Artificial Intelligence Technologies and Applications
C. Chen (Ed.)

doi:10.3233/FAIA231291

An Overview of Artificial Intelligence Security Issues

Zhihui WANG[a, 1], Yongqiang DONG[a], Zhining XIANG[a], Shaochi CHENG[b]
[a] *Army Academy of Artillery and Air Defense (Nanjing Campus), Nanjing 211132, China*
[b] *Institute of War, Academy of Military Science, Beijing 100091, China*

Abstract. Artificial intelligence (AI) technology promotes human civilization, while it also raises concerns, mainly related to security. AI plays a double-edged sword role in the network space closely related to human life. On the one hand, AI assists in network protection by intelligently detecting network intrusion, reducing missed alarm rates and false alarm rates, forecasting network threats, automatically searching for malicious code, and supporting network defense. On the other hand, AI can also assist in network attacks, such as attacking voice recognition systems and malicious software detection. AI algorithms and computational data pose many security threats, affecting the security of face recognition, speech detection, malware detection, automated driving, and other security applications based on AI technology.

Keywords. artificial intelligence (AI); network protection; network attack; self-vulnerability

1. Introduction

With the rapid development of technology and its application, cyberspace has penetrated various aspects of the national economy and social life, directly affecting social stability, economic development, political culture, and national security. In the book "Unrestricted Warfare," the following scenario is described: massive-scale network attacks that cause key infrastructure such as electricity, banks, and gas to collapse, leading to political turmoil and government regime change [1].

In recent years, emerging technologies such as cloud computing, big data, and the Internet of Things (IoT) have been applied and developed, improving social order and people's quality of life. These technologies have helped us solve network security problems. For example, cloud computing technology can help us execute cloud-based malicious code checks, big data technology can help us analyze malicious attack behavior, and IoT technology can help us improve physical-level security. These technologies enhance our security defense, but they can also be used as attacking forces. For example, in cloud-based DDoS attacks, malicious attackers can hide their intentions using big data, and using IoT technology for intelligence gathering. If these emerging technologies are used improperly, they can seriously damage our social order and cause enormous economic losses. In addition, the security of these technologies themselves is

[1] Corresponding author: Zhihui WANG, Army Academy of Artillery and Air Defense (Nanjing Campus), Nanjing 211132, China; email: zhwang66@foxmail.com

also something we need to consider, such as cloud security, data security, and security against radiation during the communication process of the IoT.

Artificial intelligence technology is closely related to these emerging technologies. It is a technical science that researches, develops, and applies theories and applications for simulating, extending, and expanding human intelligence. Artificial intelligence attempts to understand the essence of intelligence and produce an intelligent machine that can respond in a way similar to human intelligence. Research in this field includes robotics, language recognition, image recognition, natural language processing, and expert systems [2]. In recent years, artificial intelligence technology has exploded in development and has been widely applied in fields such as image processing, speech recognition, and network security. In terms of cyberspace, artificial intelligence can help network protection by intelligently detecting network attacks based on machine learning, predicting network threats, and improving intrusion detection efficiency while decreasing the rate of missed attacks and false alarms. This technology can also spontaneously search for malicious code and support network defense. On the other hand, artificial intelligence can also be used to assist network attacks, such as attacking speech recognition and malicious software recognition.

In addition, artificial intelligence is fragile. There are many security threats to its algorithms and training data. Humans can use vulnerabilities to counter AI. Therefore, there are certain security risks involved in AI-based technologies such as facial recognition, voice detection, malware scanning, and autonomous driving.

2. Artificial Intelligence Helps Network Protection

The continuously developing artificial intelligence is undoubtedly an important direction that deserves attention concerning network security in cyberspace. There are currently a variety of methods for network defense, but overall, the methods are relatively mechanical and the level of intelligence is not high enough. Artificial intelligence can provide deeper analysis and mining of data, as well as more comprehensive processing and control of processes, so it has good application prospects in the field of network defense. There are several typical applications for artificial intelligence in network defense, and the following is a brief description of how artificial intelligence helps network protection.

2.1. Artificial Intelligence Helps with Network Intrusion Detection

In a network or system, any unauthorized or unapproved activity is called an intrusion. The computer security threat monitor published by American Professor James analyzed network threats and proposed intrusion detection for the first time. The composition of the intrusion detection system is shown in Figure 1. Intrusion detection plays a vital role in network defense, and network intrusion detection systems help to detect and identify unauthorized usage, copying, modification, and destruction [3]. Network intrusion detection methods are divided into anomaly detection and misuse detection. Anomaly detection first constructs a normal model, and any access that does not comply with this model is defined as an intrusion. In [4], the use of support vector machine methods to improve the performance of intrusion detection systems is discussed, while in [5], the use of decision tree methods to improve intrusion detection is discussed. In [6], some

achievements made in improving the performance of intrusion detection systems using neural networks are mentioned and discussed.

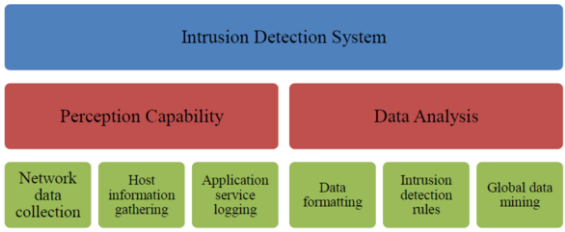

Figure 1. Intrusion Detection System

In [7], the author uses a neural network-based intrusion detection system that is based on an unsupervised neural network. The purpose is to perform intelligent real-time intrusion detection on the network. The entire system framework is shown in Figure 2. The first part of the system captures and preprocesses real-time network traffic data, extracts data features, and converts them into a binary or standardized form. The converted data is then sent to the neural network-based detection system, which uses adaptive resonance theory (ART), self-organizing maps (SOM), and neural networks. Finally, the output results are written to a log, and an alarm is issued if an abnormality is detected.

Aljurayban and Emam used a layered anomaly detection framework to effectively protect cloud network environments, which creates a data mining knowledge base for detection using an artificial neural network. Effective detection is achieved with less flow analysis resulting in increased throughput. The layered anomaly detection framework can handle large amounts of data streams and maintain the effective operation of cloud networks and services even in more demanding environments. Barollid et al. investigated the use of neural networks as an anomaly detection system solution for intrusion detection systems on the Tor network, using Tor servers and clients and a backpropagation neural network to simulate transactions and perform data acquisition on the Tor network. This system proposes training the neural network using data captured by Wireshark and comparing server and client data, with any differences being considered as an intrusion. The results of the tests were evaluated to be accurate in a testing environment.

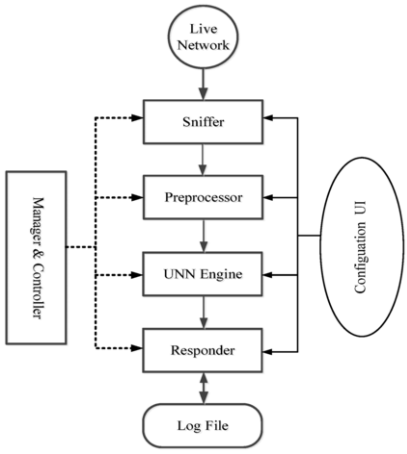

Figure 2. System Framework

2.2. Artificial Intelligence Supports Network Defense

In the field of cybersecurity, artificial intelligence is used to support network defense through the use of commonly used machine learning algorithms such as support vector machine (SVM), artificial neural network (ANN), decision tree classifiers, Bayesian classifiers, and K-means. For example, in identifying the command and control (CC) servers of a botnet using supervised learning with existing botnet network samples, scholars use unsupervised learning to cluster controlled machines with similar behavior in the botnet [8].

The traditional machine learning approach involves finding feature vectors (i.e., a set of features) based on the problem to be solved and using them to extract training data and build a learning model [9]. For example, in the process of identifying a botnet, the Netflow's timestamp, packet count, packet size, and protocol are used to construct a model. However, the quality of the features not only affects the quality of the learning but also affects the efficiency of the learning. For example, selecting a set of semantically overlapping feature vectors would lead to a high computation dimension with no increase in actual performance. To solve this problem, one approach is to use principal component analysis (PCA) to reduce dimensionality, while another approach is to use deep learning in artificial intelligence [10].

In addition, in [11] and [12], agent-based technology is used to deal with DDoS attacks, and simulation results show that a cooperative group of agents can effectively resist DDoS attacks.

3. Artificial Intelligence Assists in Network Attack

3.1. Attack Speech Recognition

In [13], the use of dolphin attacks is described to compromise speech recognition systems. This new type of hacker attack can activate a smartphone's voice assistant function using ultrasonic waves that humans cannot perceive, enabling the attacker to perform a range of sensitive operations without the victim's notice, including online shopping, making phone calls, viewing documents, and more. In the experiment, human speech was loaded as an ultrasonic wave that is inaudible to human ears and used to simulate an attack on an iPhone. Through extensive testing, researchers also discovered that "dolphin attacks" can affect the voice control systems of smart cars and smart homes, allowing attackers to remotely control the car or home by opening specific web pages and installing specific software, such as opening the car's sunroof or operating the car's navigation system.

Another attack on speech recognition is that German scholars attack speech recognition through the hidden effect in psychoacoustics. There is a weakness in human hearing, which is called the masking effect. When two kinds of sounds are introduced into the ear at the same time, human beings will be more sensitive to the louder sound wave and not sensitive to the weak one.

3.2. Attack Malicious Software Recognition

Most attacks attempt to evade classifier detection by injecting malicious data during the training process or by carefully crafting malicious data. For example, in the earliest proposed attack, the attacker injected malicious data into a spam detection system based

on the naive Bayes algorithm. Another attack method is to evade the detection of malicious programs in PDF files by applying linear kernel support vector machines.

In malicious software recognition, clustering algorithms are a typical unsupervised learning algorithm that can discover hidden patterns in data distribution. They are widely used in security areas such as malicious domain name system (DNS) detection, malicious program detection, and collecting information on sources of network attacks. The main attack method against clustering algorithms is to inject malicious data during the training process to affect the clustering results. In [14][15][16][17], the threat of malicious data injection against clustering algorithms is introduced. In addition, there are obfuscation attacks against clustering algorithms [14], in which the attacker's goal is to hide the adversarial sample by confusing the content of the adversarial sample with the other categories without changing the clustering results of other samples.

3.3. Attacking Internet of Things (IoT) intelligent systems

At the 2018 DefCon conference, a Tencent team successfully demonstrated the ability to hack into Amazon's second-generation Echo smart speaker, in only 26 seconds. The world's top-selling and supposedly most secure AI speaker was compromised and turned into a listening device.

This is a typical attack against intelligent systems. The Tencent team first removed the flash chip from the Echo and read the firmware content. They then modified the firmware and re-soldered the flash chip to gain root access, and the ability to connect the Echo device to a debugger. After further analysis of the network, they discovered a system program that opened ports allowing multiple Echo devices to connect and had root privileges. Through this system, they were able to control other Echo devices. Therefore, in a scenario where multiple intelligent devices are connected to the same WiFi, a breach in one device means that all other intelligent devices are no longer secure.

3.4. Cyber Weapon "Digital Ordnance"

The world's top security conference, NDSS, proposed the concept of "digital ordnance" in 2011 [18]. Digital ordnance has been regarded as the "ultimate weapon of the internet" by the industry. Essentially, it is a type of DDoS attack that targets not just a single server, but an entire network, area, city, or even country.

By simulating 250,000 zombie nodes and launching the digital ordnance for 20 minutes, more than 100 minutes of router processing delay can be caused. The first step is to construct a zombie network in advance, and the second step is to identify "critical pathways". Then, in the third step, the attackers launch a ZMW attack on the BGP protocols of the routers at both ends of the critical pathways, continuously disconnecting the BGP sessions between routers. Since there is a cascading effect in cyberspace, the second step is crucial. Finding and attacking the hub can potentially paralyze the entire network. How to identify critical pathways and how to use machine learning to train, is worth exploring.

4. Artificial Intelligence Self-vulnerability

4.1. Poisoning Attack

Using samples to train AI models is an important process in artificial intelligence

technology. The quality of the training process directly affects the effectiveness of the classification and prediction model used in practice. Therefore, the importance of training data to machine learning models is self-evident. It is for this reason that many attackers focus their attacks on training data, with the most common form of attack being poisoning attacks.

In [19], ATT_FLAV is introduced, which is a framework that enhances the robustness of federated learning-based autonomous driving models against poisoning attacks by using a bandit-based AttackRegion-UCB algorithm to dynamically choose the target attack label region in each round of training. In [20], scholars introduce a new type of data poisoning attack designed to preserve personal data privacy, that can also be used as a powerful clean-label backdoor attack. The attack operates by adding unnoticeable perturbations to clean data, to confuse DNNs into making incorrect classifications. Experimental results show that the proposed attack outperforms previous methods and suggests a new perspective on the role of DNNs as feature extractors.

A poisoning attack is an induced attack that mainly manipulates training data by injecting carefully crafted malicious data samples (with incorrect labels and attack properties), disrupting the probability distribution of the original training data, and reducing the classification or clustering accuracy of the trained model, to achieve the purpose of destroying the model's availability and integrity. Since the original training data used in AI algorithms are usually confidential and difficult for attackers to modify, systems using AI algorithms often need to be regularly retrained to update the models, which provides an opportunity for attackers. For example, the adaptive biometric face recognition system based on principal component analysis (PCA), malicious software classification system, and spam email detection system all require periodic retraining. An attack on an adaptive biometric face recognition system based on principal component analysis (PCA) is used as an example, where the attacker takes advantage of the system's need to update periodically. During the retraining period, the attacker injects fake adversarial samples into the training data specifically targeted at the victim, causing the centroid of the original model used for identifying the victim's features to gradually move toward the attacker's centroid. This enables the attacker to replace the victim's image with their image for authentication purposes.

4.2. Impersonate Attack

We created a pair of special glasses that can turn you into anyone when viewed by a facial recognition camera. As shown in Figure 3, a man wearing the glasses was recognized as the actress Jovovich, while his female colleague successfully portrayed a Middle Eastern man. Researchers claim that these glasses can deceive facial recognition systems based on neural network learning, and are successful in outsmarting the Face++ image recognition system at a rate of over 90%. Wearing these glasses would allow criminals to easily evade public surveillance systems or enter a company under someone else's identity. The researchers suggest that, in future security checks, ordinary items carried by passengers should also be checked, as these seemingly innocuous items have the potential to deceive artificial intelligence. This is called impersonate attack.

Figure 3. Special glasses deceiving Face++ image recognition

Impersonate attack is a type of deception attack that is similar to an evasion attack. It mainly focuses on the imitation of victim samples. It mainly occurs in machine learning-based face recognition and speech recognition systems. Attackers generate specific adversarial examples that cause the machine learning model to misclassify samples with large differences from human samples as the samples that attackers want to imitate, thus achieving the goal of obtaining victim permissions (real systems based on face recognition and voice control). Currently, this type of attack mainly occurs in DNN algorithms because DNN often recognizes goals by extracting very few features from samples. Therefore, attackers can easily achieve impersonation attacks by modifying key features.

There are many typical imitation attack examples in image imitation attacks. Nguyen et al. introduced an improved genetic algorithm, MAP-Elites, to generate multiple class images with the best evolved adversarial examples and used these adversarial examples to impersonate Google's Alex Net and Caffe-based Le-Net-5 networks, thus deceiving DNN to misclassify. For physical world imitation attacks, Kurakin et al. first generated electronic adversarial examples using the method of minimum similar classes, and then printed them out and used a mobile phone camera to take pictures to fool the real image classification system-Geek Pwn 2016. The process lost many small pixel features of the original samples in the electronic world; therefore, the success rate of physical world adversarial attacks was much lower than that of electronic world adversarial attacks, but it demonstrated the possibility of implementing real physical world imitation attacks. Sharif even proposed a method of avoiding detection by the most advanced FRS by wearing a specific pair of glasses on the attacker's face in real life and even imitating other victims in this way. This attack method has been experimentally verified to be physically feasible and challenging to detect, posing a significant security threat to FRS. In addition, the latest research shows that using algorithms that integrate multiple networks to produce transferable adversarial examples (adversarial examples generated for one DNN can threaten other DNNs). This method can produce non-targeted adversarial examples that do not migrate with the target label and targeted adversarial examples that migrate with the target label. Experiments were conducted on the large-scale dataset ILSVRC 2012 and the state-of-the-art commercial network classification system to achieve effective attacks on large-scale datasets. In addition, there are also imitation attacks on sound information.

4.3 Using the vulnerability of sensors supporting intelligent systems

Using the vulnerability of sensors to support intelligent systems, we can cheat the sensor to cheat the intelligent system. For example, in May 2016, Tesla crashed into a white box truck during autonomous driving, and the car was destroyed. Millimeter-wave radar detected a huge obstacle ahead, but because the truck's reflection area was too large and

the body was too high, millimeter-wave radar misjudged it as a traffic sign. The focal length of the camera is very narrow, and the trailer of the accident is white. In an environment with strong sunshine, the image recognition system misjudges it as white clouds.

5. Conclusion

With the development of new technologies, the issue of artificial intelligence security and its defense mechanisms has drawn great attention from both academia and industry. Artificial intelligence plays a double-edged role in the cybersecurity field, as it can assist with protection but also facilitate attacks, and its power cannot be underestimated. Currently, research on artificial intelligence technology is in an unprecedentedly hot stage, with a large number of learning frameworks, algorithms, and optimization mechanisms being proposed. However, few of these models and algorithms take their security into account, so they are inherently vulnerable. By exploiting these vulnerabilities, we can counterattack against artificial intelligence. In the future, new security threats to artificial intelligence will undoubtedly attract more attention.

References

[1] Dong, N. (2009). Cyber Warfare. 9th Zone Press. (In Chinese)
[2] Accenture. (2016). Artificial Intelligence. Shanghai Jiao Tong University Press. (in Chinese)
[3] Kabir, M. F., & Hartmann, S. (2018). Cyber Security Challenges: An Efficient Intrusion Detection System Design. In International Young Engineers Forum. https://doi.org/10.1109/IYEF.2018.8571707
[4] Lemm, S., Blankertz, B., Dickhaus, T., et al. (2011). Introduction to machine learning for brain imaging. Neuroimage, 56(2), 387-399. https://doi.org/10.1016/j.neuroimage.2010.11.004
[5] Hu, W., Liao, Y., & Vemuri, V. R. (2007). Robust Anomaly Detection Using Support Vector Machines. In International Conference on Machine Learning. https://doi.org/10.1145/1273496.1273527
[6] Li, J., Manikopoulos, C. N., Jorgenson, J., et al. (2001). HIDE: a Hierarchical Network Intrusion Detection System Using Statistical Preprocessing and Neural Network Classification. Proc IEEE Workshop on Information Assurance & Security, 85-90. https://doi.org/10.1109/IASW.2001.924949
[7] Amini, M., Jalili, R., & Shahriari, H. R. (2006). RT-UNNID: A practical solution to real-time network-based intrusion detection using unsupervised neural networks. Computers & Security, 25(6), 459-468. https://doi.org/10.1016/j.cose.2006.06.006
[8] Dilek, S., Çakır, H., & Aydın, M. (2015). Applications of Artificial Intelligence Techniques to Combating Cyber Crimes: A Review. International Journal of Artificial Intelligence & Applications, 6(1). https://doi.org/10.5121/ijaia.2015.6101
[9] Anitha, A., Paul, G., & Kumari, S. (2016). Cyber defense using artificial intelligence. International Journal of Pharmacy & Technology, 8(4), 25352-25357. https://doi.org/10.5958/0975-766X.2016.00139.1
[10] Wu, Y. L., Si, G. Y., & Luo, P. (2015). The Application of Artificial Intelligence Technologies in Cybersecurity Defense. Research on Computer Applications, 32(8), 2241-2244. https://doi.org/10.3778/j.issn.1002-8331.1508-0199
[11] Kotenko, I., & Ulanov, A. (2007). Multi-agent Framework for Simulation of Adaptive Cooperative Defense Against Internet Attacks. In International Conference on Autonomous Intelligent Systems: Agents and Data Mining (pp. 212-228). Springer-Verlag. https://doi.org/10.1007/978-3-540-74305-0_20
[12] Kotenko, I., Konovalov, A., & Shorov, A. (2010). AGENT-BASED MODELING AND SIMULATION OF BOTNETS AND BOTNET DEFENSE. (pp. 25-36). https://doi.org/10.1109/CCC.2010.58
[13] Zhang, G., Yan, C., Ji, X., et al. (2017). DolphinAttack: Inaudible Voice Commands. In ACM Sigsac Conference on Computer and Communications Security. ACM. https://doi.org/10.1145/3133956.3133994
[14] Biggio, B., Pillai, I., Ariu, D., & Giacinto, G. (2013). Is data clustering in adversarial settings secure? In Proceedings of the 2013 ACM workshop on Artificial intelligence and security (pp. 87-98). ACM.

[15] Biggio, B., Bulò, S. R., Pillai, I., & Pelillo, M. (2014). Poisoning Complete-Linkage Hierarchical Clustering. Lecture Notes in Computer Science, 8621, 42-52. doi: 10.1007/978-3-319-09891-3_4
[16] Biggio, B., Rieck, K., Ariu, D., Wressnegger, C., Corona, I., & Giacinto, G. (2014). Poisoning Behavioral Malware Clustering. In Proceedings of the 2014 workshop on Artificial intelligent and security (pp. 27-36). ACM.
[17] Zhao, W., Long, J., Yin, J., Wang, Y., & Sun, Y. (2012). Sampling Attack against Active Learning in Adversarial Environment. In M. Bichler, J. R. F. Bussjaeger, & E. P. Klement (Eds.), International Conference on Modeling Decisions for Artificial Intelligence (pp. 222-233). Springer-Verlag.
[18] Scientist, N. (2011). The Cyberweapon That Could Take Down the Internet. Nature. Retrieved from https://www.nature.com/news/2011/110126/full/469141a.html
[19] Wang, S., Li Q., Cui, Z., et al. (2023). Bandit-based data poisoning attack against federated learning for autonomous driving models. Expert Systems with Applications. doi: 10.1016/j.eswa.2023.120295.
[20] Zhang, C., Tang, Z., Li, K.. (2023). Clean-label poisoning attack with perturbation causing dominant features. Information Sciences. https://doi.org/10.1016/j.ins.2023.03.124.