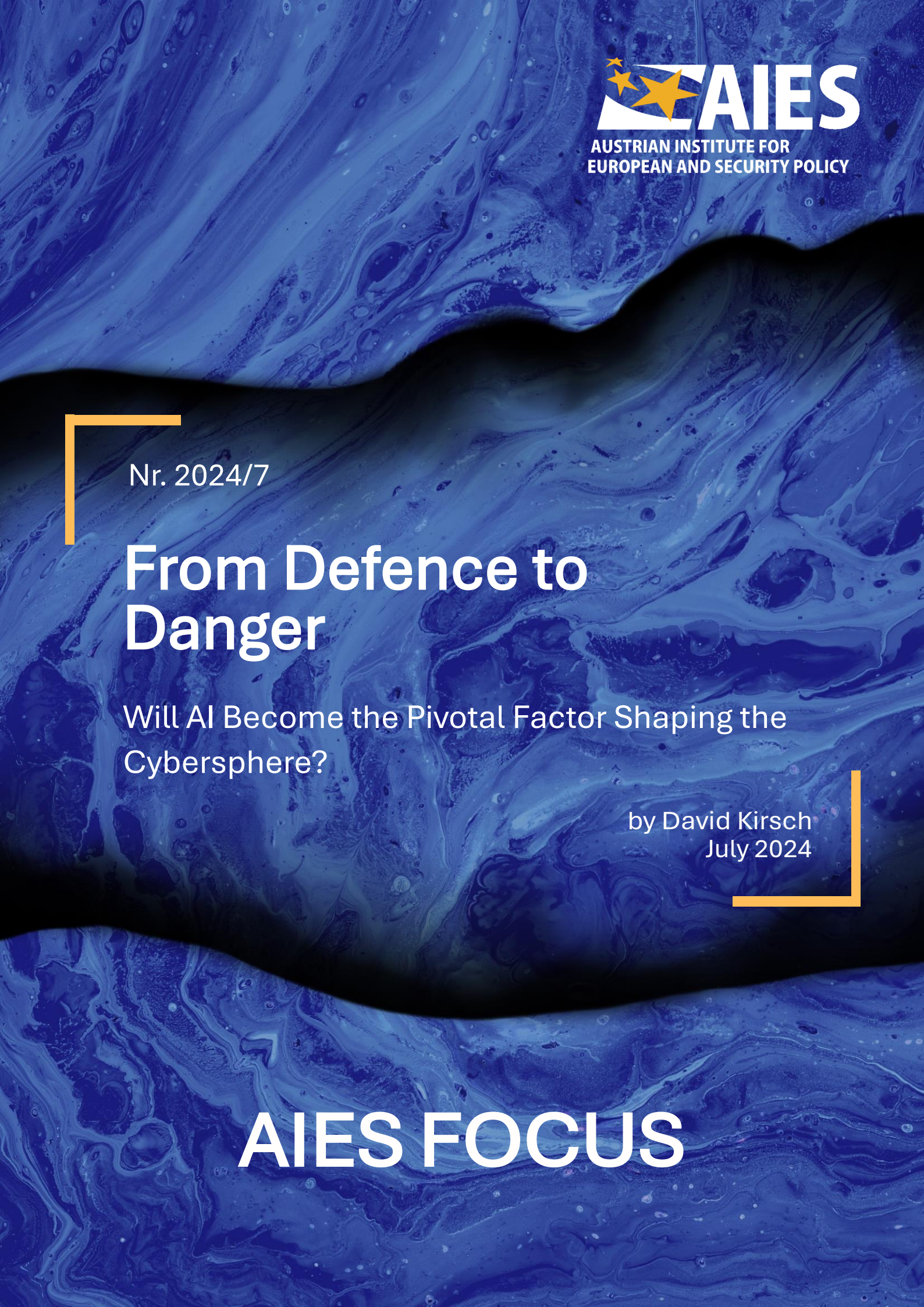

AUSTRIAN INSTITUTE FOR EUROPEAN AND SECURITY POLICY

Nr. 2024/7

From Defence to Danger

Will AI Become the Pivotal Factor Shaping the Cybersphere?

by David Kirsch
July 2024

AIES FOCUS

From Defense to Danger: Will AI Become the Pivotal Factor Shaping the Cybersphere?

Abstract

This AIES Focus explores AI's dual role in cybersecurity as both defender and aggressor. It analyzes AI integration through use cases such as threat intelligence, darknet monitoring, and behavioral analytics for defenders, and DDoS attacks, phishing, and malware design for attackers. The paper also highlights security risks posed by Large Language Models (LLMs), including prompt injection and adversarial attacks. It emphasizes the need for international cooperation and regulatory frameworks to balance AI's benefits and risks in cybersecurity, underscoring AI's significant impact on strategic cyber operations planning.

Introduction

The buzz surrounding artificial intelligence (AI) is nearing an inflection point[1], much akin to the trajectory described in Gartner's Hype Cycle[2]. Initially fueled by high expectations and broad applications, the excitement around AI is beginning to settle into a more mature phase of practical implementation. The realization that AI requires a strategic approach, rather than just piecemeal project initiatives, is becoming increasingly apparent[3]. In the long run, AI is expected to be as ubiquitous as other once-hyped innovations like machine learning—which should not be conflated with AI itself. These technologies have gradually become integral components of our digital sphere, demonstrating the normalization and assimilation of advanced technologies in everyday use.

Generative AI and Large Language Models (LLMs): Generative AI stands distinct from traditional machine learning due to its ability to create new content rather than merely analyze existing data[4]. Large Language Models (LLMs), such as OpenAI's GPT, exemplify this type of AI, demonstrating capabilities that extend beyond simple pattern recognition to include the generation of coherent and contextually relevant text based on vast datasets they have been trained on. LLMs represent a significant leap forward in how machines understand and interact with human language, offering unprecedented opportunities for automation and innovation.

The Role of LLMs in Cyber Defense and Offense: In the context of cybersecurity, LLMs hold a dual potential. On the defensive side, they can significantly enhance threat detection, automate incident responses, and streamline security operations [5]. Conversely, in offensive applications, LLMs can be employed to craft sophisticated cyber-attacks such as generating malicious phishing content or automating social engineering tactics. The strategic use of LLMs in cyber warfare underscores their potential to influence both protective and adversarial actions within the cybersphere [6].

AI for Defenders: Enhancing Cybersecurity with Large Language Models

Relevance of LLMs in Cyber Defense and the Role of SOCs: Large Language Models (LLMs) offer transformative potential for Security Operations Centers (SOCs), especially in their ability to process and analyze vast amounts of unstructured data efficiently. In a SOC, where timely and accurate threat detection is crucial, LLMs automate the analysis of threat intelligence, detect anomalies, and provide actionable insights more swiftly than traditional methods. Traditionally, SOCs serve as the nerve centers for managing security threats, with responsibilities including continuous monitoring, analysis, and response to potential security incidents. The integration of LLMs enhances these capabilities, allowing SOC teams to focus on strategic analysis and proactive threat hunting by automating routine tasks such as log analysis, alert generation, and incident categorization.

Applications of LLMs in Cyber Defense: The integration of Large Language Models (LLMs) within Security Operations Centers (SOCs) promises a transformative impact on cybersecurity. One of the most significant advancements is in automated threat intelligence analysis. LLMs can comb through extensive and diverse data sources, such as blogs, forums, and the dark web, identifying and track-

ing emerging threats with unmatched efficiency. For example, a LLM can be trained to automatically scan hacker forums for mentions of new exploits and vulnerabilities, providing SOC teams with early warnings that allow them to preempt potential threats [7]. This proactive stance is crucial in the fast-paced world of cybersecurity, where new threats can emerge and evolve rapidly.

In enhancing incident response capabilities, LLMs play a pivotal role by significantly reducing reaction times. Integrating these models allows SOCs to automate the initial steps of the incident response process[8]. When an alert is triggered, a LLM can quickly analyze it and compare it against known attack patterns. This rapid analysis enables the LLM to suggest the most effective response actions, thereby minimizing the time required to contain and mitigate the threat. Traditional methods often involve manual analysis, which can be time-consuming and prone to human error. LLMs, on the other hand, can swiftly parse through large datasets, identify patterns, and recommend immediate actions. This automation not only speeds up the response time but also improves the accuracy of the responses, leading to more effective containment and mitigation of threats.

Another critical application of LLMs is in the generation and prioritization of alerts. These models can substantially improve alert management systems by intelligently categorizing and prioritizing alerts based on their severity and potential impact. This approach ensures that SOC analysts can focus their efforts on the most critical issues first, optimizing the use of limited resources and improving overall response efficiency. In a typical SOC, analysts can be overwhelmed by the sheer volume of alerts generated by various security tools.

LLMs also provide invaluable assistance in forensic investigations[9]. They can process and correlate data from multiple sources, such as network traffic logs, past incident reports, and external threat intelligence. This comprehensive analysis allows LLMs to rapidly pinpoint the source and method of an attack, significantly aiding in understanding and mitigating breaches. By correlating data from various sources, LLMs can provide a comprehensive view of an attack, tracing its origins and identifying the methods used. This thorough understanding is crucial for developing effective countermeasures and preventing future incidents.

Moreover, LLMs enhance the continuous threat hunting capabilities of SOCs. By persistently analyzing network traffic, these models can detect subtle signs of breaches, such as unusual outbound data transfers or suspicious login attempts from rarely used locations[10]. This continuous monitoring allows for early detection and response, mitigating potential damage before it escalates. Continuous threat hunting supported by LLMs involves not just detecting known threats but also identifying anomalies that might indicate new or evolving attacks. This capability is essential in an environment where attackers are constantly adapting their tactics to evade detection. By continuously monitoring network traffic and other data sources, LLMs help SOCs stay ahead of attackers.

Predictive analytics is another area where LLMs demonstrate their value[11]. By analyzing current data patterns to forecast future attack trends, these models enable proactive defenses. SOCs can utilize these predictions to strengthen security measures in anticipation of likely attack vectors, thereby staying one step ahead of potential threats. Instead of merely reacting to incidents as they occur, SOCs can use LLMs to anticipate potential threats and prepare accordingly. This shift from a reactive to a proactive stance significantly enhances the overall security posture of an organization.

In the realm of security training and simulations, LLMs are instrumental in preparing SOC teams for various attack scenarios. These models can generate realistic cyber threat scenarios, helping to familiarize security teams with the latest tactics used by cybercriminals. This practice ensures that security personnel can respond effectively under pressure, enhancing the overall preparedness and resilience of organizations against cyber threats. By simulating various attack vectors and tactics,

LLMs help teams practice and refine their responses, improving their readiness and resilience against actual cyber threats.

In summary, the deployment of LLMs within SOCs holds the potential to revolutionize cybersecurity operations, enhancing threat detection, incident response, alert management, forensic investigations, continuous threat hunting, predictive analytics, and security training. These advancements underscore the transformative impact of LLMs in the cybersecurity domain, providing organizations with the tools they need to stay ahead of an ever-evolving threat landscape.

AI for Attackers: Leveraging LLMs in Cyber Attack Strategies

As the digital landscape evolves, so too does the sophistication of threats within it. Large Language Models (LLMs) stand at the forefront of this evolution, not only as guardians in cybersecurity defenses but also, paradoxically, as potent tools in the arsenal of cyber attackers. This dual capacity of LLMs illuminates the shadowy complexities of cyber warfare, where innovative technologies can both protect and peril. As we explore the capabilities of Large Language Models in cyber offense, it's evident that their integration into various attack strategies is refining the efficacy and sophistication of cyber threats. These advancements are reshaping the cybersecurity landscape, introducing both challenges and innovations in combating digital threats. Large Language Models power not just helpful tools like ChatGPT, but also embolden cyber attackers with their DIY version: HackGPT [12].

LLMs significantly enhance the capability of cyber attackers to develop automated exploits [13]. By analyzing vast amounts of code more efficiently than humanly possible, LLMs can quickly identify vulnerabilities. This rapid analysis accelerates the exploitation process, allowing attackers to exploit vulnerabilities before they can be patched. The implications are profound, as it shortens the time window for response by defenders and increases the success rates of cyber-attacks. With LLMs, phishing and social engineering attacks become more sophisticated and difficult to detect [14]. These models are adept at creating content that mimics legitimate user interactions, crafting messages that resonate deeply with specific targets based on their personal information or previous online behaviors. This capability makes it challenging for individuals to discern between legitimate and malicious communications, thereby increasing the success rate of these attacks.

In malware creation, LLMs contribute to developing sophisticated programs that can adapt their behaviors to evade detection tools [15]. This adaptability ensures that malware can persist within systems for longer periods, complicating detection and removal efforts. As LLMs continue to learn and evolve, the malware they help create becomes increasingly resilient and capable of bypassing advanced security measures.

LLMs also play a critical role in managing botnets, which are used to carry out large-scale automated attacks [16]. By coordinating thousands of infected devices, LLMs can execute synchronized attacks across multiple networks, increasing both the scale and impact of cyber threats. This ability to manage and direct botnets with high precision makes LLMs invaluable for executing widespread disruptions.

Through data mining and reconnaissance, LLMs enable attackers to gather extensive intelligence about potential targets. By analyzing public data and extracting actionable insights, attackers can tailor their strategies to exploit specific vulnerabilities, making each attack more precise and effective. This strategic use of data significantly enhances the potential success of cyber operations. LLMs are used to craft inputs that intentionally mislead machine learning models, a tactic known as adversarial machine learning[17]. These deceptive inputs are designed to cause security systems to misclassify or overlook malicious activities, allowing attackers to slip past defenses unnoticed. The use of LLMs in orchestrating APTs showcases their ability to conduct long-term, stealthy operations that aim to steal data or disrupt systems over extended periods. LLMs' capacity to mimic normal network traffic and user behaviors helps these threats to

remain undetected, posing significant challenges for cybersecurity teams.

LLMs assist in evading deception detection systems by simulating normal user behaviors, making it more difficult for these systems to identify malicious activities. This capability allows attackers to operate under the radar, extending the duration and effectiveness of their campaigns. Finally, LLMs facilitate the automation of exploit kits' creation and deployment. These kits are pre-packaged with code designed to automatically exploit known vulnerabilities in software systems. With LLMs, the development and update of these kits can be continuously refined, allowing attackers to exploit a broader range of vulnerabilities more effectively.

Each of these areas illustrates how LLMs, while beneficial in many respects, also pose substantial risks when used maliciously. As the digital threat landscape continues to evolve with these technologies, so too must our strategies for defense. Recognizing and understanding the potential applications of LLMs in cyber offenses will be crucial in developing more robust cybersecurity measures to counter these advanced threats.

Security Risks of LLMs: A Pandora's Box in Cybersecurity

As organizations rapidly integrate Large Language Models to enhance online customer experiences, they inadvertently expose themselves to novel cyber threats. These web LLM attacks exploit the models' access to data, Application Programming Interfaces (APIs), and user information. APIs are sets of protocols and tools for building and interacting with software applications, enabling different systems to communicate with each other. For instance, attackers might retrieve data that the LLM can access, trigger harmful actions through APIs, or launch attacks on other users and systems querying the LLM[18]. These attacks often resemble Server-Side Request Forgery (SSRF) vulnerabilities, where attackers exploit a server-side system to attack another component indirectly.

> Large Language Models ... also embolden cyber attackers with their DIY version: HackGPT.

LLMs are AI algorithms trained on vast datasets to process user inputs and generate plausible responses by predicting word sequences. They typically present a chat interface for user input, known as a prompt, which is partially controlled by input validation rules. LLMs have diverse applications, including customer service, translation, search engine optimization (SEO) improvement, and user-generated content analysis. Many web LLM attacks rely on a technique called prompt injection[19]. This involves crafting specific prompts that manipulate the LLM's output, potentially leading to unintended actions such as incorrect API calls or the generation of inappropriate content. Detecting LLM vulnerabilities requires identifying the LLM's inputs, understanding the data and APIs it can access, and probing for weaknesses in this new attack surface. Indirect prompt injection involves embedding prompts in external sources, such as webpages or emails, which the LLM processes and executes. This can result in web LLM attacks on other users, such as causing an LLM to create a malicious email-forwarding rule or embedding hidden commands in a webpage for the LLM to act upon.

LLMs are often hosted by third-party providers, and websites can describe local APIs for the LLM to use. For example, a customer support LLM might manage user accounts, orders, and inventory. The workflow for integrating an LLM with an API involves several steps, from the client calling the LLM with a user's prompt to the LLM calling external APIs and summarizing results back to the user. This process has significant security implications, as the LLM effectively acts on behalf of the user without their explicit awareness. The term "excessive agency"[20] refers to situations where an LLM has access to sensitive APIs and can be manipulated to use these APIs unsafely. Attackers can exploit this by mapping the LLM's API attack surface and sending classic web exploits to identified APIs. Even seemingly harmless

APIs can be used to find secondary vulnerabilities, such as executing a path traversal attack on an API that takes a filename as input.

Insecure output handling occurs when an LLM's output is not sufficiently validated before being passed to other systems, potentially facilitating vulnerabilities like Cross-Site Scripting (XSS) and Cross-Site Request Forgery (CSRF)[21]. For example, an LLM might return a JavaScript payload in its response due to a crafted prompt, leading to XSS when parsed by the victim's browser. Training data poisoning involves compromising the data used to train an LLM, causing it to produce misleading information. This vulnerability can arise from untrusted data sources or overly broad datasets. Attackers can also extract sensitive training data by crafting queries that prompt the LLM to reveal information about its training data, leading to privacy breaches.

To combat these vulnerabilities, organizations must implement robust security measures tailored to the unique challenges posed by LLMs [22]. This includes rigorous input validation to prevent prompt injection, careful management of API access to limit unauthorized actions, and ongoing monitoring for signs of adversarial attacks. Additionally, understanding and mitigating SSRF-like vulnerabilities in systems that integrate LLMs are crucial for securing these applications from sophisticated cyber threats.

As LLMs continue to be integrated into more aspects of digital infrastructure, their potential to both aid and undermine cybersecurity efforts becomes increasingly significant. Addressing the security challenges associated with LLMs requires a nuanced understanding of their capabilities and vulnerabilities, alongside proactive measures to secure these powerful tools against exploitation. This multifaceted approach will be essential in leveraging the benefits of LLMs while mitigating the risks they pose in the cybersecurity landscape.

Outlook: Integration and Future Implications of AI in Cybersecurity

Artificial intelligence will not only be integrated into cybersecurity but will also play a significant role in other domains such as cloud computing and process automation [23]. The deployment of AI will vary across sectors, each with its unique applications and implications. Critical infrastructures like finance and healthcare are particularly vulnerable due to their reliance on sensitive data and complex systems. These sectors have already begun to integrate AI into their operations, with notable applications in fraud detection in finance and predictive diagnostics in healthcare. This widespread integration underscores the importance of robust cybersecurity measures to protect these essential services from cyber threats.

Balancing innovation with security is a fundamental challenge for policy makers around the globe. The EU AI Act serves as a critical framework, emphasizing the importance of evaluating and managing the risks associated with AI deployment. This regulatory approach is crucial for cybersecurity [24], especially when integrating Large Language Models (LLMs). As AI technologies become embedded in digital products, it is imperative to view AI both as an opportunity and a potential risk. Strategic deployment and governance of AI [25] can ensure that it enhances cybersecurity efforts while minimizing vulnerabilities.

Looking ahead, several trends might shape the future of AI in cybersecurity. One significant trend is the increasing sophistication of AI-driven threat detection systems, which will leverage machine learning to identify and respond to threats in real time. Artificial intelligence not only simplifies the acquisition of beneficial skills but also facilitates the learning of malicious capabilities. In the context of cybersecurity, this democratization extends to cyberattacks, making them more accessible to a broader range of individuals. AI-driven tools can automate and streamline the execution of cyberattacks, lowering the barrier to entry for potential attackers. Consequently, as AI advances, it will be crucial to enhance defensive measures to counteract this increased ease of attack execution.

To future-proof cybersecurity policies for advanced AI applications, several principles must be upheld. Security by design [26] must be a core principle, ensuring that AI systems are developed with security considerations from the outset. In the rapidly evolving landscape of cybersecurity, secure software development practices are essential to minimize vulnerabilities in AI applications. This foundation ensures that AI systems are robust and less susceptible to exploitation. Regular risk assessments and continuous vulnerability management are critical, acting as proactive measures to identify and mitigate potential threats before they can be exploited.

These steps, combined with robust identity and access management mechanisms, ensure that only authorized individuals can access sensitive AI functions, safeguarding the integrity of these systems. Maintaining updated software through diligent patch management is vital for protecting against known threats. Regular updates address security flaws that could otherwise be exploited by attackers. Ensuring data backups and system redundancies is another crucial measure, providing a safety net to maintain operations during an attack. This approach helps organizations quickly restore functionality and minimize downtime, enhancing overall resilience. Improving internal capabilities through ongoing training and development is key to empowering cybersecurity teams to effectively manage AI technologies. Well-trained personnel are better equipped to handle the complexities of AI-driven cybersecurity tools and strategies. Developing comprehensive recovery plans further ensures that organizations can respond to and recover from AI-related security incidents efficiently, minimizing the impact of such events. These integrated strategies form a robust defense, balancing innovation with security in the face of evolving cyber threats.

International collaboration is essential in addressing the global nature of cyber threats. The Network and Information Security (NIS) Directive 2, which mandates the reporting of security incidents across EU member states, fosters a collaborative approach to cybersecurity. This requirement enhances transparency and coordination in responding to cyber threats. International cooperation facilitates the sharing of threat intelligence, best practices, and resources, improving the collective ability to combat cyber risks associated with AI.

Policy Recommendations

As the analysis in this paper concludes, in the light of new capabilities involving LLMs, it is imperative to implement strategic policy recommendations to strengthen cybersecurity. The following sections detail three key recommendations: increasing investments and awareness in cyber defense, addressing challenges in the public sector, and embracing Generative AI to empower CERTs and SOCs.

1. Prioritize funding for LLMs in Cyber Defense: Governments must urgently prioritize investments in cybersecurity infrastructure, particularly in AI and LLM technologies, as cyber attacks grow increasingly complex due to the use of AI by adversaries. These advanced tools are essential for real-time threat detection, automated response strategies, and enhancing overall security posture. If adversaries are leveraging AI, the state must also invest to stay competitive and secure. Funding should support continuous training for cybersecurity professionals and public awareness campaigns on cyber hygiene. Embedding AI and LLMs into cybersecurity frameworks is crucial for protecting critical infrastructure and public services from evolving cyber threats.

2. Addressing Challenges in the Public Sector: Technology alone will not suffice to ward off cyberattacks. These technologies must be integrated into the public sector, necessitating updates to the relevant authorities and agencies. Only through a holistic strategy that encompasses both technological and organizational improvements can the public sector effectively defend against modern cyber threats. The public sector faces significant challenges, including knowledge loss due to retirements, difficulty attracting young talent, and increasing workloads. Governments need to modernize their recruitment strategies to appeal to younger generations by showcasing dynamic career opportunities. Ad-

ditionally, addressing the dramatic staff shortages is critical. This requires not only improving working conditions and compensation but also investing in automation and AI technologies to alleviate the burden on existing employees.

For example, generative AI can assist in knowledge management and predictive maintenance, thus enhancing the efficiency and effectiveness of public sector operations. The public sector must confront the dual challenge of retaining institutional knowledge and attracting new talent. Pension waves are causing significant knowledge gaps, which can be mitigated by using AI to capture and retain critical information. Modernizing the perception of public sector careers to align with the expectations of younger generations is vital. Highlighting opportunities for innovation and impact within the public sector can make these careers more attractive. Additionally, addressing staff shortages through the adoption of AI-driven process automation can alleviate workload pressures, improve efficiency, and reduce burnout among employees.

3. Embracing Generative AI to Empower CERTs and SOCs: Adopting Generative AI (GenAI) can revolutionize the capabilities of Computer Emergency Response Teams (CERTs) and Security Operations Centers (SOCs). GenAI can process vast amounts of data, identify patterns, and predict potential threats more efficiently than human analysts alone. This technology can automate routine tasks, allowing cybersecurity experts to focus on more complex analyses and strategic decision-making. By integrating GenAI into cybersecurity frameworks, governments can significantly enhance their incident response capabilities and overall resilience to cyber threats.

The integration of GenAI into cybersecurity operations offers transformative potential. GenAI can significantly enhance threat intelligence by analyzing vast datasets from various sources, identifying emerging threats, and providing actionable insights. This allows CERTs and SOCs to respond more swiftly and effectively to cyber incidents. Moreover, GenAI can automate routine tasks, such as log analysis and incident categorization, freeing up human analysts to focus on more strategic activities. The predictive capabilities of GenAI enable proactive defense measures, anticipating potential attack vectors and mitigating risks before they materialize.

> ... the implications of AI must be incorporated into national cybersecurity strategies.

By implementing these policy recommendations, governments and security operations centers can build a more resilient cybersecurity framework. Increased investments, modernization of the public sector, and the adoption of advanced AI technologies like GenAI are crucial steps towards safeguarding critical infrastructure and public services in the digital age.

4. Integrate AI into the National Cybersecurity Strategy: All too often, technologies are managed in isolation, rather than being considered in combination. This fragmented approach limits their potential, as integrating multiple technologies can create synergies and enhance overall effectiveness. Artificial intelligence should not be viewed merely as a technology, but as a strategic advantage encompassing time, resources, and knowledge. In the realm of cybersecurity, whether defensive or offensive, the implications of AI must be incorporated into national cybersecurity strategies. This involves recognizing AI as more than just an application or tool, but as a strategic mindset that enhances decision-making and operational efficiency. AI will provide a strategic edge by leveraging advanced analytics, predictive capabilities, and automated responses to emerging threats. Therefore, national policies must reflect this integrated approach, ensuring that AI's potential is fully harnessed to strengthen cybersecurity frameworks. By embedding AI into the core of cybersecurity strategies, nations can stay ahead of adversaries, safeguarding critical infrastructure and maintaining robust national security. This holistic approach will facilitate a dynamic and resilient defense posture, capable of adapting to the rapidly evolving threat landscape.

Conclusion

Artificial intelligence, especially Large Language Models (LLMs), is becoming integral to cybersecurity, offering both defensive and offensive capabilities. On the defensive side, LLMs enhance threat detection, incident response, and forensic investigations, transforming Security Operations Centers (SOCs). They can predict future attack trends and provide realistic training scenarios, significantly boosting an organization's resilience against cyber threats. Conversely, AI's offensive potential is equally significant.

LLMs can automate the development of sophisticated exploits, craft highly convincing phishing schemes, and create adaptive malware. Their ability to manage botnets and conduct extensive reconnaissance further amplifies the threat landscape. The dual-use nature of AI underscores the need for robust security measures and strategic deployment. The integration of AI into critical infrastructures such as banking, finance, and healthcare necessitates stringent cybersecurity protocols. These sectors are particularly vulnerable due to their reliance on sensitive data and complex systems. The EU AI Act and similar regulatory frameworks play a crucial role in balancing innovation with security, ensuring that AI technologies are deployed responsibly. Looking to the future, AI-driven threat detection systems will become more sophisticated, automating routine security tasks and allowing human analysts to focus on complex challenges.

However, as AI capabilities advance, so will the tactics of cybercriminals. Therefore, principles such as security by design, secure software development, regular risk assessments, and robust identity and access management must be upheld. International collaboration is vital in addressing the global nature of cyber threats. The Network and Information Security (NIS) Directive 2 fosters a collaborative approach, mandating the reporting of security incidents across EU member states. This enhances transparency and coordination, improving the collective ability to combat AI-related cyber risks. In summary, as AI continues to integrate into various sectors, it is crucial to leverage its capabilities for enhancing cybersecurity while remaining vigilant about the risks. Strategic deployment, robust security measures, and international cooperation will be key in navigating the complex landscape of AI in cybersecurity.

About the Author

David Kirsch is currently a Manager at Ernst and Young in Vienna, focusing on Security and Generative AI. He holds a Master's Degree in War and Conflict Studies from the University of Potsdam and is currently finishing his MBA in Digital Transformation Management at FH BFI Wien. Formerly Director of the Data Competence Center at the Health Department of the City of Vienna and a Research Assistant at the Cyber Desk of the Institute for Counterterrorism in Herzliya, Israel, David frequently publishes and speaks on cybersecurity & public sector topics.

[1] Yao, Ricardo, The AI Hype Cycle: Are we on the Precipice of Disillusionment?, IPG Media Lab, https://medium.com/ipg-media-lab/the-ai-hype-cycle-are-we-on-the-precipice-of-disillusionment-139ab28cee77

[2] De Vynck, Gerrit, The AI hype bubble is deflating. Now comes the hard part., Washington Post, https://www.washingtonpost.com/technology/2024/04/18/ai-bubble-hype-dying-money/

[3] Kirsch, David; Bodenstorfer, Martin, KI in der öffentlichen Verwaltung – der Hofrat ohne Gesicht?, EY, https://www.ey.com/de_at/government-public-sector/ki-oeffentliche-verwaltung,

[4] Toner, Helen; What Are Generative AI, Large Language Models, and Foundation Models?, https://cset.georgetown.edu/article/what-are-generative-ai-large-language-models-and-foundation-models/

[5] Divakaran, Dinil Mon; Peddinti, Sai Teja, LLMs for Cyber Security: New Opportunities, A*STAR Institute for Infocomm Research, https://arxiv.org/html/2404.11338v1

[6] Haijzadeh, Mehrdad, Large Language Models in Cybersecurity: State-of-the-Art, https://ar5iv.labs.arxiv.org/html/2402.00891

[7] See Divakaran et al.

[8] Yamin, Muhammad Mudassar, Applications of LLMs for Generating Cyber Security Exercise Scenarios, International Journal of Information Security, https://www.researchsquare.com/article/rs-3970015/v1

[9] Aghaei, Ehsan et. al., SecureBERT: A Domain-Specific Language Model for Cybersecurity, https://ar5iv.labs.arxiv.org/html/2204.02685

[10] Ferrag, Mohamed Amine: Generative AI and Large Language Models for Cyber Security: All Insights You Need, https://arxiv.org/abs/2405.12750

[11] Siehe FN5

[12] Morrison, Ryan, How OpenAI's ChatGPT can be used to launch cyberattacks, TechMonitor, https://techmonitor.ai/technology/ai-and-automation/chatgpt-cyberattacks-openai

[13] Fang, Richard et. al., LLM Agents can Autonomously Exploit One-Day Vulnerabilities, https://arxiv.org/abs/2404.08144

[14] Heiding, Frederik, AI will Increase the Quantity – and Quality – of Phishing Scams, Harvard Business Review, https://hbr.org/2024/05/ai-will-increase-the-quantity-and-quality-of-phishing-scams

[15] Beckerich, Mika, RatGPT: Turning online LLMs into Proxies for Malware Attacks, https://arxiv.org/abs/2308.09183

[16] Charan, Sai, From Text to MITRE Techniques: Exploring the Malicious Use of Large Language Models, From Text to MITRE Techniques: Exploring the Malicious Use of Large Language Models for Generating Cyber Attack Payloads

[17] Musser, Micah, Adversarial Machine Learning and Cybersecurity, Center for Security and Emerging Technology, https://cset.georgetown.edu/publication/adversarial-machine-learning-and-cybersecurity/

[18] OWASP, Server Side Request Forgery, https://owasp.org/www-community/attacks/Server_Side_Request_Forgery

[19] Singh Sandeep, How a Prompt Injection Vulnerability Led to Data Exfiltration, hackerone, https://www.hackerone.com/ai/prompt-injection-deep-dive

[20] PortSwigger, Web LLM Attacks, https://portswigger.net/web-security/llm-attacks

[21] Yeung, Kenneth, Prompt Injection Attacks on LLMs, HiddenLayer, https://hiddenlayer.com/research/prompt-injection-attacks-on-llms/

[22] C3.ai, LLMs pose Major Security Risks, Serving as 'Attack Vectors', https://c3.ai/ai-security-vulnerabilities/

[23] Sayegh, Emil, Artificial Intelligence and Clouds: A complex Relationship of Collaboration and Concern, https://www.forbes.com/sites/emilsayegh/2023/08/23/artificial-intelligence-and-clouds-a-complex-relationship-of-collaboration-and-concern/

[24] European Commission, Cybersecurity of Artificial Intelligence in the AI Act, https://publications.jrc.ec.europa.eu/repository/handle/JRC134461

[25] Jalil, Rehan, AI Security and Governance: Navigating Complexity in the Age of AI, Forbes, https://www.forbes.com/sites/forbestechcouncil/2024/04/22/ai-security-and-governance-navigating-complexity-in-the-age-of-ai/

[26] UK Government Security, Secure by Design Principles, https://www.security.gov.uk/guidance/secure-by-design/principles/

Dr. Langweg 3, 2410 Hainburg/Donau

Tel. +43 (1) 3583080

office@aies.at | www.aies.at

Layout Design: Julia Drössler