BRIEFING NOTES

BN-70-Space and Cyberspace-Aug2021

MULTI-FACETED ISSUES OF INCORPORATING AI SYSTEMS INTO CYBERSECURITY SOLUTIONS

Authors: Mohamadreza Nematollahi[1] and Kash Khorasani[2]

1 Graduate student, Department of Electrical and Computer Engineering, Concordia University, Montreal, Canada

2 Professor, Department of Electrical and Computer Engineering, Concordia University, Montreal, Canada

SUMMARY

- The growing market for application of AI in cybersecurity has attracted the attention of both public and private sectors within the past few years. It has been estimated that this market will grow from US$1 billion in 2016 to a US$34.8 billion net worth by 2025.
- It has been argued by many experts that to trust AI in cybersecurity tasks is a double-edged sword.
- While the advancements in AI can substantially enhance cybersecurity practices, it also creates certain vulnerabilities and provides a fertile ground for emergence and growth of much more facilitated cyber-attacks and security threats.
- It has been reported recently (2019) that not only cyber-attacks have been escalating in frequency, sophistication and impact, they also tend to be more mutable and faster in reaching their targets, which ranks them among the top five most likely sources of severe, global-scale risks.
- Through advancing a system's response, AI can improve a system's capability to autonomously defeat attacks and to predict future strategies to prevent future attacks.
- With regard to the complex and dynamic nature of cyberspace, security systems often have to overcome several strategic and technological challenges. Since perpetrators leave tracks when attempting to attack a potential target system, it is possible for an integrated security approach (ISA) to gather and analyze data through a holistic view, in order to acquire cyber intelligence.
- To improve the conventional security systems and to tackle the problem of intelligence-gathering AI systems show signs of success. AI has the ability to imitate human intelligence in limited extents and the recently revived approach of deep neural networks enables systems to learn and to automatically adapt to the dynamics of changing environments.
- There are multiple potential hazards by AI systems, such as deliberate unethical actions, side effects of poor engineering and design, as well as the environment. It is important to shed more light on the nature of malicious exploits and to provide ideas on how to design tools to protect cyber infrastructures.

CONTEXT & CONSIDERATIONS

- Several governments, have already begun to rely on AI for improving the security of the critical national infrastructures such as transportation, hospitals, energy and water supply [1]. Currently, within the fast-paced context of advancing technology, conventional security systems fail to keep up with sophisticated cyber threats and are often unsuccessful in supplying practical solutions as they are slow and inadequate. Consequently, introduction of AI into the field of cybersecurity sounds quite promising and seems to be able to provide much higher levels of protection and security [4].

- It is claimed that with the advent and progress in self-driving cars and personal digital assistant systems, several unexpected, unresolved and unfavorable complications such as accidents by self-driving cars, racism and hate speech by chat-bots have appeared [2].
- Since AI systems alone have not proven a full success in this field and has instead raised certain concerns, it is advised, as a socially responsible use of AI, that a holistic view of the problem be taken into account in which both AI and human insights are involved [4].
- It is widely believed that both the frequency and seriousness of AI related failures have and will steadily increase as AI becomes more capable. At some level the occurrence of such failures within super-intelligent systems may result in catastrophic consequences that remain beyond redemption. While cybersecurity aims at reducing the number of attacks on the system, AI safety is determined to ensure zero cyber-attacks.
- However, such level of safety is unattainable and every security system will at some point could face failure [2]. “To implement resilient and continuous protection, security systems need to constantly adjust to changing environments, threats, and actors involved in the cyber play” [4]. Consequently, it has been recommended that employment of AI systems within the context of cybersecurity must come with certain levels of reliability in order to reduce the possible risks.
- AI systems have often been introduced as a tactical and strategic solution to the problem of cybersecurity through changing the dynamics that facilitate offence over defense in cyberspace. As many suggest, it can lower such risks by improving the security of systems and reducing their vulnerability to attacks. For instance, it is believed that AI is capable of performing self-testing and self-healing procedures which result in enhancing the robustness and capacity of a system to keep behaving as expected even when it processes and acts upon erroneous inputs [4].
- AI systems are capable of increasing a system’s level of resilience so that it can withstand cyber-attacks by facilitating threat and anomaly detection [1]. Cyber reality, however, tends to also depict unfavorable outcomes. As security systems lack enough flexibility and robustness they often face a number of challenging obstacles in adapting to regular alterations of known cyber-attacks. Even human interactions in such cases have been insufficient and inadequate. Therefore, AI approaches seem more suitable and promising as they are more flexible and adaptable. However, AI systems raise alarming concerns regarding existential risks for humanity, as well as ethical justifiability [4].
- Despite all the efforts on the part of individuals, private organizations, non-state actors and governmental organizations to strengthen their virtual assets against cybercrimes, their awareness of the nature of cyber-attacks and their ability to fight them is still quite limited. While the sources of cyber threats remain manifold and the nature of cyberspace is heterogeneous and dynamic, it is still possible to find resemblances among the cyber-attacks which help security systems to develop countermeasures and security frameworks.

- In short, such cyber-attack sequences can be described as a cyber kill chain that begins with a reconnaissance phase in which gaps and vulnerabilities of the target system are located. It then follows by the weaponizing phase, in which malicious codes are developed using uncovered weaknesses and vulnerabilities. The delivery phase happens next, during which the malware is transferred to the potential target. Finally, the exploit phase allows the malware to trigger the installation of and intruder's code.
- Consequently, the attacker can initiate malicious actions in the compromised host system. As a solution, the integrated security approach (ISA) provides a framework to generate early warnings, preferably before the exploit phase. Such measures include counteractions to stop the invader and also define recovery procedure to allow the system to immediately rollback to its initial state [4].
- However, the process of gaining, analyzing and employing such data remains challenging. For instance, as with the ubiquity of electronic devices the amount of data has grown significantly, that makes it challenging for ISA to implement data from all systems across entire organizations. In addition, heterogeneity of data and variety of sources lead to certain difficulties in identifying and collecting data.
- Moreover, with constant alterations occurring in the topology and behavior of systems and networks, constant adaptation may as well be required. High data velocity which refers to the high rate of data production and processing can also be problematic. Conventional security systems are unable to fully deal with problems such as low detection rates, lack of resilience and lack of automation in the continuously changing network environment [4].
- Such advancements can be advantageous in the field of cybersecurity by using autonomous AI systems. Through mitigating the above challenges in conventional cybersecurity, AI systems can improve the efficiency and utility of cybersecurity. For instance, AI systems are capable of processing vast amounts of data within feasible time periods. One favorable characteristic of an AI system is the paradigm of intelligent agents that deals with the idea that the knowledge to solve problems ought to be shared among various entities. According to this paradigm, throughout time, agents accumulate experience and self-adapt to dynamic changes in their environment. This feature can then be used in terms of defense measures but also for reconnaissance and exploitation of potential target systems [4].
- Nevertheless, AI systems can be vulnerable on many levels. Previously, cyber-attacks imposed on AI systems mostly occurred in the form of data extraction and system disruption. However, new forms of cyber-attacks tend to gain control of the targeted system through behavior modification. Since AI systems rely on datasets to train, it is possible to introduce carefully crafted, erroneous data among the legitimate data so as to make alterations to the AI system's behavior.
- It is also possible to manipulate the categorization and classifying models that AI systems rely on, which can result in unexpected and hazardous performances. Moreover, it is

often difficult to detect and explain such defective outcomes and behavior within the AI systems, as their networked, dynamic and adaptive nature make them problematic to visually see through their internal processes [1].

- For instance, some forms of cyber-attacks are deceptive such as when a backdoor is added to a neural network. In this case, the system seems to be working normally up until the trigger is activated to alter the system's behavior and even then, such skillfully crafted triggers cause minimal divergence between the actual and the expected behavior, making it difficult to identify and determine when the compromised system is depicting abnormalities.
- Such nuanced alterations may allow the attackers to achieve their goals and to interfere with the behavior of the system. "For example, it is possible to trick an AI image recognition system to misclassify subjects wearing specially crafted eyeglasses. Arguably, a similar attack could target a system that controls access to a facility and enables access to malicious actors without raising any alert for a security breach" [1].
- The exchange of such information among hackers and security experts can be beneficial and create a well-balanced cyber-ecosystem. It is suggested that mistakes and failures caused by the AI systems often have their roots in the intelligence such systems are designed to portray. To put it differently, AI failures either occur during the learning phase or the performance phase. Such systems tend to learn correlated functions but not exactly what their designers intended to teach them.
- For example, a computer vision system which is supposed to classify pictures of tanks learns instead to distinguish backgrounds of such images. So there comes the simple generalization that an AI system is designed to do X will eventually fail to do X [2].
- Therefore, it is imperative to ensure the robustness of an AI system and to develop not only a system that is safe but also a capable one. To ensure some level of certainty that the AI system will still behave according to the expected pattern even if the models and inputs have been tempered with through cyber-attacks is a crucial requirement and demand for these systems. However, assessing and maintaining such robustness require testing for all possible input perturbations that maybe practically unfeasible.
- In other words, it is impossible "to foresee exhaustively all possible erroneous inputs to an AI system, and then measure the divergence of the related outputs from the expected ones". In this case, alternative measurements are required to secure the robustness of the AI system [1]. Although experts believe that there is no such thing as perfect security, only varying levels of insecurity, it has been recommended that testing and debugging during software development can lead to production of safer codes.
- Using advanced techniques, software engineers may be able to detect and fix serious errors so that the product best matches the intended purposes. Yet, fully autonomous machines can never be assumed to be safe and there might not be a 100% safe option available [2].

- While an ideal cyber-defense system should provide its users with full protection, AI systems still has a long way to reach to the point of zero-day attack and to play a central role in cybersecurity. Consequently, cyber-defense needs to take steps toward becoming more intelligent in order to overcome its inherent and severe limitations [3].
- For this purpose, the International Organization for Standardization (ISO) has established a committee to work specifically on AI standards. In US, the Defense Advanced Research Projects Agency (DARPA) launched in 2019 a new research program, called Guaranteeing AI Robustness against Deception, to foster the design and development of more robust AI applications.
- Likewise, in 2019, the China Electronics Standardization Institute established three working groups, namely: 'AI and open source', 'AI standardization system in China' and 'AI and social ethics'. While such attempts share the same goal to secure trust in AI cybersecurity, their effectiveness remains unsettled and defining and developing such standards for generating trustworthy AI in cybersecurity could be conceptually misleading and result in further sever security risks.
- "Trustworthiness is both a prediction about the probability that the trustee will behave as expected, given the trustee's past behavior, and a measure of the risk run by the trustor, should the trustee behave differently. When the probability that the expected behavior will occur is either too low or not assessable, the risk is too high and trust is unjustified". In the case of AI in cybersecurity, the same justification is applicable since the transparency and learning abilities of AI systems are questionable. Thus, AI system can fail to ensure trust and predictability in the context of security, and to assess trustworthiness in this criterion remains problematic [1].
- It is therefore suggested that experts focus on the reliability of AI systems rather than fostering trustworthiness. "Reliability of AI implies that the technology can, technically, perform cybersecurity tasks successfully, but the risks that the technology may behave differently from what is expected are too high to forgo any form of control or monitoring over execution of the delegated task".
- Therefore, it is important to anticipate the level of control that is best applicable to the learning nature of the system and the dynamic nature of cyber-attacks, while taking lack of transparency on the part of AI system into consideration. Also, the resources such as time and computational feasibility should be taken into account [1].
- Since one of the most common cyber-attacks to AI systems occur through the commercial services that provide support for the training and development of AI solutions, it is necessary to pay specific attention to in-house development. Therefore, reliable suppliers must design and develop models for security of national critical infrastructure. Also, the data must be collected, curated and validated by reliable providers.
- Another well-known method to enhance robustness of AI systems is the adversarial training in which feedback loops enable the AI solution to adjust its own variables and

coefficients with each iteration. Moreover, parallel and dynamic monitoring can significantly boost the reliability of AI systems.

- Given that maintaining the robustness of AI systems faces several limitations and cyber-attacks are of deceptive nature, it is crucial to focus on constant monitoring in order to detect divergences early and promptly so as to address them properly.
- "To do so, providers of AI system should maintain a clone system as a control system […] The clone should go through regular adversarial exercises, simulating real world attacks to establish a baseline behavior against which the behavior of the deployed system can be benchmarked. Divergences between the clone and the deployed system should flag degrees of security alerts" [1].
- In the field of cybersecurity, artificial neural networks (ANNs) have made improvements to prevent malicious intrusions. ANNs are statistical learning models that imitate the structure and function of human brain and since their behavior is elusive they are considered undefined black-box models. While monitoring network traffic, they are able to detect malicious intrusions within the delivery phase before the actual attack occurs.
- Moreover, they have the ability and the great advantage to learn from past activities and cyber-attacks in order to identify and to hinder future cyber-attacks. They are able to recognize and distinguish between such abnormal and normal behavioral patterns automatically by using previous data which is an advantage as compared to conventional security techniques that had to be defined manually by security professionals based on their expert knowledge.
- It has been shown that by using the ANN approach "overall detection rate of attempted intrusions has improved without generating any false positive or false negative alarms […] this approach has successfully protected against instances of intrusions that were previously unknown […] In summary, ANNs are said to support a viable approach to building robust, adaptable, and accurate IDPS" [4].
- Moreover, following advancements in data processing within network infrastructure, ANNs have been used not only to prevent malicious cyber-attacks but also to predict them even up to 85 percent of the time [4].
- Despite the above-mentioned aspects of incorporating AI into cybersecurity, there remain concerns and risks. Although significant steps have been taken and approaches such as ANN have made substantial improvements, security systems are not still fully autonomous, and since they cannot replace human experts, human intervention is required.
- Moreover, in order to train AI techniques and algorithms such as ANN, large data sets are required. There is also a downside to collecting such data which is endangering the privacy of the subjects. In such cases, lack of transparency and regulations raises various legal concerns for both private and governmental parties involved.
- Due to the dynamic and unforeseeable nature of AI systems, many are concerned that in the long run human control over the AI's autonomy will be lost and it might create serious

existential threats. Last but not the least, are the ethical concerns around the use and expansion of AI.

- Although AI security systems are growing to make more decisions for human individuals, they fail to provide or follow any moral and ethical codes and the decisions that they will take may contradict human decisions on many levels. Therefore, despite the promising, adaptable, flexible and robust nature of the AI, security systems are still recognized as potential risks for human civilization [4].

REFERENCES

[1] Taddeo Mariarosaria, McCutcheon Tom and Floridi Luciano (2019) "Trusting artificial intelligence in cybersecurity is a double-edged sword", published online, nature machine intelligence, Springer Nature

[2] Yampolskiy Roman V (2015) "Artificial intelligence safety and cybersecurity: a timeline of AI failures"

[3] Morel Benoit (2011), "Artificial intelligence a key to the future of cybersecurity"

[4] Wirkuttis Nadine, and Klein Hadas (2017) "Artificial intelligence in cybersecurity" Cyber, Intelligence, and Security, Volume 1, No.1.