

SML Security



CyberAgentReview
PROTECT YOUR AGENT

Проблемы использования агентов в разработке

Максим Горшков





Почему агенты быстро вошли в разработку



Повышение скорости /
эффективности



Автономность



Риски при использовании агентов



Непрозрачность



Уязвимость к новым видам атак

- prompt-инъекции
- утечка кода
- утечка чувствительных данных и конфиденциальной информации
- ошибочные действия агентов



Излишняя трата токенов

- петли
- избыточные контексты
- анализ лишних материалов

Известные примеры инцидентов

● Удаление production базы в PocketOS

● Галлюцинации в выдаче рекомендованных библиотек

Злоумышленник называет библиотеку очень похоже на оригинальную, а LLM при выдаче рекомендации ошибается и вместо проверенной рекомендует библиотеку с уязвимостью:

npm пакет `airbnb-base-hf`

npm пакет `eslint-plugin-unicorn-ts-2`

● ИИ-ассистент
Amazon Q Developer
расширение для Visual Studio Code

Amazon Q

Amazon Web Services  amazon.com | 📦 1,801,703 installs

● Postmark-mcp

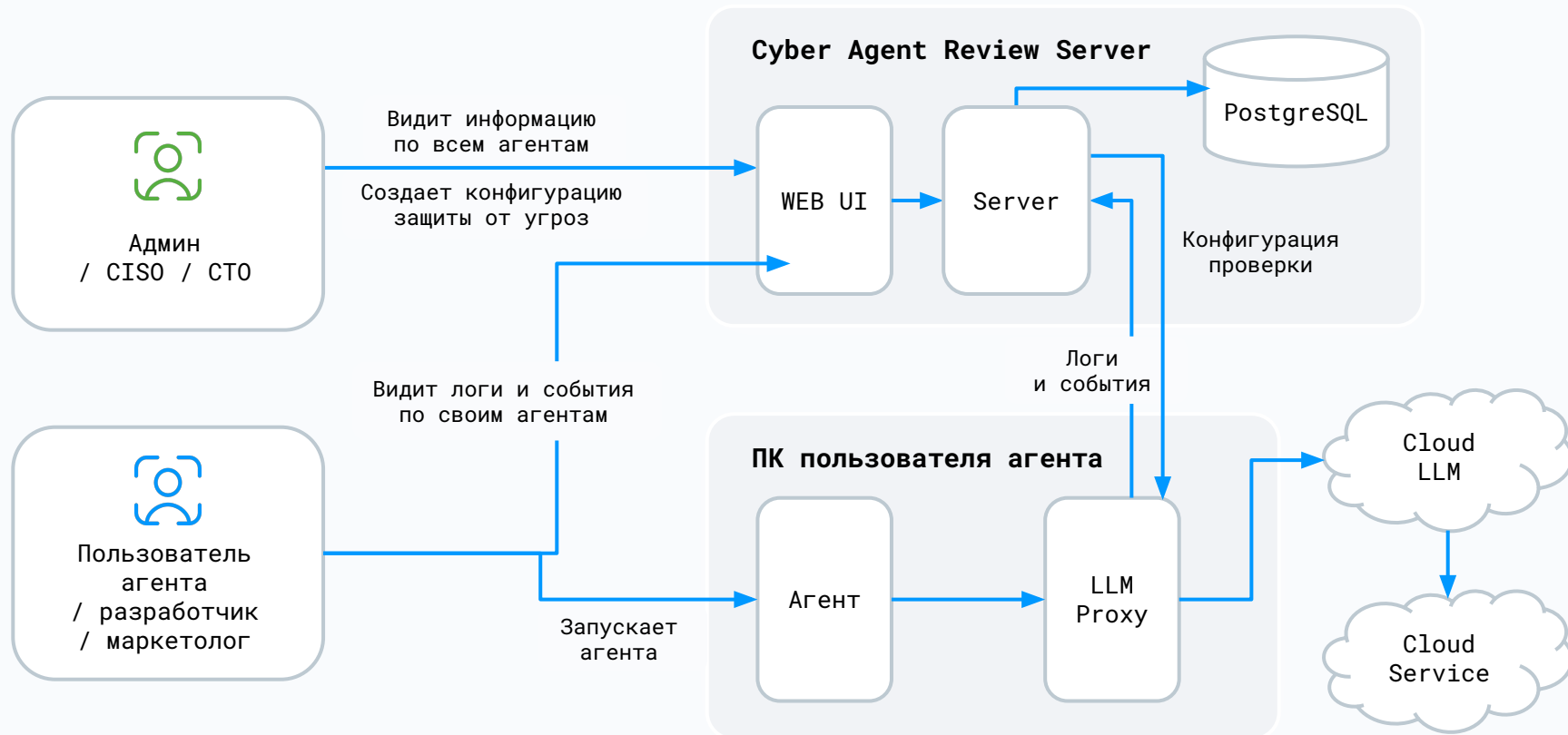
Инструмент маскировался под официальную библиотеку сервиса Postmark, которая использовалась для интеграции ИИ-ассистентов с почтовыми сервисами через MCP. Добавляла скрытую копию всех писем на адрес злоумышленника



Как снижать риски

- Инвентаризация всех используемых моделей и агентов
- Точка контроля всех запросов к модели
- Белые и черные списки для инструментов, библиотек и сетевых ресурсов
- Проверка prompt-инъекций и аномалий в поведении агента
- Маскирование чувствительных данных
- Контроль расхода токенов в рамках сессии и выдача рекомендаций по оптимизации

Как выглядит зрелая архитектура



Cyber Agent Review



Централизованный контроль, мониторинг
и защита AI-агентов в реальном времени



Отображает и логирует на едином дашборде
действия каждого агента



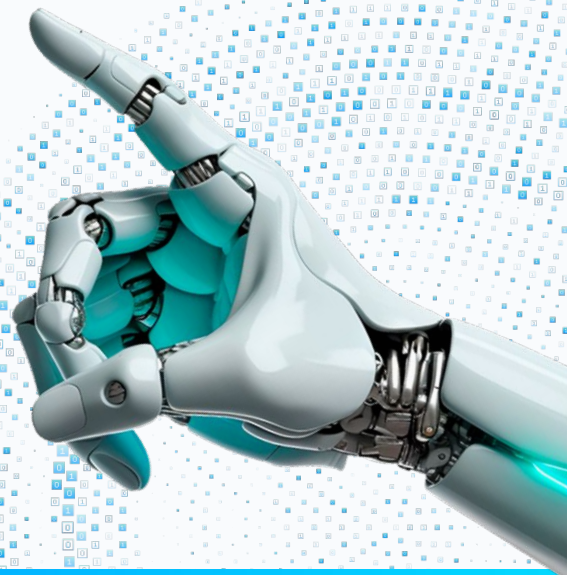
Обнаруживает и блокирует утечки данных, инъекции
промптов и другие аномалии в поведении агентов



Агент физически не может вывести конфиденциальную
информацию и использовать инструмент в обход политик



Администратор получает полный контроль
над использованием LLM в организации



Выгодно

- CISO: AI- агенты без рисков утечек и полная отчетность для аудита
- CTO: Возможность быстро внедрять новые AI-фичи без вето от отдела безопасности
- SOC: Готовые инструменты для расследования ИИ-инцидентов

Контакты



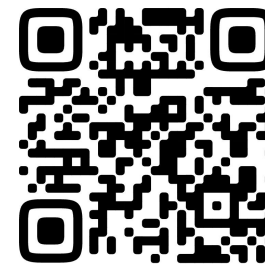
CyberAgentReview
PROTECT YOUR AGENTS

SML Security

- задать вопросы по работе системы
- заказать демо или пилот Cyber Agent Review
- вступить в наше закрытое сообщество для CIS0/AppSec

Максим Горшков

● Технический директор



➤ @MaximGorshkov
cybercodereview.ru