

Чек-лист: проверка безопасности AI-агента перед подключением к рабочим системам

Чек-лист предназначен для сохранения в PDF и использования командой разработки, ИБ и владельцем процесса при подготовке агентной системы к работе с корпоративными данными и внешними сервисами.

Данные и контекст

- Определить, какие источники данных считаются недоверенными: веб-страницы, письма, файлы, результаты поиска и ответы внешних API.
- Проверить, что внешний контент не может автоматически менять цель пользователя или вызывать привилегированное действие.
- Настроить сохранение происхождения фрагментов контекста для последующего расследования.
- Протестировать сценарии prompt injection с попыткой вызвать реальный инструмент.

Права и инструменты

- Выдать агенту отдельную учётную запись или ограниченное разрешение вместо полной сессии сотрудника.
- Ограничить разрешение задачами, ресурсом, типом операции и сроком действия.
- Разделить чтение, подготовку результата и его применение на отдельные режимы.
- Составить явный allowlist инструментов и запретить почту, shell, сеть, платежи, production-БД и секреты до отдельного разрешения.

Исполнение и подтверждения

- Ограничить файловую систему и исключить из рабочего каталога секреты.
- Закрыть исходящую сеть по умолчанию либо разрешить только необходимые маршруты.
- Настроить отдельную проверку политики до значимых вызовов инструментов.
- Показывать пользователю последствия операции перед отправкой, публикацией, удалением, изменением прав, передачей чувствительных данных и финансовыми действиями.
- По возможности использовать обратимые или безопасно повторяемые операции.

Наблюдаемость и остановка

- Фиксировать цель задачи, источники контекста, версии политик, доступные инструменты и параметры вызовов.
- Хранить сведения о токенах, правах, одобрениях и результатах действий.
- Выводить журнал в защищённое хранилище за пределами среды агента.
- Проверить, что kill switch останавливает новые задачи, отзывает токены и отменяет отложенные операции.

- Провести учение с имитацией инъекции, злоупотребления правами и опасного вызова инструмента.