О кибератаках с помощью систем Искусственного интеллекта

Д.Е. Намиот

Аннотация—**В настоящей статье рассматривается один из аспектов использования Искусственного интеллекта в кибербезопасности. Речь идет о кибератаках, которые могут совершаться с использованием систем Искусственного интеллекта (ИИ). Кибератаки с использованием ИИ можно определить как любую хакерскую операцию, которая опирается на использование механизмов ИИ. Другой используемый термин – наступательный ИИ. Кибератаки, основанные на ИИ, несомненно, меняют ландшафт кибербезопасности. В первую очередь, здесь необходимо говорить о скорости реализации атак и их масштабировании. Кибератаки, основанные на ИИ, включают использование передовых алгоритмов машинного обучения для выявления уязвимостей, прогнозирования закономерностей и использования слабых мест. Эффективность и быстрый анализ данных расширяют возможности хакеров по получению тактического преимущества, что приводит к быстрым вторжениям или уничтожению данных. Традиционных методов кибербезопасности больше недостаточно для борьбы со сложными атаками, поскольку кибератаки с использованием ИИ адаптируются и развиваются в режиме реального времени. Кроме того, внедрение систем ИИ в киберзащите порождает новые риски. Системы ИИ сами становятся объектами состязательных атак.**

В статье рассмотрены общие вопросы организации кибератак с использованием ИИ, приведены таксономия и примеры таких атак.

***Ключевые слова*—машинное обучение, глубокое обучение, кибербезопасность, кибератаки.**

I. Введение

Использование Искусственного интеллекта (ИИ) в кибербезопасности имеет несколько аспектов [1]. Следуя градации, предложенной компанией Microsoft, можно выделить следующие направления:

- ИИ в кибератаках (наступательный или атакующий ИИ)
- ИИ в защите от кибератак. Наиболее известная на сегодняшний день область применения с наибольшим количеством примеров использования
- Кибербезопасность самих систем ИИ (атаки на системы ИИ). Наиболее активно развивающаяся область
- Зловредные воздействия (например, дипфейки)

Статья получена 12 июня 2024. Д.Е. Намиот – МГУ имени М.В. Ломоносова (email: dnamiot@gmail.com)

Как обычно, под системами ИИ понимаются модели машинного обучения. В настоящей статье мы хотим остановиться на использовании ИИ в кибератаках. Очевидно, что в силу специфики, не все в данной области публикуется. Но также очевидно, что это направление получило большой дополнительный импульс к развитию с ростом популярности больших языковых моделей. Идея о том, что можно, например, автоматизировать программирование немедленно наталкивает заинтересованных лиц на мысль об автоматизации создания вредоносных программ, способность генеративных моделей создавать “человеческие” тесты порождает мысли о фишинге и т.д.

В работе представлены некоторые примеры использования ИИ в кибератаках. Материал, представленный в данной статье, является частью материалов курса по применению ИИ в кибербезопасности, который читается, по крайней мере, в двух магистратурах на факультете ВМК МГУ имени М.В. Ломоносова [2, 3].

Поскольку рассматривается очень бурно развивающаяся отрасль, материалы курса пересматриваются ежегодно. Текущая версия (2024 год) была создана при поддержке Департамента Кибербезопасности ПАО Сбербанк.

Оставшаяся часть статьи структурирована следующим образом. В разделе II рассматриваются общие положения. Раздел III посвящен таксономии наступательного ИИ. В разделе IV рассматривается пример разгадывания капчи. И раздел V содержит заключение.

II. Общие положения

В эпоху искусственного интеллекта злоумышленники используют методы, основанные на ИИ, чтобы взломать программы киберзащиты. Эти кибератаки, основанные на ИИ, несомненно, меняют ландшафт кибербезопасности. В первую очередь, здесь необходимо говорить о скорости реализации атак и их масштабировании.

Кибератаки, основанные на ИИ, включают использование передовых алгоритмов машинного обучения для выявления уязвимостей, прогнозирования закономерностей и использования слабых мест. Эффективность и быстрый анализ данных расширяют возможности хакеров по получению тактического преимущества, что приводит к быстрым вторжениям

или уничтожению данных. Традиционных методов кибербезопасности больше недостаточно для борьбы со сложными атаками, поскольку кибератаки с использованием ИИ адаптируются и развиваются в режиме реального времени [4].

Традиционная схема защиты для ИТ-организаций в начале 2000-х годов включала защиту периметра и проблемы с вредоносным ПО. Организации в те периоды также уделяли внимание безопасности программного обеспечения, но поскольку программных приложений было минимум, приоритетными были методы защиты от внешних атак. Позже появились программные приложения, помогающие решать проблемы производительности на основе пользователей, и организации создали усовершенствованные устройства защиты периметра, такие как интеллектуальные брандмауэры, маршрутизаторы и коммутаторы, для противодействия внешним сетевым атакам.

Атаки программного и аппаратного обеспечения могут представлять постоянную угрозу для бизнеса. Однако существуют эффективные способы противодействия этим угрозам. Одним из таких способов является использование модели зависимости системы. Эта модель объединяет предиктивный анализ, время отклика, тип атаки, сдерживание и киберзащиту в единую систему, а не рассматривает их как отдельные сущности. Модель зависимости системы помогает прогнозировать схемы атак и противодействовать вторжениям, особенно для персонала SOC (Security operations center) [5]. Каждый член команды имеет преимущество благодаря визуальным индикаторам и данным об угрозах, предоставляемым сетевыми устройствами безопасности. Однако кибератаки с использованием ИИ требуют от персонала SOC переоценки своей стратегии киберзащиты.

Сегодняшняя ситуация работает по-другому, поскольку кибератаки с использованием ИИ управляются (вызываются) программно и адаптируются к изменениям конфигурации. Никакие киберзащитники не могут противостоять изменениям в реальном времени, анализу и адаптивности атак, управляемых ИИ. Поскольку платформы ИИ используют машинное обучение для определения поведенческих моделей сети и уязвимых целей, они могут адаптироваться и изменять свой метод атаки.

Искусственный интеллект (машинное и глубокое обучение) все шире используется в киберзащите. При этом все подобные инструменты защиты могут быть объектами состязательных атак [6]. Такие атаки связаны с модификациями данных на разных этапах конвейера машинного обучения, относительно просты в реализации и, в большинстве случаев, не могут быть полностью исключены. Соответственно, атаки отравления, бэкдоры и, конечно, атаки уклонения, которые касаются средств защиты на основе ИИ – это типичные применения ИИ (машинного обучения) в кибератаках. NIST в своей таксономии состязательных атак [7] отдельно рассматривает состязательные атаки в области кибербезопасности. Исторически, первые состязательные атаки и начинались именно в этом домене. Первая известная отравляющая атака была разработана для генерации сигнатур червей еще в 2006 году [8]. В указанной работе рассматривались системы, которые автоматически определяют сигнатуры (признаки) программных червей. То есть, по сути, правила для сигнатур вредоносного ПО. Предложенная авторами атака с помощью загрязняла (зашумляла) трафик, используемый автоматическими генераторами сигнатур в процессе их извлечения. Атака была направлена на введение в заблуждение алгоритмов генерации сигнатур путем введения хорошо продуманного шума, предотвращающего генерацию полезных сигнатур. При этом было показано, что можно вносить шум, не зная заранее об используемой технике классификации.

Использование искусственного интеллекта привносит собственные риски, которые отличаются от традиционно рассматриваемых в кибербезопасности. Относительно этого есть много разных классификаций, одна из них приведена в работе [9]. Там перечислены 14 рисков ИИ:

1. Отсутствие прозрачности и объяснимости ИИ
2. Потеря рабочих мест из-за автоматизации ИИ
3. Социальная манипуляция с помощью алгоритмов ИИ
4. Надзорные функции, выполняемые с помощью технологии ИИ
5. Отсутствие конфиденциальности данных при использовании инструментов ИИ
6. Предвзятость из-за ИИ
7. Социально-экономическое неравенство как результат ИИ
8. Ослабление этики из-за ИИ
9. Автономное оружие на основе ИИ
10. Финансовые кризисы, вызванные алгоритмами ИИ
11. Потеря человеческого влияния
12. Неконтролируемый ИИ
13. Рост преступной активности
14. Более широкая экономическая и политическая нестабильность

Отсутствие конфиденциальности является, возможно, одной из наиболее серьезных проблем, которая, к тому же, может быть относительно легко использована через состязательные атаки, нацеленные на IP [10]. Программы устранения уязвимостей должны быть изменены, но также встают и вопросы классификации. Представьте себе утечку данных на платформе ИИ. Хотя риск основан на программном обеспечении, следует ли классифицировать его как программный риск или риск, основанный на ИИ?

Самая большая коллекция рисков ИИ содержится в проекте MIT: AI Risk repository [25]. Его описание есть в работе [26].

Помимо адаптивности и анализа в реальном времени, кибератаки на основе ИИ также могут вызывать больше сбоев в течение небольшого временного окна. Это связано с тем, как работает группа реагирования на инциденты. Когда происходят атаки на основе ИИ, есть возможность обойти или скрыть шаблоны трафика (изменении процесса анализа системного журнала или удаление данных, дающих возможность предпринять защитные действия). В системах кибербезопасности понадобятся другие алгоритмы, которые идентифицируют кибератаки на основе ИИ.

ИИ создал проблемы, в которых алгоритмы безопасности должны стать, в первую очередь, прогнозирующими и быстрыми и точными. Традиционный ИТ-ландшафт содержит множество рисков, связанных с конфиденциальностью, защитой периметра, программными приложениями или утечкой данных. Эти риски создают лазейки и ослабляют оборонную позицию организации. Тактика противодействия заключается в устранении рисков и повышении уровня киберзащиты. Внедрение ИИ в экосистему рисков и уязвимостей преобразует соответствие требованиям безопасности и киберзащиту. Поскольку ИИ использует поведенческую аналитику, машинное обучение и анализ в реальном времени, предприятия должны изучать риски на основе шаблонов и вычислительных ошибок. Именно здесь непрерывный мониторинг и ИИ будут работать лучше всего. Организации также должны определить, как должны развиваться аудиты ИТ-систем, оценки рисков и т.п., изменения конфигурации и сроки исправления.

Трансформация киберзащиты также требует разработки и внедрения средств контроля. Типичные структуры, такие как NIST 800-53 [11] или OWASP [12], структурированы на основе приложения, облака, данных, личности и инфраструктуры. Открытый вопрос - следует ли внедрять ИИ в такие же структуры контроля или следует изменить текущие средства контроля? От этого, между прочим, будут зависеть поверхности атак с использованием ИИ.

III. Таксономия наступательного ИИ

Кибератаки с использованием ИИ можно определить как любую хакерскую операцию, которая опирается на использование механизмов ИИ. Другой используемый термин – наступательный ИИ [16].

Все в научных работах начинается с некоторой классификации. Отметим сразу, что кибератаки – достаточно чувствительная область, поэтому совсем не все открыто публикуется. Тем не менее, на рисунке 1 представлена одна возможная классификация ИИ-атак:

Рис.1. ИИ-атаки [13]

В этом списке, очевидно, не хватает состязательных атак на модели машинного обучения, которые широко используются в информационных системах, в киберфизических системах и системах Интернета Вещей. К другим замечаниям можно отнести следующее:

Прослушивание нажатий клавиш – это часть более общей проблемы, которая называется атаки по побочным каналам и где широко используется ИИ.

Фишинг, в принципе, может быть отнесен к разделу атак социальной инженерии.

Дипфейки включают и клонирование голоса. Выделение клонирования голоса в отдельную категорию возможно в том случае, если речь идет о биометрической аутентификации, например. Это, традиционно, отделяется от дипфейков. Классически, дипфейк (англ. deepfake от deep learning «глубинное обучение» + fake «подделка»), изначально, понимался

как методика синтеза изображения или голоса, имитирующая человека и основанная на искусственном интеллекте. Технологии дипфейков также могут быть использованы для создания поддельных новостей и каких-либо вредоносных обманов. Дипфейки принято выделять в отдельную область использования ИИ в кибербезопасности [16], и в данной работе они рассматриваются.

Несмотря на эти замечания, по крайней мере, этот список дает представление о том, что же такое ИИ-атаки.

Из пропущенных в данной классификации элементов стоило бы добавить автоматизацию атак. На наш взгляд, это отдельная область применения ИИ в кибератаках. Например, так называемый ИИ-управляемый пентестинг. Примерами такой автоматизации пентестинга являются, например, стартапы XBOW [14] и RunSybil [15].

Рисунок 2, который взят из высоко-цитируемой работы [17], приводит классификацию описанных в научной литературе атак по типам воздействия. Атаки здесь распределены по шести этапам цепочки кибербезопасности (kill chain).

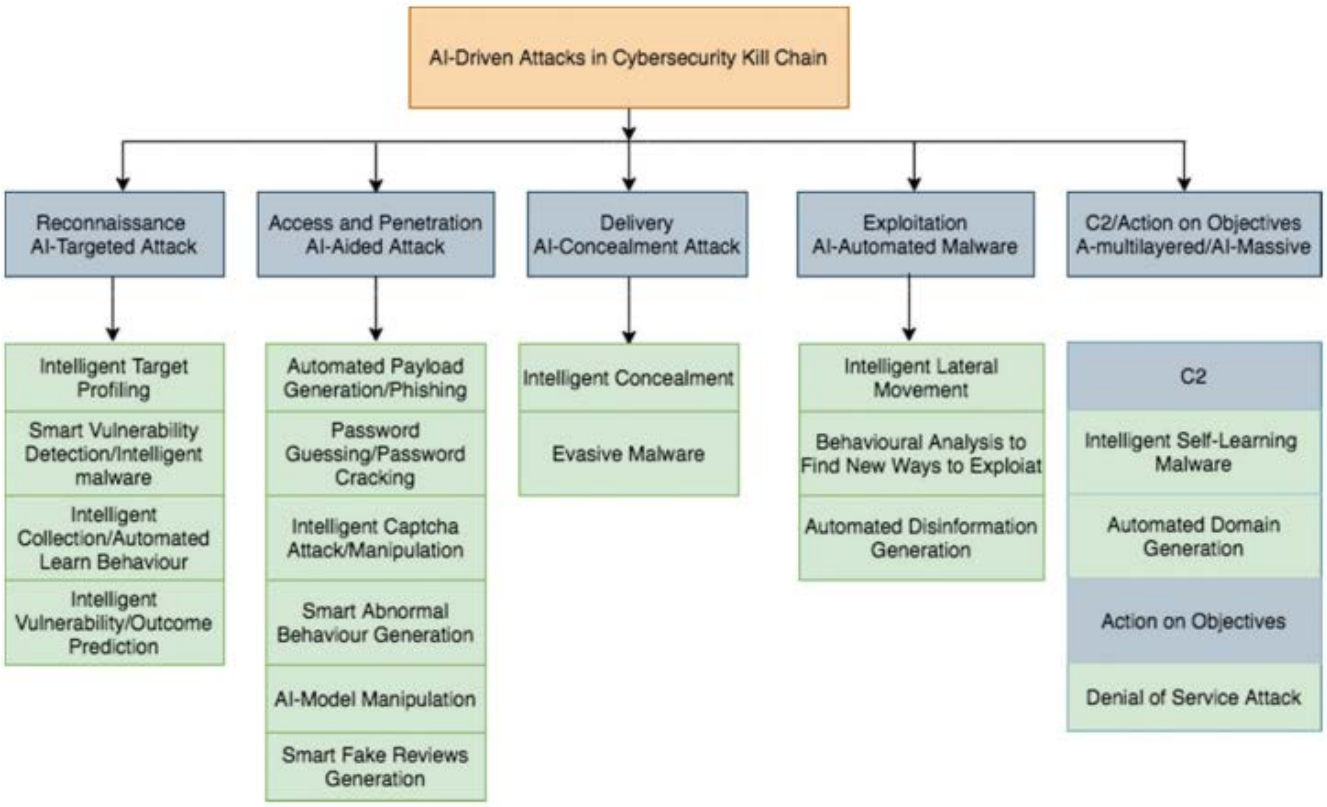

Рис.2. Атаки по этапам kill chain [17]

На этапе доступа и проникновения (access and penetration - атака с использованием ИИ) были выявлены шесть типов атак, управляемых ИИ, на этапе разведки доступа были выявлены четыре типа атак, управляемых ИИ (access reconnaissance stage - атака, нацеленная на ИИ), на этапе эксплуатации (exploitation stage - атака, автоматизированная с помощью ИИ) - три типа атак, управляемых ИИ; также были выявлены два типа атак, управляемых ИИ, на этапе доставки (AI-concealment - ИИ атака с сокрытием) и этапе C2 (Command & Control - многоуровневая атака ИИ) соответственно. Напротив, один тип атаки, управляемой ИИ, был выявлен на этапе действий по целям (AI-malware - атака с использованием вредоносного ПО ИИ).

Касательно этапа доступа и проникновения было найдено больше всего публикаций (6), за ним следует этап разведки (4), на этапе эксплуатации — три публикации, а на этапах доставки и C2 — две. Напротив, на этапе действий по целям (вредоносное ПО ИИ) было наименьшее количество публикаций (1).

Большие надежды потенциальные злоумышленники возлагают на LLM (которые вызывают, соответственно, большие опасения со стороны сообщества кибербезопасности) в плане автоматизации атак. Вот примеры использования LLM для кибератак (по состоянию на начало 2024 года) в отражении цепочки kill chain [18].

Year	MITRE Tactic(s)	Application	Model(s)
2023	Execution	Generating code to perform actions that could be malicious	GPT-3
2022	Initial Access	Generate phishing emails to bypass spam filters	GPT-2, GPT-3, RoBERTa
2022	Execution - Command & Control	Use of LLMs as plug-ins to act as a proxy	GPT-4
2023	Initial Access - Collection	Generate Phishing Website via ChatGBT	GPT-3.5 Turbo
2023	Execution	Code generation and DLL injection	GPT-3
2023	Initial Access - Reconnaissance	Collecting victim data to develop an attack email	GPT-3.5, GPT-4
2023	Initial Access - Execution - Defense Evasion	Crafting malicious scripts	GPT-3.5 Turbo, GPT-4, text-davinci-003
2018	Initial Access	Spear Phishing link	AWD-LSTM
2023	Defense Evasion	Code obfuscation, file format modification	GPT-3.5
2023	Initial Access - Credential Access	Password guessing using LLMs	GPT-2
2023	Initial Access - Reconnaissance	Impersonation for phishing aims	GPT-3.5 Turbo
2022	Initial Access	Generating content for misinformation	GPT-2

Рис. 3. LLM в атаке [18]

Следует указать, что подобного рода списки будут постоянно расти. Процесс этот абсолютно естественный. Если мы хотим обучить LLM писать код, то идеи о том, что это может быть вредоносный код или какой-либо шифровальщик данных возникают, практически, автоматически. Если мы демонстрируем возможности тех же LLM писать продающие маркетинговые предложения, то легко догадаться, что текст для фишинговых рассылок будет не сильно отличаться. И так далее.

Автоматизация (демократизация – понижение порога входа и удешевление) являются естественным процессом. То же самое, соответственно, относится и к защите: тут просто не остается другого выхода. Атакующие роботы должны встречаться такими же роботами-защитниками.

IV Разгадывание капчи

Большое количество работ посвящено такого рода задачам. Объективно – распознавание изображений есть одна из самых традиционных задач для машинного (глубокого обучения). Примеры работ – [19 – 22].

Как это выглядит, разберем на примере работы [20]. В работе описывается атака на текстовые капчи (распознавание текста на картинке), Примеры таких заданий приведены на рисунке 4. Указана длина строки в символах и вносимые модификации

Scheme	Sample image	String length	Security features	Scheme	Sample image	String length	Security features
Google		8~10	Distortion, overlapping, varied fonts	Microsoft		4~6	Hollow character, diagonal distribution
Wiki		8~10	Distortion	Apple		4~5	Background, overlapping
Baidu_1		4	Noise lines, rotation	Baidu_2		4	Rotation
Baidu_3		4	Hollow character, varied fonts	Alipay		4	Overlapping, distortion
QQ		4	Varied fonts, rotation	Bilibili		5	Distortion, noise lines, rotation
Weibo		4	Distortion	Sina		5	Noise lines, varied fonts, rotation
Csdn		5	Color background	JD_1		4	Color Background
JD_2		4	Color Background	JD_3		4	Distortion
Sohu		4	Noise lines, rotation	Douban		5~8	Color background, distortion
360_1		4~5	Noise lines, varied fonts	360_2		4~5	Color background, rotation
Baidu		2	Overlapping, varied fonts,rotation	Renmin		2	Rotation, color background
Dajie		4	Overlapping, rotation, noise lines, color background	Douban		3~5	Complex background, rotation, distortion
It168		4	Rotation, noise lines				

Рис.4. Текстовые капчи [20].

Как показано на рисунке 5, атака состоит из 3-х шагов.

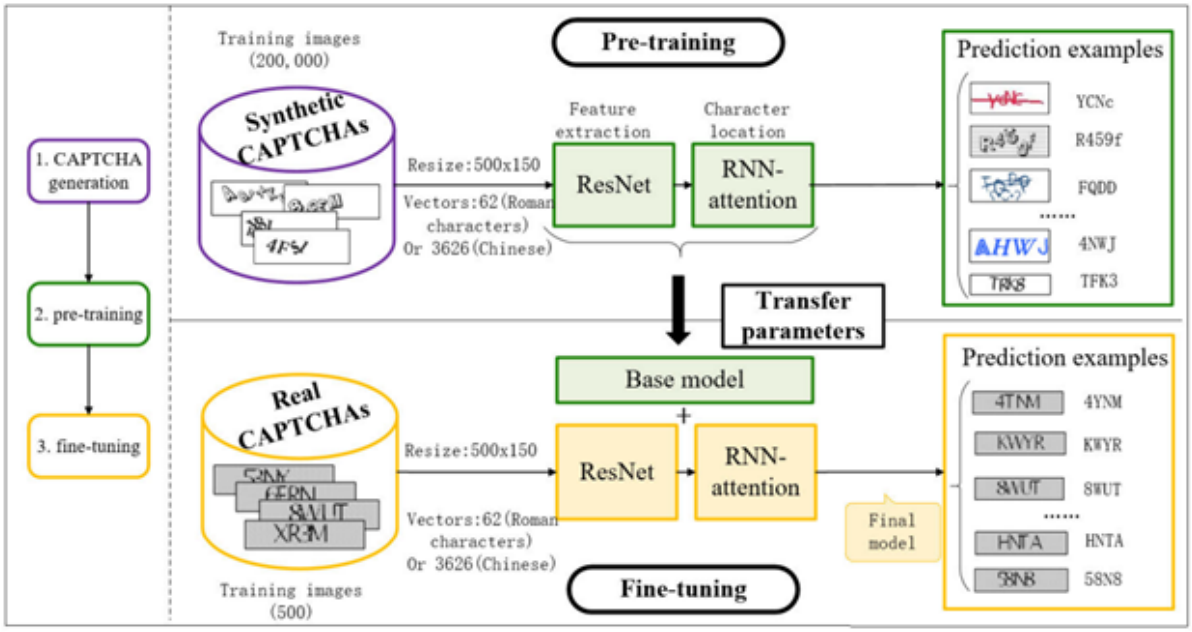

Рис.5. Схема атаки [20]

Шаг 1. Генерация CAPTCHA. На этом шаге используются алгоритмы обработки изображений для генерации CAPTCHA, не связанных с целевой схемой для обучения нашей сети распознавания. В рассматриваемой атаке все предварительные образцы генерируются совершенно случайным образом без какого-либо специального дизайна, что легко реализуется и значительно снижает усилия, затрачиваемые на сбор обучающих образцов.

Шаг 2. Предварительное обучение. После генерации синтетические CAPTCHA вводятся непосредственно в механизм распознавания без какой-либо предварительной обработки для обучения базовой модели. После предварительного обучения мы принимаем обученную модель в качестве базовой модели всех последующих схем.

Шаг 3. Тонкая настройка. Наконец, для каждой схемы были использованы 500 реальных образцов для тонкой настройки базовой модели. Этот этап был осуществлен путем повторного обучения базовой модели с использованием трансферного обучения с целью обновления параметров, соответствующих реальным признакам. Обратите внимание, что была использована только доменная адаптация трансферного обучения, и модель сохраняла согласованность на этапах предварительного обучения и повторного обучения.

Базовые архитектурные решения:

1. Чтобы сократить затраты, связанные с ручной маркировкой, для предварительного обучения были сгенерированы синтетические CAPTCHA в качестве данных предварительного обучения. Все данные обучения для базовой модели генерируются с помощью простых алгоритмов обработки изображений из библиотеки Pillow [20].

Как показано на рис. 6, все образцы предварительного обучения генерируются с черными символами на чистом белом фоне. В отличие от исходных CAPTCHA, здесь нет никаких функций безопасности в сгенерированных CAPTCHA: например, отсутствуют шумовые линии, искажения, наложения и т. д. Вместо этого образцы были сгенерированы самым простым способом, чтобы снизить стоимость генерации, поскольку этот тип CAPTCHA прост в реализации и не требует особых усилий. Сгенерированные CAPTCHA совершенно не связаны с целевыми CAPTCHA (рис. 4) и не похожи ни на одну из целевых схем.

Для схем на основе латинских символов длина текстовой строки устанавливается в диапазоне от 4 до 10; шрифты случайным образом выбираются из библиотеки шрифтов, включая как обычные, так и полые стили; все изображения имеют одинаковый размер, а угол поворота текста устанавливается от минус 45 до 45 градусов. Для китайских схем была установлена длина строки в диапазоне от 2 до 5. Было сгенерировано 500 000 изображений для предварительной подготовки базовой модели.

Рис. 6. Некоторые примеры случайно сгенерированных CAPTCHA для обучения базовой модели.

Как и для любой другой прикладной задачи в

машинном обучении, главная проблема – это данные для обучения и инженерия признаков. Все образцы были одного размера 500 × 150.

2. В предварительном обучении (базовая модель) использовалась комбинация CNN и LSTM. Чтобы распознать всю строку символов за один шаг, в качестве механизма распознавания использовалась комбинированная модель, описанная в работе [21], состоящая из CNN и модели долговременной краткосрочной памяти (LSTM). CNN отвечает за извлечение вектора признаков изображения CAPTCHA. Для этого выбрали была выбрана модель ResNet v2-101, которая разработана для решения проблемы деградации, возникающей по мере увеличения глубины сети. LSTM преобразует векторы признаков, извлеченные CNN, в одну текстовую строку; ее можно рассматривать как языковую модель на уровне символов. Решения принимаются с использованием последних состояний в ячейках памяти.

В этом эксперименте количество ячеек LSTM зависит от максимальной длины строки целевой CAPTCHA.

3. На последнем этапе (тонкая настройка) используется трансферное обучение для тонкой настройки параметров предварительно обученной модели с несколькими реальными CAPTCHA. Трансферное обучение работает следующим образом. В трансферном обучении домен D обозначается как $D = \{X, P(X)\}$, где X — пространство признаков, а $P(X)$ — предельное распределение вероятностей. Для конкретного домена задача может быть определена как $T = \{Y, f\}$, где Y обозначает пространство меток, а f обозначает целевую предсказательную функцию. В общем, полный процесс трансферного обучения включает один исходный домен (D_S) и один целевой домен (D_T), которые соответствуют одной исходной задаче (T_S) и одной целевой задаче (T_T) соответственно. Из знаний в D_S и T_S, трансферное обучение направлено на улучшение обучения целевой предсказательной функции f в D_T. В данном решателе CAPTCHA f обозначает предиктивную функцию в ResNet, а D_S и D_T имеют следующий вид:

$$D_S = \{ (x_{s1}, y_{s1}) \dots (x_{sn}, y_{sn}) \}$$
$$D_T = \{ (x_{T1}, y_{T1}), \dots (x_{Tn}, y_{Tn}) \}$$

Что касается обучающих данных, $x_{Si} \in X_S$ — это синтетическая CAPTCHA, а $y_{Ti} \in Y_T$ — соответствующая метка CAPTCHA, строка символов. Здесь x_{Ti} и y_{Ti} имеют те же значения, что и в реальных CAPTCHA. Обратите внимание, что все метки остаются теми же в D_S и D_T (62 или 3626 символов), но пространства признаков различаются, поскольку признаки синтетической и реальной CAPTCHA имеют разные детали. Для каждой схемы на основе римских символов использовались 500 вручную размеченных реальных образцов. Учитывая, что китайские CAPTCHA имеют больший набор символов, чем римские CAPTCHA, использовались 1000 реальных вручную размеченных китайских CAPTCHA на китайскую схему.

I. Заключение

В качестве заключения приведем следующие 5 пунктов, которые, по мнению авторов работы [27], определяют будущее наступательного ИИ. Они связывают это с генеративным искусственным интеллектом и большими языковыми моделями (LLM), натренированными на создание вредоносного контента (пример – FraudGPT [27]).

1. Автоматизированная социальная инженерия и фишинговые атаки

LLM, типа FraudGPT, демонстрируют способность генеративного искусственного интеллекта поддерживать убедительные сценарии для предлогов, которые могут ввести жертв в заблуждение. Один из вариантов – злоумышленники просят LLM написать научно-фантастические истории о том, как работает успешная стратегия социальной инженерии или фишинга, заставляя, таким образом, саму LLM предоставлять рекомендации по атаке. Другие варианты использования могут состоять в запросе инструкций для осуществления атак на национальных языках. В этом случае защитные фильтры, настроенные на английский язык, могут не срабатывать.

2. Вредоносные программы и эксплойты, созданные искусственным интеллектом

FraudGPT доказал свою способность генерировать вредоносные сценарии и код, адаптированные к сети, конечным точкам и более широкой ИТ-среде конкретной жертвы. Начинающие злоумышленники могут быстро освоить новейшие средства защиты, используя генеративные системы на основе искусственного интеллекта, такие как FraudGPT, для изучения и последующего развертывания сценариев атак. Вот почему организации должны сделать все возможное, чтобы обеспечить кибергигиену, включая защиту конечных точек. Вредоносное ПО, созданное искусственным интеллектом, может обходить старые системы кибербезопасности, не предназначенные для выявления и предотвращения этой угрозы.

3. Автоматическое обнаружение ресурсов киберпреступниками.

Генеративный ИИ сократит время, необходимое для проведения ручных исследований с целью поиска новых уязвимостей, поиска и сбора скомпрометированных учетных данных, изучения новых инструментов взлома и освоения навыков, необходимых для запуска сложных кампаний по борьбе с киберпреступностью. Злоумышленники всех уровней квалификации будут использовать его для обнаружения незащищенных конечных точек, атак на незащищенные поверхности угроз и запуска кампаний по атаке на основе информации, полученной с помощью простых подсказок.

Отмечается, что наряду с идентификацией, конечные точки будут подвергаться большему количеству атак. Отмечается, что самовосстанавливающиеся конечные

точки являются решающими, особенно в смешанных средах ИТ и операционных технологий (OT), которые полагаются на датчики интернета вещей (IoT). Конечная точка с самовосстановлением (self healing end-point) — это технология для автоматизации мониторинга и диагностики проблем производительности и безопасности на различных сетевых узлах или конечных точках.

Традиционное реагирование на инциденты часто требует значительного ручного вмешательства для выявления и устранения скомпрометированных систем. С другой стороны, самовосстанавливающиеся конечные точки используют алгоритмы искусственного интеллекта и машинного обучения для автоматического обнаружения, изоляции и устранения инцидентов безопасности без вмешательства человека. Эти конечные точки постоянно отслеживают и анализируют поведение системы, обеспечивая упреждающее обнаружение угроз и автономное реагирование, что приводит к сокращению времени реагирования и снижению вероятности широкого компрометации.

Эти конечные точки могут заранее обнаруживать аномалии и потенциальные угрозы безопасности, постоянно отслеживая свое поведение и сетевые коммуникации. Такой упреждающий подход не только снижает потребность в постоянном вмешательстве человека, но также помогает обнаруживать и смягчать риски, укрепляя общий уровень безопасности.

4. Уклонение от защиты с помощью ИИ только начинается, и настоящих проблем мы еще не видели.

Генеративный ИИ, вооруженный оружием, все еще находится в зачаточном состоянии, а FraudGPT — лишь первые шаги. Появляются более совершенные и смертоносные инструменты. Они будут использовать генеративный искусственный интеллект для обхода систем обнаружения и реагирования на конечных точках, а также создавать варианты вредоносного ПО, которые смогут избежать обнаружения статических сигнатур.

5. Сложность выявления и атрибуции

FraudGPT и будущие приложения и инструменты генеративного искусственного интеллекта снизят уровень обнаружения и установления авторства до уровня анонимности. Командам безопасности будет сложно отнести атаки с использованием искусственного интеллекта к конкретной группе угроз или кампании на основе криминалистических артефактов или доказательств. Большая анонимность и затрудненное обнаружение приведут к увеличению времени ожидания и позволят злоумышленникам выполнять долговременные атаки, которые типичны для атак с расширенными постоянными угрозами (Advanced persistent threat – APT [28]) на важные цели. Генеративный ИИ, снабженный оружием, в конечном итоге сделает это доступным для каждого злоумышленника.

Благодарности

Автор благодарны сотрудникам лаборатории Открытых информационных технологий кафедры Информационной безопасности факультета ВМК МГУ имени М.В. Ломоносова за обсуждения и ценные замечания. Статья продолжает серию публикаций, начатых работой об обосновании исследований, посвященных устойчивым моделям машинного обучения [29]. Традиционно отмечаем, что все публикации в журнале INJOIT, связанные с цифровой повесткой, начинались с работ В.П. Куприяновского и его соавторов [30-32].

Библиография

[1] Намиот, Д. Е., Е. А. Ильюшин, and И. В. Чижов. "Искусственный интеллект и кибербезопасность." International Journal of Open Information Technologies 10.9 (2022): 135-147.
[2] Магистратура «Кибербезопасность» МГУ-СБЕР https://cyber.cs.msu.ru Проверено 22.06.2024
[3] Магистерская программа «Искусственный интеллект в кибербезопасности» (ФГОС) https://cs.msu.ru/node/3732 Проверено 22.06.2024
[4] How AI-Driven Cyberattacks Will Reshape Cyber Protection https://www.forbes.com/councils/forbestechcouncil/2024/03/19/how-ai-driven-cyber-attacks-will-reshape-cyber-protection/ Проверено 15.08.2024.
[5] Vielberth, Manfred, et al. "Security operations center: A systematic study and open challenges." IEEE Access 8 (2020): 227756-227779.
[6] Намиот, Д. Е. "Схемы атак на модели машинного обучения." International Journal of Open Information Technologies 11.5 (2023): 68-86.
[7] NIST AI 100-2 E2023 Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations https://csrc.nist.gov/pubs/ai/100/2/e2023/final Проверено: 15.07.2024
[8] Perdisci, Roberto, et al. "Misleading worm signature generators using deliberate noise injection." 2006 IEEE Symposium on Security and Privacy (S&P'06). IEEE, 2006.
[9] 14 Risks and Dangers of Artificial Intelligence (AI) https://builtin.com/artificial-intelligence/risks-of-artificial-intelligence Проверено 15.08.2024.
[10] Song, Junzhe, and Dmitry Namiot. "A survey of the implementations of model inversion attacks." International Conference on Distributed Computer and Communication Networks. Cham: Springer Nature Switzerland, 2022.
[11] NIST SP 800-53 Rev. 5 Security and Privacy Controls for Information Systems and Organizations https://csrc.nist.gov/pubs/sp/800/53/r5/upd1/final Проверено 15.08.2024.
[12] OWASP https://owasp.org/ Проверено 15.08.2024.
[13] Defining AI Hacking: The Rise of AI Cyber Attacks https://www.sangfor.com/blog/cybersecurity/defining-ai-hacking-rise-ai-cyber-attacks Проверено 15.08.2024.
[14] XBOW https://xbow.com/ Проверено 15.08.2024.
[15] RunSybil https://www.runsybil.com/ Проверено 15.08.2024.
[16] Mirsky, Yisroel, and Wenke Lee. "The creation and detection of deepfakes: A survey." ACM computing surveys (CSUR) 54.1 (2021): 1-41..
[17] Guembe, Blessing, et al. "The emerging threat of ai-driven cyber attacks: A review." Applied Artificial Intelligence 36.1 (2022): 2037254.
[18] Motlagh, Farzad Nourmohammadzadeh, et al. "Large language models in cybersecurity: State-of-the-art." arXiv preprint arXiv:2402.00891 (2024).
[19] Derea, Zaid, et al. "Deep Learning Based CAPTCHA Recognition Network with Grouping Strategy." Sensors 23.23 (2023): 9487.
[20] Wang, P.; Gao, H.; Shi, Z.; Yuan, Z.; Hu, J. Simple and easy: Transfer learning-based attacks to text CAPTCHA. IEEE Access 2020, 8, 59044–59058.
[21] Yu, N.; Darling, K. A low-cost approach to crack python CAPTCHAs using AI-based chosen-plaintext attack. Appl. Sci. 2019, 9, 2010.

[22] Kumar, M.; Jindal, M.; Kumar, M. An efficient technique for breaking of coloured Hindi CAPTCHA. Soft Comput. 2023, 27, 11661–1168
[23] Pillow https://pypi.org/project/pillow/ Проверено 17.08.2024
[24] Wojna, Zbigniew, et al. "Attention-based extraction of structured information from street view imagery." 2017 14th IAPR international conference on document analysis and recognition (ICDAR). Vol. 1. Ieee, 2017.
[25] AI Risk Repository https://airisk.mit.edu/ Проверено 17.08.2024
[26] AI Risk Repository preprint https://cdn.prod.website-files.com/669550d38372f33552d2516e/66bc918b580467717e194940_The%20AI%20Risk%20Repository_13_8_2024.pdf Проверено 17.08.2024
[27] How FraudGPT presages the future of weaponized AI https://venturebeat.com/security/how-fraudgpt-presages-the-future-of-weaponized-ai/ Проверено 18.08.2024
[28] Постоянная серьёзная угроза https://ru.wikipedia.org/wiki/%D0%9F%D0%BE%D1%81%D1%82%D0%BE%D1%8F%D0%BD%D0%BD%D0%B0%D1%8F_%D1%81%D0%B5%D1%80%D1%8C%D1%91%D0%B7%D0%BD%D0%B0%D1%8F_%D1%83%D0%B3%D1%80%D0%BE%D0%B7%D0%B0 Проверено 18.08.2024
[29] Намиот, Д. Е., Е. А. Ильюшин, and И. В. Чижов. "Текущие академические и индустриальные проекты, посвященные устойчивому машинному обучению." International Journal of Open Information Technologies 9.10 (2021): 35-46.
[30] Цифровая железная дорога - инновационные стандарты и их роль на примере Великобритании / Д. Е. Николаев, В. П. Куприяновский, Г. В. Суконников [и др.] // International Journal of Open Information Technologies. – 2016. – Т. 4, № 10. – С. 55-61. – EDN WXBAZN.
[31] Развитие транспортно-логистических отраслей Европейского Союза: открытый BIM, Интернет Вещей и кибер-физические системы / В. П. Куприяновский, В. В. Аленьков, А. В. Степаненко [и др.] // International Journal of Open Information Technologies. – 2018. – Т. 6, № 2. – С. 54-100. – EDN YNIRFG.
[32] Умная инфраструктура, физические и информационные активы, Smart Cities, BIM, GIS и IoT / В. П. Куприяновский, В. В. Аленьков, И. А. Соколов [и др.] // International Journal of Open Information Technologies. – 2017. – Т. 5, № 10. – С. 55-86. – EDN ZISODV.

On cyberattacks using Artificial Intelligence systems

Dmitry Namiot

Abstract—**This article discusses one aspect of the use of Artificial Intelligence in cybersecurity. It is about cyberattacks that can be carried out using Artificial Intelligence (AI) systems. AI-enabled cyberattacks can be defined as any hacking operation that relies on the use of AI mechanisms. Another term used is offensive AI. AI-based cyberattacks are undoubtedly changing the cybersecurity landscape. First of all, it is necessary to talk about the speed of implementation of attacks and their scaling. AI-based cyberattacks involve the use of advanced machine learning algorithms to identify vulnerabilities, predict patterns, and exploit weaknesses. Efficiency and rapid data analysis enhance the ability of hackers to gain a tactical advantage, resulting in rapid intrusions or destruction of data. Traditional cybersecurity methods are no longer sufficient to combat sophisticated attacks, as AI-enabled cyberattacks adapt and evolve in real time. In addition, the introduction of AI systems in cyber defense creates new risks. AI systems themselves become targets of adversarial attacks.**

The article discusses general issues of organizing cyberattacks using AI and provides taxonomy and examples of such attacks.

***Keywords*— machine learning, deep learning, cybersecurity, cyberattacks.**

REFERENCES

[1] Namiot, D. E., E. A. Il'jushin, and I. V. Chizhov. "Iskusstvennyj intellekt i kiberbezopasnost'." International Journal of Open Information Technologies 10.9 (2022): 135-147.

[2] Magistratura «Kiberbezopasnost'» MGU-SBER https://cyber.cs.msu.ru Provereno 22.06.2024

[3] Magisterskaja programma «Iskusstvennyj intellekt v kiberbezopasnosti» (FGOS) https://cs.msu.ru/node/3732 Provereno 22.06.2024

[4] How AI-Driven Cyberattacks Will Reshape Cyber Protection https://www.forbes.com/councils/forbestechcouncil/2024/03/19/how-ai-driven-cyber-attacks-will-reshape-cyber-protection/ Provereno 15.08.2024.

[5] Vielberth, Manfred, et al. "Security operations center: A systematic study and open challenges." IEEE Access 8 (2020): 227756-227779.

[6] Namiot, D. E. "Shemy atak na modeli mashinnogo obuchenija." International Journal of Open Information Technologies 11.5 (2023): 68-86.

[7] NIST AI 100-2 E2023 Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations https://csrc.nist.gov/pubs/ai/100/2/e2023/final Provereno: 15.07.2024

[8] Perdisci, Roberto, et al. "Misleading worm signature generators using deliberate noise injection." 2006 IEEE Symposium on Security and Privacy (S&P'06). IEEE, 2006.

[9] 14 Risks and Dangers of Artificial Intelligence (AI) https://builtin.com/artificial-intelligence/risks-of-artificial-intelligence Provereno 15.08.2024.

[10] Song, Junzhe, and Dmitry Namiot. "A survey of the implementations of model inversion attacks." International Conference on Distributed Computer and Communication Networks. Cham: Springer Nature Switzerland, 2022.

[11] NIST SP 800-53 Rev. 5 Security and Privacy Controls for Information Systems and Organizations https://csrc.nist.gov/pubs/sp/800/53/r5/upd1/final Provereno 15.08.2024.

[12] OWASP https://owasp.org/ Provereno 15.08.2024.

[13] Defining AI Hacking: The Rise of AI Cyber Attacks https://www.sangfor.com/blog/cybersecurity/defining-ai-hacking-rise-ai-cyber-attacks Provereno 15.08.2024.

[14] XBOW https://xbow.com/ Provereno 15.08.2024.

[15] RunSybil https://www.runsybil.com/ Provereno 15.08.2024.

[16] Mirsky, Yisroel, and Wenke Lee. "The creation and detection of deepfakes: A survey." ACM computing surveys (CSUR) 54.1 (2021): 1-41..

[17] Guembe, Blessing, et al. "The emerging threat of ai-driven cyber attacks: A review." Applied Artificial Intelligence 36.1 (2022): 2037254.

[18] Motlagh, Farzad Nourmohammadzadeh, et al. "Large language models in cybersecurity: State-of-the-art." arXiv preprint arXiv:2402.00891 (2024).

[19] Derea, Zaid, et al. "Deep Learning Based CAPTCHA Recognition Network with Grouping Strategy." Sensors 23.23 (2023): 9487.

[20] Wang, P.; Gao, H.; Shi, Z.; Yuan, Z.; Hu, J. Simple and easy: Transfer learning-based attacks to text CAPTCHA. IEEE Access 2020, 8, 59044–59058.

[21] Yu, N.; Darling, K. A low-cost approach to crack python CAPTCHAs using AI-based chosen-plaintext attack. Appl. Sci. 2019, 9, 2010.

[22] Kumar, M.; Jindal, M.; Kumar, M. An efficient technique for breaking of coloured Hindi CAPTCHA. Soft Comput. 2023, 27, 11661–1168

[23] Pillow https://pypi.org/project/pillow/ Provereno 17.08.2024

[24] Wojna, Zbigniew, et al. "Attention-based extraction of structured information from street view imagery." 2017 14th IAPR international conference on document analysis and recognition (ICDAR). Vol. 1. Ieee, 2017.

[25] AI Risk Repository https://airisk.mit.edu/ Provereno 17.08.2024

[26] AI Risk Repository preprint https://cdn.prod.website-files.com/669550d38372f33552d2516e/66bc918b580467717e194940_The%20AI%20Risk%20Repository_13_8_2024.pdf Provereno 17.08.2024

[27] How FraudGPT presages the future of weaponized AI https://venturebeat.com/security/how-fraudgpt-presages-the-future-of-weaponized-ai/ Provereno 18.08.2024

[28] Postojannaja ser'joznaja ugroza https://ru.wikipedia.org/wiki/%D0%9F%D0%BE%D1%81%D1%82%D0%BE%D1%8F%D0%BD%D0%BD%D0%B0%D1%8F_%D1%81%D0%B5%D1%80%D1%8C%D1%91%D0%B7%D0%BD%D0%B0%D1%8F_%D1%83%D0%B3%D1%80%D0%BE%D0%B7%D0%B0 Provereno 18.08.2024

[29] Namiot, D. E., E. A. Il'jushin, and I. V. Chizhov. "Tekushhie akademicheskie i industrial'nye proekty, posvjashhennye ustojchivomu mashinnomu obucheniju." International Journal of Open Information Technologies 9.10 (2021): 35-46.

[30] Cifrovaja zheleznaja doroga - innovacionnye standarty i ih rol' na primere Velikobritanii / D. E. Nikolaev, V. P. Kuprijanovskij, G. V. Sukonnikov [i dr.] // International Journal of Open Information Technologies. – 2016. – T. 4, # 10. – S. 55-61. – EDN WXBAZN.

[31] Razvitie transportno-logisticheskih otraslej Evropejskogo Sojuza: otkrytyj BIM, Internet Veshhej i kiber-fizicheskie sistemy / V. P. Kuprijanovskij, V. V. Alen'kov, A. V. Stepanenko [i dr.] // International Journal of Open Information Technologies. – 2018. – T. 6, # 2. – S. 54-100. – EDN YNIRFG.

[32] Umnaja infrastruktura, fizicheskie i informacionnye aktivy, Smart Cities, BIM, GIS i IoT / V. P. Kuprijanovskij, V. V. Alen'kov, I. A. Sokolov [i dr.] // International Journal of Open Information Technologies. – 2017. – T. 5, # 10. – S. 55-86. – EDN ZISODV.