DOI 10.18699/sblai2023-43
УДК 004.056.5

Решение глобальных проблем кибербезопасности, основанных на искусственном интеллекте, с помощью искусственного интеллекта

Addressing AI-driven global cybersecurity challenges with artificial intelligence

利用人工智能應對人工智能驅動的全球網絡安全挑戰

Чжан Сяосун

Иностранный действительный член Российской инженерной академии, Москва, Россия
Университет электронных наук и технологий Китая, Чэнду, Китай
Foreign Academician, Russian Academy of Engineering, Moscow, Russia
University of Electronic Science and Technology of China, Chengdu, China
俄羅斯工程院外籍院士，俄羅斯莫斯科
中國電子科技大學, **成都**, **中國**
johnsonzxs@uestc.edu.cn

Аннотация. В работе анализируется текущее развитие ИИ в глобальной кибербезопасности и разъясняется критическая позиция ИИ в глобальных проблемах кибербезопасности. Он выдвигает идею реагирования на глобальные проблемы кибербезопасности, вызванные ИИ, с помощью ИИ. Искусственный интеллект при разработке новых сценариев угроз в кибератаках, таких как угрозы атак типа «отказ в обслуживании», угрозы атак социальной инженерии, угрозы вредоносного кода, автоматическая эксплуатация уязвимостей и другие области, имеют типичные случаи. Хакерская конференция 2018 IBM Research продемонстрировала точный выпуск вредоносного кода вредоносной программы DeepLocker. Традиционную оборону трудно предотвратить. Технология искусственного интеллекта станет важной силой в борьбе с атаками искусственного интеллекта в будущем. Проанализируйте преимущества AIGC, представленные LLM, вычислениями знаний на основе графов знаний, глубоким обучением и другими связанными технологиями, объясните текущий статус развития и будущие прорывные точки ИИ в приложениях для атак и защиты в области кибербезопасности, а также подведите итог, что технология ИИ является важной часть возможностей активной безопасности. Продемонстрируйте подлинность и эффективность защиты сетевой безопасности, достигнутой командой с использованием технологии искусственного интеллекта. С этой целью рекомендуется усилить исследования и приложения, чтобы способствовать созданию и разработке интеллектуальных сетевых систем атак и защиты, укреплять инфраструктуру данных для сетевых атак и защиты, а также поощрять практический эффект технологий сетевых атак и защиты искусственного интеллекта.

Ключевые слова: искусственный интеллект, кибербезопасность, нейронные сети, машинное обюучение, механизмы защиты

Annotation. The paper analyzes the current development of AI in global cybersecurity and

clarifies the critical position of AI in global cybersecurity challenges. It puts forward the idea of responding to the worldwide cybersecurity challenges brought by AI with AI. Artificial intelligence in developing new threat scenarios in cyberattacks, such as denial of service attack threats, social engineering attack threats, malicious code threats, automated vulnerability exploitation, and other areas have typical cases. 2018 Hacker Conference IBM Research demonstrated a precise release of malicious code DeepLocker malware. Traditional defense is difficult to prevent. AI technology will be a significant force in dealing with AI attacks in the future. Analyze the advantages of AIGC represented by LLMs, knowledge computing based on knowledge graphs, and deep learning and other related technologies, explain the current development status and future breakthrough points of AI in cybersecurity attack and defense applications, and summarize that AI technology is an essential part of active security capability. Demonstrate the authenticity and effectiveness of network security defense achieved by the team using AI technology. To this end, it is recommended to strengthen research and application to promote the construction and development of intelligent network attack and defense systems, strengthen the data infrastructure in network attack and defense, and encourage the practical effect of artificial intelligence network attack and defense technologies.

Keywords: artificial intelligence, cybersecurity, neural networks, machine learning, protection mechanisms

註解。報告分析了人工智能在全球網絡安全領域的發展現狀，明確了人工智能在全球網絡安全挑戰中的關鍵地位。 提出用AI應對人工智能帶來的全球網絡安全挑戰的理念。 人工智能在網絡攻擊中開發新的威脅場景，如拒絕服務攻擊威脅、社會工程攻擊威脅、惡意代碼威脅、自動化漏洞利用等領域都有典型案例。 2018黑客大會IBM研究院演示了精準釋放惡意代碼的DeepLocker惡意軟件。 傳統防禦很難防範。人工智能技術將成為未來應對人工智能攻擊的重要力量。 分析以LLM為代表的AIGC、基於知識圖譜的知識計算、深度學習等相關技術的優勢，闡釋人工智能在網絡安全攻防應用中的發展現狀及未來突破點，總結人工智能技術是網絡安全攻防應用的關鍵主動安全能力的一部分。展示團隊利用AI技術實現的網絡安全防禦的真實性和有效性。為此，建議加強研究應用，推動智能網絡攻防系統建設發展，強化網絡攻防數據基礎設施，鼓勵人工智能網絡攻防技術發揮實效。

關鍵字：人工智慧、網路安全、神經網路、機器學習、保護機制 註解。

1. Введение

Технология искусственного интеллекта зародилась в 1943 году, когда психолог Уоррен МакКаллок и математик Уолтер Питтс предложили первую модель нейрона [1]. В 1950 году учёный-компьютерщик Джон МакКарти предложил термин «искусственный интеллект» [2] и организовал первую конференцию по искусственному интеллекту. С 1950-х по 1970-е годы искусственный интеллект развился до такой степени, что экспертные системы начали использоваться в химическом анализе и диагностике инфекционных заболеваний. После 1980-х годов, с появлением машинного обучения и

нейронных сетей, искусственный интеллект превратился из пассивного в активный, а автономное обучение и распознавание закономерностей в данных стали реальностью. В настоящее время коммуникационные и сетевые технологии высоко развиты, а число людей, использующих Интернет и Интернет вещей, также продемонстрировало взрывной рост. Благодаря большим данным и глубокому обучению искусственный интеллект даже превзошел уровень распознавания изображений и естественного языка. Искусственный интеллект достиг человеческого уровня и может дать хорошие результаты во многих областях, таких как здравоохранение, финансы и сельское хозяйство. В будущем искусственный интеллект продолжит развиваться, и его использование также будет способствовать дальнейшему развитию человеческого общества.

В настоящее время практически все страны выработали политику поддержки развития приложений ИИ. Например, Управление научно-технической политики Белого дома [3] выпустило «Национальный стратегический план исследований и разработок в области искусственного интеллекта на 2019 год», Россия – «Текущее состояние и перспективы развития и применения искусственного интеллекта в военной сфере», а правительство Франции объявило, что в течение пяти лет инвестирует 1,5 млрд. евро в развитие стратегий ИИ. Что касается сотрудничества, то многие международные компании, такие как Microsoft, IBM и Cisco, также планируют усовершенствовать экосистему безопасности с помощью крупномасштабной технологии языкового моделирования (LLM).

Однако, в то время как искусственный интеллект быстро развивается, он также привнесет новые вызовы в различные сферы человеческого общества. С точки зрения сетевой безопасности, технология искусственного интеллекта может устранить недостатки традиционных систем безопасности, повысить их общие показатели безопасности и обеспечить лучшую защиту от все более сложных киберугроз [4].

Столкнувшись с неизбежной тенденцией участия искусственного интеллекта в сетевой безопасности, в этой статье описываются новые изменения, внесенные ИИ в цепочку сетевых атак [5], с точки зрения цепочки сетевых атак. Показаны новые преимущества новых типов сетевых атак в интеллектуальных атаках типа «DoS-атака», ликвидация вредоносного кода, автоматическое тестирование на проникновение и т.д., которое привлекло внимание людей к сетевым атакам, управляемым искусственным интеллектом. Также показано, что с точки зрения сетевой защиты искусственный интеллект также может показывать хорошие результаты, особенно при поиске сетевых уязвимостей, обнаружение уязвимостей способствует выполнению сетевые атаки, и это также способствует защите сети. Заблаговременное исправление уязвимостей в большей степени способствует поддержанию сетевой

безопасности. Наконец, в этой статье кратко излагаются проблемы развития технологий искусственного интеллекта и вызовы, которые ИИ ставит перед сетевой безопасностью, а также выдвигаются предложения по поддержанию сетевой безопасности. Цель статьи - показать, что проблемы глобальной сетевой безопасности необходимо решать совместно с ИИ.

2. Кибератаки, управляемые искусственным интеллектом

При проведении кибератаки, внедрение искусственного интеллекта в атаку имеет два преимущества. С одной стороны, технология искусственного интеллекта делает задачу кибератаки более автоматизированной и масштабируемой, а стоимость - ниже. С другой стороны, технология искусственного интеллекта может автоматически анализировать механизм защиты цели, находить слабые места, а затем генерировать новые атаки, основанные на этой слабости. Эти новые атаки могут обойти механизм безопасности и получить более высокий процент успеха атаки. Судя по модели цепочки сетевых атак, искусственный интеллект охватил весь жизненный цикл сетевых атак, как показано в таблице 1.

2.1. Интеллектуальная атака типа «DoS-атака»

Разведка и отслеживание — это первый этап цепочки поражений сети. На этом этапе ядром операции в основном являются различные средства для получения как можно большей информации о подвергшемся нападению пользователе, чтобы найти недостатки и лазейки у подвергшегося нападению пользователя, обеспечить эффективность атаки.

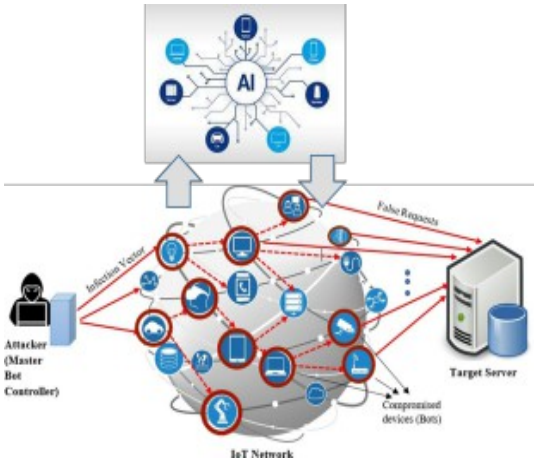

Рис. 1. Автономные, интеллектуальные и крупномасштабные атаки типа «DoS-атака»
Fig. 1. Autonomous, Intelligent and Scaled Launching of DoS Attacks
自主、智慧、規模化發動DoS 攻擊

DDoS-атаки, служащие прикрытием для атак, способствуют другим актам кражи информации и могут играть определенную роль в разведке и отслеживании.Что касается интеллектуальных атак типа «DoS-атака» (DoS) [6], то с увеличением масштаба и частоты ботнетов и распределенных атак типа «DoS-атака» (DDoS), интегрированный в DoS, искусственный интеллект станет основной формой DoS-атак в будущем.

Сегодня некоторые DDoS-атаки используют технологию искусственного интеллекта для подключения миллионов целевых устройств и создания роевой сети. Эти групповые сети с возможностями самообучения искусственного интеллекта могут применять свои собственные стратегии, основанные на информации об устройстве, без необходимости отправлять команды и другие конфиденциальные операции через терминал управления, по-настоящему реализуя децентрализованное автономное интеллектуальное сотрудничество и значительно повышая способность DDoS атаковать большее количество целей одновременно.

2.2. Применение искусственного интеллекта в социальной инженерии

Социальная инженерия — это использование человеческих уязвимых мест для получения ценной информации. Участие социальной инженерии необходимо для разведки и отслеживания. В связи с этим ключевую роль сыграло развитие технологий искусственного интеллекта, таких как перевод языка с помощью искусственного интеллекта, цифровая обработка с помощью искусственного интеллекта и генерация текста с помощью искусственного интеллекта. В частности, технология DeepFake [7] позволяет подделывать видео, аудио и картинки, а ChatGPT[8] может создавать интеллектуальных чат-ботов. Все эти приложения для искусственного интеллекта показывают, что при внедрении технологии искусственного интеллекта атаки социальной инженерии сократят затраты и значительно повысят вероятность успеха. В исследовании Darktrac, озаглавленном «Влияние создания искусственного интеллекта на кибератаки по электронной почте», количество новых атак социальной инженерии, основанных на искусственном интеллекте, увеличилось на 135% с момента выхода ChatGPT. В частности, за последние три месяца количество мошеннических электронных писем увеличилось на 70%, что заставило 82% людей беспокоиться о том, что хакеры теперь используют технологию искусственного интеллекта для автоматической и быстрой генерации неотличимых мошеннических электронных писем.

Юаньшунь Яо и другие [9] использовали технологию рекуррентной нейронной сети для генерации поддельных комментариев. Эти комментарии могут не только ускользать от обнаружения человеком, но и, как показали исследования, их распознавание людьми очень высоко.

Рис. 2. Поддельный файл, сгенерированный искусственным интеллектом
Figure 2. Fake file generated by artificial intelligence
圖2.人工智慧產生的假文件

Дж. Сеймур и др. [10] используя технологию круговой нейронной сети SNAP_R, научили создавать фишинговые посты для конкретных пользователей, а количество переходов по ссылкам, отслеживающим IP-адреса, доказывает, что размещенная поддельная информация обманула глаза людей. Эти технологии могут быть использованы в мошеннических электронных письмах для достижения эффекта подделок. Если они не будут осторожны, произойдет утечка информации.

Таблица 1 Модель сетевой цепочки атак

Модель цепочки сетевых атак (жизненный цикл сетевой атаки)								
№	Прикладные исследования	Разведка и слежение	Создание оружия	Доставка	Уязвимости	Установка и внедрение	Управление и контроль	Достижение цели
1	Освобождение от вредоносного кода		√			√		
2	Умный подбор пароля				√			
3	Новое решение кода проверки текста				√			
4	Интеллектуальная атака типа «DoS-атака»	√	√	√				
5	Применение ИИ в социальной инженерии	√		√				√
6	Методы создания и использования новых вредоносных программ		√	√	√			√
7	Автоматический инструмент для тестирования на проникновение	√			√		√	√
8	Генерация DGA на основе глубокого обучения						√	

1.1. Освобождение от вредоносного кода

Отсутствие вредоносного кода [11] является основным требованием для

создания сетевого оружия для цепного поражения. Хорошее оружие атаки должно быть трудно уничтожить. Отсутствие вредоносного кода - хороший способ защиты. В связи с этим несколько исследовательских групп предложили несколько алгоритмов, основанных на глубоком обучении с подкреплением, API [12] вставку последовательности и градиентный спуск с детализацией по байтам. Вероятность успеха этих алгоритмов в достижении статического отсутствия ошибок в моделируемых данных достигает 90%.

Гроссе К. и др. [13] использовали алгоритм прямой производной, основанный на атаке нейронных сетей, для обхода обнаружения DNN на наборе данных DREBIN Android в предположении, что атакующему известны структура и параметры нейросетевой модели, с 85%-ной вероятностью успеха. Hu W. [14] использовал альтернативные детекторы для настройки системы обнаружения «черного ящика» на наборе данных malwr в предположении, что злоумышленник знает о функциональности, используемой алгоритмом обнаружения вредоносного кода, так что детектор «черного ящика» не успевал за частотой настройки вредоносного кода и, по сути, достиг 100%-ной устойчивости, и в то же время использовал вставку нерелевантных последовательностей API для обхода RNN, основанного на последовательных признаках API, для обнаружения вредоносного кода [15]. При этом, показатели безтестовости варьировались от 96,97 до 99,56%. Колоснаджи Б и др. [16] провели тестирование на таких наборах данных, как VirusShare, и обнаружили, что алгоритм градиентного спуска, использующий гранулярность байтов, позволил получить показатель отсутствия вредоносных программ 92,83%. Андерсон Х.С. и др. [17] использовали атаку «черного ящика» на основе обучения с подкреплением против модели машинного обучения Windows PE с успешностью 99,3% в тесте kill free на основе данных Virus Share, VirusTotal. Применение искусственного интеллекта повышает эффективность средств кибернетической атаки и делает их очень удобными для использования.

1.2. Новые методы создания и использования вредоносных программ

Разумеется, эффективность защиты от кибератак не может быть полностью гарантирована только уничтожением вредоносного кода, поэтому злоумышленники начали использовать методы искусственного интеллекта для создания вредоносных программ с новыми схемами атак. В 2018 году исследователи IBM представили DeepLocker [18], который положил начало развитию вредоносного ПО на основе ИИ. В 2023 году хакеры начали использовать ChatGPT для написания вредоносных программ. Например, исследователи кибербезопасности компании CyberArk совместно с ChatGPT создали полиморфное вредоносное ПО [19], автоматизирующее эффект мутации кода. Таким образом, механизм запуска вредоносного поведения, отравляемого ИИ, представляет собой большую проблему для исследователей безопасности, поскольку традиционные методы анализа вредоносного ПО

затруднены для эффективного анализа вредоносного поведения, генерируемого ИИ, и являются устаревшими.

В работе Д. Кират и др. [20] использовалась техника DeepLocker, позволяющая скрыть атакующее поведение программного обеспечения, что позволяет избежать обнаружения и повысить эффективность работы вредоносной программы. Бахсен и др. [21] использовали модель DeepPhish для обучения искусственного интеллекта проведению атак и экспериментально доказали, что глубокий фишинг, управляемый искусственным интеллектом, превосходит традиционный фишинг. Вэйвэй Ху [14] и др. сгенерировали вредоносное ПО на основе алгоритма состязательной сети (GAN) для эффективного обхода детектора «черного ящика».

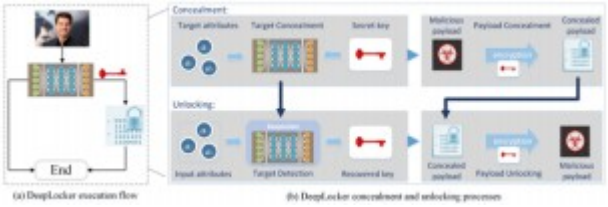

Рис. 3. ДипЛоккер, Fig.3. DeepLocker, 圖 3。深度鎖

Кейван Чунг [22] использовал метод кластеризации Kmean для создания самообучающихся вредоносных программ, способных маскироваться под непреднамеренные сбои в работе компьютерной инфраструктуры и нарушать работу системы управления компьютерной средой. Новые типы вредоносных программ, несомненно, делают оружие кибератак более мощным, и это мощное оружие внедряется в целевую систему или сеть посредством интеллектуальных сервисных атак, маскируясь под электронные письма и т.д., чтобы подготовиться к последующим поражениям.

1.3. Инструменты автоматического тестирования на проникновение

Тестирование на проникновение, как средство поиска уязвимостей, может помочь представителям безопасности устранить недостатки системы, а киберзлоумышленникам - начать атаки. Автоматизированные средства проникновения стали официально выпускаться с момента проведения конкурса CGC, организованного компанией Darpa в 2016 году. Новейшие интеллектуальные архитектуры проникновения DeepExploit, Mayhem и другие [23] позволяют проводить эффективные, качественные и недорогие атаки.

Мохамед С. Ганем и др. [24] представили интеллектуальную автоматизированную систему тестирования на проникновение (IAPTS), которая использует подход к моделированию среды и задач ПТ как частично наблюдаемого марковского процесса принятия решений, где реализуемые

стратегии решаются с помощью POMDP, а обучение с подкреплением улучшает ПТ с точки зрения затраченного времени, охваченных векторов атак, точности и аккуратности, что позволяет ей работать лучше любого человеческого эксперта по ПТ. Ц. Ли и другие [25] предложили автоматизированный подход к тестированию на проникновение, основанный на архитектуре DeepExploit с агентом обучения с подкреплением, обученным по модели A3C [26] для получения опыта в выборе точных полезных нагрузок для использования имеющихся уязвимостей с целью быстрого выявления уязвимостей сервера. Искусственный интеллект оказался более эффективным, чем человек, в решении задач поиска уязвимостей и других аспектов.

1.4. Интеллектуальное подбирание паролей

Что касается интеллектуального взлома паролей, то компания Home Security Heroes провела исследование, в ходе которого выяснилось, что 51% из 15,6 млн. паролей можно взломать менее чем за минуту с помощью различных методов искусственного интеллекта.

Если время взлома увеличить до одного часа, то можно взломать 65% паролей. Кроме того, при дальнейшем увеличении времени взлома выяснилось, что 71% паролей может быть взломан за день, а 81% - за месяц. Конкретные сценарии взлома паролей приведены в табл. 2. Кроме того, компания разработала тестовое программное обеспечение, при вводе тестового пароля он выдаст ИИ ожидаемое время взлома, конечно, скорость работы интеллектуального пароля не может быть улучшена без длительных исследований многих научных коллективов.

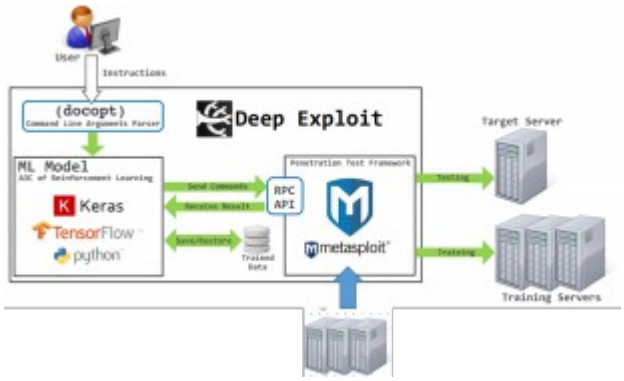

Рис. 4. DeepExploit
Fig. 4. DeepExploit
圖 4. 深度漏洞利用

Фанъи Юй и др. [27] представили GNPassGAN, когда выход Pass-GAN

комбинируется с выходом HashCat, он способен подобрать на 51%-73% больше паролей, чем при использовании только HashCat. Чжиян Ся и др. [28] решили отказаться от неэффективности HashCat и предложили GENPass. Используя генератор паролей PCFG+LSTM, сочетающий нейронную сеть и PCFG, они экспериментально доказали, что эта комбинация повышает скорость расшифровки на 16%-30% по сравнению с паролями, использующими только PCFG [29]. Дарио Паскини и другие [30] использовали глубокое обучение для улучшения обучения представлениям догадок при расшифровке паролей, что позволило ИИ получать результаты с высокой степенью дешифровки за счет быстрого обучения. Позднее Фанъи Юй и др [31] усовершенствовали генеративные адверсарные сети, которые можно использовать для автономного угадывания паролей по тралу. По сравнению с современной моделью PassGAN, GN-PassGAN смогла угадать на 88,03% больше паролей и сгенерировать на 31,69% меньше дубликатов. По сей день интеллектуальный взлом паролей продолжает развиваться вместе с развитием технологий искусственного интеллекта, и постоянно появляются новые коммерческие продукты. Интеллектуальный взлом паролей позволяет кибернетическому оружию выполнять команды в сетевой системе цели, что является важным инструментом для кражи информации и выполнения команд.

1.5. Устройство для решения текстовой капчи (CAPTCHA)

Чтобы предотвратить автоматическое выполнение высокочастотных обращений ИИ, практически все системы и сайты устанавливают CAPTCHA для верификации «человек-компьютер». В таких сценариях традиционные методы OCR [32] позволяют без помех распознавать обычные текстовые CAPTCHA, но с развитием технологий неправильно отформатированные текстовые CAPTCHA создают новые проблемы. Поэтому исследователи также разработали новый метод автоматического распознавания с использованием методов искусственного интеллекта, чтобы добиться автоматического распознавания CAPTCHA с более высокой точностью. В заключение следует отметить, что методы искусственного интеллекта могут обеспечить решения для автоматического проникновения с более высокой эффективностью и низкой стоимостью.

Дунлян Сюй и др. [33] усовершенствовали метод обучения ИИ таким образом, что он автономно выбирает информационно насыщенные образцы для обучения без ручного контроля, чтобы получить обучающие метки в реальном времени, а затем выбирает успешно распознанные образцы для обучения нейронной сети, что позволяет достичь высокой точности функции распознавания CAPTCHA и повысить скорость расшифровки CAPTCHA в случае небольшого набора данных. Тянь Цянь [34] и др. предложили новую систему распознавания CAPTCHA на основе конволюционной нейронной сети,

используя фреймворк глубокого обучения Tensorflow для применения конволюционной нейронной сети в области распознавания CAPTCHA, благодаря чему коэффициент распознавания CAPTCHA достигает 92%. Юй Н. и др. [35] использовали для взлома CAPTCHA технику TOD+CNN, которая значительно лучше, чем TOD+Peak+CNN, с точностью взлома более 90%, точностью на уровне символов и скоростью верификации более 75%. С привлечением искусственного интеллекта решение CAPTCHA сводится не просто к перебору чисел, а к автоматическому распознаванию текстовых файлов, что облегчает автоматическое развертывание инструментов атаки.

Таблица 2 График взлома паролей

Кол-во символов	**Только цифры**	**Буквы только нижнего регистра**	**Нижний и верхний регистры + прописные буквы**	**Цифры + буквы верхнего и нижнего регистра**	**Цифры + буквы верхнего и нижнего регистра + символы**
4	Мгновенно	Мгновенно	Мгновенно	Мгновенно	Мгновенно
5	Мгновенно	Мгновенно	Мгновенно	Мгновенно	Мгновенно
6	Мгновенно	Мгновенно	Мгновенно	Мгновенно	4 секунды
7	Мгновенно	Мгновенно	22 секунды	42 секунды	6 минут
8	Мгновенно	3 секунды	19 минут	48 минут	7 часов
9	Мгновенно	1 минута	11 часов	2 дня	2 недели
10	Мгновенно	1 час	4 недели	6 месяцев	5 лет
11	Мгновенно	23 часа	4 лет	38 лет	356 лет
12	25 секунд	3 недели	289 лет	2K лет	30K лет
13	3 минуты	11 месяцев	16K лет	91K лет	2M лет
14	36 минут	49 лет	827K лет	9M лет	187M лет
15	5 часов	890 лет	47M лет	613M лет	14Bn лет
16	2 дня	23K лет	2Bn лет	26Bn лет	1Tn лет
17	3 недели	812K лет	539.72M лет	2Tn лет	95Tn лет
18	10 месяцев	22M лет	7.23Bn лет	96Tn лет	6Qn лет

1.6. Генерация DGA на основе глубокого обучения

Алгоритм DGA [36] - технология, используемая вредоносными программами, которая позволяет им динамически генерировать серию доменных имен для установления связи с командно-контрольным сервером (C&C-сервером). Использование DGA позволяет обойти меры сетевой безопасности и сделать обмен данными между вредоносными программами более скрытным и трудноотслеживаемым.

В настоящее время некоторые исследователи разработали алгоритм генерации DeepDGA на основе глубокого обучения, используя в качестве обучающих данных известные доменные имена на сайте Alexa. В алгоритме

также используются LSTM [37] и GAN для построения общей архитектуры.

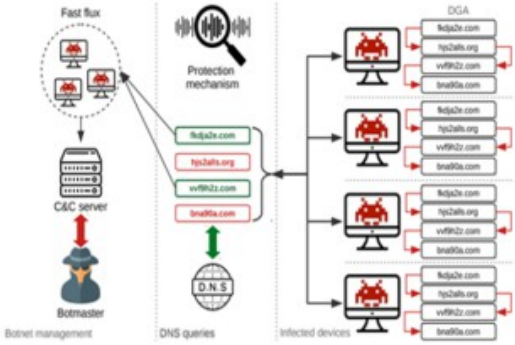

Рис. 5. Генерация доменного имени DGA с помощью GAN
Fig. 5. DGA domain name generation using GAN
圖 5. 使用 GAN 產生 DGA 域名

В результате доменные имена, генерируемые алгоритмом, очень похожи на имена обычных сайтов, которые трудно обнаружить традиционными методами обнаружения, что позволяет киберпреступникам контролировать зараженную целевую систему, отправлять команды и получать данные из удаленного места и в конечном итоге достигать цели кибератаки.

Х.С. Андерсон и др. [38] построили DGA на основе глубокого обучения, который способен обходить детекторы на основе глубокого обучения и генерировать доменные имена, которые трудно обнаружить. Й. Чэн и др. [39] разработали более угрожающий DGA ShadowDGA, использующий генеративные адверсарные сети (GAN) для имитации распределения доброкачественных доменов, и результаты показали, что домены, генерируемые ShadowDGA, труднее всего обнаружить по сравнению с существующими DGA. Технология генерации DGA на основе глубокого обучения обеспечивает скрытость и стойкость цепочки поражения сети, что является важной составляющей цепочки поражения.

3. Киберзащита с использованием искусственного интеллекта

Выше описана роль ИИ в цепочке кибератак. Хотя искусственный интеллект создал огромные проблемы для защиты сетевой безопасности, искусственный интеллект никогда не использовался для атак. Напротив, сетевая защита, управляемая искусственным интеллектом [40], может компенсировать недостатки искусственной защиты благодаря своей мощной возможности обучения и анализа данных, значительно сокращающие интервал между обнаружением угроз и реагированием на них, а также реализующие автоматическую и быструю идентификацию, обнаружение и устранение угроз безопасности. Эти функции играют важную роль в борьбе с различными

угрозами безопасности, особенно в обнаружении неизвестных и современных угроз (например, APT), которые обладают высокой эффективностью.

3.1. Интеллектуальный поиск и устранение уязвимостей

Уязвимости всегда были важной частью сетевой безопасности. Интеллектуальные методы поиска и устранения уязвимостей более эффективны, чем традиционные методы поиска уязвимостей, и могут также помочь нам проактивно обнаруживать потенциальные атаки и защищаться от них. При этом интеллектуальный поиск уязвимостей [41] может сократить время обнаружения уязвимостей и время их устранения, снижая вероятность потери активов. Кроме того, методы устранения уязвимостей с использованием искусственного интеллекта позволяют сэкономить больше времени на устранение уязвимостей по сравнению с традиционными методами.

Цзэнань Чу и др. [42] предложили алгоритм поиска и противодействия уязвимостям ботнета и его систему, объединив машинное обучение и классификационный поиск, разработали модель ботнета на основе машинного обучения и метод ее построения, основанный на характеристиках групповой сетевой атаки ботнета, управления компьютерным роем, распределенной атаки типа «DoS-атака» и т.д., а также на анализе формы уязвимости ботнета и способа распространения уязвимости. Предложен алгоритм поиска и противостояния уязвимостям ботнета, и на основе имитационных экспериментов доказано, что средняя скорость обнаружения уязвимостей алгоритмом составляет более 87%, а скорость устранения уязвимостей - более 86%. Клэр Ле Гуэс и др. [43] использовали метод GenProg, основанный на эволюционном алгоритме и вдохновленный биологической эволюцией, для устранения 55 из 105 различных уязвимостей. Обнаружение уязвимостей является ключевым моментом как для злоумышленников, так и для защитников, а ИИ позволяет устранять уязвимости, что является мощным инструментом киберзащиты.

3.2. Интеллектуальное обнаружение сетей и логов

Важной инициативой по защите цепи поражения сети является усиление защиты границ, в то время как системы обнаружения и защиты от вторжений должны ежедневно справляться с большим объемом сетевого трафика и журнальных данных при защите и обнаружении сетевой безопасности. При работе с такими массивами данных традиционные методы обнаружения сталкиваются с узкими местами, что приводит к низкой эффективности обработки данных системой обнаружения и неудовлетворительным результатам обнаружения. Напротив, интеллектуальные методы обнаружения, основанные на алгоритмах, позволяют в значительной степени уловить внутренние взаимосвязи и особенности этих массивных данных трафика и журналов. Например, методы обнаружения могут быть объединены со стратегиями на основе искусственного интеллекта, такими как мультисенсорные машины и

метаобучение, для достижения обнаружения различных или сложных сценариев атак.

А.Дж. Родригез [44] разработал интеллектуальную систему обнаружения вторжений на основе журналов, которая была обучена с помощью алгоритмов обучения без контроля, в частности алгоритма Kmeans [45] и алгоритма One Class SVM [46], и показала точность обнаружения аномальных шаблонов в тестовых журналах до 85%. Благодаря ИИ, помогающему вручную обрабатывать журналы, аномальная информация может быть обнаружена в кратчайшие сроки, что позволит проводить обработку аномалий и подавать сигналы тревоги, а также подготовиться к последующей защите от атак.

3.3. Принятие решений по обеспечению сетевой безопасности на основе ChatGPT

ChatGPT — это технология, которая может использоваться не только хакерами, но и сотрудниками сферы безопасности. В ChatGPT формально интегрировано несколько типовых сценариев [47]: обнаружение фишинговых писем, определение нагрузки проникновения, формирование документов безопасности, формирование отчетов о реагировании на инциденты, обработка данных об угрозах и т.д. ChatGPT продемонстрировал отличные результаты по снижению стоимости, сложности и оперативности выполнения операций по обеспечению безопасности для защитников.

3.4. Интеллектуальная платформа для тестирования активной безопасности, основанная на интеллектуальных

Другим показательным приложением в области кибербезопасности на основе ИИ является платформа активного тестирования безопасности на основе ИИ. Платформа основана на CDSL-Yak, больших языковых моделях и графах знаний. Она также объединяет несколько типов накопления основных данных, таких как активы, уязвимости и бизнес-информация. Обладая эффективными и экономически выгодными возможностями скрининга рисков безопасности, платформа предназначена для проведения активного тестирования безопасности с целью обнаружения потенциальных рисков безопасности, скрытых в целевой среде организации.

4. Заключение

4.1. Проблемы развития технологий ИИ

В настоящее время развитие технологий искусственного интеллекта по-прежнему сталкивается с рядом проблем: первая из них - точность предсказаний. В последнее время технологии искусственного интеллекта не могут достичь 100-процентной точности обнаружения, а в некоторых случаях имеют место ложные срабатывания. Вторая проблема - целостность данных.

Данные имеют решающее значение для использования ИИ, однако в некоторых конкретных сценариях сбор данных затруднен. Кроме того, отсутствие наборов данных, символизирующих ключевые признаки, также является проблемой для соответствующих исследований. Третья проблема - интерпретируемость. Модели ИИ, особенно основанные на глубоком обучении и использующие высокоразмерные данные, в настоящее время не могут решить такие проблемы, как интерпретируемость и неоднозначность данных. Четвертое - вычислительные ресурсы. С ростом числа моделей и технологий, связанных с LLM, потребность в вычислительных мощностях становится все выше и выше, превращаясь в реальное требование для передовых исследований в области ИИ. Для некоторых регионов с ограниченными финансовыми возможностями это еще более актуально. Пятое - мировое общественное мнение.

4.2. Проблемы в области сетевой безопасности

Одним словом, кибератаки на основе ИИ полностью трансформируют то, что раньше было трудоемкими и дорогостоящими атаками, в распределенное, интеллектуальное и автоматизированное по сравнению с традиционными кибератаками направление. В результате атаки становятся все более точными и автоматизированными. Упомянутые выше кибератаки, управляемые ИИ, являются одними из новых вызовов на пути кибератак, которые можно свести к следующим трем категориям проблем.

Во-первых, использование ИИ для изучения характеристик среды и повышения адаптивности и скрытности атаки. В сетевой среде объекта атаки данные, поведение и т.д. имеют определенные локализованные характеристики. Злоумышленники используют ИИ для сбора и моделирования характеристик данных, поведения и т.д. в целевой сети, изучения нормального содержания данных, частоты передачи, способа доставки и других характеристик среды в целевой сетевой среде. Они используют характеристики среды для выбора соответствующих средств атаки, маскируют данные атаки под обычные данные с нормальными характеристиками в целевой сети и маскируют поведение атаки под поведение обычных пользователей в целевой сети. Они реализуют адаптивное к среде поведение атаки, скрытие данных, повышают скрытность атаки и адаптивность атаки.

Во-вторых, использование искусственного интеллекта для усиления эффекта распределенного взаимодействия и повышения стойкости атаки. Атакующий внедряет алгоритмы распределенного интеллектуального взаимодействия, чтобы преобразовать традиционное единое планирование распределенных объектов атаки интеллектуальным центром для проведения совместных атак в автономное взаимодействие и групповое принятие решений распределенными многоинтеллектуальными объектами атаки без центра, что позволяет повысить эффективность взаимодействия между несколькими распределенными узлами атаки, уменьшить зависимость от централизованного

совместного планирования, снизить риск противодействия атакам и повысить их устойчивость.

В-третьих, использование искусственного интеллекта для достижения самоэволюции метода атаки и повышения ее эффективности. Атакующая сторона с помощью искусственного интеллекта анализирует эффект от различных методов атаки и возможные контрмеры защитника, а затем автоматически выбирает новый механизм атаки с учетом слабых мест защитника, реализуя тем самым интеллектуальную эволюцию метода атаки. Например, злоумышленник может принимать результаты работы системы обнаружения вторжений защитника за обратную связь, использовать технологию искусственного интеллекта для сбора и моделирования данных обратной связи, создания модели эффекта атаки и динамической настройки соответствующего метода атаки для обхода системы обнаружения вторжений.

4.3 Рекомендации по противодействию угрозе атак с использованием ИИ

Мы также можем использовать ИИ для решения глобальных проблем кибербезопасности, возникающих благодаря ИИ. Ниже приводятся некоторые рекомендации:

1. Усилить исследования и прикладные разработки. Как страны, так и университеты могут внести свой вклад в создание и совершенствование интеллектуальных киберсистем, включая атаку и оборону.

2. Расширять совместный доступ. Поощрять исследователей к решению проблемы данных искусственного интеллекта при построении технологических систем кибернетической атаки и обороны.

3. Совершенствовать меры противодействия и оценки. Мы работаем вместе, чтобы продвигать разработку технологий кибернетической атаки и защиты на основе искусственного интеллекта в практических приложениях, приносящих существенную пользу государствам, коллаборациям и отдельным людям.

1. Introduction

Artificial intelligence technology originated in 1943, when psychologist Warren McCulloch and mathematician Walter Pitts proposed the first neuron model [1]. In 1950, computer scientist John McCarthy proposed the term "artificial intelligence" [2] and organized the first artificial intelligence conference. From the 1950s to the 1970s, artificial intelligence developed to the point where expert systems were used in chemical analysis and diagnosis of infectious diseases. After the 1980s, with the emergence of machine learning and neural networks, artificial intelligence changed from passive to active, and autonomous learning and recognition of patterns in data became a reality. Nowadays, communication technology and network technology are highly developed, and the number of people using the Internet and the Internet of

Things has also shown explosive growth. With big data and deep learning, artificial intelligence has even surpassed the level of image recognition and natural language processing. Artificial intelligence has reached human level, and artificial intelligence can produce good results in many fields such as medical care, finance, and agriculture. In the future, artificial intelligence will continue to develop, and its use will also promote the continued development of human society.

At present, almost all countries have issued policies to support the development of artificial intelligence applications. For example, the White House Office of Science and Technology Policy[3] released the "2019 National Artificial Intelligence R&D Strategic Plan"; Russia released the "Development Status and Application Prospects of Artificial Intelligence in the Military Field"; the French government announced that they will invest 1.5 billion euro within 5 years to promote the development of artificial intelligence strategies. As for cooperation, many international companies, such as Microsoft, IBM and Cisco, also plan to enhance the security ecosystem through large language model technology (LLM). However, while artificial intelligence is developing rapidly, it will also bring new challenges to various fields of human society. In terms of network security, artificial intelligence technology can improve the shortcomings of traditional security systems, improve their overall security performance, and target Increasingly sophisticated cyber threats provide better protection [4]. At the same time, the use of artificial intelligence also brings risks and hidden dangers to network security.

Faced with the inevitable trend of artificial intelligence participating in network security, this article starts from the perspective of the network kill chain, describes the new changes in AI in the network kill chain [5], and demonstrates the role of new network attacks in intelligent denial of service attacks. , malicious code avoidance, automatic penetration testing, etc., have attracted people's attention to AI-driven network attacks, and then demonstrated that artificial intelligence can also show good results in network defense, especially Mining and discovering network vulnerabilities is conducive to the execution of network attacks, and is also conducive to network defense. Patch vulnerabilities in advance is more conducive to maintaining network security. Finally, this article summarizes the challenges of AI technology development and AI's challenges to network security, and puts forward suggestions for maintaining network security, aiming to show that global network security challenges need to be addressed together with AI.

2. AI-driven cyberattacks

When conducting network attacks, introducing artificial intelligence into the attack has two advantages. On the one hand, artificial intelligence technology makes network attack tasks more automated and scalable, with lower costs; on the other hand, artificial intelligence technology can automatically analyze attacks. The target's security defense mechanism looks for weaknesses, and then generates new attacks based on this weakness. This new attack can bypass the security mechanism and

achieve a higher attack success rate. From the perspective of the network kill chain model, artificial intelligence has covered the entire life cycle of network attacks, as shown in Table 1.

2.1. Intelligent Denial-of-Service (DoS) Attacks

Figure 1 Autonomous, Intelligent and Scaled Launching of DoS Attacks

Reconnaissance tracking is the first stage of the network kill chain, in this stage, the core of the action is mainly a variety of means to obtain as much information as possible about the attacked, so as to find the deficiencies and loopholes of the attacked to ensure the effectiveness of the attack.DDos attack as a front of the attack to assist the other stealing of information can play the role of reconnaissance tracking. In terms of intelligent denial-of-service (DoS) [6] attacks, DoS-integrated AI will become the mainstream form of DoS attacks in the future as botnets and distributed denial-of-service (DDoS) attacks become larger and more frequent. Today, some DDoS attacks have adopted AI techniques to connect millions of targeted devices to build swarm networks. These group networks with AI self-learning capabilities can take their own strategies based on device information, without the need to send commands and other sensitive operations through the control terminal, truly realizing decentralized autonomous intelligent collaboration, significantly enhancing the ability of DDoS to attack more targets at the same time.

2.2. AI in social engineering attack applications

Social engineering is the use of human weaknesses to obtain valuable information. Reconnaissance and tracking require the participation of social engineering. In this regard, the development of artificial intelligence technology has played a key role, such as artificial intelligence language translation, artificial intelligence number processing, and artificial intelligence text. generate. Specifically, DeepFake technology [7] can fake videos, audios and pictures, and ChatGPT [8] can realize intelligent chat robots. All of these AI applications demonstrate that social engineering attacks will be less costly and significantly more successful if AI technology is employed. In a Darktrac survey titled "The Impact of Generated AI on Email Cyberattacks," the number of new AI-based social engineering attacks increased by 135% since the release of ChatGPT. Specifically, the number of fraudulent emails has increased by 70% in the past three months, leading 82% of people to worry that hackers are now using artificial intelligence technology to automatically and quickly generate indistinguishable fraudulent emails.

Figure 2 Fake documents generated by AI

Yuanshun Yao et al. [9] used recurrent neural network technology to generate fake comments. These comments can not only escape human detection, but also survey found that people recognize them very highly. J Seymour et al. [10] used

SNAP_R recurrent neural network technology to learn to publish phishing posts for specific users. The click-through rate of IP tracking links proved that the false information released deceived people's eyes. These techniques can be used in fraudulent emails to make them look real. If you are not careful, information will be leaked.

Table 1 Cyberkill Chain Model

1.1. Malicious code exemptions

Malicious code exemptions [11] is a basic requirement for building network kill chain weapons. A good attack weapon should be difficult to destroy. Avoidance is a good protection method. In this regard, multiple research teams have proposed Some algorithms based on deep reinforcement learning, API [12] sequence insertion and byte-granular gradient descent were developed. The success rate of these algorithms in achieving static anti-killing in simulated data is as high as 90%.

Grosse K et al. [13] used the forward derivative algorithm based on the attack neural network to bypass DNN detection on the DREBIN Android data set under the assumption that the attacker knows the structure and parameters of the neural network model, and the success rate reached 85%. Hu W[14] used an alternative detector to fit the black-box detection system on the malwr data set under the assumption that the attacker knows the functions used by the malicious code detection algorithm, so that the black-box detector cannot keep up with malware adjustments. frequency, basically reaching 100% anti-virus protection. At the same time, they used irrelevant API sequence insertion to avoid RNN detection of malware based on sequential API features [15], and the test anti-virus protection rate reached 96.97%~99.56%. Kolosnjaji B et al. [16] tested on VirusShare and other data sets and found that using byte-granularity gradient descent algorithm can make the malware avoidance rate reach 92.83%. Anderson H S et al. [17] used a black box attack based on reinforcement learning for the Windows PE machine learning model, and conducted anti-virus tests under Virus Share and VirusTotal data, and the success rate reached 99.3%. The participation of artificial intelligence improves the ability of cyber-attack weapons to avoid killing, making them highly usable.

1.2. New malware construction methods and utilization

Of course, relying solely on malicious code to avoid killing cannot fully guarantee the effectiveness of network attack weapons, so attackers began to use artificial intelligence technology to build malware with new attack modes. In 2018, IBM researchers proposed DeepLocker [18], which opened the development of artificial intelligence-based malware. In 2023, hackers began to use ChatGPT to participate in writing malware. For example, cyber security researchers at CyberArk built a polymorphic malware [19] combined with ChatGPT that can automatically achieve the effect of code mutation. Therefore, the triggering mechanism of artificial

intelligence poisoning malicious behaviors brings huge challenges to security researchers, because traditional malware analysis methods are difficult to effectively analyze malicious behaviors generated by artificial intelligence and are outdated.

D Kirat et al. [20] use DeepLocker technology to hide the attack behavior of software, thereby evading detection and allowing malware to run better. Bahnsen et al. [21] used the DeepPhish model to train artificial intelligence and let it launch attacks. Experiments proved that artificial intelligence-driven deep phishing outperformed traditional phishing. The malware generated by Weiwei Hu [14] and others based on the adversarial network (GAN) algorithm can effectively bypass the black box detector. Keywhan Chung et al. [22] used K-means clustering technology to generate self-learning malware that can disguise itself as an unintentional failure of computer infrastructure and destroy the computer environment control system. New types of malware have undoubtedly made network attack weapons more powerful. These powerful weapons are deployed into target systems or networks through intelligent service attacks, disguised as emails, etc., to prepare for subsequent destruction.

Figure 3 DeepLocker

1.3. Automated penetration testing tools

Penetration testing, as a means of finding vulnerabilities, can help defenders fix system flaws and also help network attackers launch attacks. Since the CGC competition organized by DARPA in 2016, automated penetration tools have been officially produced. The latest intelligent penetration architecture of DeepExploit, Mayhem, etc. [23] can achieve efficient, high-quality, and low-cost attack effects.

Mohamed C. Ghanem et al. [24] launched the Intelligent Automated Penetration Testing System (IAPTS), which uses the PT environment and task modeling method as part of the observation Markov decision-making process, and solves feasible strategies through POMDP, while reinforcement learning can Enhances PT in terms of time consumption, attack vectors covered, accuracy and precision, making it perform beyond any human PT expert. Q Li et al. [25] proposed an automated penetration testing method based on the DeepExploit architecture, using reinforcement learning agents to train using the A3C model [26] to gain experience in selecting precise payloads to exploit available vulnerabilities for rapid identification. Server vulnerabilities. It turns out that artificial intelligence has an advantage over humans in handling things like vulnerability mining.

Figure 4 DeepExploit

1.4. Intelligent password guessing

In terms of smart password cracking, Home Security Heroes launched a study and found that 51% of 15.6 million passwords can be cracked in less than 1 minute through different types of artificial intelligence technology. If the cracking time is extended to one hour, then 65% of passwords can be cracked. Furthermore, if you

continue to extend the cracking event, you will find that 71% of passwords can be cracked within a day and 81% of passwords can be cracked within a month. The specific password cracking situation is shown in Table 2. In addition, the company has developed a test software. When you enter a test password, it will give the estimated time for AI to decipher it. Of course, the improvement of smart password speed is inseparable from the long-term research of many scientific teams.

Fangyi Yu et al. [27] introduced GNPassGAN. When the output of Pass-GAN is combined with the output of HashCat, it can match 51%-73% more passwords than using HashCat alone. Zhiyang xia et al. [28] considered giving up. For the inefficient HashCat, they proposed GENPass, which uses PCFG+LSTM password generator and combines neural network and PCFG. Experiments have shown that this combination improves the deciphering rate by 16% to 30% compared to passwords using PCFG alone [29]. Dario Pasquini et al. [30] used deep learning to improve guessing representation learning for password deciphering, allowing artificial intelligence to quickly train to produce high deciphering results. Later, Fangyi Yu et al. [31] improved the generative adversarial network, which can be used for offline trawl password guessing. Compared with the state-of-the-art PassGAN model, GN-PassGAN can guess 88.03% more passwords and generate 31.69 fewer duplicates %. To this day, smart password cracking is still developing with the development of artificial intelligence technology, and continue to introduce more commercial products, smart password cracking makes the network kill weapons can be in the target network system to further execute the instructions, is the information theft, the instructions to execute the indispensable tool.

1.5. New text verification code solver (CAPTCHA)

In order to prevent artificial intelligence from automatically performing high-frequency access, almost all systems and websites will set verification codes for human-machine verification. In these scenarios, traditional OCR technology [32] can achieve ordinary text verification code recognition without interference, but with the development of technology, malformed text verification codes bring new challenges. Therefore, the researchers also developed a new automatic identification method that uses artificial intelligence technology to achieve automatic identification based on verification codes with higher accuracy. In short, artificial intelligence technology can provide solutions for automated penetration with higher efficiency and lower costs.

Dongliang Xu et al. [33] improved the learning method of artificial intelligence, allowing it to independently select information-rich samples for learning without manual supervision to obtain real-time training labels, and then select successfully identified samples for neural network training, thereby achieving the goal of learning on a small scale. In the case of large-scale data sets, high-precision verification code recognition function is realized and the speed of deciphering verification codes is improved. Tian Qian[34] et al. proposed a new verification code recognition system

based on convolutional neural network and used the Tensorflow deep learning framework to apply the convolutional neural network to the field of verification code recognition. The verification code recognition rate reaches 92%. Yu, N et al. [35] used TOD+CNN technology to crack verification codes, and the effect was significantly better than TOD+Peak+CNN, with a cracking accuracy of more than 90%, and a character-level accuracy and verification rate of more than 75%. With the participation of artificial intelligence, verification code solving is not just about violently cracking numbers, but also automatically identifying text files, which facilitates the automatic deployment of attack weapons.

1.6. DGA generation based on deep learning

The DGA algorithm [36] is a technique used by malware that allows malware to dynamically generate a series of domain names to establish communication with a command-and-control server (C&C server). The use of DGA is designed to bypass network security measures, making the malware's communications stealthier and more difficult to track.

Currently, some researchers use well-known domain names on Alexa as training data to build a DeepDGA generation algorithm based on deep learning. The algorithm also uses LSTM [37] and GAN to build the overall architecture. Therefore, the domain name generated by this algorithm is very similar to the domain name of ordinary websites, which is difficult to detect with traditional detection methods. This is very conducive to the control of network anti-personnel weapons to infect the target system, send instructions and obtain data remotely, and ultimately achieve the purpose of network attacks.

H S. Anderson et al. [38] built a DGA based on deep learning, which can bypass deep learning-based detectors and generate domain names that are difficult to detect. Y Cheng et al. [39] A more threatening DGA, ShadowDGA uses a generative adversarial network (GAN) to simulate the distribution of benign domains. The results show that compared with existing DGA, the domains generated by ShadowDGA are the most difficult to detect. DGA generation technology based on deep learning ensures the concealment and persistence of the network kill chain and is an important part of the kill chain.

Table 2 Password cracking timetable

3. AI empowers network defense

The above describes the role of artificial intelligence in the network kill chain. Although artificial intelligence brings huge challenges to network security defense, artificial intelligence has never been used only for attacks. On the contrary, artificial intelligence-driven network defense [40] can Its powerful learning and data analysis capabilities make up for the shortcomings of manual defense, greatly shorten the interval between threat discovery and response, and realize automatic and rapid identification, detection and disposal of security threats. These functions play an

important role in responding to various security threats, especially in discovering unknown threats and advanced *threats (such as APT) with high efficiency.*

Figure 5 Generation of DGA domain name using GAN

3.1. Intelligent vulnerability mining and repair

Vulnerabilities have always been an important part of cybersecurity. Intelligent vulnerability mining and remediation technology is more efficient than traditional vulnerability mining methods, and it can also help us proactively discover and defend against potential attacks. At the same time, intelligent vulnerability mining [41] can shorten the vulnerability discovery time and vulnerability repair time, and reduce the probability of asset loss. In addition, AI-based vulnerability repair technology can save more repair time compared to traditional technologies.

ZENAN CHU et al. [42] combined machine learning with classification mining and proposed a botnet vulnerability mining and countermeasure algorithm and system. Based on the characteristics of botnet group network attacks, computer group control, and distributed denial-of-service attacks, designed a botnet model and its construction method based on machine learning. Based on the analysis of botnet vulnerability forms and vulnerability diffusion patterns, a botnet vulnerability mining and countermeasures algorithm was proposed. Through simulation experiments, the average vulnerability of the algorithm was proved the detection rate exceeds 87%, and the vulnerability repair rate exceeds 86%. Claire Le Goues et al. [43] used GenProg, an evolutionary algorithm-based method inspired by biological evolution, to fix 55 of 105 vulnerabilities. The discovery of vulnerabilities is the key to both network offense and defense. Artificial intelligence can patch vulnerabilities and is a powerful tool to assist network defense.

3.2. Intelligent network and log detection

An important measure to defend against the network kill chain is to strengthen boundary defense. Intrusion detection and prevention systems need to deal with a large amount of network traffic and log data in daily network security protection and detection. Traditional detection methods encounter bottlenecks when facing these massive data, resulting in low processing efficiency of the detection system and poor detection results. On the contrary, detection methods based on intelligent algorithms can highly capture the intrinsic relationships and characteristics in these massive traffic and log data. For example, detection methods can be combined with artificial intelligence-based strategies such as multi-perceptron and meta-learning to detect different or complex attack scenarios.

AJ Rodrigues[44] developed an intrusion detection system based on intelligent logs. The system uses unsupervised learning algorithms for training, especially the Kmeans algorithm [45] and the One Class SVM algorithm [46]. The results prove that the system detects intrusion detection in test logs. The accuracy of the exception log mode is as high as 85%. With the help of artificial intelligence in manually

processing log files, abnormal information can be detected as soon as possible, abnormal processing and alarms can be made, and preparations for subsequent attacks can be made.

3.3 Assisting network security decision-making based on ChatGPT

ChatGPT is also a technology that can be used by security guards rather than hackers. There are several typical scenarios that have been officially integrated with ChatGPT [47]: phishing email detection, penetration payload identification, security document generation, incident response report generation, threat intelligence processing, etc. ChatGPT has demonstrated outstanding performance in reducing the cost, complexity, and timeliness of defenders' security operations.

2.4 Intelligent-based active security testing platform

Another representative artificial intelligence-driven network security application is the active security testing platform based on artificial intelligence. The platform is based on CDSL-Yak, large language models and knowledge graphs. It also combines several types of core data accumulation, such as assets, vulnerabilities, and business information. The platform has efficient and low-cost security risk screening capabilities and is designed to actively carry out security testing and discover potential security risks hidden in the enterprise's target environment.

4 Conclusion

4.1 Challenges in AI technology development

At present, the development of artificial intelligence technology still faces several challenges: The first is the accuracy of prediction. Artificial intelligence technology has recently been unable to achieve 100% detection accuracy, and in some cases, there will be some false positives. Second is data integrity. Data is critical to the utilization of artificial intelligence, but in some specific scenarios, data collection is difficult. Furthermore, the lack of datasets embodying key features is also a problem in related research. The third is interpretability. Artificial intelligence models, especially models based on deep learning that utilize high-dimensional data, are currently unable to solve problems such as interpretability and data ambiguity. The fourth is computing resources. With the rise of LLM-related models and technologies, the requirements for computing power are getting higher and higher, which has become a realistic requirement for cutting-edge artificial intelligence research. This is even more critical for some financially strapped regions. The fifth is public opinion around the world.

4.2 Challenges to network security

In short, AI-based cyberattacks are completely transforming what used to be labor-intensive and costly attacks into a distributed, intelligent and automated direction compared to traditional cyberattacks. This is making attacks increasingly precise and automated. The AI-driven cyberattacks mentioned above are some of the new challenges in the way of cyberattacks, which can be summarized in the following three categories of challenges.

First, the use of AI to learn the characteristics of the environment and enhance the adaptability and stealth of the attack. In the network environment of the attack target, the data, behavior, etc. have certain localized characteristics. Attackers use AI to collect and model the characteristics of data, behavior, etc. in the target network, and learn the normal data content, transmission frequency, delivery method, and other environmental characteristics of the target network environment; they refer to the environmental characteristics to choose the appropriate means of attack, disguise the attack data as normal data with normal characteristics of the target network, and disguise the attack behavior as the network behavior of the normal users in the target network; Realize environment-adaptive attack behavior, data hiding, enhance the covertness of the attack, and enhance the adaptability of the attack.

Second, the use of artificial intelligence to enhance the effect of distributed collaboration and improve the robustness of the attack. The attacker introduces distributed intelligent collaboration algorithms to evolve the traditional unified scheduling of distributed attack entities by the intelligent center to carry out collaborative attacks into the autonomous collaboration and group decision-making of centerless distributed multi-intelligent attack entities, thus improving the collaboration efficiency between multiple distributed attack nodes, reducing the dependence on centralized collaborative scheduling, reducing the risk of attack countermeasures, and enhancing the robustness of attacks.

Third, the use of artificial intelligence to realize the self-evolution of the attack method and improve the effectiveness of the attack. The attacker uses artificial intelligence to analyze the attack effect of different attack methods and the possible countermeasures of the defender, and then automatically selects a new attack mechanism for the defender's weaknesses, so as to realize the intelligent evolution of the attack method. For example, the attacker can take the results of the defender's intrusion detection system as feedback, use artificial intelligence technology to collect and model the feedback data, establish an attack effect model, dynamically adjust the appropriate attack method, and circumvent the intrusion detection system.

4.3 Recommendations for responding to AI attack threats

We can also use artificial intelligence to address the global cybersecurity challenges posed by artificial intelligence. Some suggestions are as follows:

1. Strengthen research and application. Both the state and universities can promote the construction and enhancement of intelligent network systems that include attack and defense.

2. Strengthen sharing and utilization. Researchers are encouraged to solve artificial intelligence data problems in the construction of network attack and defense technology systems.

3. Strengthen confrontation and assessment. We work together to promote the development of artificial intelligence cyber-attack and defense technologies in practical applications, bringing substantial benefits to the country, cooperation and

individuals.

1. 簡介

人工智慧技術起源於 1943 年，當時心理學家 Warren McCulloch 和數學家 Walter Pitts 提出了第一個神經元模型[1]。 1950 年，電腦科學家約翰·麥卡錫提出「人工智慧」一詞[2]，並組織了第一次人工智慧會議。 從 1950 年代到 1970 年代，人工智慧發展到專家系統被用於傳染病的化學分析和診斷。 1980 年代後，隨著機器學習和神經網路的出現，人工智慧從被動變為主動，自主學習和識別資料中的模式成為現實。 如今，通訊技術和網路技術高度發達，使用網路和物聯網的人數也呈現爆發式成長。 憑藉著大數據和深度學習，人工智慧甚至超越了影像辨識和自然語言處理的程度。 人工智慧已經達到了人類的水平，人工智慧可以在醫療、金融、農業等多個領域產生良好的效果。 未來，人工智慧將持續發展，其運用也將推動人類社會的持續發展。

目前，幾乎所有國家都推出了支援人工智慧應用發展的政策。 例如，白宮科技政策辦公室[3]發布了《2019 年國家人工智慧研發戰略計畫》； 俄羅斯發布《人工智慧在軍事領域的發展現況及應用前景》； 法國政府宣布 5 年內投資 15 億歐元推動人工智慧戰略發展。在合作方面，微軟、IBM、思科等眾多國際公司也計劃透過大語言模型技術（LLM）來增強安全生態系統。 然而，人工智慧快速發展的同時，也將為人類社會各領域帶來新的挑戰。在網路安全方面，人工智慧技術可以改善傳統安全系統的缺點，提高其整體安全效能，並針對日益複雜的網路威脅提供更好的防護[4]。 同時，人工智慧的運用也為網路安全帶來風險和隱憂。

面對人工智慧參與網路安全的必然趨勢，本文從網路殺傷鏈的角度出發，描述了人工智慧在網路殺傷鏈中的新變化[5]，並論證了新型網路攻擊在智慧化中的作用。拒絕服務攻擊。 、惡意程式碼規避、自動滲透測試等，都引起了人們對 AI 驅動的網路攻擊的關注，進而證明人工智慧在網路防禦方面也能展現出良好的效果，特別是挖掘和發現網路漏洞有利於執行網絡攻擊，也有利於網路防禦。 提前修補漏洞更有利於維護網路安全。 最後，本文總結了人工智慧技術發展的挑戰以及人工智慧對網路安全的挑戰，並提出維護網路安全的建議，旨在表明全球網路安全挑戰需要與人工智慧共同應對。

2.人工智慧驅動的網路攻擊

在進行網路攻擊時，將人工智慧引入攻擊有兩個優點。 一方面，人工智慧技術使網路攻擊任務更加自動化、可擴展，成本更低； 另一方面，人工智慧技術可以自動分析攻擊。目標的安全防禦機制尋找弱點，然後根據這個弱點產生新的攻擊。 這種新的攻擊可以繞過安全機制並獲得更高的攻擊成功率。 從網路殺傷鏈模型來看，人工智慧已經涵蓋了網路攻擊的整個生命週期，如表 1 所示。

2.1. 智慧拒絕服務 (DoS) 攻擊

圖 1 自主、智慧、規模化發動 DoS 攻擊

偵察追蹤是網路殺傷鏈的第一階段，該階段行動的核心主要是透過多種手段獲取盡可能多的被攻擊者的信息，從而發現被攻擊者的缺陷和漏洞，從而對攻擊者發起攻擊。確保攻擊的有效性。DDos 攻擊作為攻擊的前沿，協助對方竊取訊息，可以起到偵察追蹤的作用。在智慧拒絕服務（DoS）[6]攻擊方面，隨著殭屍網路和分散式阻斷服務（DDoS）攻擊規模越來越大、越來越頻繁，整合 DoS 的 AI 將成為未來 DoS 攻擊的主流形式。 如今，一些

DDoS 攻擊已經採用 AI 技術來連接數百萬目標設備來建立群體網路。 這些具有 AI 自學習能力的群體網路可以根據設備資訊採取自己的策略，無需透過控制終端發送命令等敏感操作，真正實現去中心化自主智能協作，顯著增強 DDoS 攻擊能力同時設定目標。

2.2. 人工智慧在社會工程攻擊應用的應用

社會工程是利用人類的弱點來獲取有價值的資訊。 偵察和追蹤需要社會工程的參與。對此，人工智慧技術的發展發揮了關鍵作用，例如人工智慧語言翻譯、人工智慧數位處理、人工智慧文字等。 產生。 具體來說，DeepFake 技術[7]可以偽造視訊、音訊和圖片，ChatGPT[8]可以實現智慧聊天機器人。 所有這些人工智慧應用都表明，如果採用人工智慧技術，社會工程攻擊的成本將更低，而且成功率將顯著提高。 在一項名為「生成的人工智慧對電子郵件網路攻擊的影響」的 Darktrac 調查中，自 ChatGPT 發布以來，基於人工智慧的新社會工程攻擊數量增加了 135%。 具體來說，過去三個月詐騙電子郵件的數量增加了 70%，導致 82%的人擔心駭客現在正在利用人工智慧技術自動快速產生難以區分的詐騙電子郵件。

圖 2 AI 生成的假文檔

姚元順等。 [9]使用循環神經網路技術來產生虛假評論。 這些評論不僅能夠逃避人類的察覺，而且調查發現人們對它們的認可度非常高。 J·西摩等。 [10]使用 SNAP_R 循環神經網路技術來學習針對特定使用者發布釣魚貼文。 IP 追蹤連結的點擊率證明，發布的虛假資訊欺騙了人們的眼睛。 這些技術可用於詐騙電子郵件，使它們看起來很真實。 一不小心，資訊就會被洩漏。

表 1 Cyberkill 鏈模型

1.1. 惡意程式碼豁免

惡意程式碼豁免[11]是建構網路殺傷鏈武器的基本要求。 好的攻擊武器應該很難被摧毀。 迴避是一種很好的保護方法。 對此，多個研究團隊提出了一些基於深度強化學習、API[12]序列插入和位元組粒度梯度下降的演算法。 這些演算法在模擬資料中實現靜態反殺的成功率高達 90%。

格羅斯 K 等人。 [13]在假設攻擊者知道神經網路模型的結構和參數的情況下，使用基於攻擊神經網路的前嚮導數演算法在 DREBIN Android 資料集上繞過 DNN 偵測，成功率達到 85%。 Hu W[14]在假設攻擊者知道惡意程式碼偵測演算法使用的函數的情況下，使用替代偵測器在 malwr 資料集上擬合黑盒偵測系統，從而使黑盒偵測器無法跟上惡意軟體調整。 頻率，基本上達到 100%的防毒保護。 同時，他們利用不相關的 API 序列插入來避免基於順序 API 特徵的 RNN 對惡意軟體的偵測[15]，測試防毒防護率達到 96.97%~99.56%。 Kolosnjaji B 等人。 [16]在 VirusShare 等資料集上進行測試，發現使用位元組粒度梯度下降演算法可以使惡意軟體規避率達到 92.83%。 安德森 H S 等人。 文獻[17]對 Windows PE 機器學習模型採用了基於強化學習的黑盒子攻擊，並在 Virus Share 和 VirusTotal 資料下進行了防毒測試，成功率達 99.3%。 人工智慧的參與提高了網路攻擊武器的免殺能力，使其具有很高的實用性。

1.2. 新的惡意軟體建置方法和利用

當然，單純依靠惡意程式碼免殺並不能完全保證網路攻擊武器的有效性，因此攻擊者開始利用人工智慧技術來建構具有新攻擊模式的惡意軟體。 2018 年，IBM 研究人員提出了 DeepLocker[18]，開啟了基於人工智慧的惡意軟體的開發。 2023 年，駭客開始利用 ChatGPT 參與編寫惡意軟體。 例如，CyberArk 的網路安全研究人員結合 ChatGPT 建構了一種多態惡意軟體[19]，可以自動實現程式碼變異的效果。 因此，人工智慧中毒惡意行為的

觸發機制給安全研究人員帶來了巨大的挑戰，因為傳統的惡意軟體分析方法很難有效地分析人工智慧產生的惡意行為，而且已經過時。

D 基拉特等人。 [20]利用 DeepLocker 技術隱藏軟體的攻擊行為，從而逃避偵測，讓惡意軟體更好地運作。 巴恩森等人。 [21]使用 DeepPhish 模型來訓練人工智慧並讓它發動攻擊。 實驗證明，人工智慧驅動的深度網路釣魚優於傳統網路釣魚。 Weiwei Hu [14]等人基於對抗網路（GAN）演算法生成的惡意軟體可以有效繞過黑盒偵測器。 鐘基萬等人。 [22]使用 K-means 聚類技術產生自學習惡意軟體，該惡意軟體可以將自身偽裝成電腦基礎設施的無意故障並破壞電腦環境控制系統。 新型惡意軟體無疑讓網路攻擊武器變得更加強大。 這些強大的武器透過智慧服務攻擊、偽裝成電子郵件等部署到目標系統或網路中，為後續的破壞做好準備。

圖 3 DeepLocker

1.3. 自動化滲透測試工具

滲透測試作為發現漏洞的手段，可以幫助防禦者修復系統缺陷，也可以幫助網路攻擊者發動攻擊。 自 2016 年 DARPA 舉辦的 CGC 競賽以來，自動化滲透工具正式誕生。 DeepExploit、Mayhem 等[23]最新的智慧滲透架構可以實現高效、高品質、低成本的攻擊效果。

加內姆 (Mohamed C. Ghanem) 等人 [24]推出了智慧自動化滲透測試系統（IAPTS），該系統利用 PT 環境和任務建模方法作為觀察馬可夫決策過程的一部分，透過 POMDP 求解可行策略，而強化學習可以在以下方面增強 PT：耗時、覆蓋的攻擊向量、準確度和精確度，使其表現超越任何人類 PT 專家。 Q 李等人。 [25]提出了一種基於 DeepExploit 架構的自動化滲透測試方法，利用強化學習代理使用 A3C 模型進行訓練[26]，以獲得選擇精確有效負載以利用可用漏洞進行快速識別的經驗。 伺服器漏洞。 事實證明，人工智慧在處理漏洞挖掘等事情上比人類更有優勢。

圖 4 深度漏洞利用

1.4. 智慧密碼猜測

在智慧密碼破解方面，Home Security Heroes 進行了一項研究，發現透過不同類型的人工智慧技術，1560 萬個密碼中的 51%可以在不到 1 分鐘的時間內破解。 如果破解時間延長到 1 小時，則可以破解 65%的密碼。 而且，如果繼續延長破解事件，你會發現 71%的密碼可以在一天內破解，81%的密碼可以在一個月內破解。 特定密碼破解情況如表 2 所示。此外，該公司也開發了測試軟體。 當你輸入測試密碼時，它會給出人工智慧破解它的預計時間。 當然，智慧密碼速度的提升離不開眾多科學研究團隊的長期研究。

於方毅等人。 [27]介紹了 GNPassGAN。 當 Pass-GAN 的輸出與 HashCat 的輸出結合時，它可以比單獨使用 HashCat 多匹配 51%-73%的密碼。 夏志揚等人。 [28]考慮放棄。 針對 HashCat 效率低下的問題，他們提出了 GENPass，採用 PCFG+LSTM 密碼產生器，結合神經網路和 PCFG。 實驗表明，與單獨使用 PCFG 的密碼相比，這種組合的破解率提高了 16%至 30% [29]。 達裡奧·帕斯奎尼等人 [30]利用深度學習來改進密碼破解的猜測表示學習，使人工智慧能夠快速訓練以產生較高的破解結果。 後來，於方毅等人。 [31]改進了生成對抗網絡，可用於離線拖網密碼猜測。 與最先進的 PassGAN 模型相比，GN-PassGAN 可以多猜測 88.03% 的密碼，並減少 31.69% 的重複項。 時至今日，智慧密碼破解仍在隨著人工智慧技術的發展而發展，並不斷推出更多的商業產品，智慧密碼破解使得網路殺傷武器可以在目標網路系統中進一步執行指令，就是資訊竊取、執行指令不可缺少的工具。

1.5. 新的文字驗證碼求解器 (CAPTCHA)

為了防止人工智慧自動進行高頻訪問，幾乎所有的系統和網站都會設定驗證碼進行人機驗證。 在這些場景下，傳統的 OCR 技術[32]可以實現普通文本驗證碼的無幹擾識別，但隨著技術的發展，畸形文本驗證碼帶來了新的挑戰。 因此，研究人員也開發了一種新的自動識別方法，利用人工智慧技術實現基於驗證碼的自動識別，準確率更高。 總之，人工智慧技術可以為自動化滲透提供更高效率、更低成本的解決方案。

徐棟樑等。 [33]改進了人工智慧的學習方法，使其能夠在無需人工監督的情況下自主選擇資訊豐富的樣本進行學習，獲得即時訓練標籤，然後選擇成功識別的樣本進行神經網路訓練，從而達到學習的目的小範圍內。 在大規模資料集的情況下，實現高精度驗證碼識別功能，提高驗證碼破解速度。 田謙[34]等. 提出了一種基於卷積神經網路的新型驗證碼識別系統，並利用 Tensorflow 深度學習框架將卷積神經網路應用到驗證碼識別領域。 驗證碼識別率達 92%。 於，N 等人。 文獻[35]採用 TOD+CNN 技術破解驗證碼，效果明顯優於 TOD+Peak+CNN，破解準確率超過 90%，字元級準確率和驗證率超過 75% 。 在人工智慧的參與下，驗證碼破解不僅僅是暴力破解數字，還可以自動識別文字文件，方便攻擊武器的自動部署。

1.6. 基於深度學習的 DGA 生成

DGA 演算法[36]是惡意軟體使用的技術，允許惡意軟體動態產生一系列網域名稱以與命令和控制伺服器（C&C 伺服器）建立通訊。 DGA 的使用旨在繞過網路安全措施，使惡意軟體的通訊更加隱密且更難追蹤。

目前，有研究人員利用 Alexa 上的知名網域作為訓練數據，建構基於深度學習的 DeepDGA 生成演算法。 該演算法也使用 LSTM [37] 和 GAN 來建構整體架構。 因此，此演算法產生的網域名稱與一般網站的網域名稱非常相似，用傳統的偵測方法很難偵測到。這非常有利於控製網路殺傷武器感染目標系統，遠端發送指令並獲取數據，最終達到網路攻擊的目的。

H S.安德森等人。 [38]建構了基於深度學習的 DGA，它可以繞過基於深度學習的偵測器並產生難以偵測的網域。 Y 程等人。 [39] ShadowDGA 是一種更具威脅性的 DGA，它使用生成對抗網路（GAN）來模擬良性域的分佈。 結果表明，與現有的 DGA 相比，ShadowDGA 產生的域是最難檢測的。 基於深度學習的 DGA 生成技術保證了網路殺傷鏈的隱藏性和持久性，是殺傷鏈的重要組成部分。

表 2 密碼破解時間表

3.AI 賦能網路防禦

以上描述了人工智慧在網路殺傷鏈中的作用。 儘管人工智慧為網路安全防禦帶來巨大挑戰，但人工智慧從來都不是僅僅用於攻擊。 相反，人工智慧驅動的網路防禦[40]可以以其強大的學習和數據分析能力彌補人工防禦的缺陷，大大縮短威脅發現和回應的間隔，實現自動快速識別、偵測和處置的安全威脅。 這些功能在應對各種安全威脅，尤其是高效發現未知威脅和高階威脅（如 APT）方面發揮著重要作用。

圖 5 使用 GAN 生成 DGA 域名

3.1. 智慧漏洞挖掘與修復

漏洞一直是網路安全的重要組成部分。 智慧漏洞挖掘和修復技術比傳統漏洞挖掘方法更有效率，也可以幫助我們主動發現和防禦潛在的攻擊。 同時，智慧漏洞挖掘[41]可以縮短漏洞發現時間和漏洞修復時間，降低資產損失的機率。 此外，基於 AI 的漏洞修復技術相比傳統技術可以節省更多的修復時間。

朱澤南等人。 [42]將機器學習與分類挖掘結合，提出了殭屍網路漏洞挖掘與對策演算法及系統。 根據殭屍網路群體網路攻擊、電腦群控、分散式阻斷服務攻擊的特點，設計了基於機器學習的殭屍網路模型及其建構方法。 在分析殭屍網路漏洞形態和漏洞擴散模式的基礎上，提出了殭屍網路漏洞挖掘及對策演算法。 透過模擬實驗，證明此演算法的平均漏洞檢出率超過 87%，漏洞修復率超過 86%。 克萊爾·勒古埃等。 [43] 使用 GenProg（一種受生物進化啟發的基於進化演算法的方法）修復了 105 個漏洞中的 55 個。 漏洞的發現是網路攻防的關鍵。 人工智慧可以修補漏洞，是輔助網路防禦的有力工具。

3.2. 智慧型網路和日誌檢測

防禦網殺傷鏈的重要措施是加強邊界防禦。 入侵偵測與防禦系統在日常網路安全防護與偵測中需要處理大量的網路流量和日誌資料。 傳統的檢測方法面對這些海量資料遇到瓶頸，導致檢測系統處理效率低、偵測效果不佳。 相反，基於智慧演算法的偵測方法可以高度捕捉這些海量流量和日誌資料中的內在關係和特徵。 例如，檢測方法可以與基於人工智慧的策略（例如多感知器和元學習）相結合，以檢測不同或複雜的攻擊場景。

AJ Rodrigues[44]開發了一種基於智慧日誌的入侵偵測系統。 本系統使用無監督學習演算法進行訓練，特別是 Kmeans 演算法[45]和 One Class SVM 演算法[46]。 結果證明系統在測試日誌中偵測到了入侵偵測。 異常日誌模式準確率高達 85%。 透過人工智慧人工處理日誌文件，可以盡快發現異常訊息，進行異常處理和警報，為後續攻擊做好準備。

3.3 基於 ChatGPT 輔助網路安全決策

ChatGPT 也是一種可以被保全人員而不是駭客使用的技術。 有幾個典型場景已與 ChatGPT 正式整合[47]：網路釣魚電子郵件偵測、滲透負載辨識、安全文件產生、事件回應報告產生、威脅情報處理等。ChatGPT 在降低成本、複雜性方面表現出了出色的性能，以及防禦者安全行動的及時性。

2.4 基於智慧化的主動安全測試平台

人工智慧驅動的網路安全另一個代表性應用是基於人工智慧的主動安全測試平台。 該平台基於 CDSL-Yak、大型語言模型和知識圖譜。 它還結合了資產、漏洞、業務資訊等幾類核心資料累積。 該平台具備高效、低成本的安全風險篩選能力，旨在主動進行安全測試，發現企業目標環境中隱藏的潛在安全風險。

4。結論

4.1 人工智慧技術發展面臨的挑戰

目前，人工智慧技術的發展仍面臨幾大挑戰：首先是預測的準確性。 人工智慧技術最近還無法達到 100%的偵測準確率，在某些情況下，還會出現一些誤報。 其次是資料完整性。 數據對於人工智慧的運用至關重要，但在某些特定場景下，數據採集是困難的。 此外，缺乏體現關鍵特徵的資料集也是相關研究中的一個問題。 第三是可解釋性。 人工智慧模型，特別是基於利用高維度資料的深度學習的模型，目前無法解決可解釋性和資料模糊性等問題。 第四是計算資源。 隨著 LLM 相關模型和技術的興起，對運算能力的要求越來越高，這已成為前沿人工智慧研究的現實要求。 這對於一些財政困難的地區來說更為重要。 第五是世界輿論。

4.2 網路安全挑戰

總之，與傳統的網路攻擊相比，基於人工智慧的網路攻擊正在徹底改變過去勞動密集、成本高昂的攻擊朝向分散式、智慧化、自動化的方向發展。 這使得攻擊變得越來越精確和自動化。 上述的人工智慧驅動的網路攻擊是網路攻擊方式上的一些新挑戰，可以概括為以下三類挑戰。

首先，利用 AI 學習環境特徵，增強攻擊的適應性和隱藏性。 在攻擊目標的網路環境中，資料、行為等具有一定的局部性特徵。 攻擊者利用 AI 對目標網路中的資料、行為等特徵進行擷取與建模，獲知目標網路環境正常的資料內容、傳輸頻率、傳遞方式等環境特徵；參考環境特徵選擇合適的攻擊手段，將攻擊數據偽裝成具有目標網路正常特徵的正常數據，將攻擊行為偽裝成目標網路中正常用戶的網路行為； 實現環境自適應攻擊行為、資料隱藏，增強攻擊的隱藏性，增強攻擊的適應性。

其次，利用人工智慧增強分散式協作的效果，提高攻擊的穩健性。 攻擊者引入分散式智慧協作演算法，將傳統的由智慧中心統一調度分散式攻擊實體進行協同攻擊演變成無中心分散式多智慧攻擊實體的自主協作和群組決策，從而提高了攻擊者之間的協作效率。多個分散式攻擊節點，減少對集中式協同調度的依賴，降低攻擊對策的風險，增強攻擊的穩健性。

第三，利用人工智慧實現攻擊方法的自我進化，提高攻擊的效能。 攻擊者利用人工智慧分析不同攻擊方式的攻擊效果以及防禦者可能採取的對策，然後針對防禦者的弱點自動選擇新的攻擊機制，從而實現攻擊方式的智慧進化。 例如，攻擊者可以將防禦者入侵偵測系統的結果作為回饋，利用人工智慧技術對回饋資料進行收集和建模，建立攻擊效果模型，動態調整適當的攻擊方法，繞過入侵偵測系統。

4.3 應對 AI 攻擊威脅的建議

我們還可以利用人工智慧來應對人工智慧帶來的全球網路安全挑戰。 一些建議如下：

1、加強研究與應用。 國家和大學都可以推動包括攻擊和防禦在內的智慧網路系統的建設和增強。

2.加強共享利用。 鼓勵研究人員解決網路攻防技術體系建置中的人工智慧資料問題。

3.加強對抗和考核。 我們共同推動人工智慧網路攻防技術在實際應用中的發展，為國家、合作和個人帶來實質的利益。

Литература ***References*** **參考書目**

1. GERSTNER W, NAUD R. How good are neuron models? [J]. Science,2009, 326(5951):379-380

2. WU F, LU C, ZHU M, et al. Towards a new generation of artificial intelligence in China [J]. Nature Machine Intelligence, 2020, 2(6):312-316.

3. BAL R, GILL I S. Policy approaches to artificial intelligence based technologies in China, European Union and the United states [J]. 2020.WIRKUTTIS N, KLEIN H. Artificial intelligence in cybersecurity[J]. Cyber, Intelligence, and Security, 2017, 1(1):103-119.

4. GUEMBE B, AZETA A, MISRA S, et al. The emerging threat of AI- driven cyber attacks: A review[J]. Applied Artificial Intelligence, 2022,36(1):2037254.

5. MIRKOVIC J, REIHER P. A taxonomy of DDoS attack and DDoS defense mechanisms[J]. ACM SIGCOMM Computer Communication Review,2004, 34(2):39-53.

6. WESTERLUND M. The emergence of deepfake technology: A review[J]. Technology innovation management review, 2019, 9(11).

7. WU T, HE S, LIU J, et al. A brief overview of ChatGPT: The history, status quo and potential future development[J]. IEEE/CAA Journal of Automatica Sinica, 2023, 10(5):1122-1136.

8. YAO Y, VISWANATH B, CRYAN J, et al. Automated crowdturfing attacks and defenses in online review systems[C]//Proceedings of the 2017 ACM SIGSAC conference on computer and communications security. [S.l.: s.n.], 2017: 1143-1158.

9. SEYMOUR J, TULLY P. Weaponizing data science for social engineering: Automated e2e spear phishing on twitter[J]. Black Hat USA,2016, 37:1-39.

10. SKOUDIS E, ZELTSER L. Malware: Fighting malicious code[M].[S.l.]: Prentice Hall Professional, 2004.

11. GU X, ZHANG H, ZHANG D, et al. Deep API learning[C]//Proceedings of the 2016 24th ACM SIGSOFT international symposium on foundations of software engineering. [S.l.: s.n.], 2016: 631-642.

12. GROSSE K, PAPERNOT N, MANOHARAN P, et al. Adversarial perturbations against deep neural networks for malware classification[J]. arXiv preprint arXiv:1606.04435, 2016.

13. HU W, TANY. Generating adversarial malware examples for black-box attacks based on GAN[C]//International Conference on Data Mining and Big Data. [S.l.]: Springer, 2022: 409-423.

14. HU W, TAN Y. Black-box attacks against RNN based malware detection algorithms[C]//Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence. [S.l.: s.n.], 2018.

15. KOLOSNJAJI B, DEMONTIS A, BIGGIO B, et al. Adversarial malware binaries: Evading deep learning for malware detection in executables[C]//2018 26th European signal processing conference (EUSIPCO).[S.l.]: IEEE, 2018: 533-537.

16. ANDERSON H S, KHARKAR A, FILAR B, et al. Learning to evade static PE machine learning malware models via reinforcement learning arXiv preprint arXiv:1801.08917, 2018.

17. TADDEO M. Three ethical challenges of applications of artificial intelligence in cybersecurity[J]. Minds and Machines, 2019, 29:187-191.

18. ALRZINI J R S, PENNINGTON D. A review of polymorphic malware detection techniques[J]. International Journal of Advanced Research in Engineering and Technology, 2020, 11(12):1238-1247.

19. KIRAT D, JANG J, STOECKLIN M. Deeplocker–concealing targeted attacks with ai locksmithing[J]. Blackhat USA, 2018, 1:1-29.

20. BAHNSEN A C, TORROLEDO I, CAMACHO L D, et al. Deep- phish: simulating malicious ai[C]//2018 APWG symposium on electronic crime research (eCrime). [S.l.: s.n.], 2018: 1-8.

21. CHUNG K, KALBARCZYK Z T, IYER R K. Availability attacks on computing systems through alteration of environmental control: smart malware approach[C]//Proceedings of the 10th ACM/IEEE inter-national conference on cyber-physical systems. [S.l.: s.n.], 2019: 1-12.

22. YAMIN M M, ULLAH M, ULLAH H, et al. Weaponized ai for cyber attacks[J]. Journal of Information Security and Applications, 2021, 57:102722.

23. GHANEM M C, CHEN T M. Reinforcement learning for efficient net-work penetration testing[J]. Information, 2019, 11(1):6.

24. NHU NX, NGHIATT, QUYEN NH, et al. Leveraging deep reinforcement learning for automating penetration testing in reconnaissance and exploitation phase[C]//2022 RIVF International Conference on Computing and Communication Technologies (RIVF). [S.l.]: IEEE, 2022:41-46.

25. BABAEIZADEHM, FROSIO I, TYREE S, et al. Ga3c: GPU-based a3cfor deep reinforcement learning[J]. CoRR abs/1611.06256, 2016.

26. HITAJ B, GASTI P, ATENIESE G, et al. Passgan: A deep learning approach for password guessing[C]//Applied Cryptography and Network Security: 17th International Conference, ACNS 2019, Bogota, Colombia, June 5–7, 2019, Proceedings 17. [S.l.]: Springer, 2019: 217-237.

27. XIA Z, YIP, LIU Y, et al. Genpass: A multi-source deep learning model for password guessing[J]. IEEE Transactions on Multimedia, 2019, 22(5):1323-1332.

28. HOUSHMAND S, AGGARWAL S, FLOOD R. Next gen pcfg password cracking[J]. IEEE Transactions on Information Forensics and Security, 2015, 10(8):1776-1791.

29. PASQUINI D, GANGWAL A, ATENIESE G, et al. Improving password guessing via

representation learning[C]//2021 IEEE Symposium on Security and Privacy (SP). [S.l.]: IEEE, 2021: 1382-1399.

30. YU F, MARTIN MV. Gnpassgan: improved generative adversarial net- works for trawling offline password guessing[C]//2022 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW). [S.l.]: IEEE, 2022: 10-18.

31. MORI S, SUEN C Y, YAMAMOTO K. Historical review of OCR research and development[J]. Proceedings of the IEEE, 1992, 80(7):1029-1058.

32. XU D, WANG B, DU X, et al. Verification code recognition based on active and deep learning[C]//2019 International Conference on Computing, Networking and Communications (ICNC). [S.l.]: IEEE, 2019: 453-456.

33. TIAN Q, SONG Q, WANG H, et al. Verification code recognition based on convolutional neural network[C]//2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC): volume 4. [S.l.]: IEEE, 2021: 1947-1950.

34. YU N, DARLING K. A low-cost approach to crack python captchas using ai-based chosen-plaintext attack[J]. Applied Sciences, 2019, 9(10):2010.

35. SCHIAVONI S, MAGGI F, CAVALLARO L, et al. Phoenix: DGA-detection of intrusions and malware, and vulnerability assessment. [S.l.]: Springer, 2014: 192-211.

36. YU Y, SI X, HU C, et al. A review of recurrent neural networks: Lstmcells and network architectures[J]. Neural computation, 2019, 31(7):1235-1270.

37. ANDERSON H S, WOODBRIDGE J, FILAR B. Deepdga: Adversarially-tuned domain generation and detection[C]//Proceedings of the 2016 ACM workshop on artificial intelligence and security. [S.l.:s.n.], 2016: 13-21.

38. ZHENG Y, YANG C, YANG Y, et al. ShadowDGA: Toward evading DGA detectors with GANs[C]//2021 International Conference on Computer Communications and Networks (ICCCN). [S.l.]: IEEE, 2021:1-8.

39. JUN Y, CRAIG A, SHAFIKW, et al. Artificial intelligence application in cybersecurity and cyberdefense[J]. Wireless Communications and Mobile Computing, 2021, 2021:1-10.

40. SHAH I A, RAJPER S, ZAMANJHANJHI N. Using ml and data-mining techniques in automatic vulnerability software discovery[J]. International Journal of Advanced Trends in Computer Science and Engineering, 2021, 10(3).

41. CHU Z, HAN Y, ZHAO K. Botnet vulnerability intelligence clustering classification mining and countermeasure algorithm based on machine learning[J]. IEEE Access, 2019, 7:182309-182319.

42. LE GOUES C, DEWEY-VOGT M, FORREST Set al. A systematic study of automated program repair: Fixing 55 out of 105 bugs for 8eac*h*[c]//201234 International conference on software Engineering (IcsE).[s.l.] :IEEE,2012 ： 3 − 13.

43. RODRIGUES A. Automated log analysis using ai: intelligent intrusion de tection system[J]. Computer, 2013, 132:0886.

44. AHMED M, SERAJ R, ISLAM S M S. The k-means algorithm: A comprehensive survey and performance evaluation[J]. Electronics, 2020, 9(8):1295.

45. JOACHIMS T. Making large-scale SVM learning practical[R]. [S.l.]: Techni- cal report, 1998.

46. GUPTA M, AKIRI C, ARYAL K, et al. From ChatGPT to ThreatGPT: Impact of generative ai in cybersecurity and privacy[J]. IEEE Access, 2023.

47. WU Y. A method of character verification code recognition in network based on artificial intelligence technology[J]. International Journal of Information and Communication Technology, 2020, 16(4):325-338.

48. MUSUMECI F, FIDANCI A C, PAOLUCCI F, et al. Machine-learning-enabled DDoS

attacks detection in p4 programmable networks[J]. Journal of Network and Systems Management, 2022, 30:1-27.