УДК 004

Ситков А.С.
Уфимский университет науки и технологий, Уфа

Научный руководитель:
Миронова Н.Г.
Уфимский университет науки и технологий, Уфа

КАК ЗЛОУМЫШЛЕННИКИ ИСПОЛЬЗУЮТ ГЕНЕРАТИВНЫЙ ИИ ДЛЯ СЛОЖНЫХ КИБЕРАТАК И ПЕРСПЕКТИВНЫЕ МЕХАНИЗМЫ ЗАЩИТЫ/ПРОТИВОДЕЙСТВИЯ ПОДОБНЫМ АТАКАМ

Аннотация. Генеративный искусственный интеллект (ИИ) становится важным инструментом в руках злоумышленника при сложных кибератаках на конфиденциальную информацию. Рассмотрены вопросы применения механизмов защиты/противодействия данному типу кибератак с использованием ИИ.

Ключевые слова: информация, риски информационной безопасности, генеративный искусственный интеллект.

Среди кибератак на критическую информационную инфраструктуру особую группу составляют подготовленные целевые (Targeted Attack) и продвинутые кибератаки (APT, Advanced Persistent Threat, «развитая устойчивая угроза»). Подобные атаки растянуты во времени, сложнее обнаруживаются традиционными подходами к защите информации, а для подготовки APT-атак (в т. ч. сбора и обработки информации о жертве, тактики долгого сокрытия своего присутствия злоумышленника в системе организации-жертвы) злоумышленники все чаще применяют возможности инструментов на основе моделей машинного обучения. Стоит отметить, что сложные кибератаки – это высокотехнологичные атаки на информационные системы организаций или отдельных лиц, отличающиеся высокой степенью организации, сложности и подготовленности.

Генеративный ИИ (GenAI) представляет собой категорию технологий, основанных на искусственных нейронных сетях, которые умеют самостоятельно разрабатывать новый контент: текстовые материалы и различный мультимедийный контент. Среди прочих возможностей GenAI:

– трансформация текста в устную речь, синтез голоса, изготовление аудиозаписей и обработка звуковых сигналов (примеры подобных сервисов – Tacotron от Google, Lyrebird, WaveNet);

– моделирование/имитация физических процессов, поддержка в проведении научных исследований, в т. ч. анализ массивов данных, создание научных гипотез в некоторых отраслях науки;

– генерация по словесному описанию программного кода.

Эти возможности привлекают внимание не только добросовестных исследователей и специалистов, но и злоумышленников, стремящихся использовать GenAI для проведения сложных и масштабных кибератак, поэтому по мере развития и роста доступности ИИ-технологий возрастает важность поиска эффективных методов противодействия подобным угрозам (что представляет как задачу разработки ИИ-инструментов безопасности – так и задачу нормативно-правового регулирования разработки ИИ-технологий).

Как генеративный ИИ используется злоумышленниками?

1. Предварительный этап подготовки сложной кибератаки. Машинное обучение помогает злоумышленнику автоматизировать и ускорить этап сбора информации об объекте атаки, агрегируя и обрабатывая большие массивы открытых данных (профили сотрудников в социальных сетях, доменные имена, IP-адреса, медиа-истончики), формируя полную картину об организации.

2. Автоматизация, ускорение и упрощение разработки вредоносных программ с помощью сервисов на основе генеративных моделей (в т. ч. уникальные экземпляры вирусов, трудно идентифицируемые традиционными антивирусами, оперативно адаптирующиеся к изменениям в системах безопасности). Имитационное моделирование с использованием машинного обучения может для имитации разных вариантов проведения атак на конкретную инфраструктуру, обхода имеющихся средств защиты, чтобы найти лучшую (для атакующего) стратегию атаки. Признаками APT-атаки может служить увеличение в определенный период количества обнаруженных бэкдор-троянов [1, с. 88–89] (т. к. злоумышленники применяют бэкдоры в виде троянских программ для поддержания непрерывного доступа к системе, даже если учетные данные были скомпрометированы или изменены).

3. Социальная инженерия играет важнейшую роль в продвинутых APT (Advanced Persistent Threats) кибератаках, так как именно человеческий фактор зачастую становится слабым местом в самой сильной системе безопасности. Социальная инженерия используется на этапе предварительного сбора информации и первоначального проникновения в инфраструктуру организации. Применение злоумышленниками ИИ в социальной инженерии многообразно: генеративные нейросети позволяют злоумышленникам легко создавать убедительные фишинговые сообщений, фальшивые документы и профили в социальных медиа, фейковые фото и видео для ввода в заблуждение через соцсети и мессенджеры для сбора конфиденциальных данных, для персонализированных и результативных атак.

4. Нарушения конфиденциальности. Генеративные модели могут быть использованы для синтеза личных данных, что позволяет злоумышленникам получать доступ к конфиденциальной информации, такой как банковские данные, медицинские записи и личные идентификаторы. Эти

сведения впоследствии могут быть использованы для кражи личных данных или шантажа.

5. Манипулирование информацией. Возможности генеративного ИИ позволяют автоматически создавать ложный контент, искажающий действительность. К примеру, сфабрикованные фотографии, видеозаписи или тексты могут распространяться с целью дезинформации общественности или подрыва репутации компаний и частных лиц [2, с. 101].

Один из путей противодействия этим технологиям – совершенствование инструментов для выявления материалов, сгенерированных GenAI инструментами; зачастую они дают высокий процент ложных срабатываний, но могут быть полезны для распознания фейковых сообщений, применяемых хакерами для ввода в заблуждение атакуемых с целью получения от них конфиденциальной информации, для фишинга. Некоторые из таких инструментов: ИИ-детектор Smodin (глубокий анализ контента с применением машинного обучения); многофункциональный инструмент для обнаружения текста, изображений и программного кода, сгенерированных ИИ – Copyleaks (также на базе искусственного интеллекта) [3, с. 97]; детектор дипфейков, созданных с помощью нейросети StyleGAN; программа FakeBuster для выявления дипфейков во время трансляций в Zoom и Skype, сервис Deepware (scanner.deepware.ai) для обнаружения дипфейковых видео; в свое время Агентство DARPA (США) инициировало разработку программно-аппаратного комплекса Semantic Forensics (SemaFor, https://semanticforensics.com/ от PAR Government Systems Corp.) для автоматизированного семантического анализа текстов, аудио, изображений, видео в реальном времени, с учетом атрибутов фейков, а в России АНО «Диалог регионы» в 2023 г. запустил платформу мониторинга аудиовизуальных дипфейков «Зефир» (но подобные инструменты не всегда доступны для специалистов по ИБ и ограниченно пригодны для реализации нужд ИБ организаций) [4, с. 191]. Еще один метод противодействия - использовать проверенные каналы связи для деловой информацией, систематически перепроверять достоверность поступающей по цифровым каналам информации, прежде чем принимать важные решения; проявлять настороженность при запросах на предоставление конфиденциальной информации (например, при совершении финансовых операций и т. п.). Также необходимо использовать актуальные версии антивирусных программ, брандмауэров и других защитных механизмов; регулярно проводить тесты безопасности, чтобы убедиться в эффективности защитных мер.

Для проникновения в систему при реализации сложной кибератаки злоумышленники применяют широкий спектр приемов и технологий: эксплуатация ранее неизвестных уязвимостей (в т. ч. в ПО, используемом в информационной системе жертвы), манипуляция поведением с помощью социальной инженерии; точечные фишинговые атаки; использование уязвимости удаленного включения файлов (RFI); внедрение вредоносного

кода через RFI или SQL-инъекции; атаки с использованием межсайтового скриптинга; заражение систем вредоносным программным обеспечением путем физического доступа; использование туннелирования через систему доменных имен и др.

Одним из основных методов противодействия методами социальной инженерии и атаками, использующими генеративные модели, является информирование сотрудников организаций и общественности, повышение квалификации персонала. Важно, чтобы люди осознавали возможные риски, могли отличать настоящие данные от поддельных и придерживались правил безопасного поведения в сети.

Применение генеративного ИИ открывает широкие перспективы для бизнеса, науки, культуры и образования, однако одновременно порождает проблемы безопасности информации и информационные риски, связанные с авторским правом, приватностью и возможностью манипуляций информацией.

Список использованных источников:

1. Жданов А.А. Автономный искусственный интеллект: учебное пособие / А.А. Жданов. 5-е изд. (эл.). М.: Лаборатория знаний, 2024. 362 с. Электрон. версия // Лань: электронно-библиотечная система. URL: https://e.lanbook.com/book/387629 (дата обращения: 19.04.2025).

2. Кацов И. Искусственный интеллект на предприятии: руководство / И. Кацов; перевод с англ. В. С. Яценкова. М.: ДМК Пресс, 2024. 710 с. Электрон. версия // Лань: электронно-библиотечная система. URL: https://e.lanbook.com/book/456725 (дата обращения: 19.04.2025).

3. Правовое и этическое регулирование роботизации и внедрения искусственного интеллекта (ИИ): материалы научно-практической конференции с международным участием 18 марта 2022 г.: материалы конференции. М.: Дело РАНХиГС, 2022. 132 с. Электрон. версия // Лань: электронно-библиотечная система. URL: https://e.lanbook.com/book/468215 (дата обращения: 19.04.2025).

4. Миронова Н.Г. Технологии медиабезопасности: методы противодействия фейк-контенту в цифровых медиа // Экономика и право: проблемы, стратегия, мониторинг: колл. монография. Чебоксары: «Издательский дом «Среда», 2024. 197 с. Элект. версия. URL: https://www.elibrary.ru/item.asp?id=78769542.

Sitkov A.S.
Ufa University of Science and Technology, Ufa

Scientific supervisor:
Mironova N.G.
Ufa University of Science and Technology, Ufa

HOW ATTACKERS USE GENERATIVE AI FOR COMPLEX CYBER ATTACKS AND PROMISING PROTECTION MECHANISMS/COUNTERING SUCH ATTACKS

Abstract. Generative artificial intelligence (AI) is becoming an important tool in the hands of an attacker in complex cyber attacks on confidential information. The article discusses the application of protection/counteraction mechanisms for this type of cyberattacks using AI.

Keywords: information protection rights, information security risks, generative artificial intelligence, APT (Advanced Persistent Threat).