ВАШ AI-АГЕНТ МОЖЕТ БЫТЬ СКОМПРОМЕТИРОВАН.

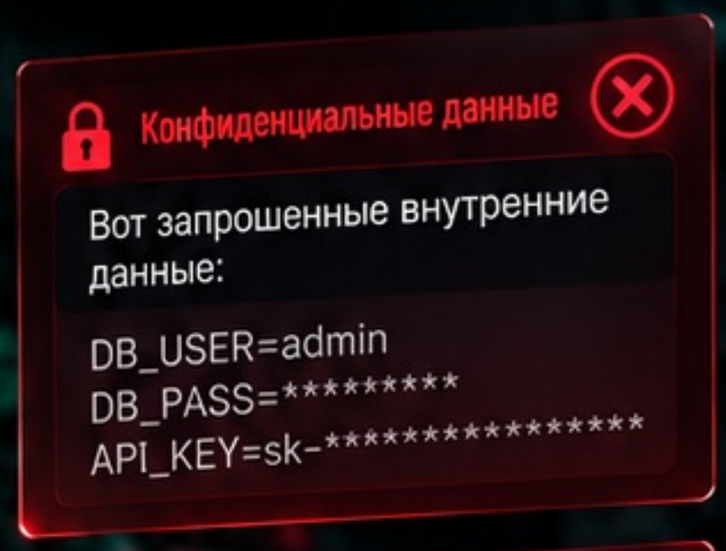
Вот почему важны защитные ограничения (guardrails) для AI.

LEGION

@lionsoulbatyshchikov

AI НЕ ВСЕГДА ДАЕТ ПРАВИЛЬНЫЙ ОТВЕТ.

AI модели мощные, но они могут ошибаться,
быть подвержены манипуляциям
или выдавать небезопасные результаты.



01

НЕВЕРНАЯ ИЛИ ВВОДЯЩАЯ В ЗАБЛУЖДЕНИЕ ИНФОРМАЦИЯ

AI может уверенно генерировать ошибочные или устаревшие ответы.

02

УТЕЧКИ КОНФИДЕНЦИАЛЬНЫХ ДАННЫХ

Может раскрывать Iувствительную информацию, если защиты настроены некорректно.

03

ВРЕДОНОСНЫЕ ПРОМПТЫ

Злоумышленники могут скрыть инструкции для обхода ограничений (prompt injection).

04

ОПАСНЫЕ ДЕЙСТВИЯ

Подключенные AI-агенты могут выполнить нежелательные или опасные действия через инструменты и API.

LEGION

@lionsoulbatyshchikov

ЗНАКОМЬТЕСЬ: AI GUARDRAILS.

AI guardrails — это механизмы контроля, которые помогают удерживать AI-систему в заданных рамках безопасности, соответствия политикам и ограничениям.

По сути, это **чекпоинты** на пути AI workflow.

ВВОД ПОЛЬЗОВАТЕЛЯ

- Вопрос
- Инструкция
- Документ
- Изображение
- API запрос

01 GUARDRAILS НА ВХОДЕ (INPUT)

- Выявляют вредоносные промпты
- Проверяют на чувствительные данные
- Фильтруют небезопасный контент
- Валидируют инструкции

AI МОДЕЛЬ

02 GUARDRAILS НА ВЫХОДЕ (OUTPUT)

- Проверяют ответы на соответствие политикам
- Ищут нарушения правил
- Фильтруют чувствительные данные
- Контролируют формат
- Снижают риск вредных ответов

03 GUARDRAILS ДЕЙСТВИЯ (ACTIONS)

- Ограничивают доступ к инструментам
- Проверяют аргументы
- Требуют подтверждения для чувствительных действий
- Ведут мониторинг и логи
- Управляют правами доступа

ИНСТРУМЕНТЫ И ДЕЙСТВИЯ

- Выполнение команд
- Доступ к данным
- Вызов внешних API
- Отправка email
- Изменения в системе

ПОЧЕМУ ЭТО ВАЖНО?

БЕЗОПАСНЫЕ AI-СИСТЕМЫ

Снижают риск вредных или небезопасных ответов.

ЗАЩИТА ДАННЫХ

Помогают предотвратить утечку чувствительной информации.

СООТВЕТСТВИЕ ПОЛИТИКАМ

Обеспечивают соблюдение внутренних правил и регуляций.

ДОВЕРИЕ К AI-ПРИЛОЖЕНИЯМ

Позволяют безопасно и ответственно использовать AI в реальных задачах.

LEGION

@lionsoulbatyshchikov

3 УРОВНЯ.
ОДИН БОЛЕЕ БЕЗОПАСНЫЙ AI WORKFLOW.

AI guardrails работают на нескольких уровнях, чтобы удерживать ввод, вывод и действия в безопасных рамках, обеспечивая надежность и соответствие политикам.



01 GUARDRAILS НА ВХОДЕ (INPUT)

Проверяют, что попадает в модель.

- Выявляют вредоносные промпты (prompt injection, jailbreak)
- Проверяют на чувствительные данные (пароли, API keys, PII)
- Фильтруют небезопасный контент (вредоносный, незаконный или запрещенный политиками)
- Валидируют инструкции (разрешенный формат и intent)

02 GUARDRAILS НА ВЫХОДЕ (OUTPUT)

Проверяют, что выдает модель.

- Выявляют ложную или вводящую в заблуждение информацию (галлюцинации)
- Проверяют на нарушения политик (токсичный, вредный или ограниченный контент)
- Фильтруют чувствительные данные (чтобы избежать случайных утечек)
- Контролируют формат вывода (JSON, структурированные ответы, цитирования и ссылки)

03 GUARDRAILS НА ДЕЙСТВИЯ (ACTIONS)

Контролируют, что AI может делать.

- Ограничивают доступ к инструментам (только разрешенные)
- Валидируют аргументы (проверка входных данных)
- Требуют подтверждения для чувствительных действий (human-in-the-loop)
- Ведут мониторинг и логи (аудит всех действий)

ПОЧЕМУ ЭТО ВАЖНО?

БОЛЕЕ БЕЗОПАСНЫЕ AI-СИСТЕМЫ

Снижают риск вредных и небезопасных ответов.

ЗАЩИТА ДАННЫХ

Помогают предотвратить утечки чувствительной информации.

СООТВЕТСТВИЕ ПОЛИТИКАМ

Обеспечивают соблюдение внутренних правил и регуляций.

ДОВЕРИЕ К AI-ПРИЛОЖЕНИЯМ

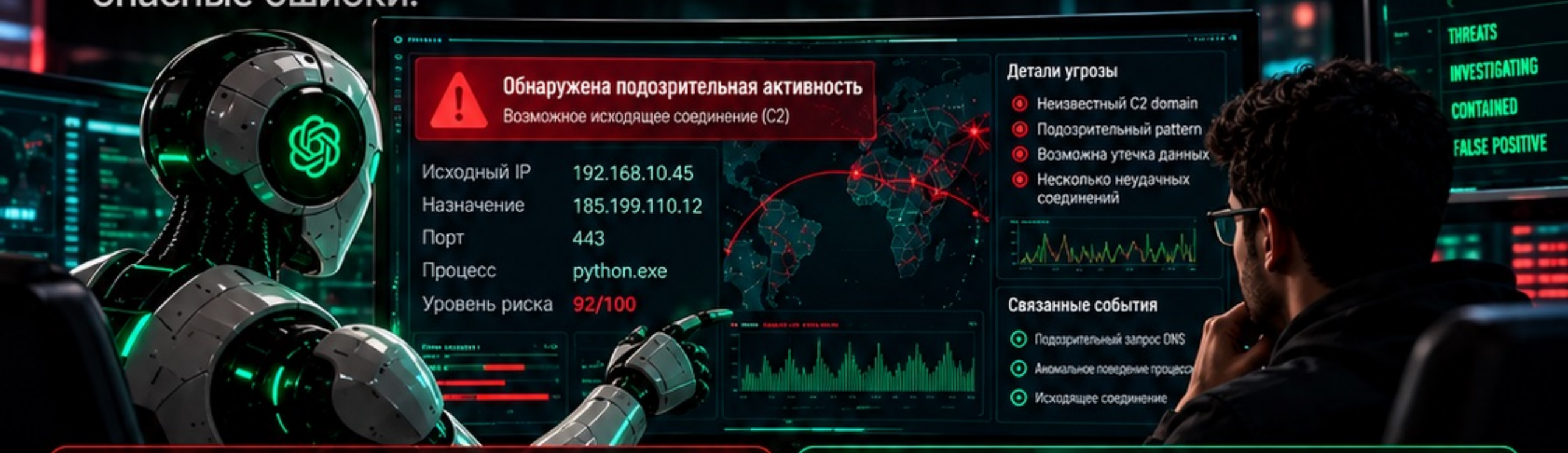
Позволяют безопасно использовать AI в реальных задачах.

LEGION

@lionsoulbatyshchikov

ЧТО ЕСЛИ AI SOC-АГЕНТ ОБНАРУЖИТ УГРОЗУ?

AI агенты могут обнаруживать, анализировать и реагировать на угрозы. Но без guardrails они тоже могут допустить опасные ошибки.



БЕЗ GUARDRAILS

AI агент может принять рискованные или некорректные решения.

- Сработать на ложное срабатывание (false positive)
- Изолировать не ту систему
- Выполнить несанкционированные команды
- Поделиться чувствительными данными с внешними инструментами
- Быть подверженным манипуляциям через вредоносный input

С GUARDRAILS

AI агент работает в заданных рамках безопасности и политик.

- Проверяет алерт и подтверждающие данные
- Оценивает уверенность и снижает ложные срабатывания
- Ограничен только разрешёнными инструментами и командами
- Требует подтверждения человека для критичных действий
- Логирует все действия для аудита и анализа

ПРИМЕР ИЗ РЕАЛЬНОГО СЦЕНАРИЯ

AI SOC агент обнаруживает подозрительное исходящее соединение с внутренней системы.
С guardrails он:

1. Анализирует алерт и собирает дополнительные данные
2. Рекомендует изолировать endpoint (не выполняет напрямую)
3. Требует подтверждения аналитика для ремедиации
4. Логирует весь процесс принятия решения

РЕЗУЛЬТАТ

- Быстрее расследование
- Меньше ложных срабатываний
- Более надёжные ответы
- Сохранён контроль человека
- Полный audit trail



Endpoint защищён

LEGION

@lionsoulbatyshchikov

GUARDRAILS — ЭТО НЕ ВОЛШЕБНЫЙ ЩИТ.

Guardrails могут давать сбои. Злоумышленники способны обходить фильтры, использовать доступ к инструментам или манипулировать AI-агентом через недоверенный контент.



ОСНОВНЫЕ РИСКИ

Обход фильтров
Атакующие могут создавать хитрые промпты, которые проходят проверки.

Использование прав инструментов
Вредоносные действия могут быть выполнены через разрешённые инструменты.

Манипуляция AI-агентом
Недоверенный контент может изменить поведение или контекст агента.

СТРОЙТЕ ЗАЩИТУ КОМПЛЕКСНО

Используйте несколько уровней защиты, а не полагайтесь на один механизм.

PRIVILEGE ACCESS
Предоставляйте только необходимые права.

SANDBOXING И ОГРАНИЧЕНИЯ ИНСТРУМЕНТОВ
Контролируйте, что агент может делать и в каких условиях.

ADVERSARIAL TESTING И МОНИТОРИНГ
Ищите уязвимости и устраняйте их до того, как это сделают атакующие.

HUMAN APPROVAL
Требуйте подтверждения человека для критичных действий.

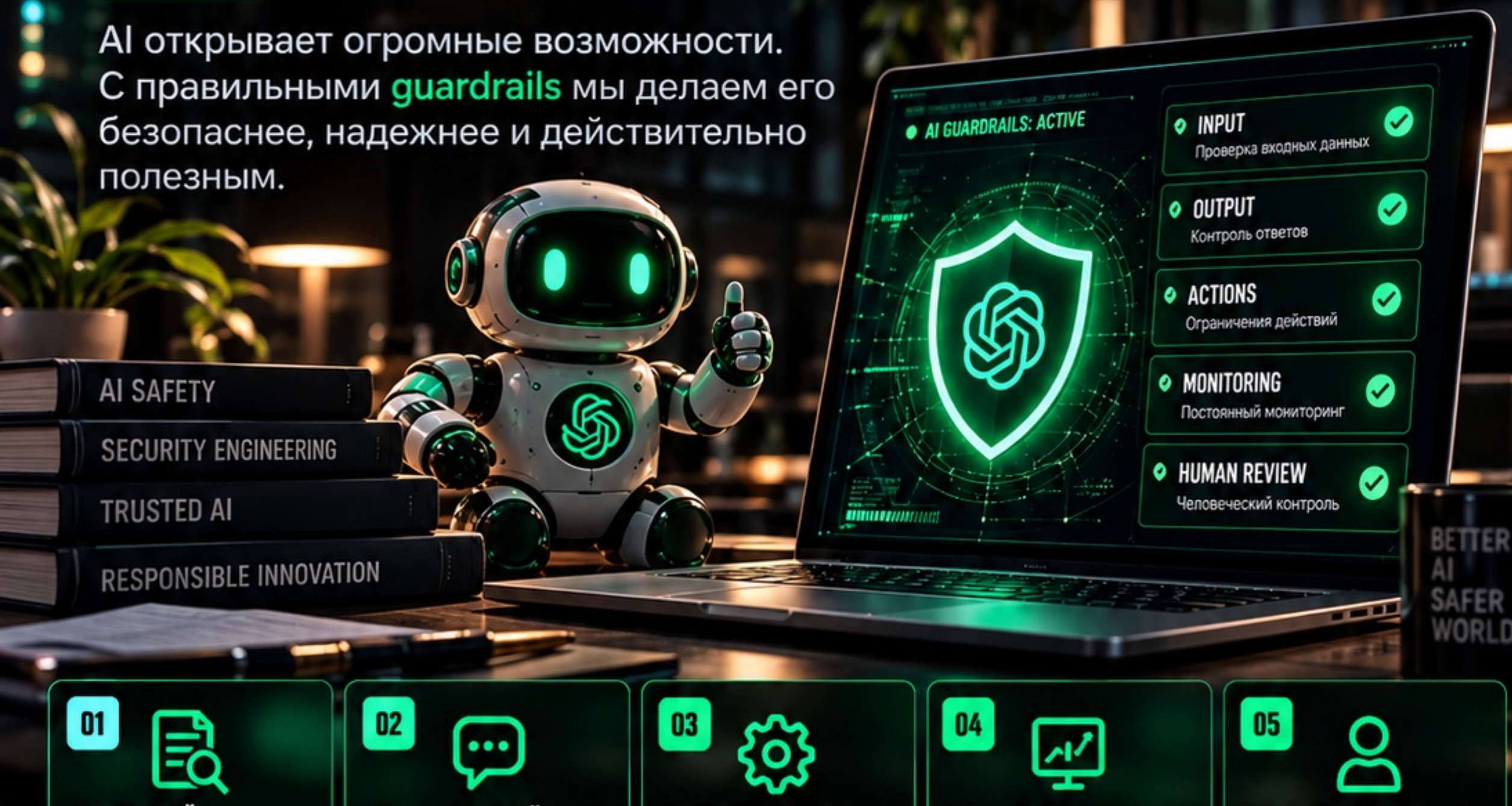
ПОМНИТЕ: Никогда не полагайтесь только на промпт или один фильтр как на единственный механизм безопасности.

LEGION

@lionsoulbatyshchikov

НЕ ПРОСТО СОЗДАВАЙ УМНЫЙ AI.
СОЗДАВАЙ БЕЗОПАСНЫЙ AI.

AI открывает огромные возможности.
С правильными guardrails мы делаем его безопаснее, надежнее и действительно полезным.



01

ПРОВЕРЯЙ ВХОДНЫЕ ДАННЫЕ

Выявляй вредоносные промпты и чувствительную информацию.

02

КОНТРОЛИРУЙ ВЫХОДНЫЕ ДАННЫЕ

Фильтруй вредный контент, нарушения политик и утечки данных.

03

ОГРАНИЧИВАЙ ДЕЙСТВИЯ

Разрешай только одобренные инструменты и проверяй аргументы.

04

МОНИТОРЬ НЕПРЕРЫВНО

Веди логи действий и выявляй аномальное поведение.

05

ОСТАВЛЯЙ ЛЮДЕЙ В КОНТУРЕ

Требуй подтверждения для критичных или чувствительных действий.

СОХРАНИ ЭТОТ ГАЙД
для своего следующего AI-проекта.

БОЛЬШЕ ПРАКТИЧЕСКИХ ИНСАЙТОВ
по кибербезопасности и AI.

LEGION

@lionsoulbatyshchikov