Проблема конфиденциальности данных в государственных информационных системах с использованием ИИ

Пермяков Руслан Анатольевич

Советник директора по инновациям

III Форум «Технологии доверенного искусственного интеллекта», Москва 20 мая 2025

Критическая проблема

- Рост роли ГИС в жизни граждан
  Государственные информационные системы стали основой предоставления услуг: медицина, образование, налогообложение, регистрация прав, социальная поддержка.
- Интеграция ИИ в ГИС
  Применение ИИ-моделей для принятия решений, обработки больших данных и автоматизации государственных процессов увеличивает зависимость качества услуг от технологий.
- Расширение объема и детализации персональных данных
  Современные ГИС аккумулируют не только идентификационные данные, но и поведенческую информацию, биометрию, финансовые и медицинские сведения.
- Уязвимость традиционных методов защиты
  Классические меры, ориентированные на защиту статичных систем, становятся недостаточными в условиях динамических самообучающихся ИИ-моделей.
- Риски снижения доверия к государству
  Утечки данных, ошибки автоматизированных решений и непрозрачность работы ИИ могут привести к потере доверия граждан к цифровым государственным сервисам.
- Необходимость обеспечения доверия к результатам работы ИИ
  Требуется не только защищать инфраструктуру, но и гарантировать объяснимость, корректность и проверяемость решений, принимаемых на основе ИИ в государственных системах.

Как ИИ меняет цифровое государство.

- Автоматизация принятия решений
  ИИ-системы участвуют в выдаче государственных услуг, валидации документов, выявлении несоответствий и рисков без непосредственного участия человека.
- Масштабная обработка персональных данных
  ИИ позволяет анализировать большие объемы чувствительных данных, строить профили граждан, выявлять скрытые взаимосвязи между событиями и субъектами.
- Динамичность моделей и данных
  ИИ-модели могут изменяться в процессе эксплуатации (самообучение, дообучение на новых данных), что затрудняет контроль, аудит и объяснение их поведения.
- Новые каналы утечки информации
  Утечки могут происходить не только через классические механизмы (доступ, хранение), но и через поведение обученных моделей (например, утечки обучающих данных через ответы моделей).
- Усиление влияния ошибок данных
  Неточности и искажения в исходных данных могут незаметно влиять на работу систем ИИ, приводя к ошибкам при принятии решений или к дискриминации граждан.
- Сложности обеспечения прозрачности и ответственности
  Автоматизация решений без возможности их интерпретации снижает контроль государства за процессами и затрудняет обоснование законности действий перед гражданами.

Ключевые угрозы для ГИС с ИИ

- Угрозы конфиденциальности:
  - Утечки персональных данных через поведение ИИ-моделей;
  - Деанонимизация обезличенных данных;
  - Ошибки обобщения;
  - Атаки на модели;
  - Data poisoning.
- Угрозы целостности:
  - Ошибки и манипуляции в данных, используемых для обучения ИИ;
  - Нарушение целостности результатов работы моделей.
- Угрозы доступности:
  - Атаки на инфраструктуру ИИ в ГИС (DDoS, сбои моделей);
  - Недоступность ключевых сервисов из-за ошибок в автоматическом принятии решений.
- Угрозы достоверности:
  - Выдача недостоверных результатов ИИ-обработки;
  - Зависимость качества решений от скрытых смещений в данных.
- Угрозы управляемости:

Традиционные меры защиты в ГИС

- Защита конфиденциальности:
  - Обезличивание персональных данных
  - Контроль доступа к информации
  - Шифрование данных при хранении и передаче
- Защита целостности:
  - Контроль версий данных
  - Аудит изменений
  - Механизмы цифровых подписей
- Обеспечение доступности:
  - Резервное копирование и восстановление данных
  - Защита от DDoS-атак
  - Мониторинг состояния сервисов
- Поддержка достоверности:
  - Проверка источников данных
  - Верификация поступающих данных вручную или автоматизированно
- Управляемость систем:

Почему этих мер недостаточно

- По угрозам конфиденциальности:
  - Обезличивание больше не гарантирует анонимность: ИИ может восстанавливать личные данные через корреляции.
  - Шифрование защищает транспорт и хранение, но не предотвращает утечки через поведение моделей.
- По угрозам целостности:
  - Классические механизмы контроля версий не учитывают динамические изменения моделей ИИ при обучении на данных.
  - Возможны атаки через незаметные изменения обучающих выборок (data poisoning).
- По угрозам доступности:
  - Рост сложности ИИ-систем увеличивает уязвимость к новым типам атак, включая целевые атаки на алгоритмы.
- По угрозам достоверности:
  - Проверка источников данных не гарантирует достоверность выводов ИИ из-за возможных смещений в данных или переобучения.
- По угрозам управляемости:
  - Стандартные механизмы разграничения прав доступа не обеспечивают контроль над «внутренней логикой» самообучающихся систем.

Доверенная информационная система

- Защита конфиденциальности:
  - Применение дифференциальной приватности и технологий защиты данных на этапе обучения ИИ.
  - Обязательное тестирование устойчивости обезличивания к деанонимизации.
- Защита целостности:
  - Контроль целостности не только данных, но и моделей ИИ после обучения.
  - Методы выявления атак на данные обучения (data poisoning detection).
- Обеспечение доступности:
  - Защита устойчивости ИИ-сервисов к новым типам атак (например, adversarial attacks).
  - Дублирование критических моделей и процедур отката.
- Поддержка достоверности:
  - Обязательная проверка моделей на наличие скрытых смещений и ошибок (bias & fairness testing).
  - Постоянная валидация результатов ИИ на актуальных данных.
- Управляемость систем:
  - Интерпретируемость и объяснимость решений ИИ (Explainable AI).
  - Прозрачность процессов обучения и функционирования моделей.
  - Полный аудит цепочки обработки данных и моделей.

Новые требования к безопасности

- Защита конфиденциальности:
  - Применение дифференциальной приватности и технологий защиты данных на этапе обучения ИИ.
  - Обязательное тестирование устойчивости обезличивания к деанонимизации.
- Защита целостности:
  - Контроль целостности не только данных, но и моделей ИИ после обучения.
  - Методы выявления атак на данные обучения (data poisoning detection).
- Обеспечение доступности:
  - Защита устойчивости ИИ-сервисов к новым типам атак (например, adversarial attacks).
  - Дублирование критических моделей и процедур отката.
- Поддержка достоверности:
  - Обязательная проверка моделей на наличие скрытых смещений и ошибок (bias & fairness testing).
  - Постоянная валидация результатов ИИ на актуальных данных.
- Управляемость систем:
  - Интерпретируемость и объяснимость решений ИИ (Explainable AI).
  - Прозрачность процессов обучения и функционирования моделей.
  - Полный аудит цепочки обработки данных и моделей.

Доверенная информационная система

Доверенная ИС отвечает на вопрос:

- «Почему можно доверять её работе и результатам?»

Ключевые компоненты:

- Прозрачность
- Обоснованность
- Защита ПД
- Проверяемость решений ИИ

Практические меры для доверия к ГИС с ИИ

- Верификация моделей ИИ
  Проведение формальных проверок корректности моделей и тестирование на устойчивость к атакам, включая проверки на утечки обучающих данных и манипуляции.
- Тестирование и проверка процедур обезличивания
  Обязательная оценка качества анонимизации данных с использованием специализированных методов анализа риска деанонимизации.
- Интеграция технологий Privacy-Enhancing Technologies (PETs)
  Применение дифференциальной приватности, федеративного обучения, гомоморфного шифрования для защиты данных на всех этапах их обработки.
- Аудит данных и процессов обучения
  Ведение полных журналов событий, контроль источников данных, отслеживание изменений в обучающих выборках и контроль среды обучения.
- Повышение интерпретируемости моделей (Explainable AI)
  Разработка и использование моделей, обеспечивающих понятное объяснение принимаемых решений для администраторов и пользователей.
- Разработка и соблюдение стандартов доверенного ИИ
  Введение обязательных требований к прозрачности, обоснованности и контролируемости систем с ИИ в государственных информационных системах.
- Управление жизненным циклом моделей
  Регулярное переобучение, тестирование и ревизия моделей ИИ в процессе их эксплуатации для предотвращения деградации качества и устойчивости.

Основные проблемы сегодня

- **Отсутствие стандартов доверенного ИИ для государственных информационных систем**
  Нет утвержденных требований к прозрачности, объяснимости, устойчивости и безопасности моделей ИИ в ГИС.
- **Недостаточная защищенность обезличенных данных**
  Обезличивание выполняется по старым методикам, которые не учитывают современные угрозы деанонимизации с помощью ИИ.
- **Сложности контроля за моделями ИИ в процессе эксплуатации**
  Отсутствие процедур аудита, мониторинга деградации качества моделей и выявления скрытых отклонений.
- **Дефицит компетенций в области приватности и безопасности ИИ**
  Недостаточное количество специалистов, способных сочетать знания в области информационной безопасности, ИИ и защиты персональных данных.
- **Рост зависимости от закрытых и непрозрачных технологий**
  Использование коммерческих ИИ-решений без полного контроля над архитектурой и данными увеличивает риски для безопасности ГИС.
- **Баланс между эффективностью ИИ и защитой конфиденциальности**
  Давление на максимизацию эффективности ИИ-моделей часто приводит к ослаблению мер защиты данных и снижению прозрачности.

Куда двигаться дальше

- Стандарты тестирования доверия
- Новые методы обезличивания
- Интеграция с технологиями гарантированной защиты данных
- Междисциплинарные команды

Спасибо за внимание.

Пермяков Руслан Анатольевич
+7(913)916-21-56, , @pransk

Институт вычислительной математики и математической геофизики Сибирского отделения РАН