Integrating Security in AI Hacking

Subin Babu[1], Nisha Mohan P.M[2]

[1]M Tech Student, APJ Abdul Kalam Technological University, Kerala, India

[2]Asst. Professor, Mount Zion College of Engineering, Kadammanitta, Kerala, India

---***---

Abstract - *The word "hacker" comes from MIT's early computer culture, which became the nucleus of the MIT Artificial Intelligence Laboratory. Originally a playful term, "hacker" now most often refers to the criminals causing hundreds of billions of dollars in damages through malicious cyber-attacks. Artificial intelligence has influenced every aspect of our daily lives. Every tech app or service we use contains at least some type of Artificial Intelligence or smart learning technology. to maintain a safe, strong and sturdy AI system, there are different measures like talk to the team, execute threat modeling, Utilize foundational security functions, take advantage of transport layer security, check the system regularly etc. It is our duty to understand how AI works and how to protect it, maximizing its application to elevate industry standards and individual quality of life. By taking the above necessary measures, we can integrate efficient, reliable and secure AI systems to improve our work and personal lives. In this paper, we report on AI hacking and the security measures for verified artificial intelligence. , we report on the state of the art of attack patterns directed against AI and ML methods. We derive and discuss the attack surface of prominent learning mechanisms utilized in AI systems. We conclude with an analysis of the implications of AI and ML attacks for the next decade of cyber conflicts as well as mitigations strategies and their limitations.*

***Key Words*: Artificial intelligence, Machine Learning, AI hijacking, Cyber Attack, Cyber Security**

1. INTRODUCTION

Artificial Intelligence (AI) is the branch of computer sciences that emphasizes the development of intelligence machines, thinking and working like humans. The natural intelligence of human beings involves emotionality and consciousness. However, Artificial intelligence works in a different mode. Artificial intelligence (AI) refers to the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions [9][10]. The term may also be applied to any machine that exhibits traits associated with a human mind such as learning and problem-solving. Artificial intelligence is a leading technology that helps companies to manage complex tasks effectively and enhance the level of productivity. In this generation, many business communities are using AI-based networks for enhancing organizational performance but they are also facing security risks and threats. Security and privacy both are major concerns linked with the AI technology that impact on the privacy of data and lead hacking issues. This proposed research focuses on the security issues of AI technology and evaluated effective countermeasures for enhancing the privacy of data. There are various methodologies adopted including qualitative design, inductive approach, content analysis method and many more. The conducted literature search helped to solve research questions and obtain reliable information about artificial intelligence security. It is found that a lack of awareness and unauthorized networks are major factors that lead to malware and Denial of Service related attacks in the AI networks. Therefore, it is suggested that companies should implement firewall and encryption based networks for protecting data against malware signals and provide complete training to the employees while using AI-based networks. While machine learning is based on the idea that machines should be able to learn and adapt through experience, AI refers to a broader idea where machines can execute tasks "smartly." Artificial Intelligence applies machine learning, deep learning and other techniques to solve actual problems. Machine learning automates analytical model building. It uses methods from neural networks, statistics, operations research and physics to find hidden insights in data without being explicitly programmed where to look or what to conclude. Machine Learning (ML) is a subset of Artificial Intelligence. ML is a science of designing and applying algorithms that are able to learn things from past cases. If some behavior exists in past, then you may predict if or it can happen again. Means if there are no past cases then there is no prediction. ML can be applied to solve tough issues like credit card fraud detection, enable self-driving cars and face detection and recognition. With society's increasing dependence on ML and AI, we must prepare for the next generation of cyber-attacks being directed against these systems. Attacking the system through its learning and automation methods allows the attackers to severely damage the system by altering its learning outcome, decision making, identification or final output. Furthermore, it is difficult to analyze AI systems post-incident and integrate real-time monitoring during their operation: much of the learning and reasoning is done in what is called a "hidden layer" and in its essence corresponding to a black box model. Therefore, the discrimination of a compromised from an uncompromised AI system in real-time is still considered very difficult. With its increasing utilization in crucial application scenarios, the security of AI systems becomes indispensable. Knowledge of AI systems the

Vulnerabilities may also become of high importance to defensive cyber operations. During 2019, we witnessed increasing weaponisation of AI, often to create "deepfakes" – artificially generated or altered media material found to impose a sincere threat to democracies [1]. The uprising of deepfakes has encouraged the U.S. DARPA to spend $68 million on the identification of deepfakes over the past four years [2]. While it is of utmost importance to identify AI-supported disinformation campaigns, identification alone will not stop such operations. The aim of this paper is to foster understanding of the susceptibility of AI systems to cyber-attacks, how incautious utilization of AI and ML may make societies vulnerable, and to transfer the value of knowing AI-/ML-system vulnerabilities within the ongoing AI arms race. Attack surface modeling is a key contribution to assessing a target's susceptibility to attacks. However, AI systems have several peculiarities, which must be addressed when deriving the attack surface. Within this article, attack surfaces of different AI systems are derived that consider systems' data assets, processing units and known attack vectors, allowing us to understand these systems' vulnerabilities. Furthermore, these attack surfaces must be discussed with the systems' societal and economic impact in mind to allow strategic and policy recommendations. At the time of writing, neither the AI systems' concrete attack surface definition nor the embedment of the different AI systems' specific operational setup has been part of the security assessment of these systems. Allowing an AI-specific, concrete attack surface discussion, which includes the operational setup associated with the AI/ML method utilized by the system, is the main contribution of this article in addition to providing insights into the role of AI systems' susceptibilities to cyber-attacks in the next decade of cyber conflicts.

2. RELATED WORK

The Role of AI in Cyber security Emerging technologies put cyber security at risk. Even the new advancements in defensive strategies of security professionals fail at some point. Besides, as offensive-defensive strategies and innovations are running in a never-ending cycle, the complexity and volume of cyber-attacks have increased. Combining the strength of artificial intelligence (AI) with cyber security, security professionals have additional resources to defend vulnerable networks and data from cyber attackers. After applying this technology, it brought instant insights, resulting in reduced response times. Cap Gemini recently released a report based on AI in cyber security, which mentions that 42% of the companies studied had seen a rise in security incidents through time-sensitive applications[5][6]. It also revealed that two out of three organizations are planning to adopt AI solutions by 2020. Data security is now more vital than ever. Updating existing cyber security solutions and enforcing every possible applicable security layer doesn't ensure that your data is breach-proof. But, having a strong support of advanced technologies will ease the task of security professionals. With the use of AI, most of the tasks are automated, and these AI-powered apps are used in every sector like healthcare, education, military, etc. The data from these smart devices is collected from a set of sensors such as heat, light, weight, speed, or noise. The machine learning technology is directly connected with the security of digital devices and our data/information. There has been an extensive progression in the artificial intelligence sector[1]. And in today's modern age, where all our devices are connected either to the internet or some other modes of networks, the risks of security issues and the need for Artificial Intelligence solutions have skyrocketed will AI affect cyber security. While cyber security experts have accepted AI as the future of the industry, finding solutions to its problems are still not adequately addressed. Apart from being a solution, it is a considerable threat to businesses. AI can efficiently analyze user behaviors, deduce a pattern, and identify all sorts of abnormalities or irregularities in the network[8]. With such data, it's much easier to identify cyber vulnerabilities quickly. Contrarily, the responsibilities which are now dependent on human intelligence will then be susceptible to malicious cyber programs imitating legitimate AI-based algorithms. Several organizations are rushing into getting their machine-learning-based products out in the market. With this behavior, they might overlook the algorithms are creating a false sense of security. Relying on "supervised learning" is another threat. Under this, the algorithms label the data sets as per their nature. It could be malware, clean data, or some other tag. Cybercriminals, if they get access to the security firm, can alter the label as per their convenience. Also, routine tasks relying on AI can be manipulated by advanced hacking campaigns through the use of machine learning. In spite of being a security risk to the businesses, AI will continue to minimize the routine security responsibilities with high-quality results. AI automation will be able to identify recurring incidents and even remediate them. It will also be able to manage insider threats and device management. Today, organizations pay close attention to their network security. They are aware of the massive impact of every small- to large-scale cyber-attack. To secure this infrastructure, organizations use multiple lines of defense [1][4]. This multi-layered security system usually starts with the best suitable firewall capable of controlling and filtering out the network traffic. After this layer, the second line of defense consists of antivirus (AV) software. These AV tools scan through the system to find and eliminate malicious codes and files. With these two lines of defense, organizations regularly run backups as a part of a disaster recovery plan. For now, setting up firewall policies, managing backups, and many such tasks require a professional, but AI will change the traditional approach. Organizations will be able to monitor and respond to security incidents by using advanced tools. The next-generation firewalls will have in-built machine learning technology that could find a pattern in network packets and block them automatically if flagged as a threat. Predictably, the natural language capabilities of AI will be used to understand the origination of cyber-attacks[3][8]. This theory can be put into practice by scanning data across the internet.

3. PROPOSED METHOD

In this paper we report on state of the art attack patterns directed against these systems and how it must be expected that these systems will become prominent targets over the next decade, derive and discuss how attack surfaces may be modeled for AI systems. Apply the previously derived attack surface model to AI systems utilizing the different methods in previously compare their susceptibility to attacks.

3.1 ML AND AI METHODS

The field of artificial intelligence and especially the sub-field of machine learning is vast. Within the scope of this article, we consider some of the prominently utilized methods with cross-domain applications. Artificial Neural Networks (ANNs) describe the basic principles of neural networks and are commonly applied to predictive modeling problems involving the analysis and classification of non-linear relationships within data sets. Convolutional Neural Networks (CNNs) are an adaptation of ANNs specifically designed to map image data to an output class[7]. CNNs are commonly applied in prediction problems involving data analyses. GANs (Generative Adversarial Neural Networks) have become publicly renowned through the emergence of “deepfakes”, which has yielded strong interest in deep learning methods. opposing to the discriminative learning of ANNs and CNNs having a clear goal, generative modeling helps with understanding data and generating hypotheses. Support Vector Machines (SVM) were the standard solution to pattern recognition tasks prior to the emergence of neural networks and were used extensively in audio, video and handwriting recognition tasks.

3.1.1 Artificial Neural Networks

ANNs provide an abstract replication of the processes existing in the human brain. These models consist of simple atomic components called neurons, which are very limited in their individual capabilities, but which may be combined to perform more complex tasks. ANNs usually do not incorporate any task-specific rules, but instead derive the correct output from examples. Similarly to the biological model that inspired ANNs, a simple neuron may only be able to decide if an input is above certain threshold or not. However, collectively, a circuit of multiple neurons is capable of performing much more complex tasks. The peculiar strengths of ANNs are scalability and flexibility, achieved through the combination of multiple neurons. The computational capabilities are achieved through the vast connections between individual neurons. However, these multiple neurons artificially expand the “parameter space” – the space of all possible parameter combinations. Hence, the enhanced flexibility and scalability come at the price of larger training sets and higher computational power being necessary to make the neural network converge towards the correct solution.

.

3.1.2 Support Vector Machine

SVMs utilize mathematical concepts to define a separating hyper plane for a given set of data. Finding a separating hyper plane for a set of linearly separable clusters can be achieved through logistic regression. In order to understand non-linear relationships or solve higher-dimensional tasks, SVMs utilize “kernel tricks”.

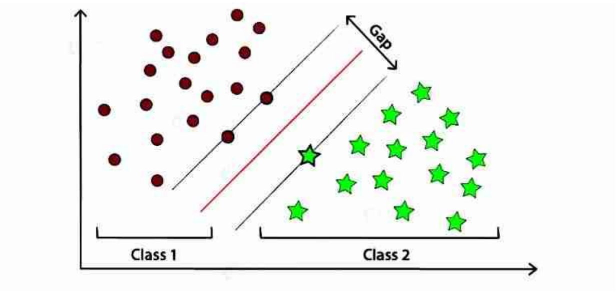

Fig1: Support Vector Machine

The results achieved by SVMs are considered to be trustworthy and robust. Some of the drawbacks of SVMs are the limitation to two-class-problems, the complexity associated with reducing multi-class problems to concurrently executable two-class-problems, the utilization of rather complex mathematical models of kernel-functions, the necessity of labelled data input and difficulties associated with the model parameter interpretation (amongst others: finding the actual kernel function). However, SVMs are still used in various application scenarios stemming from the fields of data science, data analytics and business analytics.

3.1.3 Generative Adversarial Networks

Generative Adversarial Networks (GANs) have gained much attention during the last year due to their frequent utilization in the creation of “deepfakes”. GANs consist of two competitive internal ANNs – the generator and the discriminator. These ANNs are trained in parallel in a competitive manner, which is often deployed as a zero sum or adversarial game. The discriminator tries to detect whether an input is originating from a training dataset or has been synthesized, while the generator generates adversarial samples to mislead the discriminator. As the competitive training automatically generates feedback information, GANs do not necessarily need labelled training data. However, in order to provide reasonable output, at least the discriminator should be pre-trained on labelled data. For the creation of deepfakes, conditional GANs (cGANs) are often used, which rely on labelled data to allow a target-oriented training.

3.2 ATTACK SURFACE IN AI

The security of AI systems and attacks directed against these systems are currently being neglected in public discussion, while the versatile utilization of AI in varying application contexts is widely discussed. However, within the academic and technical communities, several techniques and attack vectors directed against AI systems and methods have been reported. Currently, the most prominent attack vector categories are [3]:

- Adversarial inputs.
- Data poisoning attacks.
- Model stealing technique.

As AI is expected to become ubiquitous over the next decade, the importance of understanding the vulnerabilities of AI systems and methods becomes clear. Within the following subsections, we define how attack surface modelling for AI systems should be done to include the peculiarities of these systems.

A. Data Assets

The attack surface provides information of possible entry points for an attacker as well as exit points allowing access to the systems' data. It is the result of all possible attack vectors against a system or component. Allowing distributed data implies that the data must be kept consistent throughout the system processing entities. This is usually done by a periodic or event-triggered merging of the distributed data assets, where the data is collected from all entities. This requires authenticity of the entities involved and methods to ensure that no manipulation of the data can be performed during transportation (man-in-the-middle attacks).

B. Processing Units

Processing units within AI systems are units that are directly involved in the learning process, the data gathering or the decision making. While some attacks against the processing units will utilize data to perform the attack, other attack vectors may deploy techniques directed against the application involved (e.g. a web crawler used for data gathering is susceptible to web application vulnerabilities), the process itself or the libraries used. A specific type of attack combines the use of poisoned data and known vulnerabilities in the processing entities [4]. Previous attacks of this type have used audio/video files to hide malicious background operations in a steganographic manner to allow for the execution of arbitrary code [5]. While initially considered as an attack against a specific media player, this attack utilized Meta language library vulnerability. This attack vector could have affected other applications calling the library equally, such as AI systems processing a manipulated file.

3.3 SYSTEM VULNERABILITIES IN AI

Following the OWASP guidelines on attack surface assessments, we identify entry and exit points and briefly discuss reported and plausible attack vectors. As there are some similarities regarding the attack surfaces of ANNs, CNNs and GANs, a full explanation of an identified attack vector is given at its first encounter only.

ANNs: - The incoming data is preprocessed (reduce noise/selection of relevant material) and features are extracted. The data is labelled manually or automatically during the preprocessing. The weights of the network are adapted during the training. The final classification uses the weights derived during the training. Due to a lack of sufficient metrics for AI attack surfaces, it is difficult to derive a quantified and comparable assessment of the attack surface. However, it is observable that ANNs have a comparably large attack surface[6]. The possibility of incorporating applications for the data gathering and annotation expand this attack surface even further. Overall, ANNs appear highly susceptible to a variety of cyber security attacks due to their complex nature of internal processing units and their frequent import/export of data requiring long-term storage. The application of transfer learning expands the attack surface even further, as another entry point within the ANN is established.

CNNs: - CNNs may depict larger and more complex models as they do not have the common parameter space increase witnessed in ANNs [7]. Pre-processing and feature extraction are performed by the CNN internally. CNNs work with sensitive data assets, these are:

- The data gathered itself;
- Labelled data;
- Weights derived from training or through transfer learning;
- Classification results.

Within CNNs, the pre-processing and feature extraction are part of the network and not performed by separate application entities. Therefore, the data quality for CNN applications is of higher importance than for systems utilising ANNs. In addition to the above, further attack vectors on CNNs have been reported, amongst others utilising evolutionary computing methods, evasion attacks and side-channel attacks on CNN FPGA accelerators [31].

GAN: - The GAN is used to enhance the training of an already existing CNN (Discriminator CNN) for classification purposes. The Generator CNN creates additional training samples which are aimed to throw off the classification. The resulting Discriminator CNN after training is in general more robust against adversarial samples then the original one. Due to their composition of two CNNs operating in parallel, GANs have the same type of sensitive data assets as CNNs. Depending on the type of GAN (conditional or unconditional) labels may be present in the data (or not) and must be considered accordingly when defining the attack surface.

Most reported attacks on GANs try to reconstruct the used training data from the final model, which is called member inference attack [7]. These models can be used to generate adversarial attacks on other ML methods and also to protect them from such attacks [8].

Complicated hacking techniques, such as obfuscation, polymorphism, and others, make it a real challenge to identify malicious programs. Besides, security engineers with domain-specific workforce shortage are another issue. With AI stepping into cyber security, experts and researchers are trying to use its potential to identify and counteract sophisticated cyber-attacks with minimal human intervention. AI networks and machine learning, a subset of AI, has enabled security professionals to learn about new attack vectors. Machine learning in cyber security is much more than a mere application of the algorithms. It can be used to analyze cyber threats better and respond to security incidents. There are a few other significant benefits of machine learning, which includes

1) Detects malicious activities and stops cyber attacks
2) Analyzes mobile endpoints for cyber threats – Google is already using machine learning for the same
3) Improves human analysis – from malicious attack detection to endpoint protection
4) Uses in automating mundane security tasks
5) No zero-day vulnerabilities

4. CONCLUSIONS

In conclusion, it must be noted that AI systems are indeed susceptible to cyber-attacks and that the utilisation of AI or ML methods increases any applications' vulnerability. This necessitates more sensitive use of AI and ML methods in security- or safety-sensitive applications. The defence of AI systems is yet at its beginning and requires further investigation into the specific vulnerabilities of these systems. Furthermore, knowledge of AI systems' vulnerabilities may become crucial to defend against cyber operations which are being carried out with the aid of AI. Such operations are currently described in modern disinformation campaigns, as well as in information and hybrid warfare with only limited countermeasures currently available. In the context of political challenges and the ongoing AI arms race, a profound knowledge of AI systems' vulnerabilities must be established to uphold cyber sovereignty.

REFERENCES

[1] Keir Giles, Kim Hartmann, Munira Mustaffa, "The Role of Deepfakes in Malign Influence Campaigns", NATO StratCom COE, ISBN 978-9934-564-50-5, September 2019, https://www.stratcomcoe.org/role-deepfakes-malign-influence-campaigns.

[2] Stephanie Kampf, Mark Kelley, "A new 'arms race': How the U.S. military is spending millions to fight fake images", CBC.ca, 18 November 2018, https://www.cbc.ca/news/technology/fighting-fake-images-military-1.4905775.

[3] Elie Bursztein, Security and Anti-Abuse Research Lead at Google, "Attacks against machine learning — an overview" Personal Site and Blog featuresing blog posts, publications and talks, May 2018, https://elie.net/blog/ai/attacks-against-machine-learning-an-overview/.

[4] Jeffrey Ding, "ChinAI #47: The Sensenet Data Leak - What Actually Happened", 25 March 2019, https:// chinai.substack.com/p/chinai-47-the-sensenet-data-leak.

[5] Sam Daley, "Surgical robots, new medicines and better care: 32 examples of AI in healthcare", builtin. com, 23 September 2019, https://builtin.com/artificial-intelligence/artificial-intelligence-healthcare.

[6] Savia Lobo, "VLC media player affected by a major vulnerability in a 3rd library, libebml; updating to the latest version may help", hub.packtpub.com,25 July 2019,https://hub.packtpub.com/vlc-media playeraffected-by-a-major-vulnerability-in-a-3rd-library libebml-updating-to-the-latest-version-may-help/: CVE 2019-13615 Details, NIST National Vulnerabilities Database,16July2019,https://nvd.nist.gov/vuln/detail/CVE-2019-13615.

[7] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta and A. A. Bharath, "Generative Adversarial Networks: An Overview," in IEEE Signal Processing Magazine, vol. 35, no. 1, pp. 53-65, Jan. 2018.

[8] Valecia Maclin, "Solving the challenge of securing AI and machine learning systems", Microsoft Blog, 6 December 2019, https://blogs.microsoft.com/on-the-issues/2019/12/06/ai-machine-learning-secu

BIOGRAPHIES

Subin Babu received the B.Tech degree in Computer Science and Engineering from Mahatma Gandhi University, Kerala, India in 2012. She is currently pursuing M.Tech degree in Computer Science and Engineering from APJ Abdul Kalam Technological University, Kerala, India at Mount Zion College of Engineering, Kadammanitta, Kerala, India. Her primary research interests are in Artificial Intelligence, IOT and Cyber Security.

Nisha Mohan P.M. received the M. Tech degree in Communication and Networking from MS University, Tirunelveli, India in 2013. She is currently working as Assistant Professor in the Department of Computer science and Engineering at Mount Zion College of Engineering, Kadammanitta, Kerala, India. Her primary research interests are in Cloud Computing, Image Processing, Cyber Security and Artificial Intelligence (Machine Learning oriented programming).