AI-Enabled Network Attack Analysis and Classification

Authors: Wang Zhi

Date: 2022-01-26T00:00:37+00:00

Abstract

In recent years, alongside the continuous improvement in computer performance, artificial intelligence (Artificial Intelligence, AI) technology has experienced rapid development. AI technology, represented by deep neural networks, has achieved tremendous breakthroughs in numerous domains including autonomous driving, smart cities, and medical imaging, and has become increasingly pervasive in everyday life. While AI benefits the general public, we also note that AI can assist attackers in executing cyber attack tasks. This paper systematically reviews existing cases of AI-enabled cyber attacks, analyzes the role of AI in attack missions, classifies AI-enabled cyber attacks into online and offline categories, and provides corresponding discussions and outlook.

Full Text

Preamble

AI-Enabled Cyber Attack Analysis and Classification

Zhi WANG, University of Chinese Academy of Sciences

Abstract: In recent years, with the continuous improvement of computer performance, Artificial Intelligence (AI) technology has developed rapidly. AI technology, represented by deep neural networks, has achieved tremendous breakthroughs in multiple domains including autonomous driving, smart cities, and medical imaging, and has become increasingly ubiquitous. While AI benefits the public, we must also recognize that it can empower attackers to execute cyber attack tasks. This paper systematically reviews existing AI-enabled cyber attack cases, analyzes the role of AI in attack tasks, classifies AI-enabled cyber attacks into online and offline categories, and provides corresponding discussion and future outlook.

Keywords: Artificial Intelligence, Cyber Attack, Neural Network, Cyberspace Security, AI Security

1 Introduction

Artificial intelligence, as a strategic technology that shapes the future, has increasingly become a critical engine driving the accelerated transformation from digitalization and networking to intelligence across various economic and social sectors. In recent years, the explosive growth of data, significant enhancement of computing power, and breakthrough applications of deep learning algorithms have greatly propelled the advancement of artificial intelligence. However, technology inherently possesses a dual-use nature. When applied for beneficial purposes, it can promote scientific development and social progress; when applied for malicious purposes, it can cause developmental stagnation and social instability. Artificial intelligence is no exception. When applied in the cybersecurity domain, AI can provide powerful support for intrusion detection, malware defense, situational awareness, and other defensive measures, but it can also assist attackers in executing more efficient and difficult-to-defend cyber attacks.

In 2018, scientists from 26 different research institutions jointly released a report on the malicious use of artificial intelligence, highlighting potential risks to digital security, physical security, and political security, and calling upon AI research teams worldwide to remain vigilant against AI security risks and resist the malicious exploitation of AI. As AI security risks intensify, several prominent security vendors have also forecasted trends in AI application to cyber attacks. BeyondTrust[1] noted in its cybersecurity trend predictions that machine learning training data poisoning and the proliferation of AI weaponization will pose significant challenges to cybersecurity. Check Point[2] observed in its cyber threat trend predictions that AI technology is exhibiting a weaponization trend. Gartner[3] predicted that by 2022, 30% of cyber attacks will be related to artificial intelligence security, with AI becoming increasingly engineered. Fortinet[4] similarly concluded that the evolution of polymorphic malware, cluster attacks, and AI weaponization are becoming trends. Although AI can be applied across multiple fields (such as autonomous driving, armed equipment, smart homes, etc.), this paper focuses on its application in the cybersecurity domain, specifically emphasizing the network layer and above, with a concentration on the application layer.

In recent years, we have witnessed several cases of AI applied to cyber attacks. To better understand the characteristics of AI-enabled cyber attacks and develop more effective targeted defenses, it is essential to systematically review these cases and examine AI’ s role in cyber attacks. This paper therefore reviews relevant cases and analyzes their implications.

2 Attack Case Review

This chapter reviews representative AI-enabled cyber attack cases.

In 2016, Seymour et al.[5] presented at the renowned BlackHat conference a method for highly customized automated spear phishing on Twitter using AI technology. This approach employs clustering to identify high-value targets, combines natural language processing techniques to analyze topics of interest to targets, and constructs SNAP_R (Social Network Automated Phishing with Reconnaissance) to generate spear phishing content based on LSTM, capable of delivering content to targets according to their active hours to lure them into the trap.

Sivakorn et al.[6] introduced a method for automatically bypassing Google reCAPTCHA at BlackHat. By using descriptive words for images, the method segments given images and utilizes online platforms for retrieval. Using NLP techniques to extract information from returned pages and calculate similarity with descriptive words, it can determine whether the selected image is the target. When external networks are unavailable, offline models can be employed for judgment.

Anderson et al.[7] proposed DeepDGA, which combines Auto-Encoder and Long Short-Term Memory networks to learn features from Alexa Top 100M domain names. The encoder and decoder are reassembled in a generative adversarial network (detector + generator) to produce highly realistic DGA domain names capable of bypassing DGA detectors.

In 2017, Baki et al.[8] leveraged emails leaked in Hillary Clinton’ s “email gate” incident, using natural language processing technology to analyze grammatical features and generate (forge) a batch of emails purportedly from Hillary and Palin for others to distinguish. Results showed that most people believed the generated emails originated from Hillary and Palin, with emails from “Hillary” receiving more votes due to their colloquial language. This technology can be applied to spear phishing to generate highly realistic phishing emails.

Hu et al.[9] proposed using generative adversarial network algorithms to produce adversarial malware samples. The generative model adds perturbations to malicious samples, while a substitute detector determines whether samples are malicious. Trained on benign samples and adversarial samples generated by the generative model, the substitute detector simulates a black-box detector. Through adversarial training, the probability of generated adversarial samples being detected can be minimized.

In 2018, Kirat et al.[10] introduced DeepLocker at BlackHat for targeted covert attacks. DeepLocker trains a neural network model based on target information and uses the model’ s output as a key for symmetric cryptographic algorithms to encrypt the malicious payload. When the target is discovered, the neural network model can output the correct decryption key to unlock and execute the malicious payload. When the target is absent, due to the one-way nature and collision resistance of the neural network model, analysts cannot construct trigger conditions to obtain the key, thereby resisting analysis.

Bahnsen et al.[11] presented the automated phishing tool DeepPhish at BlackHat, which uses LSTM to learn URL features and constructs malicious TLS certificates to build phishing pages. Compared with conventional methods, DeepPhish can increase phishing success rates by 20% to 30%.

Takaesu[12] proposed DeepExploit, an automated penetration testing attack that uses reinforcement learning to scan and probe given targets and employs Metasploit for automated penetration testing based on probe results, generating test reports upon completion.

Anderson et al.[13] proposed a method to evade detection by modifying static PE malware code structure features using reinforcement learning incentive and feedback mechanisms. Using a black-box detector as feedback for sample modification results, it obtains samples capable of evading detection by adding useless function calls, modifying section names, removing signature information, and other techniques.

Ye et al.[14] introduced a novel text-based CAPTCHA cracking method. Unlike existing crackers that require large amounts of manually labeled real CAPTCHAs for learning, this approach uses a generative adversarial network to train CAPTCHA models from existing databases. When encountering new CAPTCHAs, it employs transfer learning methods, requiring only a small amount of new CAPTCHA data to achieve excellent cracking performance.

Rigaki et al.[15] proposed using generative adversarial networks to simulate Facebook chat traffic to evade detection based on traffic analysis. This method uses Stratosphere Linux IPS as a detector to identify Facebook chat traffic, employs RNN and LSTM to build generators and discriminators, and only releases traffic when the detector considers the generated traffic to be Facebook chat traffic. The trained parameters are then used for traffic generation in malware activities to bypass traffic-based detectors.

In 2020, Li et al.[16] proposed generating adversarial feature vectors in feature space and converting them into adversarial malicious PDF components, modifying features between the CRT and trailer of PDF files to generate malicious PDFs that evade PDF malware detectors. Evaluated on the PDFRate classifier, this method can attack target system knowledge under four evasion scenarios.

Novo et al.[17] introduced a flow-based C2 traffic detector and a proxy-based C2 traffic evasion detection method. Using the Fast Gradient Sign Method (FGSM) in a white-box approach, it breaks statistical features of C2 data flows while ensuring data communication functionality to counter malicious C2 traffic detectors.

Wang et al.[18] proposed DeepC2, a neural network-based C2 addressing method. Leveraging the irreversibility of neural network models, it uses neural network models to identify command and control account avatars for addressing, while employing data augmentation techniques to solve abnormal content issues in previous social network platform-based C2.

Yuan et al.[19] introduced an end-to-end byte-level black-box adversarial attack method. A generator produces a Payload appended to samples, and a discriminator is trained through feedback from the black-box detector. During use, only the generator needs to produce the Payload to counter the detector. This method can achieve 100% evasion success rate when the added Payload length is 2.5%.

XunSu et al.[20] proposed a method for batch automated detection of WAF rules using natural language processing and LSTM. Through data, algorithms, and detection, it automatically extracts text features of Payloads, employs attention mechanisms to enhance keyword learning, and uses empirical automated methods to determine different keywords for different positions. It can provide online feedback on detection boundaries and achieve WAF bypass.

In 2021, Wang et al.[21] introduced EvilModel, a method for embedding malicious code in neural network models for attack payload delivery. By analyzing neural network model structures, it replaces neurons and parameters with malicious code, enabling large-volume malicious code to be delivered to target devices without detection while maintaining model performance.

3 Attack Classification

The previous chapter reviewed representative AI-enabled cyber attack cases. This chapter analyzes the role of AI in these attacks.

Based on different attack scenarios, these cases can be categorized as follows: - Automated generation: Typical cases include Twitter and email phishing[5][8], malicious sample generation[9], and malicious traffic generation[15][17]. - Automated attacks: Typical cases include DeepExploit[12]. - Deception: Typical cases include various phishing attacks such as Twitter phishing[5], email phishing[8], and DeepPhish[11]. - Detection evasion: Typical cases include evasion of malicious domain name detectors[7], malware detectors[13][19][21], and malicious traffic detectors[15]. - Cracking: Typical cases include cracking human-computer interaction Turing tests[6][14] and firewall rules[20]. - Target identification: Typical cases include identifying attack targets[10] and identifying attackers[18].

In different scenarios, AI plays distinct roles. Minsky et al.[22] summarized AI' s role in cyber attacks based on AI capabilities as follows: - Prediction: Making predictions based on existing data. - Generation: Generating new content based on target conditions. - Analysis: Mining effective information from existing data. - Retrieval: Searching for target information from existing data. - Decision-making: Providing guidance for next actions based on existing information.

Based on different uses of AI, AI-enabled cyber attacks can be divided into “online” and “offline” categories: - “Online” refers to AI models being directly used in the attack process, with their outputs directly applied to ongoing attack tasks. - “Offline” refers to AI models operating behind the attack activity, with their outputs that may or may not be used in attack tasks.

Specifically, “online” focuses on the intrinsic characteristics of AI models and methods themselves, leveraging features that differentiate AI from other approaches and applying them to specific tasks in cyber attacks. “Offline” focuses on what AI can accomplish and which stages of cyber attacks it can be applied to, with its primary task being to automate and intelligently perform what humans can do, faster and better than manual methods. In “online” mode, the AI model resides in the target environment, while in “offline” mode, the AI model resides in the attacker’ s environment.

According to these definitions and divisions, most existing cases fall into the “offline” category: - “Online” mode typical cases: DeepLocker, DeepC2, EvilModel, swarm intelligence, etc. - “Offline” mode typical cases: Malicious content generation, deception, cracking, automated attack categories, etc.

Both “online”and “offline”modes constitute important components of AI-enabled cyber attacks, as shown in Table 1.

Table 1: Classification of AI-Enabled Cyber Attack Patterns

Mode	Definition	AI Role	Typical Cases
Online	AI model is directly used in the attack process, with its output results directly applied to ongoing attack tasks.	Generation, Retrieval, Prediction, Analysis, Decision	DeepLocker, DeepC2, EvilModel, Swarm Intelligence, etc.
Offline	AI model is behind the attack activity, with its output results that may or may not be used in attack tasks.	Generation, Retrieval, Prediction, Analysis, Decision	Malicious content generation, deception, cracking, automated attack categories, etc.

4 Discussion and Outlook

Strictly speaking, AI-enabled cyber attacks do not belong to the “artificial intelligence security” category but rather the “artificial intelligence + security” category, representing the application of AI in the cybersecurity domain (particularly in cyber attacks). Pure “artificial intelligence security” refers to the security issues of AI itself, encompassing AI model security, AI data security, AI framework security, etc. Tencent’ s related team[23] has summarized existing work, as shown in Table 2. Additionally, AI applied to cybersecurity detection and protection also falls under AI application and belongs to “artificial intelligence + security.”

Combining the attack and defense sub-directions yields a more comprehensive classification of AI security, as illustrated in the figure.

Discussing AI' s own security issues leads to another category of attacks: attacks against intelligent systems equipped with AI devices, targeting AI vulnerabilities. For instance, detection evasion cases in existing attacks target malicious object detectors deployed with AI-based methods. Attackers can also conduct model stealing, data poisoning, adversarial sample attacks, etc., against intelligent systems to disrupt normal operations. Such attacks are currently evolving. It is foreseeable that relevant achievements in this area will continue to emerge in the coming years.

Figure: Artificial Intelligence Security Architecture

In practical application, due to the substantial resources required for AI model operation (computing power, data, models, etc.), AI-enabled cyber attacks (particularly "online"mode attacks) cannot yet be deployed in real-world cyber attack tasks. With further advancement and popularization of AI, future computing devices may integrate these resources, enabling attackers to implement such attacks. Therefore, researchers in related fields must remain vigilant, proactively develop targeted defenses against such attacks, further protect and enhance the security of various information systems, and achieve effective defense when such attacks first emerge.

Table 2: AI' s Own Security Risks[23]

Stage	Security Risks
Environment Contact	Software dependency attacks, Hardware model backdoors
Data Collection	Data poisoning, Gradient recovery of data in models
Model Training	Training backdoor attacks, Initial weight modification
Model Deployment	Digital adversarial attacks, Physical adversarial attacks, Query-based architecture attacks, Side-channel architecture attacks, Model file attacks, Supply chain attacks
Decentralized Deployment	GPU/CPU overflow, Docker malicious images

5 Conclusion

This paper systematically reviews and classifies AI-enabled cyber attacks. It begins by analyzing the cybersecurity landscape, selecting representative AI-enabled cyber attack cases from recent years, and providing brief introductions to related work. It then studies and analyzes AI' s role in these cases, dividing AI-enabled cyber attacks into "online" and "offline" modes based on different uses of AI in attack tasks, and introduces both modes. Finally, it discusses and prospects the combined application of artificial intelligence and cybersecurity. Cyber attack and defense are interdependent technologies that promote each other. We have reason to believe that future information systems will integrate protection against AI-enabled cyber attacks, and future defense methods will also leverage intelligent capabilities to become more powerful and effectively safeguard cyberspace security.

During the completion of this paper, ArkTeam and Shancheng Security provided valuable assistance, for which we express our gratitude.

References

1. BEYONDTRUST. Beyondtrust releases cybersecurity predictions for 2021 and beyond[EB/OL]. https://www.globenewswire.com/news-release/2020/10/28/2115996/0/en/BeyondTrust-Releases-Cybersecurity-Predictions-for-2021-and-Beyond.html.

2. CHECKPOINT. Check Point Software´s predictions for 2021: Securing the 'next normal'[EB/OL]. https://blog.checkpoint.com/2020/11/10/check-point-softwares-predictions-for-2021-securing-the-next-normal/.

3. CEARLEY D, JONES N, SMITH D, et al. Top 10 Strategic Technology Trends for 2020[EB/OL]. 2019. https://emtemp.gcom.cloud/ngw/globalassets/en/doc/documents/432920 top-10-strategic-technology-trends-for-2020.pdf.

4. FORTIGUARD. New cybersecurity threat predictions for 2021[EB/OL]. 2020. https://www.fortinet.com/blog/threat-research/new-cybersecurity-threat-predictions-for-2021.

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

6. SIVAKORN S, POLAKIS J, KEROMYTIS A D. I' m not a human: Breaking the Google reCAPTCHA[J]. Black Hat, 2016:1-12.

7. 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, AISec@CCS 2016, Vienna, Austria, October 28, 2016. ACM, 2016: 13-21.

8. BAKI S, VERMA R M, MUKHERJEE A, et al. Scaling and effectiveness of email masquerade attacks: Exploiting natural language generation[C]//Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, AsiaCCS 2017, Abu Dhabi, United Arab Emirates, April 2-6, 2017. ACM, 2017: 469-482.

9. HU W, TAN Y. Generating adversarial malware examples for black-box attacks based on GAN[J/OL]. CoRR, 2017, abs/1702.05983. http://arxiv.org/abs/1702.05983.

10. KIRAT D, JANG J, STOECKLIN M. Deeplocker–concealing targeted attacks with ai locksmithing[J]. Blackhat USA, 2018.

11. BAHNSEN A C, TORROLEDO I, CAMACHO L D, et al. Deepphish: Simulating malicious ai[C]//2018 APWG Symposium on Electronic Crime Research (eCrime). 2018: 1-8.

12. TAKAESU I. Deep exploit: Fully automatic penetration test tool using machine learning[J]. Blackhat EUROPE, 2018.

13. ANDERSON H S, KHARKAR A, FILAR B, et al. Learning to evade static PE machine learning malware models via reinforcement learning[J/OL]. CoRR, 2018, abs/1801.08917. http://arxiv.org/abs/1801.08917.

14. YE G, TANG Z, FANG D, et al. Yet another text captcha solver: A generative adversarial network based approach[C]//Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, CCS 2018, Toronto, ON, Canada, October 15-19, 2018. ACM, 2018: 332-348.

15. RIGAKI M, GARCIA S. Bringing a GAN to a knife-fight: Adapting malware communication to avoid detection[C]//2018 IEEE Security and Privacy Workshops, SP Workshops 2018, San Francisco, CA, USA, May 24, 2018. IEEE Computer Society, 2018: 70-75.

16. LI Y, WANG Y, WANG Y, et al. A feature-vector generative adversarial network for evading pdf malware classifiers[J]. Information Sciences, 2020, 523: 38 - 48.

17. NOVO C, MORLA R. Flow-based detection and proxy-based evasion of encrypted malware c2 traffic[C]//AISec' 20: Proceedings of the 13th ACM Workshop on Artificial Intelligence and Security. New York, NY, USA: Association for Computing Machinery, 2020: 83–91.

18. WANG Z, LIU C, CUI X, et al. DeepC2: AI-powered convert botnet command and control on OSNs[J/OL]. CoRR, 2020, abs/2009.07707. http://arxiv.org/abs/2009.07707.

19. YUAN J, ZHOU S, LIN L, et al. Black-box adversarial attacks against deep learning based malware binaries detection with GAN[C]//Frontiers in Artificial Intelligence and Applications: volume 325 ECAI 2020 - 24th European Conference on Artificial Intelligence, Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020). IOS Press, 2020:

20. XUNSU, KEYUNLUO. Deep X-Ray: 一种机器学习驱动的 WAF 规则窃取器 [EB/OL]. 2020. http://t.cn/A65ZGOyL.

21. Wang Z, Chaoge Liu, and Xiang Cui. EvilModel: Hiding Malware Inside of Neural Network Models[C]. In 2021 IEEE Symposium on Computers and Communications (ISCC). IEEE, 2021:

22. MIRSKY Y, DEMONTIS A, KOTAK J, et al. The Threat of Offensive AI to Organizations[J/OL]. CoRR, 2021, abs/2106.15764.

http://arxiv.org/abs/2106.15764. https://matrix.tencent.com/