DOI: 10.2478/ie-2024-0004

sciendo

Intereconomics, 2024, 59(1), 10-13
JEL: L38, L51, Z18

Carolina Polito and Lorenzo Pupillo

Artificial Intelligence and Cybersecurity

Artificial intelligence (AI) is increasingly being incorporated into business operations and is extensively used in various applications. In the field of cybersecurity, the role of AI is becoming vital for managing cyberthreats. The AI market in cybersecurity is projected to grow at a compound annual growth rate of 21.9% between 2023 and 2028, attributed to the growing revenue of US $60.6 billion by 2028 (Marketsandmarkets, 2023). Nonetheless, AI adoption is not without its risks.

AI, being a versatile and dual-purpose technology, presents both opportunities and challenges in cybersecurity.

The role of AI in cybersecurity is akin to a double-edged sword (Taddeo et al., 2019). On the one hand, it offers advanced tools for enhancing security measures, detecting threats more efficiently and responding to attacks swiftly. These capabilities stem from the ability of AI to analyse large volumes of data quickly and identify patterns that might indicate a security breach. On the other hand, AI also empowers cybercriminals by providing them with sophisticated methods to execute attacks. Machine learning and deep learning are facilitating more sophisticated and damaging cyberattacks that are faster, more targeted and more destructive. The influence of AI on cybersecurity is expected to broaden the threat landscape, introduce new threats and change the typical nature of threats (Brundage et al., 2018). Besides, AI systems are not only vectors for attacks but are also vulnerable to manipulation.

Hence, AI is employed for offensive (supporting malicious attacks) and defensive (combating cybersecurity risks) purposes. However, while the defensive use of AI faces regulatory constraints, especially with governments, including the European Union, aiming to regulate high-risk applications and encourage responsible AI use, its offensive use is growing more prevalent. The cost of developing applications is decreasing, and the "attack surface" is expanding, making defence increasingly difficult.

Manipulation risks to AI systems

Manipulation of an AI system can occur in various forms, such as input attacks and poisoning attacks.

Input attacks focus on the data fed into the machine learning (ML) system. In these attacks, the attacker introduces an attack pattern into the input data, like placing tape on a stop sign or subtly altering the pixels in an image. These modifications alter the way the ML system interprets the data, leading to its failure.

Poisoning attacks, in contrast, target the development stage of the ML system. These attacks aim to disrupt the construction of a viable ML model by compromising its training phase. The attacker manipulates the training data or process, resulting in a deployed ML model that is fundamentally flawed from the outset. Poisoning attacks occur during the phase where the model's parameters and learning processes are being established. To poison the ML system, the attacker compromises its training and learning process so that it can perform the tasks that are requested by the attacker (Comiter, 2019).

The need for reliable AI

Against this backdrop, the distinction between trustworthiness and reliability in AI systems is of paramount importance. Indeed, it would be crucial to ensure the robustness of an AI system, so that it continues to behave as expected even when the inputs or model are perturbed by an attack. Assessing the robustness of a system would require testing for all possible input disruptions, which would present a significant challenge primarily due to the vast number of potential perturbations that can affect their performance.

The sheer number of potential disturbances in real-world applications makes exhaustive testing unfeasible. This complexity, in turn, makes it nearly impossible to fully test and validate AI systems against all possible scenarios. Such limitations hinder our ability to fully trust AI systems.

Open Access funding provided by ZBW – Leibniz Information Centre for Economics.

Carolina Polito, Luiss Guido Carli, Rome, Italy; Centre for European Policy Studies, Brussels, Belgium.

Lorenzo Pupillo, Centre for European Policy Studies, Brussels, Belgium.

As Taddeo et al. (2019) point out, the inability to thoroughly assess AI robustness directly impacts the trustworthiness of these systems. Philosophical analyses qualify “trust” as the act of delegating a task without exerting any form of control or supervision over its execution. In essence, trust implies a reliance on the system to perform its task autonomously.

Therefore, if trustworthiness is about gauging the probability that the trustee – here the AI system – will behave as expected, the unpredictable nature of attacks on these systems makes it nearly impossible to reliably determine their consistent performance in varied contexts.

Consequently, it becomes essential to implement certain control measures to ensure the reliable functioning of these systems. Policies should mandate a cautious evaluation of the level of control on an AI system based on the tasks against which it was designed. Therefore, the concept of reliability should be preferred to trustworthiness.

Ensuring effective human control over AI systems necessitates a significant degree of expertise from human operators, who must be capable of exercising effective oversight. This involves a comprehensive understanding of how to manage various degrees of automation. This task is fraught with challenges, as exemplified by the half-automation problem (Lorenzo et al., 2022), a phenomenon that arises when tasks are highly automated but not entirely so, leading human operators to overdepend on the AI system as though it were fully autonomous.

The half-automation problem highlights a critical issue in the interaction between humans and AI systems. Operators may become complacent or less vigilant when they perceive that AI is more capable or autonomous than it is. This over-reliance can lead to a decrease in human engagement and a potential increase in errors, especially in situations where human intervention is still crucial. Addressing this problem requires careful design and training strategies that emphasise the importance of maintaining human engagement and awareness in semi-automated systems.

Relevant would be the inclusion of “kill switches” in AI systems, as they provide a mechanism for humans to override automated processes and regain control, particularly in the context of automated responses. However, the feasibility of kill switches as a universal solution is debatable. In certain scenarios, such as with autonomous vehicles or aircraft autopilot systems, the decision about when and how to transfer control back to a human operator remains a complex and unresolved issue. Research into handover scenarios, where control shifts from the AI system back to the human, suggests that abrupt requests for human intervention should generally be avoided (Zhang et al., 2019). The reasoning is that in emergent situations, a driver or pilot may not have the capacity to rapidly assess the situation and take appropriate action if only given a few seconds to respond. Hence, in high-stakes environments, where split-second decisions are crucial, such as in critical medical systems or military applications, the design of kill switches must consider the operator’s ability to respond accurately under pressure. While applying kill switches in AI systems is, therefore, non-trivial, some form of control over those systems, such as security gates, must be established.

More generally, supporting the reliability of AI implies envisaging forms and degrees of operational control adequate to the learning nature of the systems and the dynamic nature of the attacks, but also attainable in terms of resources, especially time and computational feasibility. This implies ensuring a degree of control over the data fed into the algorithm and over the software. This means having cybersecure pedigrees for the data libraries used for training any machine learning algorithms used as well as cybersecure pedigrees for all software libraries linked to that code (Lorenzo et al., 2022).

Cybersecurity and generative AI

This very complex landscape of regulatory challenges has been further complicated with the advent of generative AI. Generative AI is a form of artificial intelligence technology that can produce various types of content such as text, imagery, audio and synthetic data. It starts with a prompt like a text, an image, a video, a design, musical notes, or any input that the AI system can process. Then, various AI algorithms return new content in response to the prompt (Lawton, 2023). It does this by learning patterns from existing data and then using this information to generate new and unique outputs.[1] Therefore, contrary to other forms of AI used for purposes such as analysing data or helping in airplane flight control, the purpose of generative AI is to generate new and original content.

This form of AI is called large language models (LLMs) and entails the use of machine learning to understand and generate human language. These models are trained on large amounts of text data, such as books, articles, and websites, (this is why they are called large) and are capable of generating new text that is similar to what a human would write or say.

[1] See https://generativeai.net.

In particular, LLMs use a subset of machine learning architecture deep learning techniques, specifically neural networks, to learn the patterns and structure of languages. The larger the model, the more data it has been trained on and the better it is at understanding and generating human language.

Generative AI is becoming very popular because of programmes such as ChatGPT or AI image generator DALL-E and is showing great potential – but also peril as its use in cybersecurity is a double-edged sword.

Among the promises of generative AI in cybersecurity is the possibility of automating repetitive tasks, freeing up time for cybersecurity experts to focus on more complex issues; identifying anomalies and trends that could be neglected by human analysts; creating cyberthreat scenarios for training and simulation purposes; and predicting future cyberthreats based on historical data (Boopathy, 2023).

However, generative AI with its ability to create content, from text to images and videos, introduces an entirely new dimension of cybersecurity risks. Deepfakes, or AI-generated videos that can mimic real individuals, have already highlighted the potential misuse of this technology: the use of deepfake for committing a new type of identity fraud or the use of tools like ChatGPT to create more powerful phishing programmes. Generative AI models could be used to create malware, for instance, polymorphic malware, able to change its code to bypass detection by security tools.

Furthermore, some recent research shows that the widespread integration of LLMs like ChatGPT in applications such as search engines has generated critical vulnerabilities that "when coupled with how they are developed and distributed by commercial providers and as open-source releases risk creating a systemic cybersecurity crisis" (Tsamados, 2024).

Indeed, similarly to what has been described for "classic" AI systems, two types of threats could lead to a cybersecurity crisis: those generated by LLMs' vulnerability to attacks conducted via natural language, and those deriving from how the LLM models are developed and distributed.

The first category is characterised by the natural language as a universal attack vector and entails attacks such as prompt injection attacks or attacks based on the automatic generation of adversarial suffixes. Furthermore, natural language can also be used for more complex attacks like model poisoning.

Prompt injection attacks are made possible by end-users injecting adversarial instructions – through written, audio, image, or video format – into an AI chatbot interface, or through data accessed by the model itself, for example, through web searches. The outcome is the hijacking of the model and the control of its outputs.

The attacks based on the automatic generation of adversarial suffixes are more sophisticated and require more skills and knowledge to be implemented.

Conclusions

Cyberattacks are on the rise, and they are increasingly using AI. Overall, the barrier to entry for cyberattackers has been significantly lowered, implying that even individuals with average technical skills could potentially instigate a prompt injection attack, particularly with Generative AI technologies. The gamification of such techniques on social media and forums can lead to a collaborative effort among users to explore and exploit these vulnerabilities, potentially leading to an increase in the number and variety of cyberthreats. This democratisation of hacking capabilities poses a significant challenge for cybersecurity professionals and necessitates the development of more robust defences against these emerging forms of cyberattacks.

While AI could even help companies manage cybersecurity risks, a number of conditions must be met. It is essential to develop ad hoc cybersecurity practices to mitigate the threats stemming from AI adoption, such as building AI-specific threat models based on the mapping of the different LLM vulnerabilities and threats. Furthermore, authorised and qualified auditors should be allowed to have regular access to LLM providers and data sets to avoid a single point of failure.

In terms of research, as mentioned by the latest ENISA (2023) report on AI and cybersecurity, further studies should be promoted on the application of AI and ML in cybersecurity. These include, among others, creating test beds to evaluate and enhance the performance of ML tools and technologies in cybersecurity, developing penetration testing tools that can discover and leverage security vulnerabilities through the simulation of attackers' behaviour, and creating standardised frameworks for assessing the confidentiality of information flows.

In light of these developments, achieving consensus on the regulation of generative AI systems within the trialogue negotiations of the AI Act has been key. On 9 December 2023, the Council Presidency and the European Parliament's negotiators reached a provisional agreement on the AI Act proposal.

Distinct guidelines have been established for foundation models. The provisional agreement mandates that foundation models adhere to particular transparency requirements before entering the market. A more rigorous framework is applied to "high impact" foundation models, characterised by being trained with extensive data sets, possessing advanced complexity and superior performance, and having the potential to propagate systemic risks throughout the value chain (Council of the EU, 2023).

However, the definitive measure of the AI Act's success will be observed through its execution. Although a provisional agreement on the AI Act's proposal has been established, complete with specific provisions for foundation models and stringent requirements for high-impact foundation models, the real test lies in the implementation phase. It is during this stage that stakeholders will discern the practicality and efficiency of the Act in regulating generative AI systems. The effectiveness of the AI Act will ultimately be judged by its ability to mitigate risks and ensure compliance with transparency obligations. One critical area in which the EU could have further impact on the premise of what has been established in the AI Act would be the creation of a specific unit focused on curating large-scale data sets (CEPS Think Tank, 2023). This would allow the EU to foster a more secure AI ecosystem.

References

Boopathy, J. (2023, 4 July), Generative AI in Cybersecurity, An Overview, *Medium*.

Brundage, M., S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel, A. Dafoe, P. Scharre, T. Zeitzoff, B. Filar, H. Anderson, H. Roff, G. C. Allen, J. Steinhardt, C. Flynn, S. Ó. hÉigeartaigh, S. Beard, H. Belfield, S. Farquhar, C. Lyle, R. Crootof, O. Evans, M. Page, J. Bryson, R. Yampolskiy and D. Amodei (2018), The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation, *Malicious AI Report*.

CEPS Think Tank (2023), Cybersecurity@CEPS SUMMIT 2023, https://www.youtube.com/watch?v=RnrE_ePm9bI (29 January 2024).

Comiter, M. (2019), Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do About It, Belfer Center for Science and International Affairs, Harvard Kennedy School.

Council of the EU (2023, 9 December), Artificial intelligence act: Council and Parliament strike a deal on the first rules for AI in the world, Press release, https://www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/ (29 January, 2024).

ENISA (2023), Artificial Intelligence and Cybersecurity Research, *ENISA Research and Innovation Brief*, European Union Agency for Cybersecurity.

Lawton, G. (2024), What is generative AI? Everything you need to know, *TechTarget*, https://www.techtarget.com/searchenterpriseai/definition/generative-AI (20 January 2024).

Lorenzo, P., F. Stefano, A. Ferreira and P. Carolina (2022), Artificial intelligence and cybersecurity: Technology, governance and policy challenges, Centre for European Policy Studies.

MarketsandMarkets, Artificial Intelligence in Cybersecurity Market by Offering, Security Type, Technology, Application, Vertical and Region – Global Forecast to 2028.

Taddeo, M., T. McCutcheon and L. Floridi (2019), Trusting artificial intelligence in cybersecurity is a double-edged sword, *Nature Machine Intelligence*, 1(12), 557-560.

Tsamados, A., L. Floridi and M. Taddeo (2024), *The Cybersecurity Crisis of Artificial Intelligence: Unrestrained Adoption and Natural Language-Based Attacks*, mimeo.

Zhang, B., E. S. Wilschut, D. M. C. Willemsen and M. H. Martens (2019), Transition to manual control from highly automated driving in non-critical truck platooning scenarios, Part F: Traffic Psychology Behaviour, *Transportation Research*, 64, 84-97.