Think7 ITALY 2024

G7 ITALIA 2024

Policy Brief

May 2024

Task Force 4: Science and Digitalization for a Better Future

Artificial Intelligence and Cybersecurity: Balancing Dual-Use Challenges and Embracing Opportunities for a Secure Future

Simona Autolitano, Center for Advanced Security, Strategic and Integration Studies (CASSIS), Germany

Abstract

The integration of artificial intelligence (AI) into cybersecurity frameworks offers immense potential for proactive threat detection and mitigation. However, this advancement also presents significant challenges, including the potential misuse of AI in cyberattacks. After a brief introduction to the dual-use nature of AI from a cybersecurity perspective, this policy brief argues for a collaborative approach among G7 nations to address these challenges and maximise the benefits of AI in cybersecurity. It makes recommendations for improving information sharing, establishing common standards, ensuring the trustworthy operation of AI systems, and engaging the young developer community.

1. The challenge: AI for cybersecurity or cybersecurity for AI?

The rapid evolution of large language models (LLMs) has marked a turning point in the current era, significantly increasing public interest and engagement with artificial intelligence (AI) technologies. The model's text generation capabilities have not only surprised the public but have also exceeded the expectations of experts. Since the release of the ChatGPT application by OpenAI, backed by Microsoft, in November 2022, there has been significant development in technology and expanding applications across various domains for this and similar models, including in the cybersecurity domain.

The incorporation of AI into cybersecurity frameworks offers significant opportunities to improve proactive threat detection and mitigation. However, alongside these advances come notable challenges, including the potential misuse of AI in cyberattacks.

Although the G7 countries are engaging with this topic, there is little emphasis on the benefits of AI for the cybersecurity domain. This policy brief highlights the significance of a collaborative approach among the G7 to facilitate a more promising future for AI applications in the cybersecurity sector.

1.1 Enhancing cybersecurity protection with LLMs

The use and application of LLMs has transformed the cybersecurity landscape (Motlagh et al. 2024). Indeed, from a cybersecurity perspective, predictability and the ability to learn and adapt are the most important innovations in traditional cybersecurity approaches.

AI systems are, by definition, agile, constantly learning from new data to improve their detection and response capabilities. As a result, these models offer predictive capabilities that enable the identification of potential cyber threats before they manifest into attacks, thus changing the traditional cybersecurity posture.

A number of studies are currently investigating the potential of LLM for cybersecurity. For the sake of simplicity, we can refer to the NIST framework to explore the opportunities for AI in cybersecurity. The NIST Cybersecurity Framework is a set of guidelines for mitigating organisational cybersecurity risks, published by the US National Institute of Standards and Technology (NIST) and it is based on existing standards, guidelines and practices.

It is one of the most commonly used frameworks in cybersecurity based on five pillars. LLMs have the opportunity to contribute to each of these pillars.

The first pillar involves identifying the critical functions of an organisation and the cybersecurity risks that could disrupt those functions. LLMs can play a critical role in the "Identify" function in

the context of the NIST framework by providing advanced insights and analysis, thanks to their ability to process an incredible amount of data. For example, it can support decision making by creating a risk matrix that categorises risks by severity.

The second pillar focuses on mitigating the potential impact of a cybersecurity breach. Once critical functions have been identified, cybersecurity measures are prioritised and implemented accordingly. In the "Protect" function, LLMs enable networks to anticipate and prevent problems in advance. Some experiments have used LLMs to improve web content filtering by increasing the accuracy of categorising large volumes of URLs (Vörös et al. 2023). In addition, AI can improve a system's robustness – its ability to continue to behave as expected even when it processes incorrect inputs (Taddeo et al. 2019). In practice, LLMs can automatically clean up the training data by identifying and possibly correcting corrupted codes to ensure it still meets the original security requirements. If the code fails to meet these requirements after modification, it would not be processed further. This ensures that any changes made by the LLM maintain or enhance the security of the codebase.

The third pillar is the "Detection" function, which serves to assess whether a system has been compromised so that action can be taken if necessary. As a number of academics have pointed out, LLMs can increase a system's resilience by facilitating threat and anomaly detection and assisting security analysts in retrieving information about cyber threats (Taddeo et al. 2019).

The fourth pillar is the "Respond" function which aims to minimise damage by facilitating a rapid response. LLMs can enhance a system's responsiveness, which means its ability to autonomously defeat an attack and refine future strategies based on the success achieved by generating decoys and honeypots for attackers (Taddeo et al. 2019). Furthermore, as some studies have shown, combining LLMs with honeypots – a cybersecurity mechanism designed to lure potential attackers – makes it easier to deal with computer viruses, such as malware, and other threats (Motlagh et al. 2024: 4).

Finally, the fifth pillar is the "Recovery" function, which aims to recover any data that may have been lost as a result of a breach or attack. As some researchers have pointed out, too little research has been done to really understand how LLMs can be used to improve this function (Motlagh et al. 2024). However, some experiments have already shown that LLMs can play a crucial role in improving data backup and recovery processes for businesses (Huang et al. 2024).

1.2 Strengthening LLMs with cybersecurity protection

While AI enhances cybersecurity defences, it also introduces new vulnerabilities and opportunities for exploitation by malicious actors. The dual nature of AI poses a significant risk, requiring proactive measures to mitigate potential threats.

The German Federal Office for Information Security (BSI) divides this risk into three main categories (BSI 2024).

One group of risks arises due to the probabilistic nature of LLMs, as they generate text based on so-called "stochastic correlations", which means that the prediction derives from a random distribution. This, however, may not guarantee factual accuracy. The creation of content, which is not part of the input or of the dataset used for training the model, is known as "hallucination". Hallucinations can be difficult to detect because the high linguistic quality of the generated texts makes the results convincing. Furthermore, the lack of reproducibility and up-to-date information in the output, as well as potential security gaps in the generated code, may reinforce deviations from the training data. The probabilistic nature of LLMs means that caution should be exercised when using them.

A second group of risks is based on misuse. They are useful tools for criminals due to their ability to generate output in different languages and imitate the writing styles of individuals or organizations. They can be utilised to create content for social engineering or disinformation. Furthermore, LLMs can also be used to write or improve malicious codes, in the same way they are used to improve the content of our emails. There are already a number of examples of this. For example, WormGPT is an AI-driven tool designed specifically for cybercriminals that automates the creation of personalised phishing emails. FraudGPT, meanwhile, enables attackers to create convincing content to trick users into clicking on specific links (Motlagh et al. 2024).

A third set of risks comes from attacks to the LLMs that can take the form of so-called "prompt injections" or "indirect prompt injection". In the former, the behaviour of the model, and therefore the output, can be modified by inserting specific text inputs into the model. In the latter case, if LLMs can access external content such as websites, attackers can access these to place instructions that are executed when the website is evaluated by the model, thus changing the model's behaviour.

While many LLMs or LLM-based applications have measures in place to filter and eliminate errors, these usually provide only partial protection against misuse and attack scenarios, and offer little protection against hallucinations of the LLMs.

This is where the role of the G7 should be: to ensure that new technologies and developments are embraced, while at the same time developing appropriate measures to strengthen their own cybersecurity protection, thus limiting the risks of their application.

2. Securing tomorrow: The role of the G7 and the way forward

Cybersecurity is a top priority on the political agendas of all G7 nations. It is not just important, but an absolute necessity. In today's interconnected society, where every facet of life hinges on digital infrastructure, the implementation of robust cybersecurity measures is indispensable for ensuring the seamless operation of all facets of society.

In the realm of cybersecurity on a global scale, the G7 nations stand as pivotal players: France, the United States (US), the United Kingdom (UK), Germany, Japan, Italy, and Canada, along with the European Union (EU), are at the forefront of technological innovation and possess formidable cybersecurity capabilities.

The G7 leaders have emphasised in the past years the potential of advanced AI systems also in the cybersecurity domain, while acknowledging the need to manage associated risks and protect societal values (Taddeo et al. 2019).

However, while all the G7 countries possess significant expertise and cybersecurity agencies that are doing great work studying the intersection between AI and cybersecurity, cooperation in this area has been limited.

The so-called "Hiroshima AI Process" is probably one of the most relevant steps in this direction. Initiated in May 2023, the Hiroshima AI Process seeks to establish global standards for regulating advanced AI systems. It has been successful particularly in reaching agreement among the G7 nations on International Guiding Principles alongside a Code of Conduct tailored for AI developers in October 2023 (G7 2023). Several critical facets of AI governance have been highlighted in these documents, including the need for a risk-based approach to be followed throughout the AI lifecycle. It also recognises the need for continuous monitoring, reporting and mitigation of misuse and incidents, as well as the need to establish risk management protocols, practices and robust security measures for AI systems (Habuka 2023).

A second very important development is the new "Guidelines for secure AI system development", published in November 2023, which will help developers make informed decisions about the design, development, deployment and operation of their AI systems (NCSC and CISA 2023). The Guidelines, published jointly by the UK's National Cyber Security Centre (NCSC) and the US's Cybersecurity and Infrastructure Security Agency (CISA) have been already agreed by all the G7 cybersecurity leading agencies or authorities.

In addition to these efforts, it is worth mentioning that the leading G7 cybersecurity agencies are increasingly engaging in bilateral and, in some cases, trilateral cooperation. However, this cooperation is not yet focused on AI and cybersecurity, but is much broader in nature.

Within the EU, for example, there has long been a particularly close cooperation between the German BSI and the French ANSSI (BSI 2018). The same is true for the agencies of the UK, the US and Canada. With the recent establishment of the AI Safety Institute in Japan in February 2024, international cooperation with related organisations is also expected to increase (Yomiuri Shimbun 2023).

Drawing on the work and networks of leading cybersecurity agencies, this policy brief advocates for increased collaboration among the G7 to ensure a secure future for AI applications in the cybersecurity community.

The G7 nations are well-positioned to spearhead collaborative efforts in leveraging AI for enhanced cybersecurity. By working together, they can facilitate information exchange, pool resources, and establish unified standards for the ethical and secure use of AI in cybersecurity operations. Additionally, the G7 can serve as a model for international cooperation, inspiring other nations to adopt similar approaches to address emerging cybersecurity challenges.

3. Recommendations to the G7

Each of the G7 nations has already undertaken various activities and initiatives in the field of AI. These efforts are united by a common goal: to foster greater trust in AI systems. To navigate the dual nature of AI models and ensure their effective use for cybersecurity purposes, the following steps are essential:

1. Foster informal cooperative dialogues among G7 cybersecurity bodies: It is suggested that the G7 nations promote regular exchanges between their leading cybersecurity agencies to facilitate continuous updates on the use of AI services and products. These exchanges should include the dissemination of best practices, insights into emerging threats, and joint research ventures aimed at improving AI-driven cybersecurity solutions. To maximise effectiveness, it is recommended that such exchanges take place within an informal framework. For example, in conjunction with the G7 presidency, the relevant cybersecurity authority within each nation should lead and coordinate regular meetings to address key AI cybersecurity issues. The following issues should be addressed in such meetings:

1.1 Initiate discussions around the possibility of developing a comprehensive common indicator to assess "trustworthiness" of AI services and products: In such a setting, G7 cybersecurity bodies should start a conversation to develop comprehensive indicators to assess the "trustworthiness" of AI services and products. These indicators should go beyond technical specifications and encompass broader societal, economic, and geopolitical factors. By evaluating AI systems from multiple perspectives, policymakers can better identify potential risks and ensure the responsible deployment of AI in cybersecurity operations. At present, most, if not all, of the G7 cybersecurity bodies are already working to define

"trustworthiness", but mostly in isolation, but cooperation is crucial.

1.2 Sustain efforts for securing AI operations by discussing the "reliability" of AI services and products: Sustain ongoing efforts to formulate criteria and international standards for the secure operation of AI language models and AI systems at large. Collaboration among G7 nations to collectively work on and evolve standards for secure AI services and products is crucial for maintaining a unified and robust approach to cybersecurity challenges in an increasingly interconnected world. This involves building upon existing initiatives, such as the EU's AI Act and the guidelines proposed by leading cybersecurity agencies, such as the UK's "Guidelines for secure AI systems development".

1.3 Encourage information-sharing: An essential aspect for sustaining efforts for securing AI operations is information sharing on current AI-related threats and vulnerabilities. To this end, information sharing should regularly take place among G7 cybersecurity bodies. This should include not only information on the presence and exploitation of a particular vulnerability, but also an assessment of the impact of that vulnerability on the security of society at large, including the impact on the business sector, on consumers and, where appropriate, on the government.

2. Advance joint research and innovation to better identify the needs for research on AI for cybersecurity and on securing AI: Collaboration between public and private actors in research and innovation in this area is essential. The set up of public-private laboratories for AI testing could be one way to advance research and innovation in the technological field, especially in the security realm. Such research still occurs in silos, with governments and industries developing the best possible solutions with little or no collaboration between them. AI-based cybersecurity solutions should be developed as joint projects between governments and industries, within a framework of public-private laboratories for testing. This will also ensure that any government funding for research can be applied in real-life situations.

3. Engage with the developers' community and beyond to raise awareness of developers: It is important to engage with the developer community, including those in the formation phase, from the outset. The aim is not to make developers the only responsible for ensuring the security of AI systems and applications, but rather to provide them with the necessary tools and knowledge to enable the development of secure AI applications by design. Recent research conducted by Professor Matthew Smith at the University of Bonn, in the Computer Science curriculum, has demonstrated that courses on cybersecurity and privacy positively impact students' programming skills (Gorski et al. 2023).

It is important to recognise that the differing cybersecurity views held by G7 nations present challenges to cooperation. These may include, for instance, differing security priorities, legal frameworks, technological capabilities, and geopolitical factors. To overcome these, it is necessary to engage in sustained dialogue, build trust, and align on common cybersecurity objectives, including information sharing, capacity building, and norm development.

However, as the complexity of cyber threats increases, with the advent of AI, the G7 nations must take the initiative to integrate AI into cybersecurity frameworks. By adopting a collaborative approach, the G7 nations can capitalise on the potential of AI to enhance cybersecurity while addressing concerns related to misuse and vulnerabilities. However, it is imperative that they commence these efforts without delay. The recommendations set forth in this policy brief serve as a roadmap for fostering international cooperation and establishing a secure and resilient cybersecurity ecosystem in the age of AI.

References

Ansari, Meraj Farheen, et al. 2022. The impact and limitations of artificial intelligence in cybersecurity: A literature review. *International Journal of Advanced Research in Computer and Communication Engineering* 11(9): 81-90. https://doi.org/10.17148/IJARCCE.2022.11912

Babuta, Alexander, Marion Oswald MBE and Janjeva, Ardi. 2020. Artificial intelligence and UK national security. *RUSI Occasional Papers* April. https://static.rusi.org/ai-national-security-final-web-version.pdf

Bendiek, Annegret, and Stürzer, Isabella. 2022. Advancing European internal and external digital sovereignty: The Brussels effect and the EU-US Trade and Technology Council. *SWP Comments* 20. https://doi.org/10.18449/2022C20

BSI-Bundesamt für Sicherheit in der Informationstechnik. 2018. *Joint releases by ANSSI and BSI*. https://www.bsi.bund.de/EN/Service-Navi/Publikationen/ANSSI-BSI-joint-releases/ANSSI-BSI-joint-releases.html

BSI-Bundesamt für Sicherheit in der Informationstechnik. 2024. *Artificial intelligence*. https://www.bsi.bund.de/EN/Themen/Unternehmen-und-Organisationen/Informationen-und-Empfehlungen/Kuenstliche-Intelligenz/kuenstliche-intelligenz_node.html

Calderon, Ricardo. 2019. The benefits of artificial intelligence in cybersecurity. *Economic Crime Forensics Capstones* 36. https://digitalcommons.lasalle.edu/ecf_capstones/36

G7. 2023. *G7 leaders' statement on the Hiroshima AI Process*. http://www.g7.utoronto.ca/summit/2023hiroshima/231030-ai.html

Gorski, Peter Leo, Lo Iacono, Luigi, and Smith, Matthew. 2023. Eight lightweight usable security principles for developers. *IEEE Security & Privacy* 21(1): 20-26. https://doi.org/10.1109/MSEC.2022.3205484

Habuka, Hiroki. 2023. Japan's approach to AI regulation and its impact on the 2023 G7 Presidency. *CSIS Reports* February. https://www.csis.org/node/103948

Huang, Bin, et al. 2024. FirewaLLM: A portable data protection and recovery framework for LLM services. In Tan, Ying, and Shi, Yuhui, eds. *Data mining and big data*: 16-30. Singapore: Springer Nature. https://doi.org/10.1007/978-981-97-0844-4_2

Motlagh, Farzad Nourmohammadzadeh, et al. 2024. Large language models in cybersecurity: State-of-the-art. *arXiv* 30 January. https://doi.org/10.48550/arXiv.2402.00891

NCSC-National Cyber Security Centre and CISA-Cybersecurity and Infrastructure Security Agency. 2023. *Guidelines for secure AI system development*. https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development

Novelli, Claudio, et al. 2024. Generative AI in EU law: Liability, privacy, intellectual property, and cybersecurity. *arXiv* 15 March. https://doi.org/10.48550/arXiv.2401.07348

Yomiuri Shimbun. 2023. Japan Govt to establish AI Safety Institute in January. *The Japan News* 21 December. https://japannews.yomiuri.co.jp/politics/politics-government/20231221-157027

Soni, Vishal Dineshkumar. 2020. Challenges and solution for artificial intelligence in cybersecurity of the USA. *SSRN* 15 June. https://doi.org/10.2139/ssrn.3624487

Taddeo, Mariarosaria, McCutcheon, Tom, and Floridi, Luciano. 2019. Trusting artificial intelligence in cybersecurity is a double-edged sword. *Nature Machine Intelligence* 1(12): 557-560. https://doi.org/10.1038/s42256-019-0109-1

Vörös, Tamás, Bergeron, Sean Paul, and Berlin, Konstantin. 2023. Web content filtering through knowledge distillation of large language models. *arXiv* 10 May. https://doi.org/10.48550/arXiv.2305.05027

About Think7

Think7 (T7) is the official think tank engagement group of the Group of 7 (G7). It provides research-based policy recommendations for G7 countries and partners. The Istituto Affari Internazionali (IAI) and Istituto per gli Studi di Politica Internazionale (ISPI) are the co-chairs of T7 under Italy's 2024 G7 presidency.