GCSP
Geneva Centre for
Security Policy

September 2021

Artificial Intelligence and Cyber Security: A Complex Relationship

Natalia Spinu and Gaurav Sharma

Alumni Note

Geneva Centre for Security Policy

The Geneva Centre for Security Policy (GCSP) is an international foundation established in 1995, with 53 member states, for the primary purpose of promoting peace, security and international cooperation through executive education, applied policy research and dialogue. The GCSP trains government officials, diplomats, military officers, international civil servants and NGO and private sector staff in pertinent fields of international peace and security.

Geneva Centre for Security Policy
Maison de la paix
Chemin Eugène-Rigot 2D
P.O. Box 1295
CH-1211 Geneva 1

Tel: + 41 22 730 96 00
Fax: + 41 22 730 96 49
E-mail: info@gcsp.ch

www.gcsp.ch

Published in September 2021

The ideas expressed are those of the author not the author’s affiliation or the publisher.

About the Authors

Natalia Spinu is a cyber security expert with many years of experience in governmental sectors in the Republic of Moldova. Chief of the Government Cyber Security Center (CERT-Gov) and has as responsibility the strategic planning and international and intergovernmental cooperation. Under her leadership, the CERT-Gov became actively involved in many national cyber security development processes, including national cyber security programme and policy developments. She's an active contributor to the cyber security community, promoting collaboration via knowledge sharing, organising various cyber security related workshops, events, and meetups, as well as participating as a speaker in national, regional, and international forums.

Gaurav Sharma is Advisor for Artificial Intelligence (AI) at the German Agency for International Cooperation (GIZ) and an affiliate to "The Future Society". Gaurav is the Working Group Member of the Global Diplomacy Lab. He was Ambassador, Asia and the Pacific, Global Leadership Academy (GLAC), (2019 – 2020) and was a German Chancellor Fellow (2015 – 2016). Gaurav was awarded 'New Young Leader' by Crans Montana Forum (2014). Gaurav comes with 15+ years of experience in development cooperation, international relations (policy advisory and research), private sector (information technology) and non-governmental organisations (women's health). He is also a facilitator for the "Youth Climate Leaders" (YCL) in South Asia and the founder of the Indo-Swiss Future Leaders Forum (ISFLF).

Introduction

Artificial intelligence (AI) is the buzzword in almost all discussions on technology today. It is a subarea of computer science that aims to imitate intelligent human behaviour by programming a computer so that it can solve problems independently. It is becoming pervasive and critical in almost all aspects of our everyday life. As data continues to grow in quality and quantity and computing power becomes cheaper and faster, the impact of AI on cyber security will inevitably increase.

AI comprises a superset of disciplines like machine learning (ML) and deep learning. ML is essentially the "brain" of AI, utilising a set of algorithms that automatically learn through experience and use data to build a model that enables a computer or similar device to make predictions and decisions without being explicitly programmed to do so. Throughout this strategic paper, therefore, the term AI will refer to technologies that can understand, learn, and act based on acquired and derived information.

Nowadays AI deploys three kinds of intelligence: assisted intelligence, which improves organisations' activity; augmented intelligence, which enables people to carry out operations they could not otherwise do; and autonomous intelligence, which is still developing and will allow machines to act on their own. AI can be said to possess some degree of human intelligence: a store of domain-specific knowledge; mechanisms to acquire new knowledge; and mechanisms to put that knowledge to use. Machine learning, expert systems, neural networks, and deep learning are all examples or subsets of AI technology.

The use of AI in cyber security is increasing rapidly, and is used, for example, to identify data breaches, reduce response time, automate threat detection, and make organisations compliant with security best practices such as the General Data Protection Regulation (GDPR) in Europe and National Institute of Standards and Technology (NIST) controls in the United States. The result is new levels of intelligence feeding human teams across diverse categories of cyber security, including IT asset inventory, threat exposure, effectiveness control, breach risk prediction, incident response etc. Cyber security has taken on new significance during the COVID-19 pandemic, as more and more people started using new applications such as online doctor consultations, the utilisation of apps for contact tracing and social distancing via Bluetooth, and increased use of video applications, resulting in more data entering the cloud in cyber space. But with the rise in the number of users, online utilities and applications, cyber space has become more vulnerable and therefore an easier target for hackers.

ML is a specialised branch of AI that uses algorithms to understand models of phenomena based on examples (i.e. statistical machine learning) or experience (i.e. reinforcement learning). The technological advances made by ML has greatly benefitted the field of AI due to the huge availability of all forms of datasets such as text, images, optical, voice, and earth observation. As more complex and intelligent algorithms are developed via the huge availability of datasets, large numbers of AI applications are being designed and deployed. Because cyber security is fundamentally a Big Data problem, the need for AI to process and action vast amounts of data is increasing. Security teams have utilised cyber AI for several years to identify and mitigate cyber attacks, but cyber criminals are using this same power for their nefarious purposes. Thus, the use of AI in the area of cyber security could dramatically change in the coming years.

This strategic paper discusses the pertinence of the use of AI technology in the domain of cyber security, including its perils and safeguards. The paper is divided into three sections: AI-enabled cyber security, the disadvantages of using AI in cyber security, and recommendation for the deployment of AI systems for cyber security. The focus of the paper is on AI systems that can assist or augment protective cyber services and help users to curb cyber security risks; however, the use of completely autonomous AI systems in the area of cyber security is not discussed.

AI-enabled cyber security

The ability of AI-enabled cyber security systems to secure critical infrastructure and services in high-value environments can be broken down into the following three categories:

1. **Network threat identification:** Understanding the various aspects of a network topology is a major challenge, and AI-enabled systems can determine whether a mobile or web-based application contains any malicious software during development, testing and deployment. For example, AI-based cyber security software can monitor networks by using anomaly detection software and alert authorities to the presence of a data mismatch or events linked to previous cyber threats. Thus, one of the applications of AI cyber defence is to enable the setting up of self-configuring systems that provide network resilience, prevent cyber attacks and provide protection against cyber threats.
2. **AI-based antivirus software:** AI-powered antivirus software can help to proactively detect anomalous behaviour when programmes are running. Abnormal programme behaviour identification is more effective than looking for virus or malware signatures, because it can detect new threats.
3. **Email monitoring:** Natural-language-processing-based AI technologies can read emails and identify patterns or phrases or even image size to detect whether an email is a phishing attempt, a misdirected email, or a data breach.

AI systems monitor networks to detect anomalies, use software analysis techniques to identify vulnerabilities in code, and synthesise defensive patches at the first indication of a cyber attack. Because AI systems perform analyses of this kind in seconds, in principle, cyber attacks are detected and defensive measures are put in place in real time.

Cyber security frameworks have never been 100% foolproof and never will be, because the complexity of digital infrastructure keeps evolving, giving rise to new vulnerabilities. The inclusion of new-age AI tools can, however, empower cyber security frameworks by integrating cognitive insights, contextual analytics, benchmarking and integrating security intelligence into single alerts, thus accelerating incident analysis and remediation. On the other hand, however, AI-based cyber technology can also be used by hackers and cyber criminals to carry out more effective cyber attacks. Thus, hackers' use of AI systems is extremely dangerous and can range from shutting down power plants to hacking government datasets containing citizens' personal data.

AI can complement the work of cyber security experts by undertaking tasks that require large volumes of data processing, resulting in the rapid generation of information that can be used in decision-making. An AI model picks up anomalies that reflect inconsistency with IT systems' operations, and can take corrective action and alert human operators. These applications are particularly valuable in rapid threat identification and prediction, network security, password protection and incident response management. AI systems learn from patterns of behaviour, because they are trained to analyse past records and experiences, enabling them to be highly effective defenders against cyber attacks.

AI-based user authentication technologies

Passwords have always been a huge cyber security risk. Creating strong passwords combined with an AI cyber security layer, e.g. biometric verification, would add an extra layer of security. But if we move past biometric passwords, it is not difficult to conceive that AI could identify a user more securely by using sight and sound identification methods. Rather than checking against pre-defined credentials, a machine would be able to understand and confirm whether a person was who they claimed to be by using visual and aural clues. It could also learn when to grant access, and act accordingly. Permitting access on the basis of machine learning is the logical next step on from biometric identification.

Current smart phones that utilise facial identification to identify the user and unlock the phone are a good example. Apple's "Face ID" technology works by processing the user's facial features through built-in infrared sensors and neural engines. AI software creates a sophisticated model of the user's face by identifying key correlations and patterns and detecting thousands of different reference points across the face, forming a vector model of the user's facial features. The AI software architecture can work in different lighting conditions and can compensate for changes like a new hairstyle, a newly grown beard, wearing a hat, etc. Thus, it is increasingly becoming necessary to use AI cybersecurity with biometric verifications to reduce the risk of a cyber breach. A combination of strong passwords coupled with AI cyber measures is a good start to this process. Thus, dual authentication is becoming a widely used password protection method.

AI systems could also constantly monitor users as they move around the network. Behavioural factors and real-time risk analysis can also come into play. AI systems could monitor any unusual, irrational or erratic behaviour in real time.

Early detection, identification and prediction of cyber security threats

AI-enabled surveillance systems can recognise the slightest changes in network patterns and quickly identify them. AI systems are trained to carry out faster processing and accurate anomaly detection, based on millions and millions of past records, and thus can filter information much faster and detect even the slightest variations. AI is increasingly becoming the first line of defence through its capacity to detect and report cyber breaches.

Combining traditional early warning systems with AI can create virtual sensors and sophisticated data manipulation for logic models. Because AI systems keep learning with every warning they generate or anomaly they detect, an AI system keeps improving its reliability and scalability against cyber threats. Cyber attacks are commonly not built from scratch, but are generally constructed based on behaviours, frameworks, and source codes of past attacks, which means that ML has a pre-existing path to work from. Programming based on ML can help highlight commonalities between the new threat and previously identified ones to help identify an attack.

Cyber alert identification, detection and decision-making can be integrated as part of an AI-enabled cyber set-up such as an intrusion detection system (IDS), but IDS systems were never able to identify the full spectrum of threats, because they are built on pattern detection, but cannot detect problematic behaviours and anomalies. Today’s ML-aided IDS systems are able to detect behaviours, peer-group interactions, and rule-based patterns, thus allowing the IDS to detect known and unknown cyber attacker tactics, techniques and procedures, thus enabling full-spectrum inspection of encrypted traffic. This has two advantages, in that only critical threats will be brought to the notice of the cyber expert and threat response times will be shortened. For example, ML-assisted cyber security can identify the data affected by a cyber attack and group it automatically for further analysis.

Disadvantages of the use of AI in cyber security

An AI system needs to access many different datasets of anomalies and malware codes to train the algorithm to fight against cyber attacks. This raises the question of what and how much data can be shared, given that data privacy has become a fundamental issue with regard to data protection, and affects the deployment of AI systems as a mainstream cyber security tool. Apprehension about the loss of privacy due to the amount and type of data needed to train AI systems remains a major concern.

As any other system, AI systems are also prone to manipulation, and this can lead to faulty outputs and incorrect decision-making based on these outputs. Any cyber attack that changes the algorithmic logic even minimally in an AI system can result in profound security implications for applications such as network monitoring tools, financial systems, or autonomous vehicles. There are three typical adversarial AI tactics that leave an AI system vulnerable.

In an adversarial AI attack, the attacker seeks to generate and introduce small changes to a dataset that – although imperceptible to the human eye – can cause major changes to the output of an AI system.[1] As a result, it causes ML models to misinterpret the data inputs that feed it, making it behave in a way that is favourable to the attacker. The following are the broad categories of adversarial AI attacks:

1. **Data poisoning:** Data-poisoning or model-poisoning attacks involve polluting an ML model's training data. For example, the use of AI in facial recognition technology has brought the problem of bias due to data poisoning specifically in sectors such as law enforcement, leading to the false identification of persons of interest to law enforcement. Data poisoning is considered to be an integrity attack, because tampering with the training data impacts the model's ability to output correct predictions.[2] Hackers use data poisoning to make an AI system believe that the input it is receiving for the training data is valid, resulting in the system accepting it and training on it as if it were valid data. Data poisoning can lead to bias and to inaccurate or totally false results. Adversarial data inputs could also be poisoned when a programme is run, because attackers can alter the training data used by the AI system responsible for programme security.
2. **AI model poisoning:** In AI, a “model” is a set of ideas that a machine has formed about how some part of the world works, based on its analysis of data.[3] A hacker could trick the server into weighing a model’s updates differently from the original AI model used to

[1] Fujitsu UK & Ireland, “Can Adversarial AI Threaten Our National Security?”.
[2] CSO Online, “What Is Data Poisoning? Attacks that Corrupt Machine Learning Models”.
[3] Princeton University, “Protecting Smart Machines from Smart Attacks”.

analyse the data. This would result in data of the hacker's choice being classified in the class they desire, and not the true class.

3. **Privacy attacks:** Adversaries often try to gain access to private information. This is undertaken by testing large amounts of random input data to, for example, break a password. Using AI systems makes it easier and faster for cyber criminals to do this. Hackers also use AI-based malware to develop cyber attacks that are hard to detect, and use AI systems to identify vulnerabilities in the software they are targeting. This is done by utilising a technique called neural fuzzing, which involves testing large amounts of random input data in the software in order to identify weak spots. Neural fuzzing drives AI to test large amounts of random input data faster, and combined with neural networks, is able to gather information about a targeted software package or system and learn its weaknesses.

AI models and security best practices are vitally important in order to trace the behaviour of an AI system if its decision-making system has been compromised. However, safe deployment will require understanding the multiple dimensions and implications of these adversarial AI attacks and resolving any persistent cyber security challenges that arise.

Recommendations for the deployment of AI systems for cyber security

The ideal role of AI in cyber security is that of interpreting the patterns established by ML algorithms and computing abstract datasets into framing a pattern that the human operator can utilise and act upon. The current use of AI systems involves threat data classification, data clustering and placing the clusters into threat groupings, and recommending a course of action as advisories to the cyber security expert. AI systems are currently being used to carry out predictive forecasting, which is one of the most exciting forward-thinking possibilities for any AI-enabled cyber configuration, which is good at building threat models and proactively minimising data breaches. Efforts to improve cyber security are intertwined with advances made in AI systems, as the following recommendations will show.

1. For any deployment of an AI-system in a domain-specific environment, e.g. an operational system in a hospital, it is first necessary to understand how input data is acquired, secured, maintained, and evaluated. This will help in determining any anomalies that form part of a cyber attack on the AI system. Under the new GDPR, any AI system must ensure adequate data privacy.
2. Attackers have been shown to use AI systems designed for legitimate purposes such as speech recognition and voice recordings to fake online videos of politicians and celebrities by utilising generative adversarial networks. Thus, efforts should be made to design robust

ML methods to prevent reconnaissance of AI systems and to study in greater detail data-adversarial and data-poisoning models.

3. Cyber security guarantees should be embedded into an AI model using new multistep techniques, especially in security domains such as critical infrastructure where the risk posed by a cyber attack is highest. The need to integrate AI-based threat detection with compliance and security engines to address cyber challenges in network, device and applications is one of the most pressing current needs
4. Critical thinking and creativity are vital to ensuring effective decision-making in any cyber security environment. Thus, intelligent human teams will remain essential, for they alone have this capacity. In 2017, a group of AI researchers gathered at the Asilomar Conference Grounds in California and developed 23 principles for AI use, which were later dubbed the Asilomar AI Principles. The sixth principle states that "AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible".[4] According to the study, AI best practices must ensure the end-to-end provenance of training data and the detection of data that falls outside the normal input cyber space. This requires a robust AI governance structure that ensures that data is marked and checked at multiple checkpoints and does not leave the designated data domain. It is very difficult to design data protection and data privacy regulations for AI, because data is difficult to evaluate at the pre-processing, processing and post-processing stages of any AI application. Thus, a basic guideline framework, e.g. a checklist, will enable AI developers to be cautious of the data they are using and its limitation within the cyber space in which they are working.

[4] https://futureoflife.org/ai-principles/.

Conclusion

AI systems are becoming increasingly embedded into our everyday lives, from the time we wake and tell our smart house to switch on lights, to talking to Alexa, to playing the latest song, to obtaining intelligent advice regarding our health. Similarly, the cyber security sector is increasingly utilising AI systems for pre-emptive threat detection and surveillance, and to put in place intelligent evasive measures to counter cyber attacks. But it is imperative that we realise the threats posed by these AI systems if the data it works with and the model it uses are poisoned, and should start developing countermeasures against these threats.

Considerable efforts in managing AI are needed to produce secure model training; defend models from adversarial inputs and reconnaissance; and verify model robustness, fairness and privacy. This includes secure AI-based decision-making and the development of methods for the trustworthy use of AI systems and environments in all areas of cyber space. Thus, if data security cannot be guaranteed in a particular cyber space, no valid AI forecast is possible. For AI to successfully complement cyber security and protect our critical infrastructure, a science, practice, and engineering discipline for the integration of AI into computational and cyber-physical systems are needed that include the collection and distribution of an AI corpus, including systems, models and datasets.