагностики»). Такие метаданные становятся необходимыми в условиях действия законов, таких как AI Act, GDPR и предстоящие национальные регуляции.

6. Требования к воспроизводимости и управлению жизненным циклом. Фреймворк должен обеспечивать возможность полного повторного запуска эксперимента с идентичными результатами. Это предполагает контроль версий не только моделей и данных, но и окружения, зависимостей, конфигураций оборудования и логики вычислений. Интеграция с системами управления ML-процессами (MLflow, DVC, TFX) и CI/CD-пайплайнами – необходимое условие стандарта.

7. Требования к совместимости и модульности. Стандартизация должна поощрять проектирование библиотек в виде взаимозаменяемых, слабо связанных модулей, которые могут быть независимо протестированы, сертифицированы и повторно использованы. Это создаёт основу для экосистемы доверенных компонентов и открывает путь к архитектурам безопасного композиционного ИИ.

Важным шагом на пути к стандартизации является определение процедур верификации. Одно дело – провозгласить соответствие фреймворка стандарту, и совсем другое – доказать это. Поэтому в рамках стандартизации разрабатываются тестовые пакеты, форматы деклараций, цифровые подписи, аудитные трассы и системы независимой сертификации. Возможна также интеграция с формальными методами верификации кода и моделей, а также использование политик «доверия по умолчанию» (trusted by design), если исходный код полностью открыт, задокументирован и прошёл независимую проверку [4].

Стандартизация требований может осуществляться на нескольких уровнях: (1) технические комитеты(например, ISO/IEC JTC 1/SC 42, IEEE SA), (2) индустриальные консорциумы (LF AI & Data, OpenSSF), (3) национальные регуляторы, разрабатывающие белые списки и рекомендуемые практики, и (4) академические инициативы, формирующие open benchmark-платформы и образовательные рекомендации. Эффективная стандартизация невозможна без взаимодействия всех этих уровней.

Также следует отметить, что стандартизация не должна сводиться к ограничению инноваций. Напротив, она задаёт основу для эволюционного развития экосистем, снижая транзакционные издержки на интеграцию моделей в критическую инфраструктуру, упрощая юридическое сопровождение ИИ-систем и повышая уровень доверия со стороны пользователей, клиентов и общества в целом.

В заключение, стандартизация требований к фреймворкам и библиотекам доверенного обучения – это шаг от экспериментальной инженерии ИИ к цивилизованному, регулируемому и ответственному проектированию алгоритмов, которым можно доверить принятие решений, затрагивающих жизни, ресурсы и права людей. Это не только вызов, но и историческая возможность задать архитектуру доверия для будущих поколений машинного интеллекта [5].

Список литературы:

1. ISO/IEC/IEEE 12207:2017. Systems and software engineering – Software life cycle processes. – Geneva : ISO, 2017.

2. ISO/IEC 25010:2011. Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models. – Geneva : ISO, 2011.

3. ISO/IEC 22989:2022. Information technology – Artificial intelligence – Artificial intelligence concepts and terminology. – Geneva : ISO, 2022.

4. IEEE Std 7000-2021. IEEE Standard Model Process for Addressing Ethical Concerns during System Design. – New York : IEEE, 2021.

5. OECD. Recommendation of the Council on Artificial Intelligence. – Paris : OECD, 2019. – 11 p.