

УДК 004.89

АУДИТ ПРЕДОБУЧЕННЫХ МОДЕЛЕЙ С ПРИМЕНЕНИЕМ НЕЗАВИСИМЫХ ЭКСПЕРТНЫХ СИСТЕМ

Бурняшев А.А., Оператор научной роты, III курс
Военная академия связи имени Маршала Советского Союза С. М. Буденного,
г. Санкт-Петербург

В современном технологическом ландшафте, где предобученные модели машинного обучения используются в задачах, затрагивающих жизненно важные интересы людей и организаций – от медицинской диагностики и финансового скоринга до рекомендательных систем в правовой сфере, – вопрос доверия к их выводу установится не только техническим, но и правовым, этическим и институциональным [1].

Сложность архитектур, непрозрачность данных обучения, множественные циклы адаптации и переиспользования делают предобученные модели крайне трудными для внешнего анализа и верификации. В этих условиях одной из наиболее перспективных практик обеспечения доверия становится аудит предобученных моделей с применением независимых экспертных систем, способных формализованно, воспроизводимо и когнитивно обоснованно проверять поведение моделей в критически значимых аспектах.

Аудит в данном контексте понимается как систематическая, формализованная процедура оценки поведения модели на заранее определённом наборе критериев, включая точность, устойчивость, интерпретируемость, справедливость, соответствие нормативам и допустимость применения в определённой среде. Особенность аудита предобученных моделей заключается в том, что он проводится экспертной системой, независимой от разрабатывающей и эксплуатирующей сторон, то есть не включённой в производственный цикл модели. Такая независимость необходима для предотвращения конфликта интересов, а также для получения объективной, воспроизводимой оценки, принимаемой пользователями, регуляторами и обществом.

Независимые экспертные системы представляют собой программно-аналитические комплексы, способные автономно (или при минимальном участии человека) проверять модель на соответствие критериям качества, прозрачности и безопасности. В отличие от простых тестовых наборов или ручного анализа, такие системы реализуют многоуровневые, модульные и объяснимые методы проверки, включая генеративное тестирование, формальные методы

верификации, логический анализ зависимостей, симуляцию нестандартных сценариев и интерпретируемый контроль вывода.

Они могут использовать как стандартные библиотеки (например, AIF360 для проверки справедливости, DeepChecks и Evidently AI для мониторинга дрейфа и ошибок), так и собственные фреймворки с адаптивными механизмами тестирования, адаптированными под специфику проверяемой задачи.

Процесс аудита начинается с декларации модели, включающей: описание архитектуры, спецификацию целевого поведения, информацию о данных предобучения, логику адаптации, используемые метрики и границы применимости. После этого формируется набор аудиторских требований – например, запрет на дискриминацию по полу, гарантированная точность выше определённого порога в конкретной подгруппе, ограничение на использование персональных данных, требование объяснения решения в терминах определённого домена знаний. Экспертная система затем проводит серию автоматических тестов, включая: стресс-тестирование (устойчивость к шумам и атакам), интервенционный анализ (влияние отдельных признаков), сценарные проверки (поведение на краевых случаях), сравнительный анализ с эталонными моделями, проверку на сдвиги и утечки данных, симуляцию поведения при недостоверных входах и т.д.

Одним из ключевых преимуществ использования автоматизированных экспертных систем в аудите является их способность к масштабированию и стандартизации. В условиях, когда организации используют десятки или сотни моделей, ручной аудит становится экономически и технически невозможным. При этом только формализованный подход, основанный на экспертных системах, позволяет построить процесс, удовлетворяющий требованиям нормативов, таких как AI Act, ISO/IEC 42001, NIST AI Risk Management Framework. Эти нормативы требуют не только предоставления логов и метрик, но и возможности подтверждения корректности вывода модели внешними структурами – а это возможно только при наличии прозрачной, воспроизводимой и объяснимой системы аудита [2].

Другим важным аспектом является обоснованность рекомендаций и выводов, выдаваемых аудиторской системой. Независимая система должна не только зафиксировать факт нарушения (например, предвзятости или деградации точности), но и объяснить его причину, локализовать источник, предложить корректирующие действия, и, в идеале, предсказать возможные риски при продолжении эксплуатации модели. Такие функции реализуются с помощью когнитивных модулей объяснения, которые могут включать генерацию текстовых отчётов, визуализацию значимости признаков, построение графов зависимости, логико-онтологические интерпретации и др [3].

Особо стоит выделить роль экспертных систем в судебных и правовых процедурах. Если предобученная модель используется в системах принятия решений, последствия которых могут быть оспорены в суде (например, автоматическая оценка риска рецидива, отказ в социальной помощи, диагностика заболевания), то наличие прозрачного, независимого аудита становится критически важным [4].

В таких случаях отчёты экспертной системы могут служить цифровыми доказательствами, фиксировать логи вывода, воспроизводить поведение модели на момент принятия спорного решения, а также сравнивать его с нормативной моделью поведения. Это способствует формированию электронной подотчётности и защищает как права граждан, так и ответственность разработчиков.

Существует также перспективное направление мета-аудита, при котором не только проверяется конкретная модель, но и оценивается качество процессов её создания, тестирования и применения. Это включает анализ MLOps-пайплайнов, проверку документации (Model Cards, Data Sheets), полноты логирования, воспроизводимости экспериментов, соответствия процессу валидации. В таких случаях экспертная система может выполнять не только техническую, но и управленческую функцию, выявляя нарушения стандартов разработки, недостатки в сборе данных, пробелы в контроле качества.

Таким образом, аудит предобученных моделей с применением независимых экспертных систем формирует новый уровень институционального доверия к ИИ. Он создаёт прозрачную, объяснимую и масштабируемую инфраструктуру, в которой поведение модели становится проверяемым и подконтрольным.

Это не просто средство предотвращения технических ошибок, а основа для формирования ответственных отношений между разработчиком, пользователем, регулятором и обществом.

В условиях, когда искусственный интеллект принимает решения, способные изменить человеческие судьбы, только наличие независимого цифрового арбитра – экспертной системы, фиксирующей и объясняющей поведение ИИ – способно обеспечить необходимый уровень доверия и справедливости в цифровом мире [5, 6].

Список литературы:

1. AI Risk Management Framework (AI RMF 1.0) / National Institute of Standards and Technology. – Gaithersburg : NIST, 2023. – NIST AI 100-1. – DOI 10.6028/NIST.AI.100-1.
2. ISO/IEC 42001:2023. Information technology – Artificial intelligence – Management system. – Geneva : ISO, 2023.

3. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence // Official Journal of the European Union. – 2024. – L, 12.07.2024.

4. Chorev, S. Deepchecks : A Library for Testing and Validating Machine Learning Models and Data / S. Chorev, P. Tannor, D. B. Israel [et al.] // Journal of Machine Learning Research. – 2022. – Vol. 23, № 285. – P. 1–6.

5. Bellamy, R. K. E. AI Fairness 360 : An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias / R. K. E. Bellamy, K. Dey, M. Hind [et al.] // arXiv. – 2018. – arXiv:1810.01943.

6. Свидетельство о государственной регистрации программы для ЭВМ № 2025662773 Российская Федерация. Программа для автоматического сравнения эффективности нескольких алгоритмов машинного обучения по метрике пользователя : заявл. 14.05.2025 : опубл. 23.05.2025 / Е. А. Николаева, П. А. Пылов, Р. В. Майтак ; заявитель Федеральное государственное бюджетное образовательное учреждение высшего образования «Кузбасский государственный технический университет имени Т.Ф. Горбачева». – EDN СМСРСВ.