УДК 004.89

ПРОТОКОЛЫ АУДИТА ПРИВАТНОСТИ В ПРОДУКТИВНЫХ ИИ-СИСТЕМАХ: ОТ ТЕОРИИ К ПРАКТИКЕ

Хернадез С., студент направления Computer Visions Master's Degree Programm, II курс
Открытый университет Каталонии, г. Барселона, Испания

В условиях стремительного внедрения искусственного интеллекта в критически важные сферы – здравоохранение, банковское дело, государственное управление, оборонную промышленность – проблема обеспечения приватности становится одной из центральных задач при проектировании, эксплуатации и мониторинге ИИ-систем. В частности, особое значение приобретает формирование и применение протоколов аудита приватности, которые позволяют не только оценивать соответствие модели нормативным требованиям, но и выявлять потенциальные утечки конфиденциальной информации на всех этапах жизненного цикла модели. Эти протоколы должны обеспечивать переход от абстрактных теоретических гарантий к системным и воспроизводимым механизмам проверки приватности в продуктивных условиях эксплуатации, где риски особенно высоки, а последствия утечек – необратимы [1].

Аудит приватности в ИИ-системах – это комплекс процедур, направленных на систематическую проверку того, насколько используемые модели, алгоритмы и инфраструктуры соответствуют установленным требованиям защиты персональных и чувствительных данных. В отличие от формальной верификации, которая проводится до развертывания модели, аудит ориентирован на динамическое наблюдение, диагностику и корректировку поведения уже действующей системы. Это особенно важно в случае постоянно обновляемых моделей, которые обучаются или дообучаются на новых данных, что создаёт предпосылки для накопления и потенциальной утечки чувствительных признаков.

Современные протоколы аудита приватности в продуктивных ИИ-системах строятся на сочетании теоретических моделей угроз и инструментальных средств тестирования. Теоретически аудит опирается на оценку устойчивости модели к таким типам атак, как membership inference, attribute inference, model inversion, а также их гибридные формы. Эти атаки моделируют возможное поведение злоумышленника, обладающего частичным доступом к модели (например, через API или градиенты при federated learning), и позволяют количественно оценить вероятность раскрытия тех или иных элементов обучающей

выборки. Практическая реализация этих атак в виде тестов становится частью аудиторского протокола.

Среди ключевых методов и метрик, используемых в протоколах аудита, можно выделить: [2].

- эмпирическую устойчивость к атакующим запросам, когда модель тестируется на чувствительность к специально сконструированным входам;
- анализ изменения предсказаний при модификации публичной информации, направленный на выявление перекрестной корреляции с приватными признаками;
- оценку запоминаемости обучающей выборки – через вычисление вероятности повторной генерации или предсказания уникальных обучающих примеров; [3].
- метрики информационной утечки – основанные на взаимной информации, энтропии или дивергенции Кульбака–Лейблера между пространством данных и выходами модели.

На практике протоколы аудита всё чаще реализуются в виде автоматизированных фреймворков, интегрируемых в конвейер разработки и сопровождения ИИ-систем. Примеры таких решений включают PrivacyMeter, ML Privacy Report, Opacus Auditor, TensorFlow Privacy Analysis Tools и другие. Эти фреймворки позволяют проводить регулярную проверку моделей на соответствие приватностным требованиям, генерировать отчёты, обнаруживать отклонения и автоматически инициировать процедуры коррекции (например, повторное обучение с приватностью, удаление подозрительных сэмплов, включение защиты логитов).

Отдельным направлением является внедрение формальных протоколов аудита в корпоративную и государственную регуляторную практику. В условиях действия нормативов, таких как GDPR, HIPAA, ISO/IEC 27001, NIST Privacy Framework и аналогов, возникает необходимость в стандартизированной отчётности и воспроизводимых процедурах оценки рисков. Это предполагает включение аудита приватности в политику информационной безопасности организации, регулярную аттестацию моделей, а также участие независимых экспертов и внешних аудиторов, способных верифицировать соблюдение стандартов [4].

Особую сложность представляют аудит распределённых и персонализированных ИИ-систем, таких как голосовые ассистенты, мобильные рекомендательные системы или инструменты здравоохранения, работающие с локальными данными.

Здесь аудит требует использования методов федеративного мониторинга, распределённого логирования, контроля межсессионного поведения модели и анализа устойчивости к адаптивным атакам на локальных клиентах.

В перспективе развитие протоколов аудита приватности предполагает усиление взаимодействия между техническими, юридическими и этическими направлениями. Современные подходы всё чаще дополняются интерпретируемыми формами отчётности, в которых аналитики, инженеры и юристы могут совместно анализировать поведение модели, её конфигурацию, уязвимости и воздействие на конкретные группы пользователей. В этом контексте важную роль играют и этичные аудиторы, или «privacy ethics officers», которые выступают медиаторами между разработчиками ИИ и обществом, обеспечивая прозрачность, доверие и подотчетность систем [5].

Таким образом, протоколы аудита приватности становятся неотъемлемым элементом устойчивой инфраструктуры ИИ.

Их развитие требует объединения формальных моделей приватности, практических средств тестирования и институциональных механизмов регулирования.

Только при наличии эффективного, воспроизводимого и прозрачного аудита можно гарантировать, что искусственный интеллект не просто работает корректно, но и уважает фундаментальные права пользователей на приватность, автономию и защиту информации, – то есть функционирует в соответствии с базовыми этическими и правовыми принципами цифрового общества [6].

Список литературы:

1. Пылов, П. А. Teamlead как разработчик и юридический лидер команды в одном лице / П. А. Пылов, А. В. Протодьяконов, А. И. Бобровских // Россия молодая : сб. материалов XIV Всероссийской научно-практической конференции с международным участием, Кемерово, 19–21 апреля 2022 года. – Кемерово : Кузбасский государственный технический университет имени Т. Ф. Горбачева, 2022. – С. 94406.1–94406.7. – EDN TOAHDH.

2. ISO/IEC 27005:2022. Information security, cybersecurity and privacy protection - Guidance on managing information security risks. – Geneva : ISO, 2022.

3. ISO/IEC 27557:2022. Information security, cybersecurity and privacy protection - Organizational privacy risk management. – Geneva : ISO, 2022.

4. ISO/IEC 29100:2011. Information technology - Security techniques - Privacy framework. – Geneva : ISO, 2011.

5. OECD. Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449. – Paris : OECD, 2019.

6. Council of Europe. Convention 108+: Convention for the Protection of Individuals with regard to the Processing of Personal Data. – Strasbourg : Council of Europe, 2018.