Как защитить информацию
в приложениях, использующих ИИ

Владимир МУКАСЕЕВ,
начальник отдела разработки сервисов кибербезопасности IBS

Представим, что системы контроля и анализа транзакций в банке начинают игнорировать 30% мошеннических операций. Система управления энергосетью выводит из строя ключевой узел подачи электроэнергии в город. Чат-бот службы поддержки начинает массово раскрывать персональные данные клиентов. К сожалению, это новая реальность, с которой может столкнуться любая компания, интегрирующая ИИ-системы в бизнес-процессы.

Дилемма «ИИ + ИБ или ИИ против ИБ» уже не выглядит актуальной. Вызов, который бросает новое время, звучит иначе: «Мы понимаем, что нашу ИИ-систему могут взломать. Как построить процессы, чтобы это не парализовало бизнес?»

Угрозы, специфичные для ИИ-систем – закономерный этап технологической эволюции, требующий от организаций пересмотра подходов к управлению рисками. Рассмотрим, почему классический подход «построить неприступную крепость» недостаточно эффективен в мире ИИ, и представим более прагматичную модель «устойчивого ИИ», способного обеспечить непрерывность бизнеса даже в условиях целенаправленных атак.

Иллюзия неприступности: почему могут «взломать все» и к чему готовиться

Любую сложную систему можно скомпрометировать. Цель – не стать «неуязвимыми», а сделать атаку экономически невыгодной, а ее последствия локализуемыми.

Примеры бизнес-рисков через призму реальных сценариев:

- Для финансовых организаций – процесс скоринга, оценка кредитоспособности, выявление мошеннических транзакций. Злоумышленники могут научить модель пропускать определенные шаблоны мошеннических операций, что приведет к прямым финансовым потерям и регуляторным штрафам.
- Для ритейла – системы прогнозирования спроса, оптимизация логистики, управление запасами. Нарушитель преднамеренно вносит искажения в прогноз спроса на ключевой товар. Итог: срыв контрактов, коллапс логистики, потеря доли рынка.
- Для энергетики – системы для предиктивного обслуживания оборудования и балансировки нагрузок. Атака на ИИ-систему, отвечающую за диагностику, может привести к разрушению дорогого оборудования, масштабным авариям, экологическим катастрофам.

В каждом из этих случаев – не гигабайты украденных данных, не взломанные файрволы[1] или пестрые вывески вредоносного ПО, а миллионы убытков. Защита подобных систем – не просто техническая задача, а стратегическое управление рисками бизнеса.

Специфика уязвимостей ИИ

Классическая информационная безопасность перестает быть актуальной перед угрозами ИИ. Главное отличие: атаки на «традиционные системы» происходят через «дыры» – ошибки в коде, конфигурации, незакрытые порты в инфраструктуре. Искусственный интеллект атакуют через «особенности восприятия»: это не баги, дефекты или уязвимости системы, а ее свойства.

[1] *Также брандмауэр или межсетевой экран – программные или аппаратные средства защиты сетей, которые фильтруют входящий и исходящий трафики по заданным правилам безопасности*

Представим, что обычное приложение – сейф с кодовым замком. Если код замка надежный, стены толстые – его не вскрыть или сделать это сложно. ИИ-система – не сейф, а человек, владеющий кодом от этого сейфа. Сам по себе сейф абсолютно неприступен, а код надежен. Если вы обманете владельца, сказав ему нужные слова, показав поддельные документы или усадив за неверно заполненную анкету, вам не нужно взламывать сейф – вы сразу получаете код от него. С ИИ именно так: вы не взламываете «железо», а манипулируете восприятием и логикой – ИИ-система сама выдает всю необходимую информацию.

В чем заключаются уязвимости таких систем

1. Атака происходит не там, где срабатывает защита

Классический взлом – попытка проникнуть внутрь. Сработал межсетевой экран – атака отбита.

Атака на ИИ выглядит как обычный легитимный запрос от пользователя. Злоумышленник отправляет в чат-бот фразу, и технически это просто текст. Межсетевой экран* его пропустит, файрвол веб-приложений не сработает, а модель возьмет и ответит. Защита просто «не видит», что произошел взлом.

2. «Отравление» происходит задолго до проявления симптомов

Классическая уязвимость – незакрытый порт сегодня означает взлом завтра.

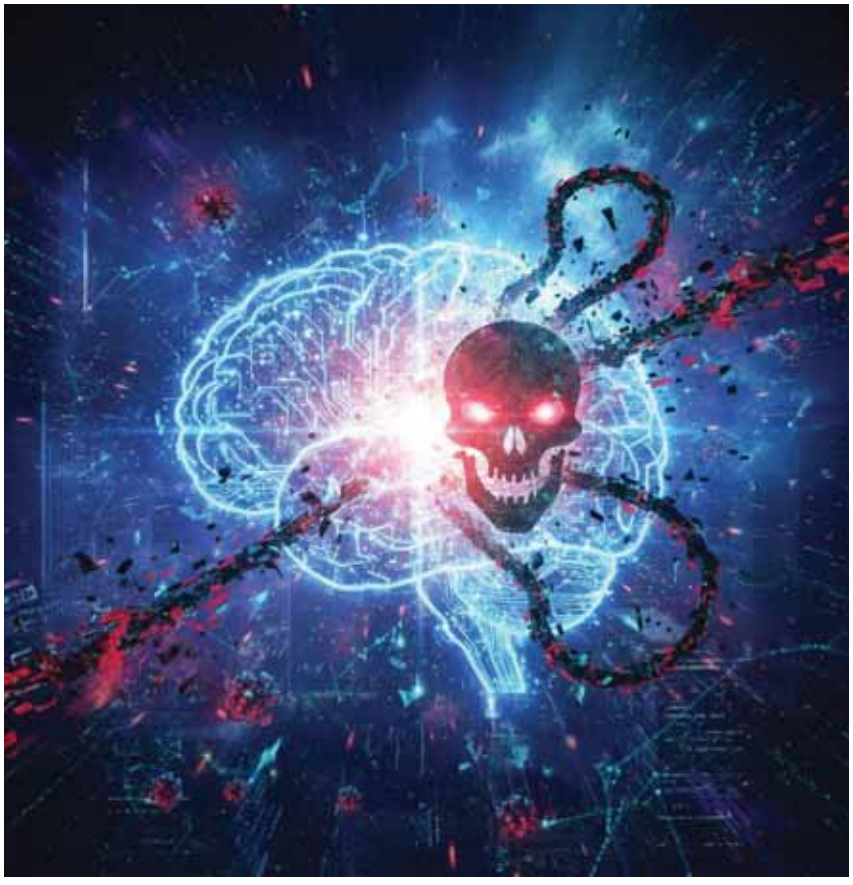
Угроза отравления ИИ: «отравленные» данные могут быть загружены на этапе обучения модели, то есть за полгода до того, как модель выйдет в промышленную эксплуатацию. Весь этот срок система просто обучается, а когда начинает работать, через месяц, два, три, год – уже несет в себе уязвимости. Например, в обучающие датасеты кредитного скоринга внедрили 1000 анкет мошенников, помеченных как «самые надежные заемщики». Модель обучилась и через год исправно выдает кредиты мошенникам. Классическая система мониторинга событий безопасности молчит: никаких вторжений не было. Атака произошла задолго до наступления этого события, а компания только начинает терять деньги.

Классическая уязвимость – незакрытый порт сегодня означает взлом завтра.

3. Атака на логику, а не на код

В информационной безопасности есть правило: если не позволять злоумышленнику запускать вредоносный код в инфраструктуре, компания находится в условной безопасности.

При взломе ИИ злоумышленнику не нужно запускать код. Ему нужно исказить входные данные так, чтобы модель ошиблась. Беспилотный автомобиль едет на знак «стоп». Хакер не трогал код, он просто наклеил несколько незаметных человеку полосок на знак. Камера «увидела» разрешение ехать дальше. С точки зрения защиты ничего не произошло, а вот в нашей реальности – авария.



4. Модель не может самодиагностировать у себя уязвимость

Классическое приложение можно просканировать автоматизированными средствами проверки кода. Они найдут некоторое количество уязвимостей и подробно их опишет: вот тут не обновлена библиотека, тут прямо в коде прописаны логины и пароли, тут потенциальный скрытый доступ для злоумышленников.

Нейросеть – это черный ящик с миллионами или миллиардами весовых коэффициентов[2]. Ни разработчик, ни создатель модели не могут точно сказать, какая комбинация нейронов отвечает за принятие решения, и где в этих нейронных связях притаилась «скрытая закладка». Нельзя просто взять и поставить патч на нейросеть, как делают с программами. Единственный способ, известный сейчас, – переобучать ее заново. Дорого, долго, а главное – непонятно, на каких данных, чтобы «закладка» не проявила себя вновь.

5. Баланс экономики атаки и защиты сломан

В классической ИБ защита дешевле взлома. Поставить патч – минутное дело, а найти уязвимость нулевого дня – долго и дорого.

В ИИ ситуация другая: атаковать дешево, а защищать дорого. Например, был случай, когда систему распознавания лиц взломали с помощью специальных очков, распечатанных на 3D-принтере. Для защиты системы от возможных вариантов таких очков нужно собрать данные обо всех мыслимых оптических искажениях и переучить модель. Бюджет атаки – 100 рублей. Бюджет защиты – сотни миллионов. Бизнес оказывается в положении, где каждая новая атака может стоить копейки для злоумышленника, а защита – целое состояние.

> Классическая ИБ ставит акценты на защите периметра. ИИ-системы не имеют периметра в привычном смысле: их уязвимость – в самом процессе принятия решений.

В чем же основная проблема? Классическая ИБ ставит акценты на защите периметра. ИИ-системы не имеют периметра в привычном смысле: их уязвимость – в самом процессе принятия решений.

> В ИИ ситуация другая: атаковать дешево, а защищать дорого.

Невозможно поставить антивирус на нейросеть; закрыть «порт», через который происходит промпт-инъекция; развернуть «патч» для весовых коэффициентов. Защита должна быть другой: необходимо строить процессы так, чтобы последствия успешной атаки были локализованы, а бизнес продолжал работать.

Стратегия устойчивости: от предотвращения к минимизации ущерба

Важно организовать процессы и системы, которые будут готовы к атакам, способны их выдерживать, восстанавливаться после атак и адаптироваться к неблагоприятным условиям. Как же обеспечить такой подход?

Есть несколько принципов, которых рекомендуется придерживаться:

1. Принцип «преднамеренно хрупких моделей» (Deliberately Fragile Models) и «человек в контуре» (Human-in-the-Loop)

Суть подхода: диагностировать и картографировать границы, за которыми система неизбежно откажет (т. е. преднамеренная хрупкость как диагностический инструмент). На границах, где система начинает отказывать, на помощь приходит человек, который контролирует действия модели при приближении к границе отказа.

Звучит сложно, поэтому разберем на аналогии: представьте двух охранников. Первый клянется, что не будет спать и никого не пропустит. Второй – более осторожный: он знает, что может ошибиться, поэтому на самые важные точки охраны ставит не себя, а видеокамеры с выводом изображения дежурному офицеру. Второй охранник понимает свои «границы отказа» и ставит контролирующий элемент над собой.

[2] *Числовые параметры, присваиваемые связям между нейронами, которые определяют силу влияния входных сигналов на выход нейрона*

В погоне за тотальной автоматизацией ИИ-системы проектируются так, чтобы они могли принимать финальное решение без участия человека. Это дает скорость, но делает атаку необратимой.

Нужно понять границы отказа и установить там средства контроля:

- Для решений с высокой ценой ошибки (одобрить кредит на 100 млн, остановить турбину станции, назначить лечение) модель не принимает решение. Она готовит проект решения, а подписывает его человек.
- Для решений с высокой неопределенностью (уровень уверенности модели ниже установленного порога) модель автоматически переводит задачу человеку, даже если это замедлит процесс.

2. Принцип резервирования

Если критический бизнес-процесс в компании связан с единственной моделью, злоумышленнику достаточно найти одну уязвимость – и процесс будет парализован.

Решением будет резервирование моделей. Необходимо запустить параллельно две разные модели, решающие одну задачу. Одну от вендора А, вторую от вендора Б; либо одну – сложную нейросеть, вторую – простую, но прозрачную модель на правилах.

3. Мониторим не действия хакеров, а бизнес-аномалии

SIEM-системы обучены искать признаки вторжения – подозрительный сетевой трафик, попытки подбора логинов и паролей, необычные запросы к базам данных. Атаки на ИИ часто не оставляют таких следов. Нельзя остановить атаку на входе, но можно поймать ее последствия на выходе.

Внедрите аналитические панели мониторинга бизнес-метрик, которые проверяет не только инженер, но и бизнес-руководитель:

- Модель кредитного скоринга[3]. Вчера одобряла 40% заявок, сегодня – 95%. Что случилось? Резко выросло качество заемщиков? Или модель отравлена и пропускает мошенников?
- Модель чат-бота. Вчера было две жалобы на некорректные ответы, сегодня – 500. Это новый релиз? Или промпт-инъекция заставила бот оскорблять клиентов?
- Модель прогноза спроса. Вчера рекомендовала закупить 1000 единиц товара, сегодня – 50 000. Это новая маркетинговая кампания? Или атака на данные с целью создать избыточные запасы?

Необходимо смотреть на цифры, которые всегда были перед глазами, но раньше интерпретировались как бизнес-показатели, а не индикаторы безопасности. Теперь любое резкое отклонение в поведении модели – инцидент, пока не доказано обратное.

4. Инцидент-менеджмент для процессов с ИИ

В каждом здании есть план эвакуации при пожаре. Есть ли план на случай, если «загорится» модель машинного обучения, обеспечивающая бизнес-процесс? В 99% компаний, которые уже используют ИИ в критических процессах, нет плана реагирования на инциденты с процессами с ИИ. Инцидент-менеджмент предполагает разработку планов реагирования на различные инциденты, связанные с ИИ в бизнес-процессах.

5. Модель «нулевого доверия»

Модель, которая общается с клиентами через чат-бот, не должна иметь доступа к внутренней CRM-системе, платежной информации, репозиторию с исходным кодом, данным обучения других моделей.

В погоне за тотальной автоматизацией ИИ-системы проектируются так, чтобы они могли принимать финальное решение без участия человека. Это дает скорость, но делает атаку необратимой.

Это кажется очевидным, но на практике чат-боту часто дают доступ «по мере необходимости». Позже выясняется, что промпт-инъекция позволила злоумышленнику не только получить доступ к истории переписки, но и поменять клиенту тарифный план.

Основные выводы

Ключевые принципы стратегии устойчивости:

- признаем, что атака произойдет;
- готовый план действий на случай инцидента (модель, алгоритм);
- смотрим не на попытки взлома, а на поведение бизнес-процессов;
- имеем инструкции в случае атак на ключевые узлы процесса в компании;
- изолируем системы: даже успешная атака не нарушит работу всей компании. ■

[3] *Автоматизированная система оценки кредитоспособности заемщика, которая присваивает баллы на основе анализа данных (кредитная история, доход, поведение) для прогнозирования вероятности возврата долга*