КИБЕРБЕЗОПАСНОСТЬ В ИСКУССТВЕННОМ ИНТЕЛЛЕКТЕ

Бердимуродов Мансур Алишерович
Исполняющий обязанности доцента
Международной исламской академии Узбекистана
m.berdimurodov@iiau.uz
Вахобов Сарварбек Музаффар угли
Магистрант Международной исламской академии Узбекистана
s.vaxobov@iiau.uz

Аннотация. В статье рассматривается двойственная роль искусственного интеллекта в информационной безопасности: с одной стороны — преимущества в обнаружении и предотвращении угроз, автоматизации аналитики и скоростном реагировании; с другой стороны — новые векторы атак, такие как adversarial-атаки, кража моделей и утечка данных через ML-системы. Изучаются методы защиты — тестирование моделей, приватность данных, устойчивость к враждебным воздействиям и организационные меры. Приведены рекомендации по безопасному внедрению AI в корпоративную инфраструктуру.

Abstract. This article analyzes the dual role of artificial intelligence in information security: AI brings significant advances in detection, automation, and threat prediction, while simultaneously introducing new attack vectors and privacy risks. The study reviews AI-driven defense techniques such as anomaly detection, automated response, and behavior analytics, alongside threats including adversarial attacks, model theft, and data leakage. Practical recommendations are provided: robust testing, privacy-preserving techniques (e.g., differential privacy), adversarial training, and governance measures to safely integrate AI into security workflows.

Ключевые слова: Искусственный интеллект(ИИ), информационная безопасность, adversarial attack, кража модели, дифференциальная конфиденциальность, обнаружение угроз, автоматизированная атака, кибербезопасность, защита данных, безопасность ML.

Введение. В последние годы стремительное развитие технологий искусственного интеллекта (ИИ) оказывает глубокое влияние практически на все сферы жизни человечества. Искусственный интеллект широко используется в автоматизированных системах управления, анализе данных, распознавании звука и изображений, а также в процессах кибербезопасности. В то же время, наряду с технологическим прогрессом, возрастают новые типы угроз, в частности, угрозы информационной безопасности[1].

Средства искусственного интеллекта могут быть использованы не только для усиления систем защиты, но и в обратном случае - для таких целей, как автоматизация атак, создание дезинформации (deepfake), оптимизация вредоносных программ. В результате вопрос обеспечения информационной безопасности становится всё более сложным[2].

Поэтому глубокий анализ взаимосвязи между искусственным интеллектом и информационной безопасностью, изучение принципов создания систем защиты на основе искусственного интеллекта и выявление их положительных и отрицательных сторон является одной из актуальных научно-практических задач сегодняшнего дня. В данной статье рассматриваются эти вопросы и анализируется влияние технологий искусственного интеллекта на системы информационной безопасности, их возможности и факторы риска.

Достижения ИИ в области информационной безопасности

1. Обнаружение и мониторинг угроз[3]

Системы обнаружения аномалий на основе машинного обучения помогают выявлять аномалии в трафике, подозрительное поведение и атаки нулевого дня.

Автоматически анализируя журналы и сравнивая поведение с нормальными условиями, системы управления информацией и событиями безопасности (SIEM) повышают свою эффективность.

2. Автоматизированное реагирование и оркестрация

Платформы SOAR (Security Orchestration, Automation, and Response) позволяют быстро принимать решения и автоматизировать повторяющиеся задачи с использованием SI.

Это снижает нагрузку на сотрудников и ускоряет реагирование в режиме реального времени.

3. Прогнозирование рисков и уязвимостей

Модели машинного обучения могут прогнозировать вероятность уязвимостей в коде, конфигурации и инфраструктуре.

Благодаря приоритизации ограниченные ресурсы могут быть направлены на устранение наиболее существенных уязвимостей.

4. Биометрическая и поведенческая аутентификация

Распознавание лиц, голоса и поведенческая аутентификация повышают безопасность и улучшают пользовательский опыт.

Угрозы и уязвимости, возникающие в результате применения искусственного интеллекта

1. Атаки с участием злоумышленников[4]

— Неверные прогнозы делаются путем внесения небольших, невидимых манипуляций в модели. Это может привести, например, к неправильной идентификации в системах классификации изображений.

2. Кража и инверсия модели

— Параметры или данные модели могут быть украдены путем запроса к модели или путем ее эксплуатации через API. Инверсия модели также может восстановить секретные обучающие данные, использованные при обучении.

3. Отравление данных

— Внедрение вредоносных образцов в обучающие данные может привести к тому, что модель будет вести себя в устаревшем или неправильном направлении.

4. Автоматизированные атаки

— ИИ используется для масштабирования и повышения сложности атак, персонализации фишинговых сообщений, управления ботнетами и автоматического обнаружения уязвимостей.

5. Конфиденциальность и требования, аналогичные GDPR

— Системы ИИ обрабатывают большие объемы персональных данных; это может привести к нарушениям конфиденциальности и проблемам с регулированием.

Ключевые риски в кибербезопасности ИИ

1. Атаки с использованием подмены (обходные) — обман модели. Атаки с использованием подмены — это небольшие, бессмысленные манипуляции входными данными модели, направленные на то, чтобы заставить модель принять неверное решение (например, незначительное изменение изображения или текста для получения неправильной классификации)[5].

В результате системы обнаружения на основе ИИ могут быть обойдены; системы распознавания лиц, спам/непатентованные фильтры или детекторы вредоносных программ могут быть обмануты. Эти атаки трудно обнаружить, и их можно автоматизировать.

2. Отравление данных — повреждение обучающих данных. Неправильное обучение модели путем добавления опасных или некорректных

примеров в обучающую базу данных. Эта атака происходит во время обучения или онлайн-обучения.

В результате изменяется все поведение модели — она принимает неверные решения, делает замечания или, при определенных обстоятельствах, дает результат, полезный для злоумышленников.

3. Кража/извлечение модели. Злоумышленник пытается восстановить функциональность или параметры модели, отправляя запросы к модели через API (запрос «черного ящика»). Таким образом, коммерческую модель можно скопировать или перенаправить на достижение собственных целей.

Результатом является раскрытие коммерческой тайны, использование модели для дополнительных атак, потеря лицензии/уникальной интеллектуальной собственности.

4. Инверсия модели и утечка конфиденциальной информации (утечка конфиденциальной информации / инверсия модели). При атаке с инверсией модели злоумышленник пытается восстановить конфиденциальную информацию (персональные данные), использованную при обучении, используя выходные данные модели.

Это может привести к раскрытию частных или конфиденциальных записей в обучающих данных — нарушению требований GDPR/PDPL/местного законодательства.

5. Внедрение подсказок и атаки на основе входных данных (специфичные для LLM). При работе с LLM (большими языковыми моделями) внедрение подсказок — злоумышленник манипулирует правилами бэкэнда модели или результатами пользовательских запросов посредством текста (или других входных данных), введенных в модель.

Результатом является извлечение из модели неподходящих или вредоносных данных, нарушение внутренних правил или дезориентация агентов. Эти атаки особенно опасны в агентных системах (агентический ИИ).

6. Рост генеративного ИИ и социальной инженерии (дипфейки, автоматизированный фишинг). С помощью генеративного ИИ (LLM, TTS, дипфейк-видео) злоумышленники быстро завоевывают доверие посредством высоко персонализированного фишинга, клонирования голоса (вишинг) или дипфейков видео. ИИ может автоматически генерировать множество вариантов и масштабировать кампанию.

В результате получается социальная инженерия, ведущая к хищениям, финансовому мошенничеству и принятию неверных решений. Британский NCSC и CrowdStrike также предупреждают об этой тенденции.

7. Автоматизированные и масштабируемые атаки. ИИ помогает автоматизировать атаки: сканирование на наличие уязвимостей, масштабирование фишинговых сообщений и быстрое создание сценариев атак. Эти атаки будут намного быстрее и масштабнее, чем при использовании традиционного человека. Анализ CrowdStrike за 2024–2025 годы показал эту тенденцию. Временное окно для защитников сокращается — атаки будут распространяться быстрее, нанося ущерб до обнаружения.

8. Объяснимость моделей и ошибочные решения. Многие модели ИИ становятся «черными ящиками»; непонятно, почему модель приняла именно такое решение. Становится сложно выявлять опасные или ошибочные решения, а также затрудняется отслеживание и анализ инцидентов безопасности. Это иногда приводит к позднему обнаружению ошибок или последствий атаки.

9. Правовые, этические и нормативные риски. Системы информационной безопасности могут приводить к дезинформации, дискриминации или нарушению конфиденциальности; они могут не соответствовать нормативным требованиям на государственном и коммерческом уровнях.

Меры защиты и рекомендации

1. Устойчивость к атакам и обучение на основе состязательных примеров. Повышение устойчивости путем обучения моделей на основе состязательных примеров и их стресс-тестирования [6].

2. Дифференцированная конфиденциальность и защита персональных данных. Индивидуальные данные могут быть защищены с помощью методов дифференцированной конфиденциальности во время обучения.

3. Тестирование и верификация моделей. Внедрение строгих практик MLOps: контроль версий, объяснимость (например, SHAP, LIME) и мониторинг моделей. Обязательные тесты безопасности и конфиденциальности перед выпуском модели.

4. Защита данных от отравления. Проверка надежности обучающих данных, аутентификация источника данных и применение фильтров.

5. Организационные и политические меры. Разработка политики использования информационных систем, оценка рисков и распределение обязанностей. Проведение обучения сотрудников по вопросам информационной безопасности.

6. Планирование действий против сценариев атак. Регулярное проведение учений «красной команды»/«синей команды», пентестов и тестирования на основе моделей.

Практические рекомендации

— Интегрируйте безопасность с самого начала разработки программного обеспечения, используя принцип «сдвига влево» при разработке систем информационной безопасности.

— Внедрите строгую аутентификацию и ограничения на конечные точки API и моделей.

— Настройте системы мониторинга и оповещения во время обучения и эксплуатации.

— Минимизируйте и анонимизируйте информацию в соответствии с требованиями конфиденциальности.

Заключение

Искусственный интеллект (ИИ) открыл большие возможности и создал новые риски для информационной безопасности. Достижения ИИ в области обнаружения угроз, автоматизации и прогнозирования значительно повысили эффективность операций по обеспечению безопасности, но также возникли новые риски, такие как атаки злоумышленников, захват моделей, отравление данных и нарушения конфиденциальности. Поэтому при внедрении решений на основе ИИ организациям необходимо сбалансировать технические и организационные меры: повышение устойчивости моделей к атакам злоумышленников, защита обучающих данных, обеспечение безопасности инфраструктуры для разработки проектов ИИ или машинного обучения (МО) и ограничение API/конечных точек — все это должно осуществляться одновременно. Также важно применять принципы дифференциальной конфиденциальности и конфиденциальности по умолчанию, а также объяснять решения моделей с помощью инструментов прозрачности и объяснимости. Самое важное — это человеческий контроль и культура безопасности: помимо технических мер, необходимо обеспечить выявление рисков и быстрое реагирование посредством обучения сотрудников, учений «красной команды»/«синей команды» и постоянного мониторинга. Короче говоря, для полного использования положительного потенциала информационной безопасности новый стандарт безопасности должен стать интегрированным подходом, сочетающим технологии, политику и человеческий фактор.

Использованная литература

[1] Гудфеллоу, И., Бенджио, Й., и Курвиль, А. (2016). Глубокое обучение. Издательство MIT.

[2] Карлини, Н., и Вагнер, Д. (2017). Состязательные примеры нелегко обнаружить: обход десяти методов обнаружения. Труды 10-го семинара ACM по искусственному интеллекту и безопасности.

[3] Папернот, Н., МакДэниел, П., Гудфеллоу, И., Джха, С., Челик, З. Б., и Свами, А. (2016). Практические атаки «черного ящика» против машинного обучения. Труды Азиатской конференции ACM по компьютерной и коммуникационной безопасности 2017 года.

[4] Дворк, К., и Рот, А. (2014). Алгоритмические основы дифференциальной конфиденциальности. Основы и тенденции® в теоретической информатике.

[5] Скулли, Д., Холт, Г., Головин, Д., Давыдов, Е., Филлипс, Т., Эбнер, Д., Чаудхари, В., и Янг, М. (2015). Скрытый технический долг в системах машинного обучения. Труды NIPS.

[6] Хаднаги, К. (2018). Социальная инженерия: наука о взломе человека. Wiley. (для контекста о поведении пользователей и социальной инженерии)