Ответы на вопросы с вебинара “Управление ИИ-системами (AI Governance) в критических отраслях: как управлять рисками, соответствовать требованиям регуляторов и эффективно внедрять ИИ”

Отвечает Николай Павлов, архитектор MLSecOps и управления ИИ-системами (AI Governance), Академия АйТИ (ГК Softline)

Вопрос 1

Бизнес готов передавать критичные данные и процессы в ИИ, и мы планируем глубокую интеграцию ключевых бизнес-систем с облачными ИИ-моделями из-за отсутствия собственных мощностей. Вопрос в комплексной архитектуре доверия: как организовать пайплайн данных, чтобы соблюсти требования регуляторов и при этом исключить попадание конфиденциальных сведений в контур обучения внешней модели?

Существуют ли проверенные сценарии безопасного шаринга данных с облаком без потери контроля над ними?

Да, это абсолютно решаемая задача, и многие российские компании уже её решают. Постараюсь ответить так, чтобы большинству участников вебинара это было доступно.

Главная идея тут проста - ваши данные могут использоваться для получения ответа от ИИ-системы, но при этом они не должны становиться частью его памяти или использоваться для обучения (или дообучения).

Чтобы этого добиться, нужно выстроить многослойную защиту, начиная с юридической части. При заключении договора с провайдером облачных услуг необходимо **прописать жесткие условия**: запрет на использование ваших данных для дообучения моделей и обязательство **удалить их сразу же после** обработки, сохранить об этой операции запись (логи), хранить их и быть готовыми в любое время дня и ночи предоставить эти логи по вашему требованию.

Это база, на которую накладываются и технические меры. Сам пайплайн данных строится по принципу фильтра. То есть перед отправкой в облако ваша информация должна проходить через специальный шлюз на вашей стороне, который автоматически находит и маскирует конфиденциальные сведения, например, заменяет фамилии, номера счетов, номера банковских карт или паспортные данные, СНИЛС и другие персональные, коммерческие данные на специальные коды

В результате в облако уходит уже обезличенный запрос, безопасный. И модель выдает ответ, с этими же кодами обезличенными, а на вашей стороне эти коды снова заменяются на исходные данные. Для такой тактики можно использовать шаблоны промптом, протестированные заранее, чтобы точно знать, что мы отправляем допустим паспортные данные, серия, номер (обезличенные конечно) и модель точно их вернет при определенной структуре шаблона, не изменит, не потеряет цифры на конце и т.п. Таким образом, в этом сценарии провайдер технически уже не увидит критичной, конфиденциальной информации.

При этом чтобы соблюсти требования регуляторов, важно вести полный журнал событий, хранить и проверять логи. Нужно фиксировать, какие данные, в каком виде и когда передавались, и гарантировать, что серверы физически находятся в российской юрисдикции.

В итоге базовая формула здесь: юридический запрет на обучение + техническая изоляция данных через шлюзы и шифрование + постоянный аудит, фиксация действий в логах. Это все позволяет пользоваться облачным ИИ, не передавая ему контроль над вашей информацией.

Подчеркну здесь, что это лишь общая картина при ответе на Ваш вопрос. Он не является однозначной рекомендацией к действию. Зная специфику ваших систем, можно предложить более прицельные и эффективные меры.

Отвечаю на вторую часть вопроса – проверенные сценарии безопасного шаринга.

На практике для безопасного шаринга часто используется сценарий, когда конфиденциальные данные вообще никогда не покидают локальный контур. И в облачную модель передаются только или обезличенные эмбеддинги, или же текстовые запросы без чувствительного контекста.

То есть сначала ваши внутренние документы индексируются в локальной векторной базе данных. При поступлении запроса пользователя система ищет релевантные фрагменты локально.

А в облачную LLM отправляется только промпт типа: «На основе предлагаемого контекста ответь на вопрос...», где контекст уже отфильтрован от персональных данных и коммерческой тайны.

Затем ответ модели возвращается, на всякий случай проходит проверку и доставляется пользователю.

Важные принципы такого сценария:

1. Данные физически не должны передаваться провайдеру. Передаются только запросы и сгенерированные ответы.
2. Важно также обсудить с провайдером фильтры, механизмы, тарифы, предполагающие отсутствие конфиденциальных данных при взаимодействии с ним. В случае если такие данные все же начнут поступать, провайдер обязан: а) Остановить их поступление. Б) Проинформировать вас. в) Зачистить их у себя.
3. При этом локальное хранение эмбеддингов также исключает их попадание в публичные датасеты.

И еще один возможный сценарий – это прямой вызов облачной модели через API, но с юридическими и техническими гарантиями от провайдера, что ваши данные не сохраняются и не используются для дообучения его моделей.

Сценарий предполагает, что Вы используете выделенный enterprise-эндпоинт (не публичный API). И уже вместе с запросом провайдеру передаются служебные заголовки (типа x-retention-policy: none, x-training-opt-out: true). Тогда провайдер технически может отключить логирование входных данных и изолирует ваши запросы от общих датасетов.

Все такие сессии аутентифицируются через mTLS, ключи ротируются, а трафик идёт через приватные каналы.

Вопрос 2	**Можете поделиться примером документа стратегии по развитии ИИ компании (направление MLSecOps), из чего она должна состоять, страт проекты, эффекты для бизнеса (возврат инвестиций, увеличение выручки и т. д.)? Содержание документа и методология ее разработки.**

Важно! Ниже – обезличенный примерный шаблон. Вы можете адаптировать, переделать его под свою компанию. **Не является однозначной рекомендацией к действию.**

Скачать файл

Вопрос 3	**Как связать ущерб от ИИ-инцидентов с бизнес-стратегией бизнеса (стратегические цели/проекты) и возвратом инвестиций (ROSI или ROI)?**

Чтобы связать ущерб от ИИ-инцидентов с бизнес-стратегией компании и рассчитать возврат инвестиций (так называемый ROSI), нужно уметь говорить с руководством в первую очередь на языке денег и целей, а не только в части технических уязвимостей. И нужно оценивать это с позиции уже работающей ИИ-системы, исходить из того, что будет **если она остановится**, сколько денег мы тут потеряем, а не только анализировать последствия инцидента.

Здесь можно выделить такие шаги:

Шаг 1. Берем стратегическую ИИ-систему. Например: «Запустить ИИ-помощника для клиентов и увеличить продажи на 10%».

Шаг 2. Оцениваем, что будет, если такая ИИ-система сломается. Не «хакеры атакуют модель», а вообще, что будет если «помощник начнет выдавать ошибки»? Сколько мы потратим на восстановление, на ликвидацию угрозы. Сколько потеряем потом денег в результате последствий – какой будет мультипликативный ущерб? Например, клиенты уйдут, значит, мы потеряем около 5% выручки (50 млн.) + заплатим штрафы + рухнет цена акций на бирже, значит, придется повысить дивиденды.

Шаг 3. Считаем риск. Умножаем вероятность этого сбоя (например, 20% в год) на сумму всех потерь (50 млн.). Получаем «ожидаемый ущерб» или «потенциальный ущерб» - 10 млн. в год.

Шаг 4. Оцениваем защиту. Если решение за 4 млн. в год предотвращает 80% таких инцидентов, мы этим «спасаем» 8 млн. тыс. (80% от $10 млн.). И экономим этим условно тоже 4 млн.

Шаг 5. Считаем ROSI. Формула простая: (Спасенные деньги − Стоимость защиты) / Стоимость защиты. В нашем примере: (8 млн. – 4 млн.) / 4 млн. = 100%. То есть защитная мера очень эффективна и окупится в течение полугода.

Вопрос 4	**В чем заключается обучение обычных пользователей при работе с ИИ в AI Governance. Чему их учить?**

Вопрос очень актуален и компании всё активнее обращают внимание именно на обучение пользователей при работе с ИИ-системами. И это не просто очередной курс по промпт-инжинирингу или базовый ML, а целенаправленное формирование культуры ответственного и безопасного использования искусственного интеллекта. В таком обучении нужно заложить и этику, и выгоды компании, и немного подсветить тренды, и описать инциденты, и реагирование на них, и дать сотрудникам чек-листы, инструкции на закрепление, и многое другое.

Вам нужно, сделать так, чтобы каждый сотрудник, работающий с вашими ИИ-системами, понимал не только как работать, но и какие риски это несет, какой будет ущерб и как действовать корректно, правильно, чтобы не навредить с одной стороны, а с другой – чтобы выполнить работу.

У программы обучения могут быть такие разделы:

1. Базовая грамотность.

Расскажите про то, что такое ИИ и как он работает. Пользователи должны понимать фундаментальные ограничения технологии, особенно на реальных примерах вашей компании. Очень важно учить людей не просто слепо доверять результату, а критически оценивать его, сформировать понимание, что у ИИ есть галлюцинации, что важно все же проверять ключевые выводы, относиться критически. Нужно помнить, что ответственность за принятое решение всегда остаётся на сотруднике, а не на модели.

2. Безопасность данных.

Это самый важный, на мой взгляд, блок. Ваши сотрудники должны чётко знать, какую информацию можно, а какую категорически нельзя вводить в публичные или внешние ИИ-инструменты и даже во внутренние ИИ. Нужно учить распознавать типы конфиденциальных данных (персональные данные – их все возможные виды и типы, коммерческая тайна, куда могут относиться и технологические решения, врачебная тайна и другие). Надо понимать принципы классификации таких данных и сформировать понимание, куда обращаться, если случайно отправил чувствительные данные.

Здесь же можно объяснить, как правильно маскировать информацию перед запросом, если уж нужно что-то отправить критичное в модель, какую-то чувствительную информацию, чтобы при этом и данные отправить и ответ получить эффективный. Сотрудникам нужно также рассказать, как работать с внутренними ИИ-системами, если есть какие-то специфичные моменты по безопасности.

3. Этика и предвзятость.

Также пользователей нужно обучить замечать признаки предвзятости в результатах, например, при подборе кандидатов, оценке кредитоспособности и генерации контента. Ваши сотрудники должны изначально хотя бы примерно понимать, почему это может происходить, и знать, как эскалировать такие случаи, куда обращаться.

4. Правильное формулирование запросов и работа с контекстом (промпт-инжиниринг и контекст-инжиниринг).

Качество ответа ИИ-системы напрямую зависит от качества промпта и поданного контекста. Пользователей необходимо учить базовым принципам промпт-инжиниринга, то есть как чётко поставить рабочую задачу, как предоставить нужный контекст, как разбивать сложные запросы на отдельные шаги, как валидировать ответы модели и как проверять их полноту. Это не только повышает эффективность работы, но и снижает риск получения некорректного или опасного результата из-за двусмысленной формулировки.

5. Распознавание манипуляций и угроз.

При этом сотрудники должны уметь идентифицировать некоторые современные попытки атак через ИИ, особенно промпт-инъекции (когда злоумышленник через текст пытается «взломать» модель и заставить её выдать запрещённую информацию), социальную инженерию с использованием дипфейков или сгенерированного контента, а также фишинговые письма, созданные ИИ, и другие виды атак. Про это нужно говорить, приводя примеры и необходимо учить сотрудников базовой «цифровой гигиене» при работе с ИИ. Например, проверять источники, которые подсветила модель хотя бы выборочно. И сообщать в службу безопасности о любом странном, подозрительном, необычном поведении ИИ-систем.

При разработке обучения в крупнейших компаниях учитываем, что не всем нужно учить всё и в одинаковом объёме. Так разработчикам и аналитикам данных требуется глубокое погружение в технические аспекты безопасности моделей, тестирование на уязвимости, работу с MLOps-пайплайнами. Менеджерам и руководителям нужен фокус на управленческих рисках, принятии решений на основе ИИ, интерпретации метрик и этических дилеммах. Рядовым пользователям необходимы простые правила безопасного использования, распознавания рисков и алгоритмы действий в нестандартных ситуациях. Всем этим категориям, конечно, нужен слайд с контактами техподдержки в случае инцидентов ИИ-систем.

При этом мы понимаем, что сфера ИИ меняется стремительно, и эти изменения только ускоряются. Ежемесячно, а то и еженедельно появляются новые модели, поступает информация о новых уязвимостях, возникают новые регуляторные требования. Обучение пользователей в AI Governance не может быть разовым мероприятием.

Эффективная учебная программа включает регулярные обновления, доработки как основной учебной программы, так и сопутствующие рассылки, микро-курсы, вебинары, новости, внутренние чек-листы. Важно создать и еженедельно поддерживать культуру, в которой Ваши сотрудники не боятся задать вопрос или сообщить об ошибке ИИ в техподдержку и понимают, что это вклад в общую безопасность и стратегический актив компании, которым и являются ИИ-системы.

Академия АйТи

Партнер для корпоративного обучения

Telegram

MAX

Николай Павлов

Архитектор MLSecOps и AI Governance

Telegram

VK