Artificial intelligence and cybersecurity

Artificial intelligence (AI), which is being integrated into our daily lives at an overwhelming pace, has the potential to shape our digital landscape. As it can influence everything – from personal data security to national defence strategies – the issue of cybersecurity is becoming increasingly critical.

Three distinct dimensions of AI use

The relationship between AI and cybersecurity has three dimensions:

- **the cybersecurity of AI**, which covers AI standardisation;
- **the use of AI to support cybersecurity**, which empowers cybersecurity defenders;
- **the use of AI for malicious purposes**, which explores AI's potential to create new threats.

While the cybersecurity aspects of AI are receiving a lot of attention, evidence shows that the use of AI for and against cybersecurity is rapidly developing.

Cybersecurity of AI

The increasing integration of AI into our daily lives requires that we pay special attention to protecting models, data, training and deployments related to its use. Cybersecurity is the precondition for reliable, secure and resilient AI models and algorithms. However, cybersecurity of AI is not just about protecting AI systems against threats such as poisoning and evasion attacks. It also involves ensuring they have **trustworthiness** features such as human oversight and **robustness** – the ability to resist cyber-attacks, as required by the EU's AI Act for high-risk AI systems. The need for human oversight of AI has also been emphasised by experts.

Standards could play a crucial role in ensuring that security requirements – on matters such as data quality, risk management and conformity assessment – are integrated into the entire life cycle of AI systems. While they provide guidelines on safe, ethical and responsible AI development, the development of AI-specific technical European standards has only just begun, and EU stakeholders are eagerly awaiting their adoption. However, creating standards for a wide range of AI systems, which are essentially black boxes, is a challenging task that requires more work. In May 2023, the European Commission asked the European Committee for Standardisation and the European Committee for Electrotechnical Standardisation (CEN-CENELEC) to develop standards in support of the AI Act, with a deadline set for April 2025. Alongside CEN-CENELEC (JTC 13 and JTC 21 groups), several standards-developing organisations, including the European Telecommunications Standards Institute (ETSI) and the International Organization for Standardization (ISO), are also working on developing AI standards. Although most harmonised AI-specific standards have yet to be established, general-purpose standards for information security (such as ISO/IEC 27001 and ISO/IEC 27002) and quality management (such as ISO/IEC 9001) are transposed and can be partially applied to AI systems.

In the absence of standards specific to AI cybersecurity, several governmental agencies have published voluntary **AI security frameworks** to assist stakeholders in securing their AI systems, operations and processes. For example, the EU Agency for Cybersecurity (ENISA) has published a multilayer security framework for good AI cybersecurity practices (FAICP). The FAICP provides a gradual approach to enhancing the trustworthiness of AI activities. It consists of three layers: the groundwork of cybersecurity, focusing on the ICT infrastructure used; AI-specific aspects, focusing on the specificities of the AI components deployed; and sectorial AI, which is specific to the sector in which AI is being used.

Similarly, the US National Institute of Standards and Technology (NIST) has published an AI risk management framework to help organisations involved in the design, development, deployment or use of AI systems to better mitigate the risks associated with AI and contribute to its trustworthy and responsible

Author: Polona Car with Tristan Marcelin, Members' Research Service
PE 762.292 – April 2024

development and use. While some think tanks appreciate the framework and its non-binding nature, others recommend establishing a regulatory framework to enhance US national security.

In addition, the AI Safety Summit in November 2023 resulted in the publication of a short guide for secure AI system development, providing developers with a set of recommendations for all steps of the AI development cycle: design, development, deployment, operation and maintenance. In parallel, private companies such as Google, IBM (for generative AI only), Open AI, Amazon and KPMG have published their own frameworks for securing AI systems.

AI in support of cybersecurity

An increasing number of companies, such as IBM, Google and Microsoft, have started advertising and showcasing ways in which AI can be used to enhance cybersecurity. Advertised use cases fall into four categories: detection, prediction, analysis and threat mitigation.

AI systems can **detect** threats and vulnerabilities. In terms of threats, research shows that machine learning technologies are capable of detecting malware. Google suggests that AI could reverse the dynamic known as the Defender's Dilemma, meaning that AI can help cybersecurity professionals to scale their work in threat detection. Google also claims that generative AI has contributed to a '51 % time saving and higher quality results in incident analyst output to their internal detection and response efforts'. In terms of vulnerabilities, Google claims that its generative AI model, Gemini, has significantly helped in detecting new vulnerabilities.

In addition to detection, AI systems are capable of **predicting** threats and risks. ENISA suggests using AI to report risks of service outage in the context of the internet of things (IoT). Google reports that AI can make predictions based on a dataset of malicious uniform resource locators (URLs) linked to its enhanced safe browsing tool.

AI technologies are also capable of analysing code and classifying malware. VirusTotal, a popular malware multi-scanning tool owned by Google, has illustrated how generative AI can improve the understanding of a given malware and possibly prevent false-positive detections (i.e. files inaccurately flagged as malware by antivirus programmes). Improving accuracy in threat detection contributes to efficiency in responding to actual threats. Additionally, ENISA suggests that genetic algorithms, a type of evolutionary AI algorithms, could be used for malware classification.

Lastly, AI can assist in **threat mitigation**. AI-powered solutions automate incident response capabilities, thus speeding up the response time. They can prioritise threats, identify trends and contribute to predicting future threats. For example, Google claims that Gemini has successfully fixed 15 % of discovered bugs.

Malicious use of AI

However, AI can also be used for cyber-attacks. It can assist malicious actors in performing known attacks such as disinformation campaigns and malware coding. According to ENISA, AI systems are becoming particularly powerful in social engineering techniques thanks to their ability to mimic human interaction.

Microsoft recently reported that threat actors are actively using available large language models (LLM) to design their attacks. While Microsoft's and Open AI's threat intelligence systems have not identified 'novel or unique AI-enabled attack or abuse techniques resulting from threat actors' usage of AI' so far, they highlight the importance of continued close monitoring to detect any incremental attempts and stay ahead of evolving threats. Despite stating that GPT-4 'offers only limited, incremental capabilities for malicious cybersecurity tasks beyond what is already achievable with publicly available, non-AI powered tools', the March 2023 GPT-4 technical report identified the potential for 'risky emergent behaviours'. The report reveals that these novel capabilities are characteristic of more powerful models and range from phishing attacks to using humans – TaskRabbit workers – to complete simple tasks (e.g. solve a CAPTCHA). According to experts, they exhibit a power-seeking behaviour, which could be detrimental to cybersecurity.

In addition to known attacks, AI can be leveraged to create powerful **new types of attacks**. Google's researchers have shown how AI can be used to understand advanced cryptographic patterns and predict information that can be exploited. They have published a research paper presenting a powerful AI tool capable of attacking multiple cryptographic algorithms. The paper concludes by emphasising the 'pressing need to devise new protections that are resilient to deep-learning attacks'.