Виткова
Лидия Андреевна

К.т.н., Начальник АЦКБ Газинформсервис,
с.н.с. ЛПКБ СПб ФИЦ РАН

GIS ГАЗИНФОРМ СЕРВИС

Санкт-Петербургский Федеральный исследовательский центр Российской академии наук

Оценка декларируемых показателей моделей машинного обучения в средствах защиты информации на объектах КИИ

GIS ГАЗИНФОРМ СЕРВИС

Санкт-Петербургский Федеральный исследовательский центр Российской академии наук

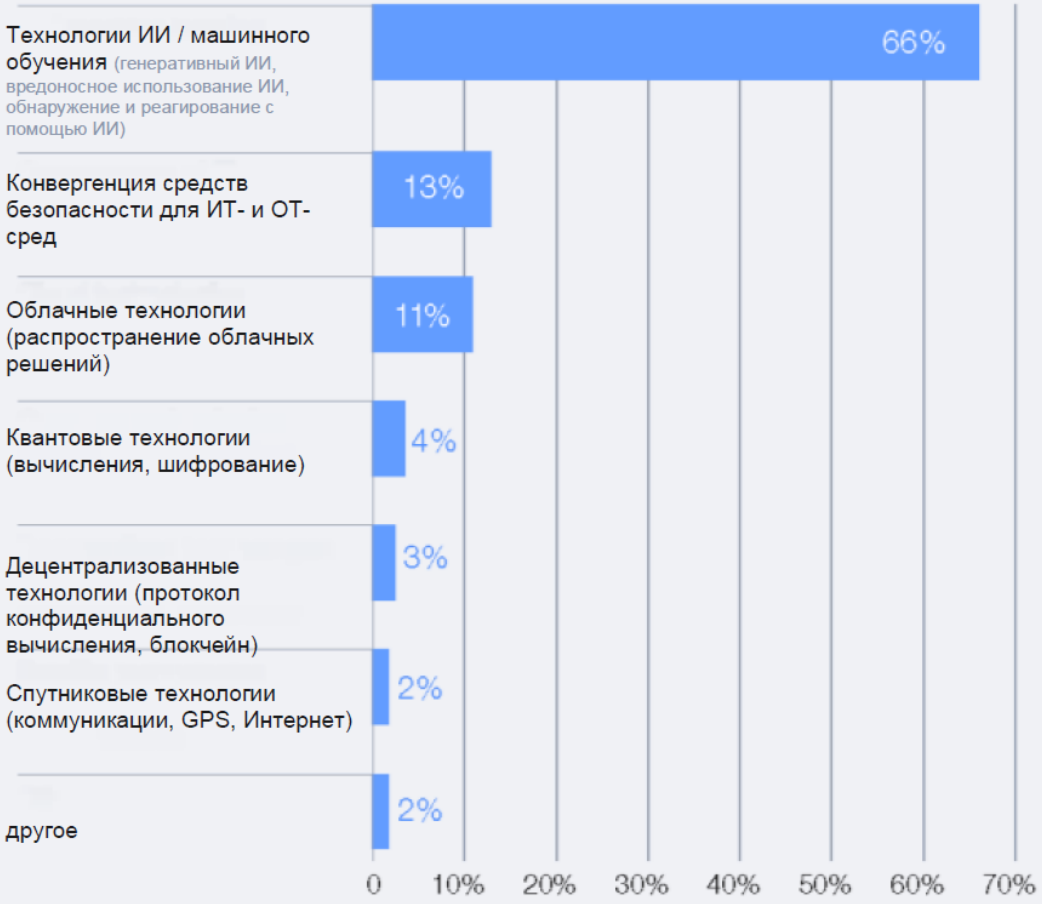
План

1. Актуальность
2. Декларируемые показатели
3. Угрозы и жизненный цикл
4. Проблемы оценки
5. Возможные решения

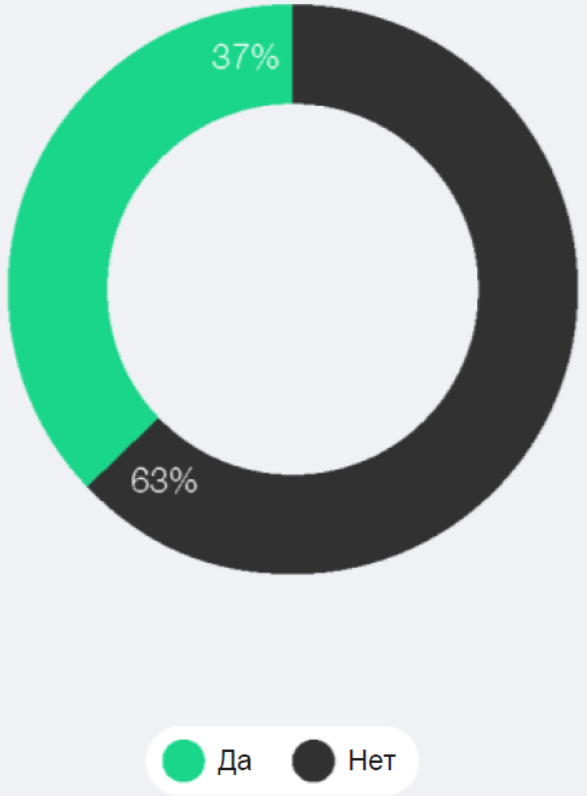
Главные уязвимости в 2025 году

GIS ГАЗИНФОРМ СЕРВИС \ Санкт-Петербургский Федеральный исследовательский центр Российской академии наук

Какие технологии, по вашему мнению, больше всего будут влиять на кибербезопасность в течение следующего года?



Предусмотрены ли в вашей организации процессы для оценки безопасности ИИ-инструментов до их внедрения?



Отчет
World Economic Forum

Где есть ИИ в СЗИ

GIS ГАЗИНФОРМ СЕРВИС \ Санкт-Петербургский Федеральный исследовательский центр Российской академии наук

Risk management

Asset management

NGFW

Контроль сети и анализ сетевого трафика

Анализ защищенности и ограничение среды

Защита конечных устройств

DLP

VM

SOC

Управление событиями ИБ

Threat Intelligence

AI-ассистенты

Защита периметра

SOAR

от 20-50%

решений содержат элементы технологий ИИ

Декларируемые показатели

ГАЗИНФОРМ СЕРВИС \ Санкт-Петербургский Федеральный исследовательский центр Российской академии наук

Хороший товар

Мамой клянусь

Лучше всех

Надо же как интересно получилось

Accuracy

Precision

Recall

F1-score

TPR

FPR

EPS

Mean Squared Error (MSE)

Root Mean Squared Error (RMSE)

Mean Absolute Error (MAE)

Ложные срабатывания

Пропуски

Общепринятых в ИБ метрик нет, в основном технологии наследует требования для КИИ или ГИС от ФСТЭК

Угрозы и жизненный цикл

ГАЗИНФОРМ СЕРВИС \ Санкт-Петербургский Федеральный исследовательский центр Российской академии наук

Описание бизнес-задачи и разработка ТЗ	Сбор данных	Предобработка и очистка
Некорректное определение бизнес-задачи	Сбор низкокачественных, устаревших, избыточных/недостаточных или вредоносных данных	Неполная или отсутствующая маскировка конфиденциальных данных
Неправильный выбор или описание данных	Нелегитимное изменение собираемых данных	Злонамеренное умышленное или неумышленное изменение существующего набора данных
Нарушение требований качества данных, надежности моделей и ПО	Использование нелегитимных источников данных для сбора данных	Неполная или некачественная очистка данных
Предвзятость на этапе формулирования задачи	Избыточная зависимость от конкретного источника данных	Отсутствие версионирования данных
	Невозможность определения источника данных	Утечка конфиденциальных данных
	Использование недостоверных данных	Использование вредоносного/недоверенного ПО для предобработки и очистки
	Утечка конфиденциальных данных	
	Использование вредоносного/недоверенного ПО для сбора данных	

Угрозы

Угрозы, актуальные для всех секторов

- Угрозы, связанные с ML
- Угрозы, специфичные только для LLM
- Угрозы, связанные с данными
- Угрозы, связанные с инфраструктурой
- Угрозы, связанные с организационной частью

Угрозы и жизненный цикл

ГАЗИНФОРМ СЕРВИС \ Санкт-Петербургский Федеральный исследовательский центр Российской академии наук

Угрозы

Разметка	Извлечение признаков	Обучение модели
Манипулирование метками размеченных данных	Злонамеренное умышленное использование искаженных признаков	Некачественная проверка способа случайной выборки данных, подаваемой на вход в пайплайн обучения модели
Неполная или некачественная разметка	Использование вредоносной модели для извлечения признаков	Вывод на этап тестирования предвзятой или уязвимой модели
Отсутствие версионирования данных	Нелегитимное изменение извлеченных признаков или изменение их весов	Нелегитимное изменение гиперпараметров модели или использование специально сформированных входных данных
Утечка конфиденциальных данных	Отсутствие версионирования данных	Использование конфиденциальных данных в системных инструкциях
Использование вредоносного/недоверенного ПО для разметки данных	Утечка конфиденциальных данных	Использование в коде обслуживания модели компонентов, имеющих или уязвимости, или вредоносные функции или налагающих на разработчика различные юридические требования
	Использование вредоносного/недоверенного ПО для извлечения признаков	Внедрение уязвимых функций в код обслуживания модели, в том числе небезопасной сериализации и десериализации *и, как следствие, возможность подмены злоумышленником сериализованного файла модели вредоносным объектом*
		Переобучение модели
		Использование вредоносного/недоверенного ПО для разработки моделей
		Утечка конфиденциальных данных

Угрозы, актуальные для всех секторов

- Угрозы, связанные с ML
- Угрозы, специфичные только для LLM
- Угрозы, связанные с данными
- Угрозы, связанные с инфраструктурой
- Угрозы, связанные с организационной частью

Угрозы и жизненный цикл

ГАЗИНФОРМ СЕРВИС \ Санкт-Петербургский Федеральный исследовательский центр Российской академии наук

- Угрозы, связанные с ML
- Угрозы, специфичные только для LLM
- Угрозы, связанные с данными
- Угрозы, связанные с инфраструктурой
- Угрозы, связанные с организационной частью

</>

Проблемы оценки

Проблемы оценки

GIS ГАЗИНФОРМ СЕРВИС \ Санкт-Петербургский Федеральный исследовательский центр Российской академии наук

1. Оценка метрик происходит на стороны разработчика
2. В ходе разработки решения используются наборы данных или открытые или синтетические (что не всегда соответствует реальности)
3. В ходе внедрения могут поменять входные данные от источников инженеры Заказчика
4. Нет встроенных механизмов проверки метрик на этапе эксплуатации
5. Нет требований к тестированию ИБ решений ИИ
6. Нет стандартных, утвержденных методик
7. Нет подготовленных кадров

GIS ГАЗИНФОРМ СЕРВИС

Санкт-Петербургский Федеральный исследовательский центр Российской академии наук

</>

Возможные решения

Возможные решения

GIS ГАЗИНФОРМ СЕРВИС \ Санкт-Петербургский Федеральный исследовательский центр Российской академии наук

1. Обязательно встраивать механизмы проверки данных на вход в модель в СЗИ
2. Обязательно тестировать модель на дрифт данных (в режиме черный ящик)
3. Создавать эталонные наборы данных для проверки и сравнения моделей в СЗИ
4. Создавать тестовые киберполигоны и испытательные лаборатории
5. Создавать требования и стандарты
6. Готовить кадры в ИБ с профилем MLSecOps
7. Готовить курсы ДПО для действующих специлистов ИБ
8. Мониторить поведение моделей в СЗИ
9. Собирать информацию об инцидентах

Виткова
Лидия Андреевна

К.т.н., Начальник АЦКБ Газинформсервис,
с.н.с. ЛПКБ СПб ФИЦ РАН

GIS ГАЗИНФОРМ СЕРВИС

Санкт-Петербургский Федеральный исследовательский центр Российской академии наук

Спасибо за внимание