Еще раз про искусственный интеллект в контексте новой угрозы: от дипфейков до автоматизированных кибератак

Ещё недавно искусственный интеллект воспринимался преимущественно как инструмент повышения эффективности. Сегодня генеративные модели создают реалистичные изображения, видео и аудио, ведут диалог, пишут код и анализируют информацию, и эти возможности всё чаще используются злоумышленниками. Современная кибератака уже не выглядит как примитивное письмо с ошибками: с помощью генеративных моделей готовятся персонализированные обращения, имитируется стиль конкретного человека, создаются поддельные голос и изображение.

По мнению доцента кафедры «Безопасность жизнедеятельности» Финансового университета при Правительстве Российской Федерации, к.э.н., доцента Рожкова Р.С., фишинг остаётся одним из основных способов первоначального проникновения в системы, при этом отдельно необходимо отметить рост использования больших языковых моделей для совершенствования фишинга и автоматизации социальной инженерии. ИИ не всегда создаёт принципиально новый вид атаки: во многих случаях он делает существующие атаки быстрее, дешевле, масштабнее и убедительнее.

Наиболее заметное проявление – это дипфейки: синтетические изображения, аудио и видео перестали требовать специальных навыков, достаточно открытых фотографий, видеозаписей или образцов голоса. Технология используется для социальной инженерии: сотрудник может получить видеозвонок или голосовое сообщение от «руководителя» с просьбой срочно перевести средства, передать конфиденциальные данные или предоставить доступ к системе. Особенно опасна комбинация технологий, когда злоумышленник собирает открытые сведения о сотруднике, составляет персонализированное сообщение, применяет синтез голоса или изображения и дополняет атаку поддельной веб-страницей.

ИИ меняет и принцип подготовки фишинговых сообщений: большие языковые модели позволяют создавать тексты на разных языках, менять стиль и адаптировать сообщение под должность или сферу деятельности конкретного человека, а значительная часть подготовительной работы автоматизируется. Применение ИИ не ограничивается социальной инженерией. Языковые модели помогают писать код, анализировать

техническую информацию и искать решения, что снижает порог входа в «киберпреступную» деятельность.

В отчёте ENISA (Агентство Европейского союза по кибербезопасности) за 2025 год отмечается использование коммерческих больших языковых моделей группами угроз для повышения эффективности операций, разведки, разработки вредоносных инструментов и создания дипфейков. Меняется и экономическая модель атак: автоматизация позволяет одному оператору работать с гораздо большим числом целей. При этом ИИ представляет угрозу не только как инструмент злоумышленников, зачастую сами системы ИИ становятся объектом атаки.

Организации внедряют языковые модели в корпоративные процессы, поддержку, разработку, документооборот и анализ данных, и чем больше функций передаётся таким системам, тем важнее вопрос их защищённости.

Одна из проблем – это манипулирование входными данными и инструкциями для модели, когда злоумышленник формирует специальный запрос или размещает вредоносную инструкцию в анализируемых данных; последствия могут включать раскрытие данных, выполнение нежелательных действий или обход ограничений. Другое направление – это «отравление» данных, при котором в обучающий или анализируемый массив внедряется специально подготовленная информация, и система начинает выдавать некорректные результаты.

Доцент кафедры «Безопасность жизнедеятельности» Финансового университета при Правительстве Российской Федерации, к.и.н., доцент Фадеев М.К. подчёркивает двойственную природу технологии: ИИ расширяет возможности атакующих, но и сами системы ИИ имеют уязвимости и становятся частью поверхности атаки организации. Также особое значение приобретает проблема доверия к цифровой информации: если пользователь не может быть уверен, что услышанный голос принадлежит руководителю, а видео показывает реальное событие, проверять необходимо не только пароль, адрес отправителя или цифровую подпись, но и контекст запроса. В корпоративной среде это означает переход от принципа «доверять знакомому человеку» к принципу дополнительной проверки критических действий: запрос на перевод средств, изменение реквизитов, выдачу привилегированного доступа или передачу конфиденциальных сведений должен подтверждаться независимым каналом связи. Важную роль играют многофакторная аутентификация, управление привилегиями и постоянный мониторинг

аномального поведения. Здесь ИИ может стать не только источником угрозы, но и инструментом защиты: алгоритмы машинного обучения анализируют большие массивы событий, выявляют закономерности и отклонения, классифицируют инциденты и приоритизируют их. Однако полностью передавать принятие решений ИИ опасно, так как ошибка алгоритма может привести к блокировке легитимного пользователя или пропуску реальной атаки, поэтому в критических сценариях должен сохраняться контроль специалиста.

Таким образом, появление ИИ не отменяет существующие принципы информационной безопасности, но требует их пересмотра и усиления. Главное изменение – это скорость и масштаб процессов, а сама инфраструктура ИИ становится частью защищаемой среды. Современная система безопасности должна учитывать два направления риска: использование ИИ против организации и использование ИИ внутри самой организации. В условиях, когда голос, изображение и текст могут быть сгенерированы мгновенно, главный механизм защиты – это построение многоуровневой системы проверки, в которой ни один отдельный сигнал не считается достаточным основанием для доверия.

Рожков Р.С.,
кандидат экономических наук, доцент,
доцент кафедры «Безопасность жизнедеятельности»
Финансового университета при Правительстве РФ

Фадеев М.К.,
кандидат исторических наук, доцент,
доцент кафедры «Безопасность жизнедеятельности»
Финансового университета при Правительстве РФ